AN ABSTRACT OF THE
DISSERTATION OF

Safaa R. Amer for the degree of Doctor of Philosophy in Statistics presented on June 7, 2004.
Title: Neural Network Imputation: A New Fashion or a Good Tool.

Abstract approved:

— Redacted for Privacy —

Most statistical surveys and data collection studies encounter missing data. A common solution to this problem is to discard observations with missing data while reporting the percentage of missing observations in different output tables. Imputation is a tool used to fill in the missing values. This dissertation introduces the missing data problem as well as traditional imputation methods (e.g. hot deck, mean imputation, regression, Markov Chain Monte Carlo, Expectation-Maximization, etc.). The use of artificial neural networks (ANN), a data mining technique, is proposed as an effective imputation procedure. During ANN imputation, computational effort is minimized while accounting for sample design and imputation uncertainty. The mechanism and use of ANN in imputation for complex survey designs is investigated.

Imputation methods are not all equally good, and none are universally good. However, simulation results and applications in this dissertation show that regression, Markov chain Monte Carlo, and ANN yield comparable results. Artificial neural networks could be considered as implicit models that take into account the sample design without making strong parametric assumptions. Artificial neural networks make few assumptions about the data, are asymptotically good and robust to multicollinearity and outliers. Overall, ANN could be time and resources efficient for an experienced user compared to other conventional imputation techniques.

Neural Network Imputation:
A New Fashion or a Good Tool

By

Safaa R. Amer

A DISSERTATION

Submitted to

Oregon State University

In partial fulfillment of

the requirements for the

degree of

Doctor of Philosophy

Presented June 7, 2004

Commencement June 2005

Doctor of Philosophy dissertation of Safaa R. Amer
Presented on June 7, 2004.

APPROVED:

Redacted for Privacy

Major Professor, representing Statistics

Redacted for Privacy

Chair of the Department of Statistics

Redacted for Privacy

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Redacted for Privacy

Safaa R. Amer, Author

ACKNOWLEDGEMENTS

First, I would like to thank God, who provided me with the strength to carry on during my study. I also want to thank my family, who offered me unconditional love, guidance, support and trust through the course of my study and my life. Specially, to my parents who taught me the value of hard work by their example and rendered me enormous support and motivation during the whole tenure of my research. To my father, Prof. Rabie Amer, who has been a source of inspiration to me throughout my life, and to my mother, Wasfia, for her love and support in my determination to find and realize my potential. I owe them everything I am today. Many thanks are due to my older sister Abir and her family for love, advice and support; to my younger sister Sawsan, and to my brother Mohamed for compassion and motivation to be a good example.

I extend my sincere gratitude and appreciation to many people who made this PhD dissertation possible. First and foremost, I would like to thank my advisor Dr. Virginia Lesser, who helped me mature as a student and as a researcher, for offering financial as well as emotional support and for her patience. I would like to acknowledge with much appreciation the crucial role of Dr. Robert Burton for providing me constant encouragement and for his valuable comments and suggestions without which this research could not be produced in the present form.

Many thanks are due to Dr. Fred Ramsey, Dr. Dave Birkes, Dr. Paul Murtaugh, and Dr. Robert Duncan for their time and caring support. I would also like to thank Dr. Justus Seely for helping me to come to OSU and for a warm welcome and understanding. I am fortunate to have had the opportunity to study and work in the Department of Statistics at Oregon State University under the guidance of Dr. Robert Smythe, with such dedicated educators, staff and in a friendly and welcoming environment.

I would like to express my sincere thanks to Dr. AlSaffar who constantly provided encouragement and support and believed in my potentials. Many thanks are

due to Dr. Dennis Oehler who always stood by me in hard times and helped me in many ways. I would also like to thank, Dr. Bassiouni, Dr. Azam, Dr Riad and Dr. Hussaini for their advice and efforts to bring me to Oregon State University. I would further like to acknowledge Osama and his family who were a second family for me during my stay in the United States and took care of me in many aspects. Special thanks are due to Anwar, Anfal, Reem, Lydia, Ahmed, Amr, Mohamed, and all my friends who provided emotional support and memories. Finally, I would like to thank Kent and Maya for helping me polish my writing.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

LIST OF TABLES

To Dad and Mom
With Love

Neural Network Imputation: A New Fashion or a Good Tool

1. INTRODUCTION

1.1. Background

Missing data is a common and persistent problem encountered by many data collection studies. Incomplete data occur for a variety of reasons, such as, interruption of experiments, equipment failure, measurement limitation, attrition in longitudinal studies, censoring, use of new instruments, changing methods of record keeping, lost records, and nonresponse to questionnaire items (Hopke, Liu and Rubin, 2001; Fichman and Cummings, 2003). The data could also be missing by design (Graham, Hofer and MacKinnon, 1996; Fetter, 2001).

The literature on the analysis of missing data goes back to the 1950s. One of the first incomplete data problems to receive attention in the statistics literature is that of missing data in designed experiments. For example, in agricultural trials, analyses of experimental designs have studied the missing plot problem (Johnson and Kotz, 1970). In psychological as well as clinical research, missing data is also a problem. For example, panel studies, cohort studies, and clinical trials with repeated measures usually suffer from attrition or drop-out resulting in incomplete records (Fitzmaurice and Clifford, 1996). Similarly, in longitudinal studies, missing data occurs for many reasons, such as subjects missing visits during the study or subjects dropping out (Liu, Tseng and Tsao, 2000). Many chemical and environmental datasets are problematical because of missing values or censored values. Censoring occurs when the outcome variable measures time to an event and the times for some events are not recorded because the experiment is terminated before the event occurs (Cox, 1972; Mann, Schafer and Singpurwalla, 1974). Time series is another area where missing values received attention (Jones, 1980; Shumway, 1984; Hopke, Liu and Rubin, 2001).

In survey sampling, the non-sampled values are regarded as missing data. In this case, inferential statistics are used for extending inferences from the sample to the

population (Kotz, Read and Banks, 1998). Early literature in survey research was mainly concerned with small numbers of variables and simple patterns of missing data (Afifi and Elashoff, 1966; Anderson, 1957). Larger survey data sets with different patterns of missing data have also been addressed (Beale and Little, 1975; Buck, 1960; Little, 1976; Trawinsky and Bargmann, 1964).

In a survey-based political science article, approximately half the respondents to surveys do not usually answer one or more questions (King, Honaker, Joseph and Scheve, 2001). Missing data is encountered in census data as well. For example, the 2000 U.S. census had 5.8 million missing data points, representing 2 percent of the population and in the 2001 British census 8.4 million values were missing, representing an average missing rate of 1 percent per question (Citro, Cork and Norwood, 2002).

1.2. Motivation and objectives

Depending on the type of research, experimental or survey design, follow-up and many other procedures are used to reduce missing data. The effort to reduce missing data is due to the potential risks that emerge when the dataset is incomplete. Missing data can lead to incorrect conclusions, bias in the parameter estimates, inflation of type I and type II error rates, degradation in the performance of confidence intervals, loss of power, and threaten the validity of statistical inference.

Discussions of determining methods to handle missing data have been initiated in the early 1950s. Some researchers consider the missing data as uninterpretable and just ignore the missing values. This is due to the uncertainty that the missing data introduces in the dataset. Another approach to handle missing data is to substitute an imputed value for the missing values. Imputation is a term used for the substitution of estimated values for missing or inconsistent data (Little and Rubin, 2002). Imputing incomplete observations as well as editing inconsistent entries has become an intermediate phase between data collection and analysis (Fellegi and Holt, 1976).

However, finding ways to deal with missing data does not diminish the responsibility of the researcher to investigate ways of reducing missing data during data collection.

An early review paper by Afifi and Elasoff (1966) summarized the literature on handling multivariate data with missing observations. They discussed approaches for estimation of means, variances, correlations and linear regression functions from incomplete data using least squares, maximum likelihood, Bayesian techniques and order statistics. Their review was restricted to missing observations that are independently and identically distributed. The authors noted that the majority of writers on the missing data problem thus far had only considered data from the multinormal distribution. Some of the methods they investigated involved mean imputation and average of simple linear regression imputations. They concluded that classical least squares and maximum likelihood do not generally yield the same estimators when there are missing values problems.

Nordbotten (1963) and Fellegi and Holt (1976) presented different ways of automating the imputation procedure. The reason behind suggesting automation is the amount of effort required for imputation. While the development of ways to handle missing data has been extensive, they are not able to substitute critical thinking. If the different approaches for imputation do not work, then a manual imputation via a human expert (researcher) is always considered. For this reason, researchers have tried to build different types of expert systems as part of the automated imputation procedure for missing data. Hand (1984) defined an expert system as "an intelligent computer program that uses knowledge and inference procedures in order to solve difficult problems, which require significant human expertise for their solution." The problem in applications of expert systems has been how to capture the human knowledge needed in a set of conditions.

Two trends emerged in the 1980's. The first trend was Artificial Intelligence (AI), a powerful inference machine based on the principles of expert systems. The artificial intelligence and expert systems were very useful in some cases but they failed to capture key aspects of human intelligence. That failure was attributed to the

fact that in order to reproduce intelligence, it was necessary to build systems with architecture similar to the brain. This led to the second trend, artificial neural networks (ANN), which grew from research in AI. Artificial neural network is a model which attempts to mimic the learning power of the human brain (Patterson, 1996). A model could be defined as physical representation of a system under study. Therefore, a network can be regarded as an implicit modeling technique. The goal for both AI and ANN is to capture the knowledge of experts (i.e. humans) and represent it in computer systems that have the capability to solve non-trivial mental tasks (Nordbotten, 1995).

Artificial neural networks, also referred to as neural networks (NN), are a recent automated imputation techniques. Neural networks have been used successfully in medicine, engineering, finance and other fields, including statistics. ANN were used to solve statistical problems such as prediction, classification, discriminant analysis, density estimation, and have been recently suggested for use in imputation. Overall, neural networks are applicable in situations in which there is a relationship between the predictor variables and predicted variables, even when that relationship is very complex. For example, health-related indices (e.g., heart rate) could be monitored and NN applied to recognize the predictive pattern to help prescribe appropriate treatments (Furlong, Dupuy and Heinsimer, 1991; Healey, 1999). Fluctuations of stock prices are another example of a complex and multidimensional phenomenon where NN are used for predicting stock prices (O'Sullivan, 1994). Artificial neural networks were also used to impute missing data in the 1990 Norwegian census and planned for the 2001 British Census imputation (Norbotten, 1996; Teague and Thomas, 1996).

The missing data for any variable could be either continuous or categorical. Evaluation of missing data for continuous variables has been discussed most often in statistical literature. Categorical variables have been the focus of ANN and classification trees in data mining, discriminant analysis and logistic models for classification. The focus of this research will be on the case of continuous variables with missing values (Breiman, Freidman, Olshen and Stone, 1984; Heitjan and Rubin, 1991; Molenberghs and Goetghebeur, 1997).

Both NN and other statistical techniques deal with missing data, but there are differences between the two approaches. Little interaction occurred between the statistical community and researchers in ANN (Cherkassky and Mulier, 1998). Overall, NN could be considered an extension of other statistical techniques (Chatfield, 2002).

Solutions to the missing data problem will be discussed with a main focus on imputation techniques. We will introduce the theoretical background of ANN and link it to statistical theory and methods. The effect of model misspecification and distributional assumptions of the missing data on the imputation accuracy of the results will be examined. We will also explore methods to integrate the sample design structure in the network and study the effects of ANN imputation on the variance. Finally, an overall evaluation of ANN in imputation will be provided.

1.3. Basic concepts

This section, combined with the glossary of terms (Appendix), is presented to help the reader achieve a better understanding of the terminology used in this research.

1.3.1. *Unit and item nonresponse*

Missing data refers to those values which are not available. In survey research missing data are also known as nonresponse. Nonresponse is the failure to obtain the information needed. Nonresponse can take one of two forms: unit nonresponse and item nonresponse. If the respondent is unwilling or unable to answer the entire questionnaire, we encounter unit nonresponse. Unit nonresponse includes undeliverable surveys or not at home, refused surveys, or if the respondent is incapable of answering the survey due to language barriers or mental and/or physical illness. In the case of unit nonresponse, it is necessary to understand the source of nonresponse in order to control it or reduce it and to estimate its effect on the survey. The nonresponse rate needs to be measured and reported. In order to measure the nonresponse rate, the researcher needs to account for all eligible units in the sample.

The ineligible respondents are not part of the unit nonresponse and should be excluded from the analysis and reported separately (Kish, 1965).

In unit nonresponse, the entire survey instrument is missing. However, if a respondent answers several questions but not others, then the missing items are referred to as item nonresponse. Item nonresponse could occur if the respondent refuses to answer specific questions, is unable to answer the question, records an unusual answer, or simply forgets to answer it.

Weights are customarily used to adjust for unit nonresponse while imputation is typically used for item nonresponse (Little, 2002). Response weights are usually estimated as the inverse of the response rate. Weighting is made under the assumption that both respondents and nonrespondents have the same characteristics within the same weighting class, which is not always a valid assumption (Lohr, 1999).

Imputation -substitution for the missing data- is not always the best solution. Imputation may make matters worse if not performed carefully. Researchers can contaminate their data by filling in the missing data resulting in a potential source of bias (King, Honaker, Joseph and Scheve, 2001). Therefore, the quality of the imputation is as important as the quality of the survey itself (Meng, 1994). Ideally, we hope to minimize missing data. However, a researcher needs to provide a tool that generates valid inferences in real applications. The reduction of the percentage of nonresponse or its effect is an attempt to reduce the bias caused by the differences between respondents and nonrespondents. The major concern is to provide a tool that generates valid inferences in real applications.

1.3.2. *Mechanisms of missing data*

Sometimes there is partial (e.g. censored data)or no information about the missing data. Any analysis of incomplete data requires certain assumptions about the distribution of the missing values, and in particular how the distributions of the observed values and the missing values of a variable are related.

The missingness pattern affects the imputation procedure since not all imputation techniques are applicable with non-ignorable missing data. Afifi and Elashoff (1966) described the pattern of missingness and selected methods for handling it. Rubin (1976) formalized the mechanism of missingness by providing definitions and terminology widely adopted in the survey literature. Rubin classified the missing data into three different types. The missingness pattern could be missing completely at random (MCAR), where the respondents are considered representative of the sample. Each value in the data set is considered equally likely to be missing (Anderson, 1957; Wilkinson, 1958; Afifi and Elashoff, 1966; Hocking and Smith, 1968; Hartley and Hocking, 1971; Orchard and Woodbury, 1972; Rubin 1972, 1976; Dempster, Laird and Rubin, 1977; Little, 1982). Therefore, the concept of missingness could be ignored and assumed to be accidental under MCAR. A second case is the case of missing at random (MAR), where we can model the non-response as a function of auxiliary information, in order to proceed with imputation or weighting (Little, 1982). The more complicated case is when the pattern of data missingness is not random and is not predictable from other variables. This type of missing data pattern implies the missing data mechanism is related to the missing values of the imputed variable (y) and other variables (x covariates). In such a case, the missingness cannot be ignored and should be handled with extreme care (Little and Rubin, 2002).

1.3.3. *Approaches to handling missing data*

Little and Rubin (2002) summarized the following classification of the most commonly used imputation techniques.

1.3.3.1. Procedures based on complete records

Procedures based on complete records are the most common methods that depend only on available data with no missing values. Some of these methods are subjective and biased, and most of them do not offer a reliable estimate of the variance. The most common procedures are complete case analysis (list-wise deletion)

and available case analysis (pair-wise deletion), which are considered safe and conservative (Allison, 2000).

Complete case analysis makes use of records with no missingness occurring for any of the variables. If a variable is missing for a certain unit, then it leads to the entire record being disregarded. This allows comparison between univariate statistics because the calculations are based on the same records. For example, we might have three variables, age, gender and weight. If any of the three variables is missing for respondent "A", then this respondent is not accounted for in the analysis.

Available case analysis makes use of all the available data. This is an advantage in univariate statistics, because this procedure retains the record if the variables of interest are not missing, although other variables in the same unit could be missing. Using the same previous example, if gender is missing for respondent "A", then a correlation coefficient between age and weight would still include information from this respondent. One major problem with available case analysis is that the sample size changes from variable to variable depending on the missingness pattern. This variability in the sample size leads to difficulty in comparing the results across all variables.

These complete record techniques assume that complete cases have the same characteristics as incomplete cases. In addition, some valuable data are lost. The amount of loss depends on the percent of missingness and the size of the dataset. The loss of data (information) reduces the statistical power during the analysis leading to a loss of precision, as well as increased bias (Collins, Schafer and Kam, 2001). These methods can, however, provide adequate results if the missing values are assumed to be MCAR. Under MCAR with modest correlation, Kim and Curry (1977) believe that available case method yields more efficient results than complete case analysis. Researchers can occasionally use the discarded information from incomplete cases to study the validity of the assumption of the records in a random sub-sample. This could be done by comparing the distribution of a particular variable based on complete cases and incomplete cases.

1.3.3.2. Weighting procedures

In sample survey data without nonresponse, randomization inference uses sampling weights. The sampling weight, the reciprocal of the probability of selection, can be considered to be the number of units in the population represented by the sample unit. Weighting procedures are often used to handle unit nonresponse as well. While sample weights are known, the nonresponse rate is unknown and needs to be calculated.

Weighting is based on assigning a specific weight to each observation. Estimates of the population characteristics are obtained using the weighted values instead of the values themselves. An easy way to compute weights is post-stratification. In post-stratification, the population is divided into different strata after the sample selection. Observed and unobserved survey elements share similar characteristics when every stratum is homogeneous with respect to the target variable of the survey. As a result, estimates of stratum characteristics will be less biased and can be used to estimate population parameters. All observations within a stratum are assigned the same weight. The weight is calculated such that the weighted sample distribution of the auxiliary variables matches the population distribution of these variables. If the strata are well constructed, the weighted sample distribution of the target variable will be similar to the population distribution (Lohr, 1999).

Another way to compute weights is by using class adjustment. In class adjustment, the respondents and non-respondents are classified into adjustment cells that cross-classify on the survey design variables (Little, 1986). Respondents in each cell are then weighted by the inverse of the response rate in the cell. The adjustment cells need to be predictive of response in order to reduce bias resulting from complete case analysis (Vartivarian and Little, 2003). They are used mostly when there are few covariates and with larger sample size. A major advantage of this type of procedure is the avoidance of model specification.

1.3.3.3. Imputation-based procedures

Imputation-based procedures simply edit an incomplete data set by filling in the missing values. Imputation allows the use of standard data analysis methods that require no missing data. Imputation-based procedures include substitution of means, regression predictions (conditional means), or single imputation techniques (e.g. "hot deck"). The performance of each method is poor except under very restrictive or special conditions (Little and Rubin, 1987).

Mean imputation is based on the idea of using the mean of a variable to fill in the missing values. The main assumption for mean imputation is the normality of the imputed variable. Although the procedure is simple, it yields inconsistent parameter estimates even under MCAR assumptions (Little, 1992). This is mainly due to a resulting superficial peak at the mean of the distribution. In this case, the variance is underestimated due to imputing the missing values at the center of the distribution. Under mean imputation, the data distribution is distorted and non linear estimates (e.g. variances, percentiles, etc.) from the data are not consistent under regular complete data procedures.

Regression imputation, a form of conditional mean imputation, is an improvement over mean imputation (Schafer and Schenker, 2000). Regression imputation makes use of auxiliary information in the dataset as well as researcher knowledge to build a model used for imputation purposes. This helps maintain the association between missing and observed variables. The estimated average from data imputed with this procedure is consistent under MCAR. In the case of MAR, additional assumptions about the moments of the distribution are required to obtain a consistent mean estimate. However, regression imputation could cause over-fitting and runs the risk of extrapolation beyond the range of the complete data. The same problem of variance underestimation occurs as in the case of unconditional mean imputation but with a slight improvement of precision.

Hot deck imputation tries to match cases by using common characteristics (Rubin, 1987). In hot deck imputation, missing values are imputed using similar responding units in the sample. The main advantage of hot deck is preserving the distribution of the sampled values as opposed to mean imputation. It yields unbiased estimators under the MCAR assumption. Hot deck is commonly used for its simplicity but with large data sets it becomes burdensome and may not properly simulate the distribution characteristics of the data. One major problem with the imputation-based procedures mentioned in this section is that they do not account for the uncertainty resulting from imputation.

1.3.3.4. Model-based procedures

Model-based procedures use a defined model for the observed data while basing inferences on the likelihood (or posterior distribution) under the suggested model. The parameters are estimated by techniques such as maximum likelihood. This approach is characterized by its flexibility and avoidance of ad-hoc methods. Model-based methods offer the availability of variance estimates that account for missingness in the data and the presence of uncertainty due to imputation.

This approach using Maximum Likelihood (ML) is a fully parametric technique. Principles for applying likelihood-based procedures to incomplete data problems were first described by Rubin (1976). Rubin showed through an example that Bayesian techniques and likelihood inferences are less sensitive to missing data in case of data missing at random. Little and Rubin (1987) provided an overview of the theory of ML estimation, both with and without missing values. ML imputation is favored over regression imputation because of the consistency and efficiency of the estimates under MAR conditions. Multiple imputation (MI) makes use of ML techniques.

Rubin (1978b) proposed the idea of multiple imputation (Rubin, 1986; Herzog and Rubin, 1983; Rubin and Schenker, 1986; Rubin, 1987). He explained the theoretical basis for multiple imputation (MI) and basic rules for combining

imputations (Rubin, 1977, 1987; Little and Rubin, 2002). However, computation strategies for generating the imputations were provided by Schafer (1997). Multiple imputation is based on imputing the missing values several times, resulting in multiple complete data sets. Regular analysis run on these data sets yield estimates that are subsequently combined to get the final results. The combined estimate from a multiply imputed data set is the average of the estimates resulting from the analysis of each completed data set separately. However, the variance of this estimate is divided into two components, the average within imputation variance and the between imputation component. The total variance is then a weighted sum of these two variance components. Inferences resulting from combining the imputations reflect the uncertainty due to non-response. Overall, the process of MI regards missing data as random variables and removes them from the inferential system by averaging. In real data analyses, MI may not result in good performance if it is not applied properly or if the mechanisms generating either the data or the missing values depart substantially from the underlying statistical assumptions (Collins, Schafer and Kam, 2001).

Multiple imputation is an attractive but a laborious process. The MI process includes multiple steps. It introduces random error into the imputation procedure in order to get approximately unbiased parameter estimates and the standard errors that account for the imputation. MI has the luxury of offering adequate variance estimates by including the error due to imputations which is not possible in any single imputation method. It has been shown that MI inferences are statistically valid from both Bayesian and frequentist perspectives (Schafer, 1997). As opposed to the traditional frequentist approach, the Bayesian perspective has been useful in providing prior information about the missing data and including it in the imputation step. However, Robins and Wang (2000) show that MI confidence intervals are not always conservative in cases of misspecification in imputation and/or analysis modeling (King, Honaker, Joseph and Scheve, 2001). Multiple imputation has successfully been applied to data sets such as the National Health and Nutrition Examination Survey (NHANES), randomized pharmaceutical trials presented to the US-FDA, and

marketing in business surveys. More recent applications involve the use of MI to address noncompliance in human randomized trials of anthrax vaccines (Rubin, 2002).

Schafer (1997) offers an extensive overview of imputation techniques associated with MI procedures. Expectation Maximization (EM) algorithm, Gibbs sampling (Hopke, Liu and Rubin, 2001), Data Augmentation (Tanner and Wong, 1987), Markov Chain Monte Carlo (MCMC) method (Rubin, 1996), Bootstrap (Efron, 1994), and other iterative techniques have been implemented with MI. These iterative techniques are used as tools for increasing the performance of the imputation.

One of the most widely known techniques is the EM algorithm proposed by Dempster, Laird and Rubin (1977), Little and Rubin (1987), and Little and Schenker (1995). The EM algorithm is an iterative algorithm that consists of two main steps. The Expectation step (E-step) computes the expected values of the missing data. The maximization step (M-step) uses the completed data to maximize the likelihood function. The new parameter estimates are substituted back into the E-step and a new M-step is performed. This procedure iterates through the two steps until convergence of the estimates is obtained. The main advantages of the EM algorithm are its generality, stability and ease of implementation. Two drawbacks of the EM algorithm are a slow rate of convergence and the lack of providing a measure of precision for the estimates (Schafer, 1997).

Another useful approach is the MCMC method which is based on a sequence of iterates forming a Markov chain (Gelman, Carlin, Stern and Rubin, 1995). The MCMC procedures are a collection of methods for simulating random draws from the joint distribution of $\left(Y_{\text{missing}}, \theta \middle| Y_{\text{observed}}\right)$, which is assumed to be a multivariate normal distribution. These draws result in a sequence of iterates form a Markov chain, which are used for imputation (Geyer, 1992; Smith and Roberts, 1992). The goal of MCMC is to sample values from a convergent Markov chain in which the limiting distribution is the joint posterior of the quantities of interest (Schimert, Schafer, Hesterberg, Fraley

and Clarkson, 2000). In practice, the major challenge of MCMC is the difficulty of assessing convergence (Gelman and Rubin, 1992).

Imputation using propensity scores is another model-based imputation. This method applies an implicit model approach based on propensity scores and approximate Bayesian bootstrap to generate the imputations (Rubin 1985a; Little 1986). A propensity score is the estimated probability that a particular element of data is missing. In order to calculate a propensity score for variables with missing values, a logistic regression is used to model the missingness. Based on the logistic regression, the propensity that a subject would have a missing value is calculated. Subjects are grouped based on quintiles of the propensity score. Then within each quintile, a posterior predictive distribution of observed data is created by taking random samples equal to the number of observed values. Finally, a value is randomly sampled from the posterior predictive distribution to impute each missing value.

1.3.3.5. Automated techniques

Increased computer power and decreased cost have encouraged more research into the automated edit and imputation techniques. Automated techniques are not part of the Little and Rubin (2002) classification. These techniques are an addition as a result of development in expert systems, machine learning, computational technology, and data mining.

Further advances in computer technology have resulted in data mining techniques. Data mining is often regarded as a merger of statistics, artificial intelligence and database research. Data mining is an analytical process intended to explore data to search for constant patterns and relationships between variables, and then validate the findings by applying the identified patterns to new subsets of data or, in our case, to apply these patterns for imputation purposes (Pregibon, 1997). Data mining is focused on the accuracy of prediction/imputation, regardless of interpretability or explanation of models or techniques used for prediction. Traditional statistical data analysis is usually concerned with the estimation of population

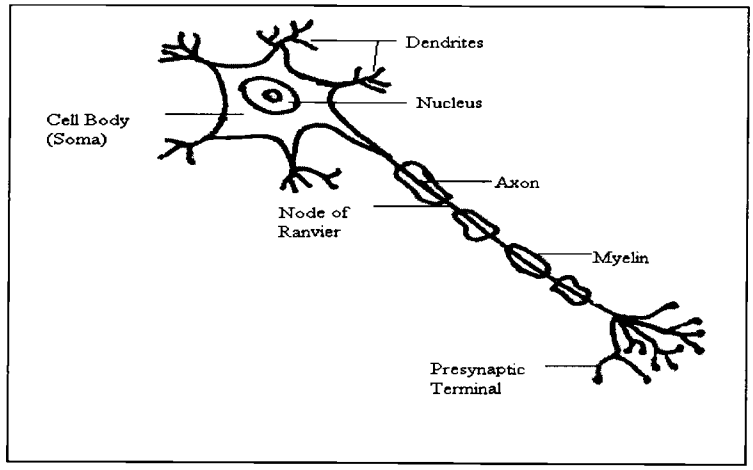
parameters by statistical inference while emphasizing the interpretability of the results. Data mining has increased recently in popularity among statisticians due to its important applications.

Several data mining techniques are being suggested lately as potential imputation techniques. The EUREDIT project, a large-scale research project that involves twelve participating organizations in seven countries over a period of three years, has been investigating and evaluating edit and imputation methods that emerged from data mining. These methods include different machine-learning techniques compared to standard statistical imputation procedures. Neural networks, which will be the focus of this research, are an example of a technique often applied to data mining.

1.3.4. *Artificial neural networks*

The field of neural networks has five decades of history, but evolved in the last fifteen years and is still developing rapidly. McCulloch and Pitts (1943) first presented ANN in a paper with the aim of explaining neuro-physiological phenomena. ANN were presented as a system modeled on the structure of the human brain. One of the most basic elements of the human brain is a specific type of cell, which provides us with the abilities to remember, think, and apply previous experiences to our actions. These cells are known as neurons (Figure 1). Each neuron is a specific cell that can generate an electrochemical signal. The power of the brain comes from the numbers of these basic components and the multiple connections between them. All natural neurons have four components, which are dendrites, soma, axon, and synapses. The synapse is the junction between the axon terminals of a neuron and the receiving cell. Basically, a biological neuron receives inputs from other sources, combines them, performs a generally nonlinear operation on the result, and then outputs the final result. When a neuron is activated, it fires an electrochemical signal. This signal crosses to other neurons, which then fire in turn. A neuron fires only if the total signal received at the cell body exceeds a certain level.

Figure 1. Biological neurons



Research in cognitive science has demonstrated the capacity of the human brain to learn simple input-output covariations from extremely complex stimuli (Lewicki, Hill, and Czyzewska, 1992). Therefore, from a very large number of simple processing units, the brain manages to perform complex tasks.

An artificial neuron is simpler than a biological neuron. A number of components are needed to specify ANN: interconnection architecture, activation function, and training algorithm that changes the interconnection parameters in order to minimize the error. Haykin (1994) defines neural network as "a massively parallel distributed processor that has a natural propensity for storing experiential knowledge and making it available for use." He compares it to the brain in two aspects: knowledge acquired from learning, and the way this knowledge is stored. Knowledge in these networks is presented as the strengths of connections between neurons. ANN do not represent knowledge that is easily interpreted, and they cannot explain the results.

A simple network has a feed-forward structure where signals flow from inputs, forward through any hidden units, eventually reaching the output units. Such a structure is chosen for the stability of its behavior. The input layer simply introduces the values of the input variables. The hidden and output neurons are connected to all

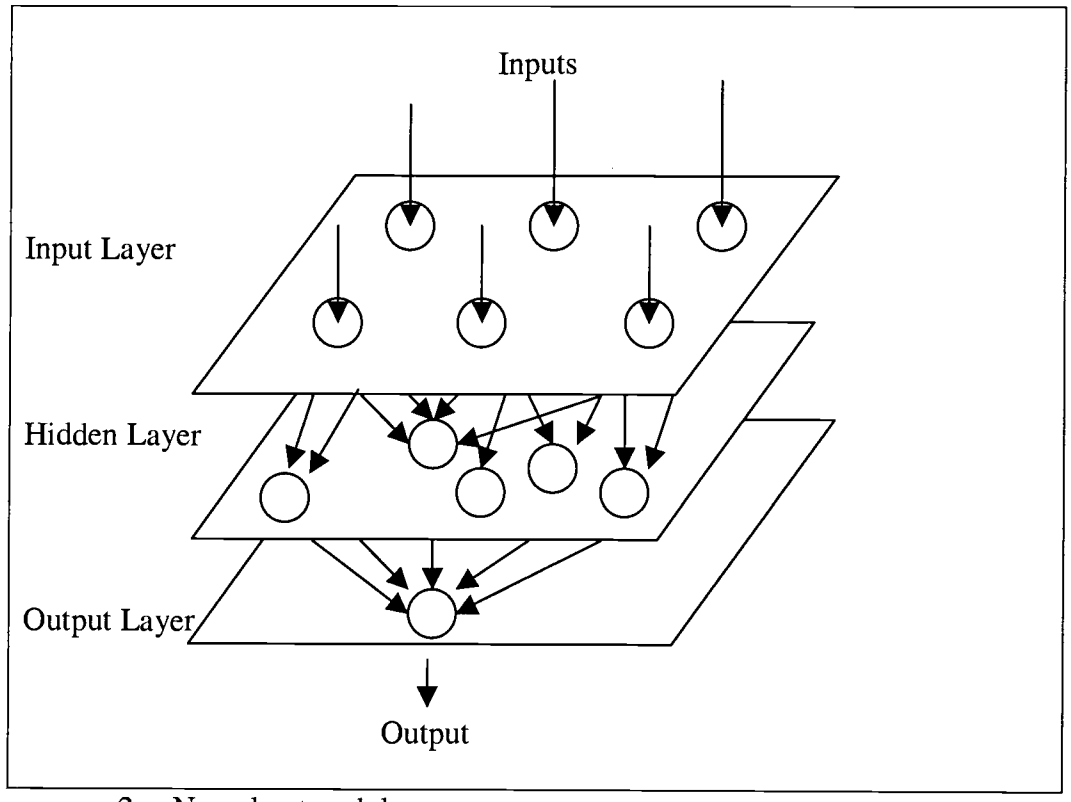
of the units in the preceding layer. The input variable values are placed in the input units, and then the hidden and output layer units are gradually executed. Each layer calculates a value by taking the weighted sum of the outputs of the units in the preceding layer. This value is passed through the activation function to produce the output of the neuron. The activation function is a function used by a node in a neural network to transform input data from any domain of values into a finite range of values. When the entire network has been executed, the output layer acts as the output of the entire network, see Figure 2. Information about errors is also filtered back through the system and is used to adjust the connections between the layers, therefore improving performance. This procedure of feedback is the key issue in the network (Haykin, 1994).

Feed-forward, multilayer networks can be considered as examples of logistic discriminant functions or non-linear regression functions (Nordbotten 1997). A feed-forward neural network with a single hidden layer could approximate any given function under certain conditions (Burton and Dehling, 1997). This is a key advantage of neural network imputation because it allows a larger choice of functions to represent the relationships between the variables.

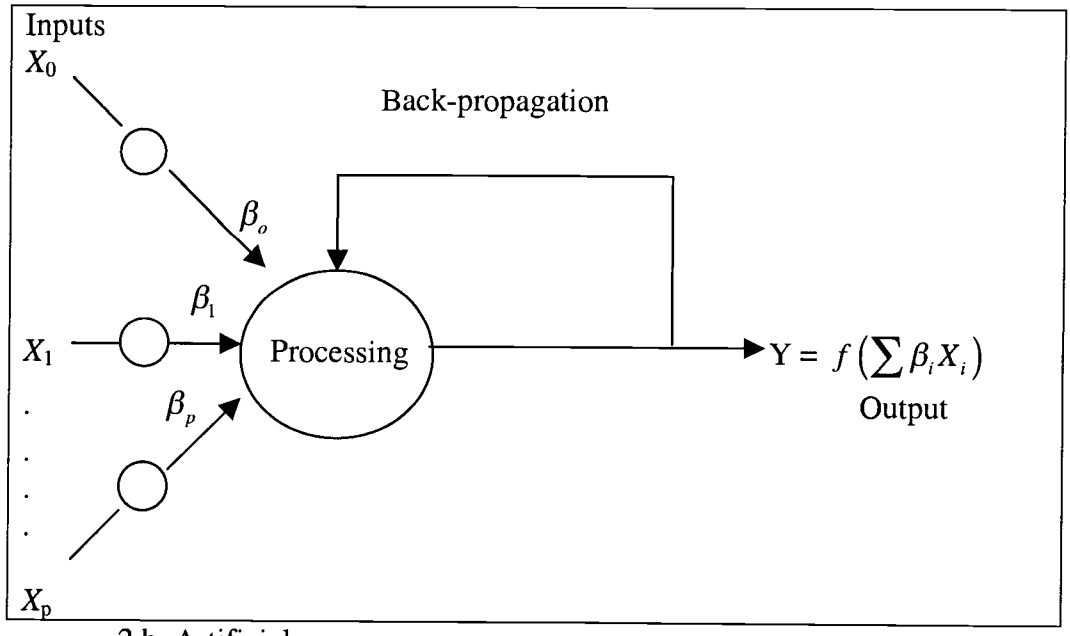
The first step in building a neural network is to design specific network architecture. This design stage is not easy because it often involves trial and error. In the second phase, the network built during the design stage is trained. Neurons apply an iterative process to the number of inputs to adjust the parameters of the network in order to predict the sample data on which the training is performed. Neural networks learn the input/output relationship through training.

There are two types of training used in neural networks. They are supervised and unsupervised training, of which supervised is the most common. In supervised learning, the analyst assembles a set of training data. The neural network is then trained using one of the supervised learning algorithms (e.g. the back propagation algorithm devised by Rumelhart, Hinton and Williams, 1986), which use the data to

Figure 2. Artificial neural networks



2.a. Neural network layers



2.b. Artificial neurons

adjust the network's weights in order to minimize the error in its predictions on the training set. After the phase of learning from an existing data set, the new network is ready and it can then be used to generate predictions. The training data contains examples of inputs together with the corresponding outputs, and then the network learns to infer the relationship between the two. The other learning technique is the unsupervised training, where the network is not presented with a final output and attempts to find patterns in the data.

The back-propagation algorithm is a well-established iterative learning algorithm that adjusts parameters according to the gradient descent principle (Werbos, 1974; Rumelhart, 1986; Kohonen, 1982; Parker, 1985; Patterson, 1996; Haykin, 1994; Fausett, 1994). The gradient descent principle is an optimization technique for non-linear functions, which attempts to move certain steps to successively lower points in the search space, in order to locate a minimum error. The difficult part is to decide how large the steps ought to be. Large steps may converge more quickly, but may also surpass the solution or move in the wrong direction. On the other end of the scale, very small steps may go in the correct direction but they require a large number of iterations. The algorithm therefore progresses iteratively, through a number of epochs (i.e. a single pass through the entire training set). For each epoch, the training cases are each submitted in turn to the network and estimated and actual outputs are compared to calculate the error. This error is used to adjust the parameters, and then the process repeats. The training stops after a certain number of epochs elapse and the error reaches an acceptable level, or when the error stops improving. For smaller networks, modern second-order algorithms such as Levenberg-Marquardt are preferred as a substantially faster alternative (Bishop, 1995; Shepherd, 1997).

All of the stages mentioned above rely on one key assumption that the training, verification, and test data must be representative of the underlying model. A neural network can learn only from cases that are present, because extrapolation beyond the training data might be incorrect. A network learns the easiest features. The resulting network represents a pattern detected in the data. Thus, the network is comparable to a

model in the traditional modeling approach. Like other imputation tools (e.g. regression), there are situations where a neural network is appropriate and others where its use is questionable. For example, an important requirement for the use of a neural network, as well as for regression, is that the user knows there is a relationship between the proposed known inputs and unknown outputs. This relationship may be unknown but it must exist. In general, a neural network is used when the exact nature of the relationship between inputs and outputs is unknown. If the researcher knew the relationship, statistical modeling would be a better technique. However, unlike in the traditional statistical modeling process, relationships in the network cannot be explained in the usual statistical terms. Neural networks can help explore data sets in search of relevant variables or groups of variables, which could facilitate model building. One major advantage of neural networks is that, theoretically, they are capable of approximating any continuous function. Thus, the researcher does not need to have any hypotheses about the underlying model. One disadvantage is that the final solution depends on the initial conditions of the network. ANN models can be designed for prediction of a single dependent variable as well as for simultaneous prediction of a number of variables (White, 1989, 1992).

Nordbotten (1963, 1997) investigated two issues when he attempted to use ANN in imputation as an automated editing procedure. He first studied the possibility of training a neural network from a sample of edited records to function like a human expert. Second, he studied the possibility of a neural network being trained on a sample of deficient raw records and then matching edited records to perform the same corrections as humans would. The ANN approach has the advantage that it does not need specific detailed editing rules or explicit assumptions about imputation. ANN are also easier and quicker than other multiple imputation models. Problems inherent in ANN are related to the specification of initial values for the parameters, learning rate, and the number of neurons of the hidden layer (Nordbotten, 1995). For more information on neural networks see Haykin (1994), Masters (1995), Ripley (1996), Welstead (1994), and Warner and Misra (1996).

1.4. Evaluation techniques

Following imputation, an essential aspect of the analysis is to test the appropriateness of the imputation procedures. For example, if a linear regression model is specified for imputation, but the relationship is intrinsically non-linear, then the parameter estimates and the standard errors of those estimates may be unreliable. When a model is drastically misspecified, or the estimation procedure converges to a local minimum in non-linear analysis, the standard errors for the parameter estimates can become very large.

For a few variables, plots are useful to investigate the type of relationships between the variables. Plots represent the most direct visual check of whether or not a model fits the data, and whether there are apparent outliers. However, in large datasets with many variables, plots might not reveal the true effect. In this thesis, several measures of accuracy are used for comparing the results from different imputation techniques. These include mean absolute percentage error (MAPE), mean absolute deviation (MAD), and mean squared deviation (MSD). These measures of accuracy are based on the average error.

Let y be the actual value, $\hat{y}$ the imputed value, and m the number of imputed values. Mean absolute percentage error measures the accuracy of imputed values as a percentage:

$$MAPE = \frac{\sum_{t=1}^{m} \left| \left(y_t - \hat{y}_t\right) / y_t \right|}{m} \times 100 \quad \left(y_t \neq 0\right).$$

Mean absolute deviation, accuracy is expressed in the same units as the data, which helps conceptualize the amount of error.

$$MAD = \frac{\sum_{t=1}^{m} \left| y_t - \hat{y}_t \right|}{m}.$$

Mean squared deviation (MSD) is similar to mean squared error (MSE) a commonly used measure of accuracy. Mean squared deviation is computed using the same denominator, m, regardless of the model, in order to compare MSD values across all models. Mean squared error deviations are computed with different degrees of freedom for different models, so MSE values are not easily compared across models.

$$MSD = \frac{\sum_{t=1}^{m} (y_t - \hat{y}_t)^2}{m}.$$

1.5. Thesis overview

This thesis is composed of five chapters as follows: Chapter 1 provides an introduction to the missing data problem and some proposed solutions as well as definitions and literature review for imputation methods and ANN. Chapter 2 presents the simple case of feed-forward neural networks with a linear relationship and no hidden layers and its similarities to and differences from simple linear regression. Then a more complex network is presented corresponding to multiple linear regression where explanatory variables could be correlated. Both simulated and real data examples are used to verify the theoretical findings. Chapter 3 extends the previous results to non-linear cases with more complex networks, allowing more hidden layers. In addition, different patterns of missingness are considered and investigated. A simulated data and a real data case are used to verify the results. Chapter 4 extends the ANN to integrate sampling design structure in the network, which are investigated and compared with traditional imputation techniques. Additionally, the variance resulting from ANN imputation is investigated. Chapter 5 concludes the thesis and gives directions for future work.

1.6. References

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

Afifi, A.A. and Elashoff, R.M. (1967). Missing Observations in Multivariate Statistics II, Point Estimation in Simple Linear Regression. Journal of American Statistical Association, 62, 10-29.

Allison, Paul D. (2000). Multiple Imputation for Missing Data: A Cautionary Tale. Sociological Methods and Research, 28, 301-309

Anderson, T.W. (1957). Maximum Likelihood Estimates for a Multivariate Normal Distribution When Some Observations are Missing. Journal of the American Statistical Association, 52, 200-203.
Beale, E.M.L. and Little, R.J.A. (1975). Missing Values in Multivariate Analysis. Journal of Royal Statistical Society, Series B, 37, 129-146.

Bishop, C.M. (1995). Neural Networks for Pattern Recognition. New York: Oxford University Press.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). Classification and Regression Trees. Wadsworth.

Buck, S.F. (1960). A Method of Estimation of Missing Values in Multivariate Data Suitable for Use with an Electronic Computer. Journal of Royal Statistical Society, Series B, 22, 302-306.

Burton, R.M. and Dehling, H.G. (1997). Mathematical Aspects of Neural Computing. Department of Mathematics, Oregon State University.

Chatfield, C. (2002). Confessions of a Pragmatic Statistician. The Statistician, 51(1), 1-20.

Cherkassky, V. and Mulier, F. (1998). Learning from data. New York: Wiley.

Citro, C.F., Cork, D.L. and Norwood, J.L. (Eds.) (2002). The 2000 Census: Interim Assessment. Washington, D.C.: National Academy Press.

Cochran, W. (1963). Sampling Techniques. (2nd edition) New York: Wiley and Sons.

Collins, L.M., Schafer, J.L. and Kam, C-M. (2001). A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures. Psychological Methods, 6 (4), 330-351.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, Series B, 39(1), 1-38.

Efron, B. (1994). Missing Data, Imputation, and the Bootstrap. Journal of the American Statistical Association, 89, 463-474.

Fausett, L. (1994). Fundamentals of Neural Networks: Architectures, Algorithms and Applications. Prentice Hall.

Fellegi, I. P. and Holt, D. (1976). A Systematic Approach to Automatic Edit and Imputation. Journal of the American Statistical Association, 71, 17-35.

Fetter, M. (2001). Mass Imputation of Agricultural Economic Data Missing by Design: A Simulation Study of Two Regression Based Techniques. Federal Committee on Statistical Methodology Research Conference.

Fichman, M., and Cummings, J. (2003). Multiple Imputation for Missing Data: Making the Most of What You Know. Organizational Research Methods, 6(3), 282-308.

Fitzmaurice, G.M., Heath, A.F. and Clifford, P. (1996). Logistic Regression Models for Binary Panel Data with Attrition. Journal of the Royal Statistical Society, Series A (Statistics in Society), 159 (2), 249-263.

Furlong, J., Dupuy, M. and Heinsimer, J. (1991). Neural Network Analysis of Serial Cardiac Enzyme Data. American Journal of Clinical Pathology, 96, 134-141.

Gelman, A., Carlin, J.B., Stern, H.S., and: Rubin, D.B. (1995). Bayesian Data Analysis. Chapman.

Gelman, A.E.; and Rubin, D.B. (1992) Inference from Iterative Simulation Using Multiple Sequences, Statistical Science, vol. 7, pp 457-472.

Geyer, C.J. (1992) Practical Markov Chain Monte Carlo, Statistical Science, vol. 7, No. 4.

Graham, J.W., Hofer, S.M., and MacKinnon, D.P. (1996). Maximizing the Usefulness of Data Obtained with Planned Missing Value Patterns: An Application of Maximum Likelihood Procedures. Multivariate Behavioral Research, 31, 197-218.

Hand, D.J. (1984). Statistical Expert Systems: Design. The Statistician, 33, 351-369.

Hartley, H.O. and Hocking, R.R. (1971). The Analysis of Incomplete Data. Biometrics, 27, 783-823.

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

Healey C. (1999). Semi-Continuous Cardiac Output Monitoring using a Neural Network. Critical Care Medicine, 27(8), 1505-1510.

Heitjan, D.F. and Rubin, D.B. (1991). Ignorability and Coarse Data. Annals of Statistics, 19, 2244-2253.

Herzog, T.N. and D.B. Rubin (1983). Using Multiple Imputations to Handle in Nonresponse in Sample Surveys. Incomplete Data in Sample Surveys, Vol.II: Theory and Annotated Bibliography (W.G.Madow, I.Olkin, and D.B.Rubin, Eds.). Academic Press.

Hocking, R. and Smith, W.R. (1968). Estimation of Parameters in the Multivariate Normal Distribution with Missing Observations. Journal of the American Statistical Association, 63, 159-173.

Hocking, R.R. and Smith, W.B. (1972). Optimum Incomplete Multinormal Samples. Technometrics, 14, 299-307.

Hopke, P. K., Liu, C., and Rubin, D.B. (2001). Multiple Imputation for Multivariate Data with Missing and Below-Threshold Measurements: Time Series Concentrations of Pollutants in the Arctic. Biometrics, 57, 22-33.

Johnson, N.L. and Kotz, S. (1970). Distributions in Statistics: Continuous Univariate Distributions I, New York: Wiley.

Jones, R. H. (1980). Maximum Likelihood Fitting of ARMA Models to Time Series with Missing Observations. Technometrics, 22 (3), 389-395.

Kim, J.O. and Curry, J. (1977). Treatment of Missing Data in Multivariate Analysis. Sociological Methods and Research, 6, 215-240.

King, G., Honaker, J., Joseph, A., and Scheve, K. (2001). Analyzing Incomplete Political Science Data: An Alternative Algorithm for Muliple Imputation. American Political Science Review, 95 (1), 49-69.
Kish, L. (1965) Survey Sampling, New York: Wiley.

Kohonen, T. (1982). Self-Organized Formation of Topologically Correct Feature Maps. Biological Cybernetics, 43, 59-69.

Kotz, S., Read, C. B. and Banks, D. L. (Eds.) (1998). Encyclopedia of Statistical Sciences. Wiley-Interscience.

Lewicki, P., Hill, T., and Czyzewska, M. (1992). Nonconscious Acquisition of Information. American Psychologist, 47, 796-801.

Little, R.J.A. (1986). Survey Nonresponse Adjustments for Estimates of Means. International Statistical Review, 54, 139-157.

Little, Roderick J.A. (1992). Regression With Missing X's: A Review. Journal of the American Statistical Association, 87(420), 1227-1237.

Little, R.J.A. and Rubin, D.B. (2002). Statistical Analysis with Missing Data. (2nd edition). New York: John Wiley and Sons.

Little, R.J.A. and Rubin, D.B. (1987). Statistical Analysis with Missing Data. (1st edition). New York: John Wiley and Sons.

Little, R.J. and Schenker, N. (1995). Missing data. In: G. Arminger, C.C. Clogg and M.E. Sobel (Eds.) Handbook of Statistical Modeling for the Social and Behavioral Sciences. New York: Plenum Press.

Little, R.J.A. (1976). Inferences about Means from Incomplete Multivariate Data. Biometrika, 63, 593-604.

Little, R.J.A. (1982). Models for Nonresponse in Sample Surveys. Journal of the American Statistical Association, 77, 237-250.

Liu, H.-M., Tseng, C.-H., and Tsao, F.-M. (2000). Perceptual and Acoustic Analyses of Speech Intelligibility in Mandarin-Speaking Young Adults with Cerebral Palsy. Clinical Linguistics and Phonetics, 14, 447-464.

Lohr, Sharon L. (1999). Sampling: Design and Analysis. Duxbury Press.

Mann, N.R., Schafer, R.E., and Singpurwalla, N.D. (1974). Methods for Statistical Analysis of Reliability and Life Data. New York: Wiley.

Masters, T. (1995). Advanced Algorithms for Neural Networks: A C++ Sourcebook. New York: John Wiley and Sons.

McCulloch, W.S. and Pitts, W.H. (1943). A Logical Calculus of Ideas Immanent in Nervous Activity. Bulletin of Mathematical Biophysics, 5, 115-133.

Meng, X.-L (1994). Multiple-Imputation Inferences with Uncongenial Sources of Input. Statistical Science, 9, 538-558.

Molenberghs, G. and Goetghebeur, E. (1997). Simple Fitting Algorithms for Incomplete Categorical Data. Journal of the Royal Statistical Society, Series B, 59, 401-414.

Nordbotten, S. (1963). Automatic Editing of Individual Observations. Conference of European Statisticians. U.N. Statistical and Economic Commission of Europe.

Nordbotten, S. (1995). Editing Statistical Records by Neural Networks. Journal of Official Statistics, 11 (4), 391-411.

Nordbotten, S. (1996). Neural Network Imputation Applied to the Norwegian 1990 Population Census Data. Journal of Official Statistics, 12 (4), 385-401.

Nordbotten, S. (1997). New Methods of Editing and Imputation. (http://www.unece.org/stats/)

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

O'Sullivan, J. W. (1994) Neural Nets: A Practical Primer, AI In Finance, Spring.

Parker, D.B. (1985). Learning Logic, Technical Report TR-47, Center for Computational Research in Economics and Management Science. MA: MIT, Cambridge.

Patterson, D.W. (1996). Artificial Neural Networks: Theory and Applications. Singapore: Prentice Hall.

Pregibon, D. (1997). Data Mining. Statistical Computing and Graphics, 7(3), 8.

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

Robins, J.M. and Wang, N. (2000). Inference for Imputation Estimators. Biometrika, 87, 113-124.

Rubin, D.B. (1972). A Non-iterative Algorithm for Least Squares Estimation of Missing Values in any Analysis of Variance Design. Applied Statistics, 21, 136-141.

Rubin, D.B. (1976) Noniterative Least Squares Estimates, Standard Errors and F-tests for Analyses of Variance with Missing Data. Journal of Royal Statistical Society, Series B, 38, 270-274.

Rubin, D.B. (1977). Formalizing Subjective Notions about the Effect of Nonrespondents in Sample Surveys. Journal of the American Statistical Association, 72, 538-543.

Rubin, D.B. (1978b). Multiple Imputations in Sample Surveys a Phenomenological Bayesian Approach to Nonresponse. Proceedings of the Survey Research Methods Section, American Statistical Association, 20-34.

Rubin, D.B. (1985a) The Use of Propensity Scores in Applied Bayesian Inference, in Bayesian Statistics 2 (J.M. Bernardo, M.H. De Groot, D.V. Lindley, and A.F.M. Smith, eds.), Amsterdam: North Holland, 463-472.

Rubin, D.B. (1986). Basic Ideas of Multiple Imputation for Nonresponse. Survey Methodology, 12, 37-47.

Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: Wiley.

Rubin, D.B. (1996). Multiple Imputation After 18+ Years. Journal of the American Statistical Association, 91, 473-489.

Rubin, D.B. and Schenker, N. (1986). Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse. Journal of the American Statistical Association, 81, 366-374.

Rumelhart, D.E., Hinton, G.E. and Williams, R. J. (1986). Learning Internal Representations by Error Propagation, in D. E. Rumelhart and J. L. McClelland (Eds.). Parallel Distributed Processing: Explorations in the Microstructures of Cognition, 1, 318-362.Cambridge, MA: MIT Press/

Schafer, J.L. (1997). Analysis of Incomplete Multivariate Data. London: Chapman and Hall.

Schafer, J.L. and Schenker, N. (2000). Inference with Imputed Conditional Means. Journal of the American Statistical Association, 95, 144–154.

Shepherd, A.J. (1997). Second-Order Methods for Neural Networks. New York: Springer.

Schimert, J., Schafer, J.L., Hesterberg, T.M., Fraley, C. and Clarkson, D.B. (2000). Analyzing Data with Missing Values in S-Plus. Seattle: Insightful Corp.

Shumway, R.H. (1984). Proceedings of the Symposium on Time Series Analysis of irregularly Observed data, E. Parzen, (Ed.) Lecture Notes in Statistics. New York: Springer-Verlag.

Tanner, M.A. and Wong, W.H. (1987). The Calculation of Posterior Distributions by Data Augmentation. Journal of the American Statistical Association, 82, 528-549.

Teague, A. and Thomas, J. (1996). Neural Networks as a Possible Means for Imputing Missing Census Data in the 2001 British Census of Population. Survey and Statistical Computing, 199-203.

Trawinski, I.M. and Bargmann, R.E. (1964). Maximum Likelihood Estimation with Incomplete Multivariate Data. Annals of Mathematical Statistics, 35, 647-657.

Vartivarian, S. and Little, R. (2003). On the Formation of Weighting Adjustment Cells for Unit Nonresponse, The Berkeley Electronic Press, working paper 10. (http://www.bepress.com/umichbiostat/paper10)

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

Welstead, S.T. (1994). Neural Network and Fuzzy Logic Applications in C/C++. New York: Wiley.

Werbos, P.J. (1974). Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences, PhD thesis, Harvard University.

White, H. (1989). Learning in Artificial Neural Networks: A Statistical Perspective. Neural Computation.

White, H. (1992). Artificial Neural Networks Approximation and Learning Theory. Cambridge, MA: Blackwell Publishers.

Wilkinson, J. (1958). The Calculation of Eigenvectors of Co-diagonal Matrices. The Computer Journal, 1, 90-96.

2. LINEAR NEURAL NETWORK IMPUTATION

2.1. Abstract

For a skilled user, the neural networks (NN) imputation technique offers an advantage of speed over other traditional imputation methods. Several theoretical results on the use of NN in missing data imputation for linear cases are investigated in this chapter. In addition, an evaluation of NN with statistical imputation procedures, such as mean, regression and hot deck imputations, is performed using simulations and real-world datasets. Simulations investigate the effect of sample size as well as assumptions about linearity and data distribution on the efficiency of the imputation methods. Results show that neural networks and linear regression provide similar estimates which are superior to the other imputation techniques under investigation.

Keywords: missing data, imputation, neural networks

2.2. Introduction

Missing values represent a common problem in most real-world datasets for governmental agencies as well as academic research that use statistics. Missing data is also referred to as nonresponse (NR). Nonresponse takes one of two forms: unit NR and item NR (Little and Rubin, 2002). Unit NR occurs when we fail to obtain any survey information on one or more selected sampling units. In this case, weighting is customarily used to adjust for the non-respondents under specific assumptions. Item NR occurs when a unit response is obtained but missing values are present on a particular survey item(s) or question(s) (Dillman, Eltinge, Groves and Little, 2002). Item NR can be adjusted by using weights or using imputation techniques (Lessler and Kalsbeek, 1992). Imputation is the substitution of the missing value by a best guess. The impact of NR on the results depends on the mechanism that causes it and on the way the missing values are handled during the analysis.

Let Y be the partially unobserved variable and X a matrix of completely observed covariates. According to Little and Rubin (2002), there are three mechanisms that generate missing data: non-ignorable (NI), missing completely at random (MCAR), or missing at random (MAR). Non-ignorable missing data implies the missing data mechanism is related to the missing values of the Y and X variables. The missing data pattern is not random and can not be predicted from other variables in the dataset. When NI missing data occur, complete case analysis can provide biased results. This bias may occur because complete case analysis implies that complete cases have the same characteristics as incomplete cases which is not the case under NI mechanism. Missing completely at random implies that the missing data mechanism is not related to the values of any variable. It assumes that respondents are representative of the sample and complete case analysis is generally used to analyze this data. When MCAR missing data occur, complete case analysis provides unbiased results. However, deleting cases with missing values could result in a loss of valuable data. The loss of data reduces the sample size, lowers the statistical power during the analysis and could lead to biased results (Collins, Schafer and Kam, 2001). Finally, MAR assumes that the missing data is unrelated to the missing values of Y but may relate to other observed values of Y or X. In case of MAR, cases with incomplete data are different from cases with complete data, but the missing data pattern is predictable from other variables.

To adjust for item nonresponse, imputation techniques are generally used. The adequate imputation method depends on the missing data mechanism detected or assumed. In case of NI missing data, a super-population model describing the behavior of the data and the response model is generally used for imputation (Woodruff, 1988). Other imputation techniques (e.g. mean, hot deck and regression imputation) are more suitable for MAR and MCAR missing data. Because handling NI missing data is complex and usually data dependent, we will restrict our research to imputation methods for the MAR and MCAR cases.

Dempster and Rubin (1983) affirm that imputation is an alluring and risky method to fill in the missing values in order to create a complete data set. Analysis using imputed data is alluring because the analysis is based on a larger data set which contains more information than the complete case analysis. The user of an imputed dataset may be inclined to use it the same way as a complete data set to draw inferences and conclusions. The use of imputed data in the analysis is considered risky if not performed with caution. Imputation could result in biased parameter estimates and lead to an underestimate of the standard errors resulting in smaller p-values.

Briefly, imputation makes use of a predictive distribution of the missing values based on the observed data. Some imputation techniques, such as mean imputation and regression imputation use modeling. Other common imputation techniques, such as hot deck imputation use a non-parametric approach based on algorithms. Analysts may apply more than one imputation technique to yield better results, which requires more effort.

Traditionally, human experts (data analysts/researchers) play a key role in determining if the imputation provides reasonable results. Recently, automation seemed necessary for large data set imputation in order to reduce human interference and make imputation more feasible. Automated editing and imputation techniques were introduced in the 1960s (Nordbotten, 1963). Researchers have proposed different types of expert systems as part of an automated imputation procedure for missing data. An expert system is an intelligent computer program that uses knowledge and inference procedures to solve complex problems, rather than significant human expertise for their solution (Hand, 1984).

Artificial intelligence (AI) techniques have been used to capture human knowledge and represent it in computerized systems in order to solve non-trivial mental tasks (Nordbotten, 1995). Artificial intelligence can be classified as a data mining technique. Data mining tools emerged as a data analyst tool due to increased computer power, improved data collection/management, and both statistical and

machine-learning algorithms. Machine-learning is the attempt to enable a computer to generate rules based on raw data that has been fed into it, learn from experience, and improve over time. Machine learning can be considered an intelligent application of statistical processes. Artificial neural networks are machine-learning algorithms and one of a class of multivariate regression models used for prediction methods (White, 1989, 1992). Linear ANN can be regarded as a two-stage regression or a classification model where the output is a linear combination of the inputs (Hastie, Tibshirani and Friedman, 2001). For more extensive details, the reader may refer to Principe, Euliano and Lefebvre (2000) for an introduction in the case of regression and to Ripley (1994) for a discussion of classification methods.

Advances in computer software and increased memory accelerated the development of artificial neural networks (ANN), as part of an effort to automate multiple tasks performed by human experts. More recently, computational techniques have been proposed by both statisticians and researchers in ANN. Unfortunately, there has been little interaction between the two communities, even though neural networks could be considered as an extension of the usual techniques in statistical pattern recognition (Cherkassky and Mulier, 1994). Imputation, along with many other analytical methods, has been influenced by the ANN trend. Bishop (1995) provided some theoretical background and discussed the accuracy of general ANN results. Artificial neural networks have been introduced by Nordbotten (1995) as a new imputation technique to account for missing data. Attempts to use ANN techniques in imputation were introduced in the Norwegian 1990 population census and the planning for the British 2001 population and housing census (Nordbotten, 1996; Clarck, 1996). These attempts were purely applied and did not provide the theoretical background of ANN.

A number of articles have been presented in the literature comparing NN and statistical tools. Cheng and Titterington (1994) offered a general review of ANN from a statistical perspective. Other review articles of the statistical aspects of neural networks were presented by Geman, Bienenstock and Doursat (1992), White (1992),

Murtagh (1994), and Michie, Siegelhalter and Taylor (1994). White (1992) concluded that NN are more efficient than statistical tools because ANN are a self-learning technology that does not require programming skills. Computer scientists have advertised ANN as the future trend in computing (Anderson and McNeill, 1992). There is no consensus on which of these approaches may perform better than the other in data analyses. Some statisticians believe when the comparisons between ANN and traditional statistical techniques are made, statistical models are shown to out-perform advanced ANN (Ripley, 1993). From the point of view of some statisticians, ANN are considered statistics for beginners (Ripley, 1993; Anderson, Pellionisz and Rosenfeld, 1990).

Two recent papers compared ANN to hot deck imputation as an evaluation effort for ANN use in imputation. In the first paper, Curdas and Chambers (1997) compared the use of ANN imputation in the U.K. census data with a hot deck imputation. Their results showed that although ANN provided more accurate imputed values, hot deck outperformed ANN by preserving the distribution and consistency of the data. Curdas and Chambers (1997) believe that the neural network approach is useful in imputation and requires more research. The second paper comparing hot deck and ANN imputation was performed by Wilmot and Shivananjappa (2001). They artificially generated missing data and used hot deck and artificial neural network imputation procedures to fill in these values and compare them to the original known values. Artificial neural network produced similar or more accurate imputed values than hot deck imputation. The authors concluded that ANN are a feasible alternative for imputation. However, they did not show that either one outperformed the other consistently. Currently, a number of European countries are evaluating ANN in the Euredit project. This is a large-scale three-year research project that involves twelve participating organizations located in seven countries. The objective of the program is to investigate and evaluate new editing and imputation methods. This project is evaluating different machine learning techniques such as ANN compared to standard statistical imputation procedures.

In this chapter, we plan to investigate imputation using linear ANN and compare it with different imputation techniques. The reason behind choosing linear ANN is their simplicity and efficiency which can often be generalized to the non-linear case. This investigation provides the reader with theoretical background for the ANN and links it to the statistical theory of regression. This will be followed by both a simulation study and a real-data application for evaluation.

2.3. Methods

Different methods for dealing with missing data are presented and their performance will be compared with ANN in this chapter. These methods include unconditional and conditional mean imputation as well as hot deck imputation.

2.3.1. *Missing data and imputation*

Missing data is a statistical problem because many traditional procedures require all values for each variable. The standard treatment for missing data in most statistical software (e.g. SAS proc glm and SPSS descriptive statistics and regression) is complete case analysis (listwise deletion) and available case analysis (pairwise deletion) where cases with missing values are not included in the analysis either completely or partially. This has been the easiest way to deal with missing data and is generally used as a baseline method for comparisons of more complicated ways to handle the missing values. However, some of the statistical software packages offer other alternatives to complete case analysis (e.g. SAS proc mixed and SPSS general linear models procedures). These procedures use maximum likelihood methods to handle missing data. Some statistical packages provide the user with specific modules for imputation techniques as well (e.g. SAS proc standard for mean substitution and proc MI for multiple imputation, SPSS mean substitution and missing value analysis module, and Solas as a specific software built to handle missing data).

When imputation is used to handle missing data, the full data set is analyzed as if the imputed values are observed data. Single imputation techniques used in this

chapter are unconditional mean imputation, hot deck and regression imputation. Unconditional mean imputation fills in the missing values of a variable with the mean of the observed data for the same variable. Hot deck imputation replaces the missing value with a randomly chosen value from an individual who has similar characteristics for other variables. Hot deck imputation is usually preferred over unconditional mean imputation due to the consistency of its results and its ability to preserve the distribution of values compared to mean imputation. Regression, also referred to as conditional mean imputation, makes use of complete cases to build a prediction model. This model uses the covariates to predict the missing response variable. Conditional mean imputation is a form of parametric imputation since it requires the specification of a model and of the conditional distribution. Therefore, more consideration is given to conditional mean imputation as compared to unconditional mean imputation, and hot deck imputation since model misspecification can affect the results (Meng, 1994).

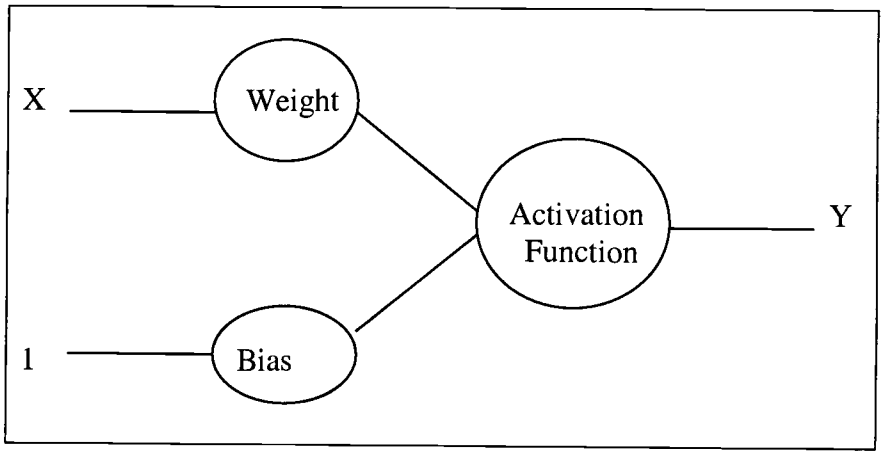
2.3.2. *Linear neural network*

Neural network research was initiated in the 1940's. The first NN were motivated by the mechanism of brain physiology. McCulloch and Pitts (1943) were the first to present ANN as a link between neurology and mathematical biophysics and the use of the NN to represent logical expressions satisfying certain conditions. Neural networks have the ability to learn by trial and error using training data sets. A neuron, the basic unit of a neural network, is an operator which performs the mapping from R^p to R. In a supervised training phase, the network receives input vectors X and corresponding desired output Y from a set of training examples.

A number of ANN models and architectures have been developed, starting with the single perceptron and evolving into multilayer perceptron (MLP). A perceptron is the simplest, non-trivial neural network used as a linear classifier, a linear discriminant function, with an input layer of p neurons and an output layer with one neuron (Rosenblatt, 1962). The perceptron, shown in Figure 3, consists of the

input vector X, and output vector Y, a weight parameter representing the strength of the relationship of the input with the output, a bias component representing the constant part of Y not attributed to X, and an activation function that represents the functional form of the relationship between X and Y.

Figure 3. Perceptron

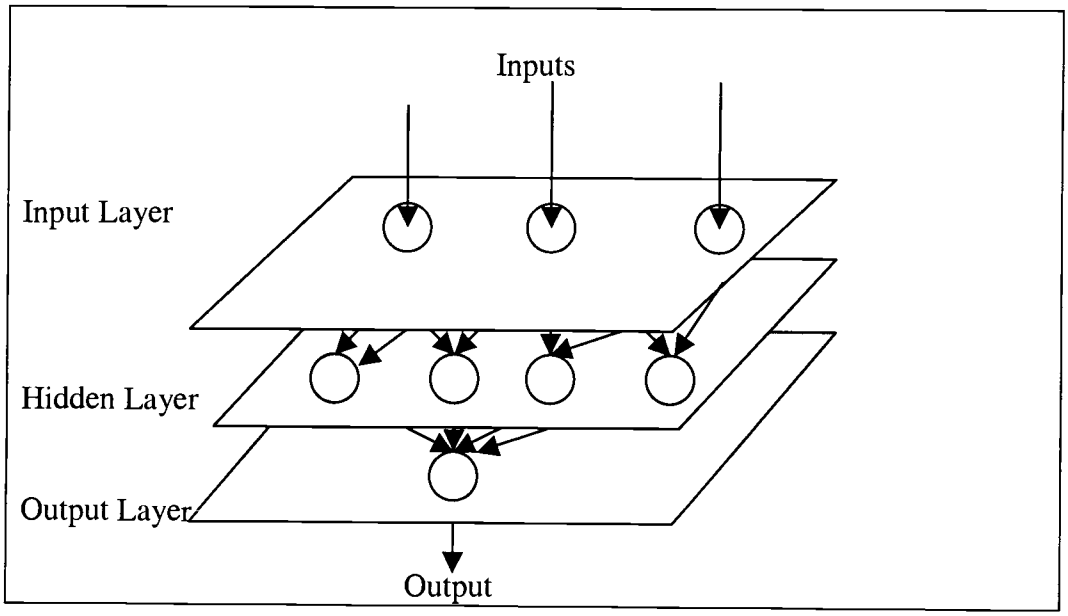


Perceptrons have several limitations. First, the output can only take a value of 0 or 1. Second, the number of inputs and outputs is restricted. Third, perceptrons can only classify linearly separable sets of vectors. The input vectors are said to be linearly separable when, in the case of a straight line or a flat subspace with many input vectors, a line can be drawn to separate them into their correct categories.

The type of artificial neural network illustrated in this chapter is feed-forward networks. Multilayer perceptrons, also known as "feed-forward" networks with hidden (intermediate) layers, are the most popular networks in use today (Rumelhart and McClelland, 1986). In a feed-forward network, the information flows from inputs through any hidden units, eventually reaching the output units as shown in Figure 4. In feed-forward networks, the input neurons receive values of explanatory variables and the output provides the dependent variable. A feed-forward network with one hidden layer containing a suitable number of neurons (processing units) is able to mimic complex functions. Burton and Dehling (1997) showed that neural networks with

enough hidden neurons could approximate any practical function. A better fit of the data can be obtained by increasing the complexity of the network.

Figure 4. Feed-forward network



Multilayer feed-forward networks distinguish themselves by the presence of one or more hidden layers of neurons. The role of the hidden neurons is to mediate between the external input and the network output. Inputs and outputs are connected through neurons, which transform the sum of all received input values to an output value according to an activation function and connection weight. The connection weights represent the strength of the connection between the neurons. The network weights (parameters) are randomly initialized and then changed in an iterative process to reflect the relationships between the inputs and outputs. The NN compute a function in the form $f_\theta : R^p \rightarrow R$, where p is the number of input variables and θ is the vector of parameters consisting of network weights and thresholds (biases).

In order to help the reader understand the similarities between ANN and statistical techniques, consider the following example. This example illustrates a

simple network with one input variable X and one output variable Y, and no hidden layers:

$y_i = \beta x_i + \alpha + \xi_i = \tilde{y}_i + \xi_i$, $i = 1,\ldots,n$, where $\xi_i = y_i - \tilde{y}_i$, such that α is the network bias, β is the network weight, and ξ is the network error. The NN predict outputs (predicted values of Y) and compare them to a set of desired outputs (observed values of Y) in order to adjust the parameters (i.e. α and β) to yield the smallest error ξ. Training is based on the idea of minimizing the output error using a pre-specified objective function (i.e. the function that needs to be minimized, for example the mean squared error or likelihood function) through adjustment of the network parameters.

Analytical solutions require knowledge of the error surface/distribution, which is not available to the neural network. Instead, a numerical solution using the gradient descent algorithm for function minimization is used. This is a form of an iterative algorithm starting with the initial values of network weights/parameters, which are adjusted during the training phase in response to training errors using the delta rule (Widrow and Hoff, 1960). This rule can be expressed as: $u_k = u_{k-1} - \eta_k \nabla J$, where u is the network parameter, η_k is the step size upon iteration k, and ∇J is the rate of change of the objective function. Note that the step size should be greater than zero for the training to occur.

To measure the divergence between observed and predicted values, different objective functions could be used. The mean square error function is often adopted due to its mathematical convenience (Hastie, Tibshirani and Friedman, 2001). In this case we have the error function J as the mean sum of squared errors:

$$J = \frac{1}{2n}\sum_{i=1}^{n}\xi_i^2$$
$$= \frac{1}{2n}\sum_{i}\left(y_i - \tilde{y}_i\right)^2$$
$$= \frac{1}{2n}\sum_{i}\left(y_i - \alpha - \beta x_i\right)^2$$
$$= \frac{1}{2n}\sum_{i}\left(y_i - \beta x_i\right)^2 \qquad \text{w.l.o.g. assume } \alpha=0$$

$$J = \frac{1}{2n}\sum_{i}\left(y_i^2 - 2y_i x_i \beta + x_i^2 \beta^2\right)$$

This derivation is identical to simple linear regression procedure (Principe, Euliano, and Lefebvre, 2000). Taking the derivative with respect to β, we have

$$\nabla_\beta J = \nabla J$$
$$= \frac{\partial J}{\partial \beta}$$
$$= \frac{1}{n}\left(-\sum_i y_i x_i + \beta \sum_i x_i^2\right) \overset{set}{=} 0$$

The solution to these equations yields the following parameter estimate

$$\beta^* = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i}^{n} x_i^2}.$$

In order to use the gradient decent procedure, at step k, we need

$$\nabla J(k) = \frac{\partial J}{\partial \beta(k)}$$
$$= \frac{\partial}{\partial \beta(k)} \frac{1}{2n} \sum_i \xi_i^2$$
$$\approx \frac{1}{2} \frac{\partial}{\partial \beta(k)} \left(\xi^2(k)\right)$$
$$= -\xi(k) x(k)$$

Therefore, using the delta rule, the parameter estimate at iteration $(k+1)$ can be written as $\beta(k+1) = \beta(k) + \eta \xi(k) x(k)$ where η = step size.

In order to achieve convergence, the largest step size is needed and can be found through the relationship between J and $J_{\min}$:

$$J_{\min} = \frac{1}{2n} \left[\sum y_i^2 - \frac{\left(\sum y_i x_i\right)^2}{\sum x_i^2} \right]$$

The objective function J is a function of $J_{\min}$ as follows:

$$J = \frac{1}{2n} \sum (y_i - \beta x_i)^2$$
$$= \frac{1}{2n} \sum \left(y_i^2 + \beta^2 x_i^2 - 2\beta x_i y_i\right)$$
$$= \frac{1}{2n} \left[\sum y_i^2 + \beta^2 \sum x_i^2 - 2\beta \sum x_i y_i \right]$$

By adding and subtracting $\frac{\left(\sum y_i x_i\right)^2}{\sum x_i^2}$, we get

$$J = \frac{1}{2n}\left[\sum y_i^2 - \frac{\left(\sum y_i x_i\right)^2}{\sum x_i^2} + \beta^2 \sum x_i^2 - 2\beta \sum x_i y_i + \frac{\left(\sum y_i x_i\right)^2}{\sum x_i^2}\right]$$

$$= \frac{1}{2n}\left[\sum y_i^2 - \frac{\left(\sum y_i x_i\right)^2}{\sum x_i^2} + \left(\beta - \frac{\sum y_i x_i}{\sum x_i^2}\right)^2 \sum x_i^2\right]$$

Substituting in the equation with $\beta^* = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i}^{n} x_i^2}$, we get

$$J = \frac{1}{2n}\left[\sum y_i^2 - \frac{\left(\sum y_i x_i\right)^2}{\sum x_i^2}\right] + \frac{1}{2n}\left(\beta - \beta^*\right)^2 \sum x_i^2$$

$$J = J_{\min} + \frac{1}{2n}\left(\beta - \beta^*\right)\left(\sum x_i^2\right)\left(\beta - \beta^*\right)$$

This relationship can be re-written as

$$J = J_{\min} + \frac{v}{2}\left(\beta - \beta^*\right)^2 \qquad \text{where} \quad v = \frac{1}{n}\left(\sum_i x_i^2\right)$$

The rate of change in J is $\nabla J = v\left(\beta - \beta^*\right)$. Therefore, using the gradient descent formula, the iteration that updates the parameter can be written as

$$\beta(k+1) = \beta(k) - \eta_{k+1} \nabla J_k$$
$$\beta(k+1) = (1 - \eta v)\beta(k) + \eta v \beta^*$$

The above is a difference equation that can be solved by induction. First subtract β^* from both sides of the equation to get

$$\beta(k+1) - \beta^* = (1 - \eta v)\beta(k) + \eta v \beta^* - \beta^*$$

By using $\beta(0)$ as a starting value and use induction, we get

$$\beta(k) = \beta^* + (1-\eta v)^k \left(\beta(0) - \beta^*\right)$$
$$\beta(k+1) = \beta^* + (1-\eta v)^{k+1} \left(\beta(0) - \beta^*\right)$$

Since η>0, we need $(1-\eta v)^k \leq 1$ to guarantee convergence. As a result, we have to satisfy the following condition: $|1-\eta \upsilon| < 1$. Therefore, the maximum step size is obtained through the relationship $\eta < \eta_{\max} = \frac{2}{\upsilon}$. As a result, J can be expressed as

$$J = J_{\min} + \lambda (1-\eta v)^{2k} \left(\beta(0) - \beta^*\right)^2.$$

Alternative functions may be used instead of the MSE to determine the parameters. For example, the NN can be considered a conditional probability distribution, $P(Y|X,\theta)$, parameterized by an n-dimensional real parameter vector (Bishop, 1995). In this case, we assume the error is based on maximum likelihood and is i.i.d. N (0, σ^2). This yields the same result using either maximum likelihood estimation or MSE as follows:

$$P(X,Y) = P(Y|X)P(X)$$

$$L = \prod_i P(x_i, y_i) = \prod_i P(y_i|x_i)P(x_i)$$

When (x_i, y_i) are drawn independently from the same distribution, ANN could be regarded as framework for modeling $P(Y|X)$. In order to find the maximum likelihood estimator for β, set the error function to be

$$E = -\ln(L)$$
$$= -\sum_i \ln\left(P(y_i \mid x_i)\right) - \underbrace{\sum_n \ln\left(P(x_i)\right)}_{\text{doesn't depend on network parameters} \Rightarrow \text{can be dropped}}$$
$$E = -\sum_i \ln\left(P(y_i \mid x_i)\right)$$

If we assume that the network error is i.i.d. N (0, σ^2), we get:

$$P(y|x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\hat{y}(x,\beta)-y)^2}{2\sigma^2}}$$

Therefore by substituting in the error function with the actual $P(Y|X)$, given above, we get

$$E = \underbrace{\frac{1}{2\sigma^2}}_{\text{constant}} \sum_{i=1}^{n}\left[\hat{y}(x_i,\beta) - y_i\right]^2 + \underbrace{n\ln\sigma + \frac{n}{2}\ln(2\pi)}_{\text{doesn't depend on network parameters} \Rightarrow \text{can be omitted}} = \sum_{i=1}^{n}\left[\hat{y}(x_i,\beta) - y_i\right]^2$$

We need to find β^* that minimizes the error using MLE. We obtain the same estimate $\beta^* = \frac{\sum_i x_i y_i}{\sum_i x_i^2}$, as we obtained using MSE. The gradient descent is used to adjust the network parameters. This shows that we are able to reach the same results while using different objective functions.

Application of the gradient descent to the error function for adjusting parameter estimates results in a back-propagation (BP) algorithm (Rumelhart, 1986). Back-propagation is the basis for supervised training of a neural network. A feed-forward network with BP consists of using the data to estimate the parameters in an iterative procedure. This iterative procedure consists of a sequence of forward and backward passes. During the forward pass, the parameters of the network are set and

fixed to produce an output. Once the output is estimated, the backward pass computes the errors from the output layer, compared to the real value of y, and then propagates these errors to provide an estimate of the hidden layer error. The forward and backward pass calculations are called the back-propagation process. First, the network is set with arbitrary initial parameters. Second, the errors are calculated to compute the gradient in order to adjust the estimates of the parameters. The aim of this algorithm is to produce an output value with minimum mean square error averaged over the inputs.

A feed-forward network with no hidden layers is identical to a multiple linear regression with more than one explanatory variable for predicting one output. The corresponding model is as follows (Principe, Euliano. and Lefebvre, 2000):

$$\hat{y} = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

$$\begin{aligned}
\xi_i &= y_i - \left(\beta_o + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip}\right) \\
&= y_i - \left(\beta_o + \sum_{k=1}^{p} \beta_k X_{ik}\right) \qquad \text{where } p\text{=number of explanatory variables} \\
&= y_i - \sum_{k=0}^{p} \beta_k X_{ik} \qquad \text{i=1,...,n} \; ; X_{io} = 1
\end{aligned}$$

with $\beta = \left(\beta_o, \ldots, \beta_p\right)^T$.

The objective function, also called the performance function because it is used to test the performance of the neural network, can be written as:

$$J = \frac{1}{2n} \sum_{i=1}^{n} \left[y_i - \sum_{k=0}^{p} \beta_k X_{ik} \right]^2 .$$

In order to find the parameter estimates, we need to minimize the objective function. Taking the derivative with respect to the parameters we get the normal equations

$$\frac{\partial J}{\partial \beta_k} \Rightarrow (p+1) \text{ equations in } (p+1) \text{ unknowns.}$$

For the m^{th} parameter β_m, we have

$$\frac{\partial \mathrm{J}}{\partial \beta_m} = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \sum_{k=0}^{p} \beta_k x_{ik} \right] x_{im} \overset{set}{\equiv} 0$$

$$\frac{1}{n}\sum_{i=1}^{n} x_{im} y_i = \frac{1}{n}\sum_{k=0}^{p} \beta_k^* \sum_{i=1}^{n} x_{im} x_{ij} \qquad ; \ \mathrm{j}=0,1,\ldots,p$$

$$P_m = R_{mj} \sum_k \beta_k^* \qquad ; \mathrm{j}=0,1,\ldots,p$$

The above can be written in a matrix format as $P = R\beta^*$ where β is a vector with $(p+1)$ connections, β^* represents the connection weights for the optimal solution, and R and P are correlation matrices with the following elements:

$$R_{mj} = \frac{1}{n}\sum_i x_{im} x_{ij} \qquad \text{correlation between the } x\text{'s}$$

$$P_m = \frac{1}{n}\sum_i x_{im} y_i \qquad \text{correlation between } y \text{ and } x\text{'s}$$

These correlations are important because they determine the relationships between the variables in the model and affect the parameters estimates. A weak correlation structure in the data could lead to fitting more complex networks with extra intermediate layers and neurons to account for the complex relationship between the variables. The objective function can be re-written as a function of these correlations as follows

$$J = \left[\frac{1}{n} \beta^T R \beta - P^* \beta + \sum_i \frac{y_i^2}{2n} \right]$$

$$\nabla J = R\beta^* - P \Rightarrow \beta^* = R^{-1} P$$

where $\nabla J = \left[\frac{\partial J}{\partial \beta_o}, \ldots, \frac{\partial J}{\partial \beta_p} \right]^T$

The solution is derived through solving $|R-\lambda I|=0$ to find the eigenvalues resulting in $\beta(k+1) = \beta(k) - \eta \nabla J(k)$, where $\beta(k) = \left[\beta_o(k), \ldots, \beta_p(k) \right]$. The largest step size η is part of the least squares solution as follows: $\beta(k+1) = (I - \eta R)\beta(k) + \eta R \beta^*$. The least mean square solution is $\beta(k+1) = \beta(k) + \eta \xi(k) x(k)$.

A feed forward network with one hidden layer is suggested to account for the correlation structure of the variables (Principe, Euliano and Lefebvre, 2000). A linear activation function, on both the hidden and output layers, results in a linear neural network, and is considered one of a class of linear models, which is the main focus of this chapter. Linear ANN typically produce results with high bias and low variance. Complex non-linear networks, with an increased array of activation functions, offer flexibility in modeling the data structure. Non-linear models have low bias and high variance and are the focus of Chapter 3. Linear neurons could be trained to learn an affine function of the inputs, of the form $f(x) = \beta.x + \alpha$, or be used to find a linear approximation to a non-linear function. An affine function is a linear function plus a translation, where a translation is a transformation in which the origin of the coordinate system is moved to another position while the direction of the axes remains the same. Linear neurons have limited capabilities and are restricted to linear computations only.

A simple linear network constructed with few network weights may not be flexible enough to model the original function. Larger networks with more weights create a more complex network, but with a risk of over-fitting. Over-fitting is a serious

problem if not handled correctly (White, 1992). A network needs to represent the data used in training and be able to generalize to new data points.

The dimensionality of data, in other words the degrees of freedom, in NN is not usually a problem. Instead, it is the complexity of the data that leads to the need for a more complex network with more neurons. The larger number of neurons leads to a larger number of parameters to be estimated in the neural network. From a statistical standpoint, by increasing the number of parameters in the model, fewer degrees of freedom are available for the estimation of the mean square error. However, this is not the case in ANN since the main concern is the complexity of the network and not the degrees of freedom. Artificial neural networks can be considered to be statistically over-parameterized, due to the number of parameters that are estimated. However, the degrees of freedom have an effect only on the step size in the gradient descent algorithm and not on the possibility of estimating more parameters. The step size may be doubled with the increase of the sample size but, at the same time, the choice of the step size is arbitrary. Therefore, the step size could change during the different iterations of the gradient descent algorithm. This is shown in the following result.

Result 1:

The aim of this result is to show that the degrees of freedom affect only the gradient descent step size. The definition of *MSE* is $MSE = \frac{SSE}{n-p} = \frac{\sum \xi^2}{n-p}$ where p=2 in the case of simple linear regression with an intercept. Set $MSE=J$, we have:

$$
\begin{aligned}
J &= \frac{1}{n-p}\sum_{i=1}^{n}\xi_i^2 \\
&= \frac{1}{n-p}\sum_{i}\left(y_i-\beta x_i\right) \\
&= \frac{1}{n-p}\sum_{i}\left(y_i^2-2y_i x_i\beta+x_i^2\beta^2\right)
\end{aligned}
$$

Taking the derivative with respect to β,

$$\begin{aligned}\frac{\partial J}{\partial \beta} &= \frac{1}{n-p}\left(-2\sum_i y_i x_i + 2\beta\sum_i x_i^2\right)\\ &= \frac{2}{n-p}\sum_i\left(\beta x_i^2 - y_i x_i\right)\\ &= \frac{-2}{n-p}\sum_i x_i\left(y_i - \beta x_i\right)\\ &\overset{set}{=} 0\end{aligned}$$

As a result, we have $\beta^* = \frac{\sum_i x_i y_i}{\sum_i x_i^2}$. This is the same estimate as in the case where we had $J = \frac{1}{2n}\sum_{i=1}^{n}\xi_i^2$.

Therefore, $J_{\min} = \frac{1}{n-p}\left[\sum y_i^2 - \frac{\left(\sum y_i x_i\right)^2}{\sum x_i^2}\right]$

Following the same steps used while proving the similarity between ANN and regression, we intend to show that the degrees of freedom do not affect the parameter estimate.

$$\begin{aligned}J &= \frac{1}{n-p}\sum\left(y_i - \beta x_i\right)^2\\ &= \frac{1}{n-p}\sum\left(y_i^2 + \beta^2 x_i^2 - 2\beta x_i y_i\right)\\ &= \frac{1}{n-p}\left[\sum y_i^2 + \beta^2\sum x_i^2 - 2\beta\sum x_i y_i\right]\\ &= \frac{1}{n-p}\left[\sum y_i^2 - \frac{\left(\sum y_i x_i\right)^2}{\sum x_i^2} + \beta^2\sum x_i^2 - 2\beta\sum x_i y_i + \frac{\left(\sum y_i x_i\right)^2}{\sum x_i^2}\right]\end{aligned}$$

$$J = \frac{1}{n-p}\left[\sum y_i^2 - \frac{\left(\sum y_i x_i\right)^2}{\sum x_i^2} + \left(\beta - \frac{\sum y_i x_i}{\sum x_i^2}\right)^2 \sum x_i^2\right]$$

$$= \frac{1}{n-p}\left[\sum y_i^2 - \frac{\left(\sum y_i x_i\right)^2}{\sum x_i^2}\right] + \frac{1}{n-p}\left(\beta - \beta^*\right)^2 \sum x_i^2$$

$$= J_{\min} + \frac{1}{n-p}\left(\beta - \beta^*\right)\left(\sum x_i^2\right)\left(\beta - \beta^*\right)$$

At step k, we have

$$\nabla J(k) \approx \frac{n}{n-p}\frac{\partial}{\partial \beta(k)}\left(\xi^2(k)\right)$$

$$= -\frac{2n}{n-p}\xi(k)x(k)$$

As a result,

$$\beta(k+1) = \beta(k) + \frac{2n}{n-p}\eta\xi(k)x(k)$$

$$= \beta(k) + \eta^*\xi(k)x(k)$$

where the new step size $\eta^* = \frac{2n}{n-p}\eta$, in the gradient descent algorithm, is the only term of the equation influenced by the number of parameters in the model. The degrees of freedom affect only the step size during the estimation procedure. Since the step size is arbitrarily chosen during ANN estimation, our concern is creating a network that is not too complex.

A method for reducing the model complexity is ridge regression controlled by a single parameter (Bishop, 1995). Ridge regression is a method for analyzing multiple regression data that suffer from multicollinearity by adding a degree of bias to the estimates to achieve more reliable estimates as a result. When multicollinearity exists, least squares estimates are unbiased, but their variances are large. We hope to

achieve more precise estimates in ANN estimation by using ridge regression. The ridge penalty does not minimize the mean squared error (MSE). The ridge regression method consists of adding a penalty before the minimization process. In the ridge regression method, the objective function, J, to be minimized takes the form $J = \frac{1}{2n}\sum_{i=1}^{n}\xi_i^2 + \lambda\beta^2$ where n is the number of observations, ξ is the network error, λ is the ridge biasing constant, and β network parameter.

Result 2:

Using the ridge penalty method to reduce the network complexity provides the following results. First we determine the penalty to be added and proceed with the minimization of the objective function J, such that

$$
\begin{aligned}
J &= \frac{1}{2n}\sum_{i=1}^{n}\xi_i^2 + \lambda\beta^2 \\
&= \frac{1}{2n}\sum_{i=1}^{n}\xi_i^2 + \frac{\lambda^*}{2n}\beta^2 \\
&= \frac{1}{2n}\sum_{i}\left(y_i - \tilde{y}_i\right)^2 + \frac{\lambda^*}{2n}\beta^2 \\
&= \frac{1}{2n}\sum_{i}\left(y_i - \beta x_i\right)^2 + \frac{\lambda^*}{2n}\beta^2 \\
&= \frac{1}{2n}\sum_{i}\left(y_i^2 - 2y_i x_i\beta + x_i^2\beta^2\right) + \frac{\lambda^*}{2n}\beta^2 \\
&= \frac{1}{2n}\sum_{i} y_i^2 - \beta\frac{\sum_i y_i x_i}{n} + \beta^2\frac{\sum_i x_i^2}{2n} + \frac{\lambda^*}{2n}\beta^2 \\
&= \frac{1}{2n}\sum_{i} y_i^2 - \beta\frac{\sum_i y_i x_i}{n} + \frac{\beta^2}{2n}\left(\sum_i x_i^2 + \lambda^*\right)
\end{aligned}
$$

Taking the derivative with respect to β, we have

$$\nabla_\beta J = \nabla J$$
$$= \frac{\partial J}{\partial \beta}$$
$$= -\frac{\sum_i y_i x_i}{n} + \frac{\beta}{n}\left(\sum_i x_i^2 + \lambda^*\right) \overset{set}{=} 0$$

The solution to these equations yields the following parameter estimate

$$\beta^* = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_i^n x_i^2 + \lambda^*}$$

Using the above estimate, we get $J_{\min}$ as follows

$$J_{\min} = \frac{1}{2n}\left[\sum y_i^2 - \frac{\left(\sum y_i x_i\right)^2}{\sum x_i^2 + \lambda^*} + \frac{\left(\sum y_i x_i\right)^2 \left(\sum x_i^2 + 2n\lambda^*\right)}{\left(\sum x_i^2 + \lambda^*\right)^2}\right]$$
$$= \frac{1}{2n}\left[\sum_i y_i^2 - \frac{\left(\sum_i x_i y_i\right)^2}{2\left(\sum x_i^2 + \lambda^*\right)}\right]$$

We had,

$$J = \frac{1}{2n}\sum\left(y_i - \beta x_i\right)^2 + \lambda\beta^2$$

Following similar steps as in page 41, it could be shown that

$$J = J_{\min} + \frac{1}{2n}\left(\beta - \beta^*\right)\left(\sum x_i^2 + \lambda^*\right)\left(\beta - \beta^*\right)$$

At step k,

$$\nabla J(k) = \frac{\partial J}{\partial \beta(k)}$$
$$= \frac{\partial}{\partial \beta(k)} \frac{1}{2n} \left[\sum_i \xi_i^2 + \lambda^* \beta^2 \right]$$
$$= \frac{1}{2}.2 \left[\sum_i \beta(k) x_i(k) - y_i \right] \frac{\partial}{\partial \beta(k)} \left(\sum_i \beta(k) x_i(k) - y_i \right) + 2\lambda^* \beta$$
$$= \left(\sum_i \beta(k) x_i(k) - y_i \right) x_i(k) + 2\lambda^* \beta$$
$$= -\xi(k) x(k) + 2\lambda^* \beta(k)$$

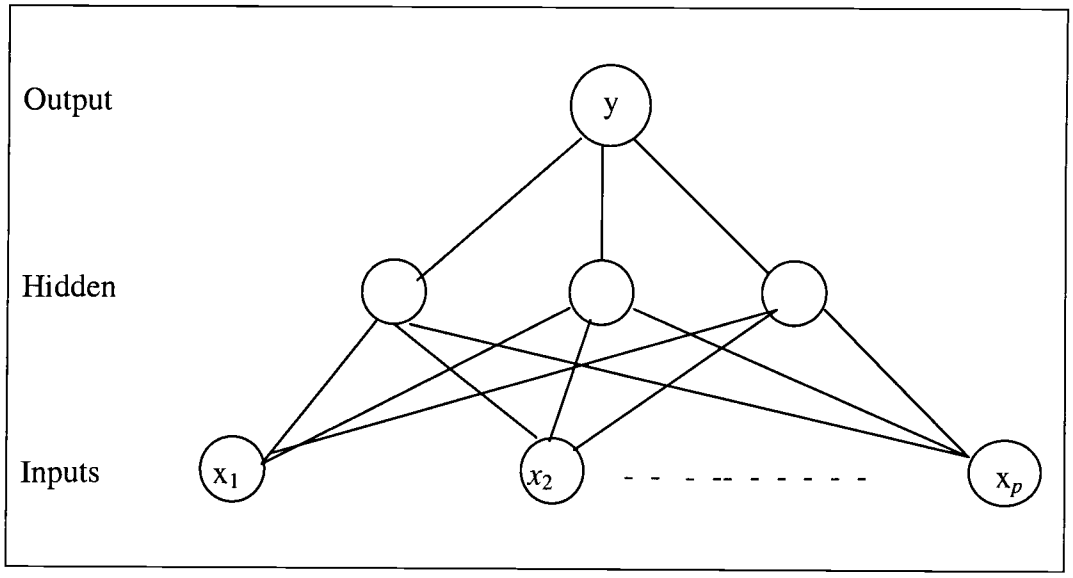
This yield $\beta(k+1) = \beta(k) + \eta\left(\xi(k) x(k) - 2\lambda^* \beta(k)\right)$. This result shows that the parameter estimates are reduced by the amount $2\eta\lambda^*\beta$, at each iteration, compared to the estimate without the ridge penalty which could help reach convergence faster.

The goal of ANN is to construct a network and then estimate the parameters by minimizing the error to get the best prediction for the output while accounting for the interrelationships between the variables in the model. Initial values of network parameters are set randomly. By using an iterative training algorithm, the network parameters are updated until the global minima for the error is obtained. A multilayer perceptron linear ANN model with inputs matrix X=[x_1, x_2,...,x_p] and output vector y and one hidden layer (Figure 5) takes the following form:

$$y = f\left[\sum_{j=1}^{k} \alpha_j y_j + \alpha_o \right] \text{, where } y_j = g\left(\sum_{i=1}^{p} \beta_{ji} x_i + \beta_o \right)$$

This results in $y = f\left[\sum_{j=1}^{k} \alpha_j g\left(\sum_{i=1}^{p} \beta_{ji} x_i + \beta_o \right) + \alpha_o \right]$ (*) where f and g are linear functions.

Figure 5. Multilayer perceptron



Beginning with initial random values of weights and thresholds (NN bias) and using a training algorithm, we seek the global minimum. Assume the simplest case and therefore set $f(u)=u$, $g(v)=v$, using equation (*) we have $y_t=\alpha_o+\sum_{j=1}^{k}\alpha_j\left(\sum_{i=1}^{p}\beta_{ji}x_{ti}+\beta_o\right)$ and the error for observation t takes the form $\xi_t=y_t-\left[\alpha_o+\sum_{j=1}^{k}\alpha_j\left(\sum_{i=1}^{p}\beta_{ji}x_{ti}+\beta_o\right)\right]$.

In order to minimize the performance function, we write the objective function as

$J=\frac{1}{2n}\sum_{t=1}^{n}\left[y_t-\alpha_o-\sum_{j=1}^{k}\alpha_j\left(\sum_{i=1}^{p}\beta_{ji}x_{ti}+\beta_o\right)\right]^2$. Taking the derivatives with respect to the different parameters, we get:

$$\frac{\partial J}{\partial \beta_{ji}} \Rightarrow (kp+1) \text{ equations in } (kp+1) \text{ unknowns}$$

$$\frac{\partial J}{\partial \alpha_{j}} \Rightarrow (k+1) \text{ equations in } (k+1) \text{ unknowns}$$

In order to get parameter estimates, this system of equations needs to be solved by setting the derivatives equal to zero. Let $j = m$ and $i = \text{n}$, we have:

$$\frac{\partial J}{\partial \alpha_o} = \frac{1}{n}\sum_{t=1}^{n}\left[y_t - \alpha_o - \sum_{j=1}^{k}\alpha_j\left(\sum_{i=1}^{p}\beta_{ji}x_{ti} + \beta_o \right)\right]^{set} = 0$$

$$\frac{\partial J}{\partial \alpha_m} = \frac{1}{n}\sum_{t=1}^{n}\left[y_t - \alpha_o - \sum_{j=1}^{k}\alpha_j\left(\sum_{i=1}^{p}\beta_{ji}x_{ti} + \beta_o \right)\right]\left(\sum_{i=1}^{p}\beta_{mi}x_{ti} + \beta_o \right)^{set} = 0$$

$$\frac{\partial J}{\partial \beta_o} = \frac{1}{n}\sum_{t=1}^{n}\left[y_t - \alpha_o - \sum_{j=1}^{k}\alpha_j\left(\sum_{i=1}^{p}\beta_{ji}x_{ti} + \beta_o \right)\right]\left(\sum_{j=1}^{k}\alpha_j \right)^{set} = 0$$

$$\frac{\partial J}{\partial \beta_{mn}} = \frac{1}{n}\sum_{t=1}^{n}\left[y_t - \alpha_o - \sum_{j=1}^{k}\alpha_j\left(\sum_{i=1}^{p}\beta_{ji}x_{ti} + \beta_o \right)\right]\alpha_m x_{tn}^{set} = 0$$

Actually, the parameters are estimated using the Backpropagation (BP) algorithm, applied to the *MSE* performance function by means of the gradient descent algorithm. For the simple case, where $f(u) = u$, $g(v) = v$, the results are similar to a multiple linear regression model. In this case, the estimates of the parameters are:

$$\alpha_o(k+1) = \alpha_o(k) - \eta(k)\frac{\partial J}{\partial \alpha_o(k)} \qquad \beta_o(k+1) = \beta_o(k) - \eta(k)\frac{\partial J}{\partial \beta_o(k)}$$

$$\alpha_m(k+1) = \alpha_m(k) - \eta(k)\frac{\partial J}{\partial \alpha_m(k)} \qquad \beta_{ml}(k+1) = \beta_{ml}(k) - \eta(k)\frac{\partial J}{\partial \beta_{ml}(k)}$$

where η is the network learning rate/step size, and α and β are the network parameters.

This model has two error sources: one at the output layer and the second at the hidden layer. These errors occur in the backward pass of the BP algorithm, where the

errors from the output layer are calculated by comparing the network output to the true value of y. In addition, these errors are propagated to create the hidden layer error. One property of feed-forward ANN is that the error is local, i.e., it relates only to a specific neuron and its connections, which is discussed in more details in chapter 5.

In the ANN data analysis, the data is usually divided into several parts. Some observations are used for training and estimating the network parameters, while others are used to check the performance of the network. A third set of cases, a test set, is used to evaluate the imputation error. The test set data is used to make sure that the results of the selection and training set are representative of the population of interest.

2.4. Results

This section contains simulations and real data results that compare the applied performance of linear ANN in imputation to mean imputation, hot deck and regression. This comparison is performed under several conditions (e.g. linear and several non-linear relationships between the variables of interest) and assumptions about the data distribution of the data (normal, cauchy, and gamma).

2.4.1. *Simulation*

A simulation study was done to compare the performance of imputation using ANN compared to other traditional imputation techniques. Data for this simulation was generated using a complex survey design. The design used is a stratified simple random sample from two strata with equal allocation. For this simulation, each of the two strata (z=1,2) had three variables x_1, x_2 and y. Using Matlab software, the x's were generated with a normal distribution in each stratum. y was generated as a function of all the x's with some random error. The relationship between the x's and y was set to be linear in some cases and non-linear in other cases for the purpose of generating simulated data. The linear model takes the form $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \xi$. While in the non-linear case, three different models were used to generate the data in order to explore examples of different data patterns:

Model 1: $y = \alpha_o + \frac{\beta_3}{\beta_4 + \beta_5 x_1} + \gamma_o e^{\gamma_1 - \gamma_2 x_2} + \xi$ (exponential)

Model 2: $y = \beta_6 x_1^3 - \beta_7 x_2^3 + \beta_8 x_1 x_2 + \xi$ (Cubic)

Model 3: $y = \beta_9 + \beta_{10} x_1 x_2 + \beta_{11} \sin(x_1 - x_2) + \xi$ (Cyclic)

The parameters in the chosen models for data simulation (α, β, and γ) were set arbitrarily and separately for each stratum. However, the random error term, ξ, in the models was simulated from three different distributions (normal, cauchy, and gamma) to study the effect of the data distribution on imputation. The distributions of the error term were chosen to allow normal as well as non-normal skewed data to be investigated. Three sample sizes (small = 50, moderate = 500, large = 1000 in each stratum) were studied to assess the effect of sample size on the results.

Once the complete data set was simulated, a random number was generated to select observations with the missing Y value. The number of missing observations, in each simulation run, represented 10 percent of the sample size in each stratum. In addition, these observations were used for evaluating the performance of the different imputation techniques.

Several measures of accuracy were used for comparing the results from the different imputation techniques. These included mean absolute percentage error (MAPE), mean absolute deviation (MAD), and mean squared deviation (MSD). Let y be the actual value, $\hat{y}$ the imputed value, and m the number of imputed values. The mean absolute percentage error measures the accuracy of imputed values as a percentage:

$$MAPE = \frac{\sum_{t=1}^{m} \left| (y_t - \hat{y}_t) / y_t \right|}{m} \times 100 \quad (y_t \neq 0).$$

The mean absolute deviation (MAD) expresses the accuracy in the same units as the data, which helps conceptualize the amount of error.

$$MAD = \frac{\sum_{t=1}^{m} |y_t - \hat{y}_t|}{m}.$$

The mean squared deviation (MSD) is similar to mean squared error (MSE) and is a commonly used measure of accuracy. Mean squared deviation is computed using the same denominator, m, regardless of the model, in order to compare MSD values across all models. Mean squared error deviations are computed with different degrees of freedom for different models, so MSE values are not easily compared across models.

$$MSD = \frac{\sum_{t=1}^{m} (y_t - \hat{y}_t)^2}{m}.$$

Numerous software packages were utilized in the data analysis to make use of the available imputation techniques (as explained below). Imputation was performed separately in each stratum. SOLAS software was used for mean imputation as well as for hot deck. SAS was used for regression imputation. X_1 and X_2 were kept in the regression model that was used for imputation. Stochastic regression imputation was run using the regression algorithm in the missing values module in SPSS. This algorithm adds a normally distributed residual to the imputed values from regression. The residual used in stochastic regression imputation reflects the uncertainty in the predicted value. Finally, the Matlab Neural Network toolbox was used for ANN imputation where a feed-forward network with back-propagation and two layers (one hidden layer and one output layer) was used for each stratum. For the ANN imputation, online training was used. The network was assigned one observation for each pass. This is different than batch training used in statistical methods where all observations are presented at once during the analysis. Initial parameter values were randomly assigned. Several networks were observed to select and investigate the

effect of increasing the number of nodes in the hidden layer. Increasing the number of nodes in the hidden layer beyond three did not reduce the network error. To improve consistency and comparability, the hidden layer was set to three nodes for all networks. With one response variable, *y*, the output layer consisted of only one node.

Tables 1a, 1b, and 1c show the imputation results. For evaluation purposes, the performance of the imputation techniques was judged based on the agreement of at least two out of the three evaluation measures (MAPE, MAD, and MSD). The lowest error detected with each of the accuracy methods was marked with a * within each of the sample sizes for each distribution.

Hot deck imputation under the linear model resulted in lower error with the gamma distribution and a moderate sample size as compared to small and large sample sizes under gamma, and all sample sizes at normal and Cauchy distributions (Table 1a.). However under non-linear model 1, hot deck gave lower error as compared with both normal and gamma distributions in the case of moderate sample size. For non-linear models 2 and 3, hot deck performed well under gamma distribution with small sample size. Overall, hot deck imputation, which is the most commonly used imputation technique, had the best results in only 2% of all the imputation cases. Hot deck appeared to yield better results with a moderate sample size and with data generated from non-linear models.

Table 1. Linear imputation simulation

a. Linear model

MAPE

	Normal			Cauchy			Gamma		
n	***50***	***500***	***1000***	***50***	***500***	***1000***	***50***	***500***	***1000***
Hot Deck	37.6	54.1	66.6	63.9	44.9	59	82.9	25.9	28.7
Mean	19.6	36	53.1	26	36.4	51.7	66.6	31.7	32.2
Regression [a]	0.08	1^{E-2}	8^{E-2}	12.8	28.9^{*}	18.7^{*}	3.2^{*}	1.9^{*}	1.7^{*}
Stochastic Regression	9.9	8.7	16.1	12.7^{*}	209.6	31.4	17.7	9.6	8.9
Linear ANN	0.004^{*}	6^{E-4*}	5^{E-4*}	11.8	36.2	18.7^{*}	21.9	36	1.7^{*}

MAD

	Normal			Cauchy			Gamma		
n	***50***	***500***	***1000***	***50***	***500***	***1000***	***50***	***500***	***1000***
Hot Deck	15.4	14.5	14.5	21.9	14.9	34	16.9	10.8	12.3
Mean	8.3	9.6	9.8	10.6	10.8	28.4	13.9	10.6	10.7
Regression [a]	0.04	5^{E-3}	3^{E-2}	4.9	8.7^{*}	21.8	0.9^{*}	0.7^{*}	0.6^{*}
Stochastic Regression	3.3	2.7	3.5	4.7^{*}	93.4	24.4	3.9	3.3	3.3
Linear ANN	0.002^{*}	2^{E-4*}	2^{E-4*}	4.8	13.1	21.7^{*}	5.6	12.8	0.6^{*}

MSD

	Normal			Cauchy			Gamma		
n	***50***	***500***	***1000***	***50***	***500***	***1000***	***50***	***500***	***1000***
Hot Deck	427.9	335.5	338.2	985	380.8	7075.2	431.3	228.2	273.4
Mean	142.4	146.9	159	157.1	202.2	6215.8^{*}	258.8	171.1	204.6
Regression [a]	0.002	4^{E-6}	1^{E-3}	45.8	146.9^{*}	6308.2	1.2^{*}	0.8^{*}	0.7^{*}
Stochastic Regression	18.4	11.2	19.6	40.8	6.09^{E+5}	6373.6	21.2	16.7	17.7
Linear ANN	0^{*}	2.2^{E-7*}	0^{*}	38.8^{*}	285.9	6297.4	45.3	286.6	0.7^{*}

a. The R^2 from the regression models ranged between 0.41 and 1 except for the Cauchy distribution with n=500

Table 1. Simulation results (Continued)

b. Non-linear model 1

MAPE

	Normal			Cauchy			Gamma		
n	*50*	*500*	*1000*	*50*	*500*	*1000*	*50*	*500*	*1000*
Hot Deck	47.9	110.7*	71.3	55.7	372.5	94.3	222.4	10.5	17.5
Mean	39.4	48465.7	53.2*	29.9	52	96	19.3	15	12.8
Regression [a]	39.3	1.94E+5	55.1	28.4*	51.5*	60*	16.8*	693.4	12.3*
Stochastic Regression	454.1	57145.7	281.4	267.7	255.6	1492.6	105.5	116	122.7
Linear ANN	39.1*	592.6	54.5	49.3	54.0	63.4	17	12.3*	12.4

MAD

	Normal			Cauchy			Gamma		
n	*50*	*500*	*1000*	*50*	*500*	*1000*	*50*	*500*	*1000*
Hot Deck	1.1	74.5*	1.2	5.3	61.1	17.1*	12.1	0.7*	1.1
Mean	0.8	446.6	0.9*	1.3	54.6*	17.3	1*	0.8	0.9
Regression [a]	0.8	1445.2	0.9*	1.2*	54.6*	17.1*	1*	33	0.8*
Stochastic Regression	8.9	1218.8	6.4	6.4	61.1	43.7	4.5	7.1	6.9
Linear ANN	0.7*	80.77	0.9*	1.6	54.7	17.2	1*	0.7*	0.8*

MSD

	Normal			Cauchy			Gamma		
n	*50*	*500*	*1000*	*50*	*500*	*1000*	*50*	*500*	*1000*
Hot Deck	2	5.2E+5*	2.3	50.1	1.14E+5	7401.8	149.8	1*	2.1
Mean	0.8	7.1E+5	1.3	2.6	1.12E+5*	7392*	1.5	1.3	1.3
Regression [a]	0.9	7.5E+6	1.2*	2.4*	1.12E+5*	7423.6	1.5	1175.6	1.1*
Stochastic Regression	100.6	2.5E+6	62.6	65.7	1.14E+5	8330	32.7	78.6	70.1
Linear ANN	0.7*	5.3E+5	1.3	4	1.12E+5*	7.4E+5	1.3*	1*	1.1*

a. The R^2 from the regression models ranged between 0.04 and 0.98

Table 1. Simulation results (Continued)

c. Non-linear model 2

MAPE

	Normal			Cauchy			Gamma		
n	***50***	***500***	***1000***	***50***	***500***	***1000***	***50***	***500***	***1000***
Hot Deck	8.1^{E+3}	6.3^{E+3}	1.7^{E+4}	9.1^{E+3}	2911.6	9.8^{E+3*}	9.9^{E+3}	4.7^{E+3}	2.5^{E+3*}
Mean	6.3^{E+3}	5^{E+3}	2.1^{E+4}	1.3^{E+4}	1620.5	1.6^{E+4}	1^{E+4}	5.6^{E+3}	4.9^{E+3}
Regression [a]	6^{E+3}	5^{E+3}	2.1^{E+4}	1^{E+4}	1566.3^{*}	1.3^{E+4}	1.2^{E+4}	3.7^{E+3*}	4.3^{E+3}
Stochastic Regression	6.1^{E+3}	6.2^{E+3}	2^{E+4*}	9.3^{E+3}	1895.5	2.7^{E+4}	3.7^{E+3*}	5.1^{E+3}	6^{E+3}
Linear ANN	2.2^{E+3*}	2.2^{E+3*}	2.1^{E+4}	3^{E+3*}	1691.2	1.2^{E+4}	4.2^{E+3}	1.6^{E+4}	4.2^{E+3}

MAD

	Normal			Cauchy			Gamma		
n	***50***	***500***	***1000***	***50***	***500***	***1000***	***50***	***500***	***1000***
Hot Deck	3.7^{E+5}	3.9^{E+4}	4.7^{E+4}	3.7^{E+4}	3.9^{E+5}	3.4^{E+4}	2.2^{E+4}	3.8^{E+4}	4^{E+4}
Mean	2.6^{E+4}	3.6^{E+4}	3.5^{E+4}	2.8^{E+4}	2.7^{E+4}	2.6^{E+4}	2.3^{E+4}	3.4^{E+4}	3.1^{E+4}
Regression [a]	3.7^{E+4}	2.1^{E+4*}	1.5^{E+4}	1^{E+4*}	1.3^{E+4*}	1.3^{E+4*}	1.5^{E+4}	1.5^{E+4*}	1.6^{E+4*}
Stochastic Regression	2.2^{E+4}	3.1^{E+4}	2.5^{E+4}	1.5^{E+4}	2.2^{E+4}	3.8^{E+4}	1^{E+4*}	2.7^{E+4}	2.6^{E+4}
Linear ANN	1.7^{E+4*}	4.2^{E+4}	1.5^{E+4*}	3^{E+4}	3.7^{E+4}	1.3^{E+4*}	4.2^{E+4}	7^{E+4}	1.6^{E+4*}

MSD

	Normal			Cauchy			Gamma		
n	***50***	***500***	***1000***	***50***	***500***	***1000***	***50***	***500***	***1000***
Hot Deck	1.5^{E+9}	4^{E+9}	5.8^{E+9}	1.6^{E+9}	4^{E+9}	3.6^{E+9}	7.8^{E+8}	4^{E+9}	5^{E+9}
Mean	1.1^{E+9}	3^{E+9}	3.1^{E+9}	1.2^{E+9}	1.4^{E+9}	1.3^{E+9}	7.6^{E+8}	3.3^{E+9}	3^{E+9}
Regression [a]	2.5^{E+9}	1^{E+9*}	6.8^{E+8*}	2.3^{E+8*}	3.9^{E+8*}	4.4^{E+8*}	5.3^{E+8}	8.9^{E+8*}	7^{E+8*}
Stochastic Regression	8.7^{E+8}	2.2^{E+9}	1.4^{E+9}	3.8^{E+9}	1^{E+9}	2.9^{E+9}	2.7^{E+9}	1.8^{E+9}	2^{E+9}
Linear ANN	4.6^{E+8*}	4.2^{E+9}	7.1^{E+8}	1.8^{E+9}	2.9^{E+9}	4.4^{E+8*}	2.6^{E+9*}	9.4^{E+9}	7^{E+8*}

a. The R^2 from the regression models ranged between 0.75 and 0.87

Table 1. Simulation results (Continued)

d. Non-linear model 3

MAPE

	Normal			Cauchy			Gamma		
n	***50***	***500***	***1000***	***50***	***500***	***1000***	***50***	***500***	***1000***
Hot Deck	189.2	179.1	639.9	258.8	329.7	832.8	122	728	328.1
Mean	176.2	190.3	1377.3	136.8	353.7	536.3	80.2	417.9	229.5
Regression [a]	54.3*	37.6	690.7	17.5	779	272.3*	21.2*	219.4	140.8*
Stochastic Regression	68.8	91.2	485.9*	53.4	158.1*	328.4	112.4	224.5	170.3
Linear ANN	63.4	36.6*	696.1	17.1*	174.9	351.4	19.9	218.8*	144.5

MAD

	Normal			Cauchy			Gamma		
n	***50***	***500***	***1000***	***50***	***500***	***1000***	***50***	***500***	***1000***
Hot Deck	111.7	86.8	90.2	112.7	93.4	210.9	69.1	196.5	88.9
Mean	78.9	67.6	72.2	78.9	84.2	70.7	51.6	86.8	69.3
Regression [a]	18*	13.9*	21.1	14.9*	425.9	22.4*	14.6	24.6*	21.6*
Stochastic Regression	39.2	28.1	44.3	39.2	39.7	39.8	84.9	45.7	40.4
Linear ANN	22.3	14	21*	17.2	23.2*	63.7	14*	26.9	21.8

MSD

	Normal			Cauchy			Gamma		
n	***50***	***500***	***1000***	***50***	***500***	***1000***	***50***	***500***	***1000***
Hot Deck	1.3^{E+4}	1.1^{E+4}	1.3^{E+4}	1.7^{E+4}	1.5^{E+4}	6.8^{E+4}	6.8^{E+3}	5.4^{E+4}	1.1^{E+4}
Mean	7.7^{E+4}	7.2^{E+4}	8.8^{E+3}	8^{E+3}	1.1^{E+5}	8.1^{E+3}	3.8^{E+3}	1.2^{E+4}	7.5^{E+3}
Regression [a]	6.9^{E+2*}	4.8^{E+2*}	1.2^{E+3}	5.4^{E+3*}	2.7^{E+5}	1.1^{E+3*}	1.3^{E+2*}	1.8^{E+3*}	1.2^{E+3*}
Stochastic Regression	2.2^{E+5}	1.67^{E+3}	3.8^{E+2}	2.3^{E+3}	3.4^{E+3}	2.9^{E+3}	1.1^{E+4}	4.2^{E+3}	3.5^{E+3}
Linear ANN	9^{E+2}	4.9^{E+2}	1.2^{E+3*}	5.9^{E+2}	1.7^{E+2*}	9.2^{E+3}	3^{E+2}	2^{E+3}	1.2^{E+3}

a. The R^2 from the regression models ranged between 0.46 and 0.89

Results from mean imputation were not consistent in most of the cases. The evaluation error was not stable except in the case of the linear model with normal distribution and small sample size, and non-linear model 1 with gamma distribution for all sample sizes. Mean imputation resulted in a larger error than hot deck imputation in most of the sample sizes and distribution cases, and therefore is not recommended for imputation.

Regression imputation resulted in the most consistent error and accuracy measures under all models with different distributions and sample sizes. The R^2 (percent of variation explained by the regression model) of the regression models were checked to assess the fit of the models. Regression imputation offered a large percentage of the smallest imputation error compared to other imputation techniques. The only exception to this performance was under non-linear model 2, with normally distributed data where results were not as satisfactory.

Stochastic regression did not provide small error. The results from the stochastic regression imputation were even worse than mean imputation and not very consistent across measures in most of the cases. Under the linear model, it gave relatively good results with small to moderate sample size and normal distribution, and moderate to large sample size with gamma distribution. In case of non-linear models 1 and 2, results improved slightly under gamma distribution with small sample sizes. However in non-linear model 3, it improved slightly under the normally distributed data with large sample size. A key reason that could explain the failure of stochastic regression to provide good results is that the software does not allow the flexibility n determining the regression model used in stochastic regression.

Finally, linear ANN yielded consistent results similar to linear regression. Under the linear model, the ANN provided the smallest measures of accuracy in all cases except for Cauchy distribution with moderate sample size. In non-linear model 1, ANN gave best performance under gamma distribution with all sample sizes. In

non-linear model 2, normally distributed data helped ANN yield the best results. In non-linear model 3, ANN performed well in moderate to small sample sizes.

Overall, the simulation results indicate that both simple linear ANN and regression provided the smallest measures of accuracy compared to the other imputation techniques tested in this example. It has been determined that linear perceptrons are equivalent to simple linear regression under certain conditions (Principe, Euliano. and Lefebvre, 2000). The results show that ANN gave better results than linear regression in some of the simulation cases, which may be due to the adaptive nature of ANN. The best results with ANN were achieved under the normality condition in the case of linear model similar to linear regression.

We were expecting that with larger sample sizes the imputation techniques used would be able to adapt to the data and present better results. Contrary to what was expected, increasing the sample size did not always improve the performance of the different imputation techniques. The distribution of the error affected the overall performance of the imputation techniques. Although normally distributed data were expected to offer the best results, gamma distribution offered most of the consistent results especially for regression imputation. Cauchy distribution was the most difficult to impute, especially under moderate sample size.

Both ANN and regression imputation results were robust to model misspecification. While assuming the linearity of the relationship, ANN and regression imputation performed well even when the relationships between the variables were originally non-linear.

2.4.2. *Application*

To illustrate the imputation technique, a public-use data set from the National Health Interview Survey (NHIS) was used. The National Health Interview Survey is a health survey conducted by the National Center for Health Statistics (NCHS), Centers for Disease Control (CDC), and is the principal source of information on the health of

the civilian, non-institutionalized population in the United States (National Center for Health Statistics, 2002). The NHIS is a cross-sectional household, interview survey. The data investigated is from the last quarter of 2001. The NHIS data is obtained through a complex sample design involving stratification, clustering, and multistage sampling designed to represent the civilian, non-institutionalized population of the United States.

Researchers found that the prevalence of obesity has increased rapidly in the United States. This increase in obesity has a major impact on type II diabetes and other chronic diseases. Body mass index (BMI =weight/height^2) was used as the response variable for this analysis as an indicator of obesity. The other variables in the data were gender, age, income, family size and respondents' self-conscience perceptions (i.e. desired weight, frequency of happiness, sadness, etc).

The data was filtered to generate a set of 7543 records with no missing data in any of the 27 variables studied. Data was assumed to be a simple random sample for simplicity, further details on sampling design features will be discussed in Chapter 4. Missing data, representing 10 percent of the records, were generated at random. Imputation was performed using the imputation techniques discussed in Section 2.3.1. The results of the imputation are shown in Table 2.

Table 2. Linear imputation application

	MAPE	*MAD*	*MSD*
Hot deck	20.6	5.8	57.8
Mean	16.2	4.3	31.1
Regression [a]	5.7	1.6*	5.4*
Stochastic Regression	9.4	2.6	17.2
Linear ANN	5.6*	1.6*	5.9

a. The R^2 from the regression model was 0.82

The regression model in this analysis resulted in an R-squared of 0.823. As a consequence, both ANN and regression had nearly the same values for accuracy measures for the imputation evaluation. Linear ANN and linear regression provide the lowest values of each of the three accuracy measures. Stochastic regression – where

random error was added to the regression imputation- resulted in high error but was smaller than mean imputation. Hot deck provided the largest measures of accuracy as compared to other imputation tools investigated for these data.

2.5. Conclusion

Imputation procedures should be theoretically proven, appropriate for the analysis, consistent, and make use of all available data. Under neural networks imputation, the model is more exploratory and non-parametric. Whereas with traditional techniques, the model is parametric and linear. The automation level in ANN distinguish them from traditional techniques. As discussed in Section 2.3.2, neural networks have less constraining hypotheses than other statistical methods used for imputation. The advantages of ANN include flexibility, simplicity, and non-parametric nature. From a conceptual viewpoint, ANN have the ability to account for any functional dependency, handle complex data structures, are insensitive to moderate noise and multicollinearity, and are easy to handle with no conditions on the predicted values.

Some imputation techniques, such as regression, use auxiliary variables. First, auxiliary variables are evaluated to determine if they are correlated to the missingness pattern. Second, it is possible to include variables that are simply correlated with the variables that have missing values, whether or not they are related to the mechanism of missingness (Collins, Schafer and Kam, 2001). With increasing the number of variables in the data, and by including categorical and numerical variables, regression requires some effort in determining a model and assumption checking. The artificial neural network offers a simple alternative that does not require parametric assumptions. In other words, when the regression models fail, ANN offer an alternative analytical method providing similar results.

Hot deck requires matching cases and was found to be time-consuming when used with large datasets. The artificial neural network can produce more accurate

imputed values than hot deck as shown by Wilmot and Shivananjappa (2001) in their analysis. In spite of the effort required for hot deck imputation, and the improved accuracy using ANN in imputation, the hot deck imputation method was found to be superior to ANN in maintaining the distribution of the data (Curdas and Chambers, 1997). Mean imputation is not recommended because it does not conserve the distribution of the data. Furthermore, mean imputation does not generally provide valid estimates (Little and Rubin, 2002).

Single imputation techniques do not account for the imputation uncertainty, which results in an underestimation of standard error. For example, single imputation does not reflect the sampling variability under a specific nonresponse model or the uncertainty about the correct model for nonresponse. Artificial neural network depends on an adaptive iterative process that could account for imputation uncertainty. Little and Rubin (2002) advise that imputations should, in general, be conditional on observed variables to reduce bias, improve precision, and preserve relationship between variables. The artificial neural network appears to meet these requirements and provides similar accuracy compared to other imputation techniques.

Artificial neural networks avoid the predetermination of a specific model for the imputation process. In this chapter, we imposed a linearity constraint in the neural network. No assumptions concerning the data distribution are needed for ANN, as they are for linear regression. Considering the flexibility of ANN, these results support the idea that ANN should be considered as an imputation method.

2.6. References

Anderson, J.A., Pellionisz, A. and Rosenfeld, E. (Eds.). (1990). Neurocomputing 2: Directions for Research. Cambridge, Massachusetts: MIT Press.

Anderson, D. and McNeill, G. (1992). Artificial Neural Network Technology, A DACS State-of-the-Art Report. Kaman Sciences Corporation.

Bishop, C.M. (1995). Neural Networks for Pattern Recognition. Oxford University Press.

Burton, R.M. and Dehling H.G. (1997). Mathematical Aspects of Neural Computing, Oregon State University.

Cheng B. and Titterington D.M. (1994). Neural Networks: A Review From a Statistical Perspective.

Cherkassky, V. and Mulier, F. (1994). Statistical and Neural Network Techniques for Non-Parametric Regression. In: Cheeseman, P. and Oldford, R.W., Selecting Models From Data: Artificial Intelligence and Statistics IV. New York: Springer.

Clarck, Alex (1996). Planning for the 2001 Census of the United Kingdom. (http://www.census.gov/prod/2/gen/96arc/xastreet.pdf)

Collins, L.M., Schafer, J. L. and Kam, C-M. (2001) A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures, Psychological Methods, vol. 6 (4), 330-351.

Curdas, Marie and Chambers, Ray (1997). Neural Network Imputation: Statistical Evaluation. Conference of European Statisticians.

Dempster, A.P. and Rubin, D.B. (1983). Incomplete Data in Sample Surveys: Theory and Bibliography. New York: Academic Press.

Dillman, D.A., Eltinge, J.L., Groves, R.M. and Little, R.J.A. (2002). Survey Nonresponse in Design, Data Collection, and Analysis. In Survey Nonresponse by Groves, R.M., Dillman, D.A., Eltinge, J.L. and Little, R.J.A. (Eds.). New York: John Wiley and Sons.

Geman, S., Bienenstock, E. and Doursat, R. (1992). Neural Networks and the Bias/Variance Dilemma. Neural computation, 4, 1-8.

Hand, D.J. (1984). Statistical Expert Systems: Design. The Statistician, 33, 351-369.

Hastie, T., Tibshirani, R. and Friedman, J (2001). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer.

Jacobs, R.A., Jordan, M.I., Nowlan, S.J. and Hinton. G.E. (1991). Adaptive Mixtures of Local Experts, Neural Computation, 3(1), 79-87.

Lessler, J.T. and Kalsbeek, W.D. (1992). Nonsampling Error in Surveys. John Wiley and Sons.

Little, R.J.A. and Rubin, D.B. (2002). Statistical Analysis with Missing Data. New Jersey: John Wiley and Sons.

National Center for Health Statistics (2002). Data file Documentation, National Health Interview Survey, 2001 (machine readable file and documentation). National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, Maryland.

Nordbotten S. (1963). Automatic Editing of Individual Statistical Observations Statistical Standards and Studies, Handbook No. 2. New York.

Nordbotten, S. (1995). Editing and Imputation by Means of Neural Networks. Statistical Journal of the UN/ECE,12.

Nordbotten S.(1996). Neural Networks Imputation Applied to the Norwegian 1990 Census Data. Journal of Official Statistics, 12(4), 38-401.

Nordbotten, S. (1996). Editing Statistical Records by Neural Networks. Journal of official Statistics, 11(4), 391-411.

McCulloch, W.S. and Pitts, W. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. Bulletin of Mathematical Biophysics, 5, 115-133.

Meng, XL. (1994). Multiple Imputation Inferences with Uncongenial Sources of Input. Statistical Science, 10, 538-573.

Michie, D.J., Siegelhalter, D.J. and Taylor, C.C. (1994). Machine Learning, Neural and Statistical Classification. New York: Ellis Horwood.

Murtagh, F. (1994). Neural Networks and Related Massively Parallel Methods for Statistics: a Short Overview. International Statistical Review, 62, 275-288.

Principe, J.C., Euliano, N.R. and Lefebvre, W.C. (2000). Neural and Adaptive Systems: Fundamentals Through Simulation. New York: John Wiley and Sons.

Ripley, B.D. (1993). Statistical Aspects of Neural Networks. Networks and Chaos: Statistical and probabilistic Aspects (U. Bornndor-Nielsen, J Jensen, and W. Kendal, (Eds.). Chapman and Hall.

Ripley, B.D. (1994). Neural Networks and Related Methods for Classification, Journal of Royal Statistical Society B, 6(3), 409-456.

Rosenbaltt, F. (1962). Principles of Neurodynamics. Spartan Books, New York.

Rumelhart, D.E., McClelland, J.L. (1986). Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vols. 1 and 2. Cambridge: MIT Press.

White, H. (1992). Artificial Neural Networks: Approximation and Learning Theory. UK: Oxford and USA: Cambridge.

White, H. (1989). Learning in Artificial Neural Networks: A Statistical Perspective. Neural Computation, 1(4), 425-464.

Wilmot, C.G. and Shivananjappa, S. (2001). Comparison of Hot Deck and Neural-Network Imputation.

Woodruff, Stephen M. (1988). Estimation in the Presence of Non-Ignorable Missing Data and a Markov Super-population Model. Washington: D.C. Bureau of Labor Statistics. (http://www.amstat.org/sections/srms/Proceedings/papers/1988_114.pdf).

3. NON-LINEAR NEURAL NETWORK

3.1. Abstract

Regression imputation is a common technique used to estimate missing values. Non-linear models offer a flexible realistic way of imputing missing data. The theory behind non-linear models and techniques, including neural networks, in missing data imputation are investigated. An evaluation of the performance of a class of feed-forward non-linear neural networks in imputation is presented. The evaluation was performed by comparing non-linear neural networks and several statistical imputation methods, including the Expectation Maximization (EM) algorithm and multiple imputation using Markov chain Monte Carlo (MCMC) and propensity scores. The effect on the distribution of the missing data pattern is explored. Results suggest that artificial neural networks (ANN) are similar to multiple imputation using MCMC, and outperform the EM algorithm and propensity score imputations.

Keywords: missing data, imputation, neural networks, non-linear models

3.2. Introduction

Linear modeling has been a common technique for dealing with missing data problems due to its well-known optimization strategies such as the maximum likelihood estimation procedure (Schafer and Schenker, 2000). The objective of imputation by modeling is to obtain a useful approximation of the true relationship between the variable to be imputed and the other variables in the dataset. However, when a linear approximation is not valid, the model breaks down and the imputation is unreliable (Meng, 1994). Most relationships between variables in datasets are unlikely to be linear. This is the case in the physical, chemical, biological, social, and engineering sciences, where relationships between variables in the data are described in terms of non-linear equations (Gallant, 1975).

Many statistical methods depend on assumptions made about the distribution of the population or a particular mathematical model. Due to the increase in the power and speed of computers, it becomes more feasible to consider a wider class of linear and non-linear models in missing data imputation. In the majority of cases, the non-linearity in the relationship derives from highly irregular behavior among the variables. The non-linearity of the relationships between the variables in the data set could be described by existing parameterized classes of non-linear models. Determining the most appropriate non-linear model can be difficult particularly with large data sets and a broad range of explanatory variables. This leads to the need to investigate self-modeling techniques to avoid model misspecification. Spline regression functions are methods of allowing the data to determine the model (Lawton, Sylvester and Maggio, 1972; Treiman, Bielby and Cheng, 1993). Similar to spline functions, artificial neural networks (ANN) provide another type of non-linear modeling, where no pre-determined relationship is imposed upon the data (Amari, 1993).

Artificial neural network is an iterative system modeled on the structure of the human brain. The most basic element of the human brain is a specific type of cell known as a neuron. The power of the brain comes from its large number of neurons and the connections between them. Basically, a neuron receives input from other sources, combines them in some fashion, performs a generally non-linear operation, and then passes on a final result (Lewicki, Hill and Czyzewska, 1992). Artificial neurons are simpler than biological neurons. Artificial neurons are arranged and connected together to form a network. This network and its architecture represent a data pattern in the population. Thus, an artificial network is comparable to a model in the traditional modeling approach. With a sufficient number of artificial neurons, ANN can model a high order of complexity within a population. In addition, minimal a priori information about the nature of the population distribution is needed to design the network. Artificial neural networks learn by example and are supported by statistical and optimization theories (Hastie, Tibshirani and Friedman, 2001). Artificial

neural network have arisen recently as an alternative methodology for missing data imputation.

Imputation techniques vary depending on the assumptions related to the variable with the missing data (Y) as well as other predictors (X), and both the missingness mechanism and patterns. Chapter 2 discussed solutions and alternatives when linearity of the relationships was assumed. This chapter focuses on non-linear relationships. Several non-linear imputation models are considered.

3.3. Missingness mechanisms and patterns

The impact of missing data on the final analysis results depends on the mechanism that causes the missingness and the way that the missing data is handled during the analysis. The missing data could be missing completely at random (MCAR), missing at random (MAR), or non-ignorable (NI) missing data. Missing completely at random implies that the mechanism behind the missing data is not related to the values of any variable. It also assumes that the respondents are representative of the sample and that complete case analysis could be used to analyze the data. Missing at random is assumed if the missing data is unrelated to the missing values of Y. The nonresponse depends only on the observed values of the explanatory variables. Non-ignorable missing data implies that the missing data mechanism is related to the missing values of the variable to be imputed, Y, as well as other variables, X covariates. Imputation techniques are used to adjust for the nonresponse for both MAR and NI. In case of NI missing data a model is required for the missing data imputation (Little and Rubin, 2002).

The choice of analytical method to account for nonresponse depends on both the missing data mechanism and the pattern of missingness. The missing data mechanism describes the relationship between the missingness and the values of the variables in the dataset. The pattern of missingness examines which values are missing. In general, statistical literature has focused on general missing patterns

without specific concern to the distribution of the missing data. The processes that caused missing data were considered explicitly by some researchers during missing data analysis (Cochran, 1963; Trawski and Bargmann, 1964; Hocking and Smith, 1972, Wachter and Trussell, 1982).

In MCAR and MAR, the mechanism causing missing values is random. Nevertheless, the distribution of the missing values may have an effect on the imputation results particularly with MAR. Researchers have often ignored this factor. The imputed missing values are commonly predicted from patterns in the non-missing data. This procedure does not consider the possibility of differences between the distribution of the missing values and the non-missing values. Specifically, the distribution of the missing data may or may not be the same as the distribution of the non-missing data. This chapter includes a simulation study to investigate the impact of assuming various distributions for the missing values. In order to account for the missing data pattern, we need to investigate the process that resulted in missing data. In general, the process causing the data to be missing is not obvious in real datasets and is case dependent. This makes the procedure of accounting for missing data a difficult process.

3.4. Imputation

Imputation is the procedure of editing an incomplete data set by filling in the missing values. In order to predict missing values, imputation relies on the predictive distribution estimated using the observed data. The most widely-used conditional mean imputation method uses least squares regression, but it can often be unsatisfactory for non-linear data and be biased if misspecification of the model occurs. Non-parametric approaches using the expectation maximization algorithm (EM) and Markov Chain Monte Carlo (MCMC) are used as alternative methods. In addition, other parametric approaches such as logistic regression or non-linear regression imputation have been used to predict missing data.

Imputation can be performed as single imputation, or repeated several times resulting in multiple imputations (Fellegi and Holt, 1976). One drawback to single imputation is the unaccounted uncertainty attributed to the imputation from the filled-in data. Multiple imputation (MI), as proposed by Rubin (1977), replaces the missing value by a vector of imputed values to obtain a number of complete data sets. Regular analysis run on these data sets yield estimates that are subsequently combined to get the final results (Little and Rubin, 2002).

The EM algorithm is used in this chapter as an example of single imputation technique (Schafer, 1997). The EM is a likelihood-based approach for handling missing data. This procedure is iterative and uses a two-step mechanism (Dempster, Laird and Rubin, 1977). First, in the expectation step (E-step), the expected value of the complete data log likelihood is calculated. Second, in the maximization step (M-step), the expected values for the missing data obtained from the E-step are used and the likelihood function is maximized, as if no data were missing, to obtain the new parameter estimates. The new parameter estimates are substituted back into the E-step and a new M-step is performed. This procedure iterates through these two steps until convergence is obtained. Convergence occurs when the change of the parametric estimates from iteration to iteration becomes negligible. The main advantages of the EM algorithm are its generality, stability and ease of implementation. Two drawbacks of the EM algorithm are a typically slow rate of convergence and the lack of providing a measure of precision for the estimators.

Many single imputation techniques can be repeated several times resulting in MI. Multiple imputation offers the advantage of calculating the variance of the imputed values and therefore estimating the total variance due to both within and between imputation variances (Rubin, 1976a, 1978b, 1996; Little and Rubin, 2002). However, multiple imputation is difficult to implement in large data sets, due to the amount of computer memory needed to store the different, multiply-imputed data sets and the time required to run the analysis. Allison (1999), and Horton and Lipsitz

(2001) offer a review of MI. Schafer (1997) offers an extended review of techniques used for MI. In this chapter, data augmentation techniques such as MCMC and propensity score will be used as examples of these techniques.

Data augmentation algorithms are often used for parameter estimation as well as imputation (Tanner and Wong, 1987; Schafer, 1997). The MCMC procedures are a collection of methods for simulating independent random draws of the missing data from the joint distribution of $(Y_{mis}, \theta | Y_{obs})$ where Y_{mis} are the missing values of Y, Y_{obs} are the observed values of Y, and θ is the distribution parameter. This conditional distribution is assumed to be a multivariate normal distribution (Geman and Geman, 1984; Ripley, 1977). These random draws of the missing data result in a sequence of values that form a Markov chain, which are used for imputation (Gelman, Carlin, Stern and Rubin, 1995; Geyer, 1992; Smith and Roberts, 1992).

A Markov chain is a sequence of random variables where the distribution of each element depends only on the value of the previous one and the iterative procedure consists of two steps. The first step is an imputation step (I-step), which is a draw Y_{mis} from the conditional predictive distribution $P(Y_{mis} | Y_{obs}, \theta)$ given a value for θ. The second step is a posterior step (P-step), given Y_{mis}, draw θ from its complete data posterior $P(\theta | Y_{obs}, Y_{mis})$. The goal of MCMC procedure is to sample values from a convergent Markov chain in which the limiting distribution is the joint posterior of the quantities of interest (Schimert, Schafer, Hesterberg, Fraley and Clarkson, 2001). In practice, the major challenge in using MCMC is the difficulty for the user to assess convergence, as described above (Gelman and Rubin, 1992).

Multiple imputation using propensity scores is another method that will be used in comparison to ANN. This method applies an implicit model approach based on propensity scores and approximate Bayesian bootstrap to generate the imputations (Rubin 1985a, Little 1986). A propensity score is the estimated probability that a particular element of data is missing. In order to calculate a propensity score for

variables with missing values within each stratum, a logistic regression is used to model the missingness. Based on the logistic regression, the propensity that a subject would have a missing value is calculated. Subjects are grouped based on quintiles of the propensity score. Then within each quintile, a posterior predictive distribution of observed data is created by taking random samples equal to the number of observed values. Finally, a value is randomly sampled from the posterior predictive distribution to impute each missing value. The MI represents independent repetitions of the imputation from a posterior predictive distribution for the missing data given the observed data.

In non-linear regression imputation, the problem of model specification for an imputation procedure represents a major challenge. The artificial neural network, an alternative to conventional non-linear modeling, does not make any assumption about the model or distribution of the data. Therefore, both ANN and non-parametric methods based on algorithms do not require a model, which is advantageous in large data set imputation. Advances in computer software and increased memory have made the use of both MI and ANN more practical.

3.5. Non-linear models

Pioneering work in non-linear modeling started in the 1920s by Fisher, who showed that non-linear models could result in better predictions of the response value in some instances when the data relationship is non-linear (Fisher, 1922; Fisher and Mackenzie, 1923). Further major development of non-linear models occurred with increasing availability of computing power in the 1960s and 1970s. At this time, Box and other researchers (Box and Lucas, 1959; Box, 1960; Box and Hunter, 1962, 1965; Box and Hill, 1967, 1974; Box, Hunter and Hunter, 1978) published a series of papers presenting an intensive investigation into the properties and use of non-linear models. The researcher is responsible for considering various non-linear models reflecting the observed relationships between dependent and independent variables.

3.5.1. *Non-linear estimation procedures*

In general, all regression models may be stated in the form:

$y = f(x_1, x_2, \ldots, x_n; \theta) + \xi$, where x_1, ..., x_n are the explanatory variables, y the dependent variable, θ a vector of parameters, f a linear or non-linear function in θ and ξ the model error (Draper and Smith,1981) . The function $f(x_1, x_2, \ldots, x_n; \theta)$ specifies a relationship between the y and the x's. The function $f(x_1, x_2, \ldots, x_n; \theta)$ could be either linear or non-linear. However, non-linear estimation allows for the specification of any type of continuous or discontinuous regression model. Two of the most common non-linear models are logit and exponential growth models. The researcher can also define any type of regression equation to fit the data under investigation. These models can only be estimated using non-linear estimation procedures (e.g. non-linear least squares).

Consider the general model formula of the form $y_i = f(x_i, \theta) + \varepsilon_i$, for i = 1, 2, ..., n where θ is a vector of parameters and f is a non-linear function in θ. The least squares estimator in this case involves the minimization of the sum of squares of residual (SS_{Res}) given by $SS_{Res} = \sum_{i=1}^{n} \left[y_i - f(x_i, \hat{\theta}) \right]^2$. If the model errors are normal, independent, and have common variance, σ^2, then the resulting estimator is a maximum likelihood estimator. Therefore, a likelihood function is a function of the parameters $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ for fixed values of the random variables:

$\prod_{i=1}^{n} h(\varepsilon_i) = \frac{1}{(2\pi)^{n/2} \sigma^n} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left[y_i - f(x_i, \theta) \right]^2 \right\}$. Hence, the maximum likelihood estimator, $\hat{\theta}$, maximizes $\prod_{i=1}^{n} h(\varepsilon_i)$.

Similar to a linear model, the likelihood is maximized when the exponent of the previous equation is minimized. As a result, the least squares estimator, which minimizes $\sum_{i=1}^{n}\left[y_i - f\left(x_i,\hat{\theta}\right)\right]^2$ is also the maximum likelihood estimator. However, we only report the asymptotic properties of parameter estimates resulting from the non-linear estimation procedure. Linear regression computes the parameter estimates directly. In non-linear regression an iterative procedure is required for calculating the parameter estimates (Gallant, 1975). Non-linear least squares are optimization problems where an optimization algorithm such as the gradient descent could be used to solve non-linear estimation (Chambers, 1977). More specialized methods such as Gauss-Newton algorithm are favored due to their special properties as will be discussed in the next section (Kotz, Read and Banks, 1998).

3.5.2. *The Gauss-Newton procedure*

A common aspect of non-linear estimation procedures is that they require the user to specify starting values, initial step sizes and a criterion for convergence. The non-linear estimation procedure begins with specifying a particular set of initial parameter estimates. These initial parameter estimates change from iteration to iteration. In the first iteration, the step size determines how much the parameter will change. These iterations continue to adjust the parameter estimates until convergence is attained. The convergence criterion determines when the iteration process ends. In spite of the difficulties of non-linear models relative to linear models, they offer greater modeling flexibility.

Usually, exact properties of non-linear regression estimates cannot be derived (Neter, Wasserman and Kutner, 1985). Therefore, inferences from the non-linear model are generally based on a linear approximation. Gauss-Newton is one of the most commonly used methods in statistical software algorithms for finding the least squares estimator $\hat{\theta}$ in non-linear models (Bard, 1974; Draper and Smith, 1981; Kennedy and Gentle, 1980; Bates and Watts, 1988). This procedure is iterative and

requires starting values for the parameters. Taylor series expansion of the non-linear function around $\theta = \theta_o$ is used to find the estimates for θ, the p-dimensional vector of parameters. Only linear terms are considered in the Taylor series. The following expansion is considered a linearization of the non-linear form $f(x_i,\theta)$ where:

$$f(x_i,\theta) \cong f(x_i,\theta_o) + (\theta_1 - \theta_{1,0})\left[\frac{\partial f(x_i,\theta)}{\partial \theta_1}\right]_{\theta=\theta_o} + (\theta_2 - \theta_{2,o})\left[\frac{\partial f(x_i,\theta)}{\partial \theta_2}\right]_{\theta=\theta_o}$$
$$+ \ldots + (\theta_p - \theta_{p,o})\left[\frac{\partial f(x_i,\theta)}{\partial \theta_p}\right]_{\theta=\theta_o} \qquad i=1,2,\ldots,n$$

By subtracting $f(x_i,\theta_o)$ from both sides, we get

$$f(x_i,\theta) - f(x_i,\theta_o) = \tau_1 z_{1i} + \tau_2 z_{2i} + \ldots + \tau_p z_{pi} + \zeta_i$$
$$= y_i - f(x_i,\theta_o)$$

where $z_{ji} = \left[\frac{\partial f(x_i,\theta)}{\partial \theta_j}\right]_{\theta=\theta_o}$ is the derivative of the non-linear function with respect to the j^{th} parameter, and $\tau_j = \theta_j - \theta_{j,o}$ is the difference between the parameter initial value and the estimated value. In the above expansion, z_{ji} are known and considered as the explanatory variables in linear regression, while τ_i represent the regression coefficients. As a result, the Gauss-Newton procedure builds a linear regression. Estimation of τ_i produces estimates of θ_j, which can be viewed as more accurate then the initial value of θ_o. This procedure continues iteratively until convergence. Therefore, the vector of estimated parameters at iteration, s, is:

$$\hat{\theta}_s = \hat{\theta}_{s-1} + \left(Z_{s-1}' Z_{s-1}\right)^{-1} Z_{s-1}'\left[y - f\left(x,\hat{\theta}_{s-1}\right)\right]$$
$$= \hat{\theta}_{s-1} + \hat{\tau}_{s-1}$$

In this case, an estimate of the asymptotic variance-covariance matrix of $\hat{\theta}$ is given by

$$\hat{\text{var}}\left(\hat{\theta}\right) = s^2 \left(Z'Z\right)^{-1} \text{ where } s^2 = \sum_{i=1}^{n} \frac{\left[y_i - f\left(x_i, \hat{\theta}\right)\right]^2}{n-p}.$$

3.5.3. *Levenberg-Marquardt algorithm*

In order to improve convergence, there are many modifications of the Gauss-Newton method. The Levenberg-Marquardt algorithm is one of the modifications that we use in this chapter (Levenberg, 1944; Marquardt, 1963; Bishop, 1995; Press, Flannery, Teukolsky and Vetterling, 1992). The Levenberg-Marquardt (LM) algorithm provides a compromise between gradient descent and the Gauss-Newton method. In the LM algorithm, the structure of the vector of increments for the s^{th} iteration is given by the solution $\hat{\tau}_s$ to the following equation:

$$\left(Z_s'Z_s + \lambda I_p\right)\hat{\tau}_s = Z_s'\left[y - f\left(x, \hat{\theta}_s\right)\right] \qquad \text{for } \lambda > 0 \qquad (1)$$

$$\hat{\theta}_s = \hat{\theta}_{s-1} + \left(Z_s'Z_s + \lambda I_p\right)^{-1} Z_{s-1}'\left[y - f\left(x, \hat{\theta}_{s-1}\right)\right]$$

$$= \hat{\theta}_{s-1} + \hat{\tau}_{s-1}$$

The LM algorithm has been viewed as a ridge regression approach. Ridge regression is a method for dealing with correlated explanatory variables by adding a degree of bias to the estimates. The LM algorithm was designed specifically to minimize the sum of squares error function, using a formula that approximates the function modeled by a linear network. The Levenberg-Marquardt algorithm is a compromise between a linear model and a gradient-descent approach (Press, Flannery, Teukolsky and Vetterling, 1992). A change in the parameter is only considered if it reduces the error. The step size is chosen to be sufficiently small in the gradient descent model to allow small movements towards the minima.

The first term in the LM formula for estimating in the left hand side of equation (1) represents the linear assumption. The second term in the same formula

represents a gradient-descent step. The ridge parameter, λ, affects the relative influence of the linear assumption and the gradient descent. Each time LM lowers the error, it decreases the ridge parameter. As a result of decreasing the ridge parameter the linear assumption is strengthened and an attempt is made to jump directly to the minimum. However, if LM fails to lower the error, the ridge parameter increases. The increase in the ridge parameter gives more influence to the gradient descent step and makes the step size smaller to guarantee downhill progress at some point (Bishop, 1995).

Under the LM algorithm, the estimate of the asymptotic variance-covariance matrix of $\hat{\theta}$ is given by:

$$\hat{\text{var}}\left(\hat{\theta}\right) = s^2\left(Z'Z + \lambda I_p\right)^{-1} Z'Z\left(Z'Z + \lambda I_p\right)^{-1} \text{ where } s^2 = \sum_{i=1}^{n} \frac{\left[y_i - f\left(x_i, \hat{\theta}\right)\right]^2}{n-p}.$$

3.6. Non-linear neural network

Artificial neural network models have been extensively studied with the aim of achieving "human-like" performance. The origin of neural networks (NN) research dates back to the 1940's, when McCulloch and Pitts (1943) proposed a computational model based on a simple element, the neuron. The first neural network was motivated by the knowledge of brain physiology. Meanwhile, Hebb (1949) devised a learning rule for adapting the connections within artificial neurons. Rosenbaltt (1958) developed the perceptron. Minsky and Papert (1969) provided extensive analysis of the perceptron. Grossberg (1974) proposed new architectures and learning systems based on biological and psychological systems. This development led to the introduction of ANN.

Artificial neural networks surfaced as a machine-learning algorithm and a semi-parametric non-linear regression model that can be used for imputation (White, 1989, 1992). The NN can be considered a non-linear generalization of the linear

model. ANN represent a modeling technique capable of fitting extremely complex functions. The importance of ANN lies in the way in which they deal with the problem of scaling with dimensionality. Generally, ANN models represent non-linear functions of a single variable, which are called hidden units. The hidden units are themselves developed during the training process. Therefore, the number of such functions only needs to grow as the complexity of the problem itself grows, and not simply as the dimensionality of the input grows (Bishop, 1995). This type of model can be regarded as a two-stage regression or classification model, where the output is a non-linear combination of the inputs. Statistics and probability theory are well founded and provide ANN with powerful methodology for the estimation procedure and inference methods (Hastie, Tibshirani and Friedman, 2001). Artificial neural networks provide Statistics with a new type of non-linear modeling (Amari, 1993).

Neural networks have the ability to learn the structure of the data by trial and error through the training data set. A neuron, the basic unit of NN, is an operator that performs the mapping from R^p to R. In the supervised training phase, the network receives input vectors, x, and corresponding desired output, y, using the complete cases. Unlike linear NN, which have a record of successful applications, non-linear NN have only been recently developed. This is largely because the most important NN algorithm, back-propagation (BP), did not become widely known until 1986, by Rumelhart and McClelland.

Neural networks offer a powerful and generalized framework for representing non-linear mappings from several input variables to one or several output variables (Bishop, 1995). For instance, a network with two layers, where the first layer is sigmoid and the second layer is linear, can be trained to approximate any function with a finite number of discontinuities arbitrarily well.

The type of ANN used in this chapter for imputation are called feed-forward, where input terminals receive values of explanatory variables X, while the output provides the imputed variable Y. Multilayer feed-forward networks consist of one or

more hidden layers. The role of the hidden layer of neurons is to intervene between the external input and the network output. Inputs and outputs are connected through neurons that transform the sum of all received input values to an output value, according to connection weights and an activation function. The connection weights represent the strength of the connection between the neurons. The network weights (parameters) are randomly initialized and are then changed in an iterative process to reflect the relationships between the inputs and outputs.

An activation function is a function used by a node in a neural network to transform input data from any domain of values into a finite range of values. Almost any non-linear function is a suitable candidate for a network activation function. However, for gradient-descent learning algorithms, the activation function should be continuous and differentiable. The function is preferably bounded (Mandic and Chambers, 2001). It is typically chosen to be a sigmoid function having the following properties:

$(i)\sigma_i(x_i)$ is a continuously differentiable function

$(ii)\sigma_i'(x_i)=\dfrac{d\sigma_i(x_i)}{dx_i}>0$ for all $x_i\in\mathbb{R}$

$(iii)\sigma_i(a)=(b_i,c_i), a,b_i,c_i\in\mathbb{R}, b_i\neq c_i$

$(iv)\sigma_i'(x)\to 0$ as $x\to\pm\infty$

$(v)\sigma_i'(x)$ takes a global minimum value $\max_{x\in\mathbb{R}}\sigma_i'(x)$ at a unique point x=0

(vi)a sigmoidal function has only one inflection point, preferably at x=0

(vii) from (ii), the function σ_i is monotonically increasing; i.e. of $x_1<x_2$ for $x_1,x_2\in\mathbb{R}\Rightarrow\sigma_i(x_1)<\sigma_i(x_2)$

Examples of sigmoidal functions are:

$\sigma_1(x)=\dfrac{1}{1+e^{-\beta x}}, \quad \beta\in\mathbb{R}$ (This is the most commonly used activation function)

$\sigma_2(x)=\tanh(\beta x)=\dfrac{e^{\beta x}-e^{-\beta x}}{e^{\beta x}+e^{-\beta x}}, \quad \beta\in\mathbb{R}$

$$\sigma_3(x)=\frac{2}{\pi}\arctan\left(\frac{1}{2}\pi\beta x\right), \quad \beta\in\mathbb{R}$$

The most commonly used activation function is logistic.

3.7. Neural network imputation

Let X_{nxp} be an input matrix, Y the output vector, β the first layer of network weights and α as the second layer of network weights. For each input case, i.e. for each output value of X, we get the following NN function:

$$y=g\left(\sum_{h}^{n_h}\alpha_h g^h\left(\sum_{i}^{n_i}\beta_{hi}x_i+\beta_{ho}\right)+\alpha_o\right)$$

where g^h is the activation function at hidden layer for input node h, and g is the activation function at the output layer. The non-linearity of the network is attributed to the activation functions (g and g^h) of the neurons within NN. Therefore, the activation functions are used to introduce non-linearity into the network (Mandic and Chambers, 2001).

Let $T_n=\{(x_1,y_1),\ldots,(x_n,y_n)\}$ be the training set, which is a sample of n i.i.d. observations from $P(x,y;\theta)=P(x)P(y|x,\theta)$, where $P(x)$ does not depend on θ. The parameter used is not the maximum likelihood estimate, but rather it is obtained by learning during the training phase of the NN. The purpose of the minimization algorithm is to determine the minimum of the objective function in the parameter space. Starting from an initial parameter set, the minimum is determined iteratively by evaluating the value and/or gradient of the objective function, and performing small steps towards the minimum using the LM algorithm.

The most common NN architectures have outputs defined within a limited range (e.g., (0, 1) for the logistic activation function). This limited range for the output works in case of classification problems, where the desired output falls inside this

range. In the case of continuous variables, a scaling algorithm is sometimes applied to ensure that the NN output will be in the required range (Bishop, 1995). The simplest scaling function is minimax, which finds the minimum and maximum values of the response variable in the training data. A linear transformation then converts the values into the target range. If this is used on a continuous output variable, then all training values will be converted into the range of possible outputs of the NN. However, using minimax is restrictive because the curve is not extrapolated beyond the range of the data and does not estimate the mean. In addition, the model is saturated at either the minimum or maximum, depending on whether the estimated curve was increasing or decreasing as it approaches the region containing the minimum (Statsoft, 2004).

One approach to correcting the restrictive output of NN is to replace the logistic activation function at the output layer with a linear activation function. A linear activation function can cause numerical difficulties for the BP algorithm, in which case, a reduced learning rate must be used (Principe, Euliano. and Lefebvre, 2000). In the case of a feed-forward network with one hidden layer, if we set the activation function at the hidden layer as a logistic function and the activation function at the output layer as a linear function, the imputation of a continuous variable is possible. In this case, we have

$$y = f\left(x;\underline{\alpha},\underline{\beta}\right)+\zeta = \alpha_o + \sum_{h=1}^{q}\left[\frac{\alpha_h}{1-\exp\left(-\beta_{ho}-\sum_{i=0}^{r}\beta_{hi}x_i\right)}\right].$$

Then the error takes the form

$$\xi = y-\left[\alpha_o + \sum_{h=1}^{q}\frac{\alpha_h}{1-\exp\left(-\beta_{ho}-\sum_{i=1}^{r}\beta_{hi}x_i\right)}\right]$$

The estimation problem could be solved using one of the following two methods:

Method 1: Minimizing the performance function:

$$J = \frac{1}{2n}\sum_{t=1}^{n}[y_t - \hat{y}_t]^2$$

$$= \frac{1}{2n}\sum_{t=1}^{n}\left[y_t - \alpha_o - \sum_{h=1}^{q} \frac{\alpha_h}{1-\exp\left(-\beta_{ho} - \sum_{i=1}^{r}\beta_{hi}x_{it}\right)} \right]^2$$

To find the minimum we need to take the derivatives with respect to the parameters.

$$\frac{\partial J}{\partial \beta_{li}} \Rightarrow q(r+1) \text{ equations in } q(r+1) \text{ unknowns}$$

$$\frac{\partial J}{\partial \alpha_l} \Rightarrow (q+1) \text{ equations in } (q+1) \text{ unknowns}$$

This results in a system of equations that could be solved by setting the derivates equal to zero. For h=m and i=l, we have

$$\frac{\partial J}{\partial \beta_{ho}} = \left[\alpha_o + \sum_h \frac{\alpha_h}{1-e^{-\beta_{ho}-\sum_i \beta_{hi}x_{it}}} - y_t \right] \left[-\sum_h \frac{\alpha_h e^{-\beta_{ho}-\sum_i \beta_{hi}x_{it}}}{\left(1-e^{-\beta_{ho}-\sum_i \beta_{hi}x_{it}}\right)^2} \right]^{set} = 0$$

$$\frac{\partial J}{\partial \beta_{ml}} = \left[\alpha_o + \sum_h \frac{\alpha_h}{1-e^{-\beta_{ho}-\sum_i \beta_{hi}x_{it}}} - y_t \right] \left[\frac{\alpha_m x_{lt} e^{-\beta_{ho}-\sum_i \beta_{hi}x_{it}}}{\left(1-e^{-\beta_{ho}-\sum_i \beta_{hi}x_{it}}\right)^2} \right]^{set} = 0$$

$$\frac{\partial J}{\partial \alpha_o} = \left[\alpha_o + \sum_h \frac{\alpha_h}{1 - e^{-\beta_{ho} - \sum_i \beta_{hi} x_{it}}} - y_t \right]^{set} = 0$$

$$\frac{\partial J}{\partial \alpha_m} = \left[\alpha_o + \sum_h \frac{\alpha_h}{1 - e^{-\beta_{ho} - \sum_i \beta_{hi} x_{it}}} - y_t \right] \left[\frac{1}{1 - e^{-\beta_{mo} - \sum_i \beta_{mi} x_{it}}} \right]^{set} = 0$$

The parameters are estimated using the BP algorithm, in conjunction with the mean square error performance function by means of the gradient descent algorithm. In this case, the estimates of the parameters are:

$$\left.\begin{aligned}
\alpha_o(k+1) &= \alpha_o(k) - \eta(k) \frac{\partial J}{\partial \alpha_o(k)} \\
\alpha_m(k+1) &= \alpha_m(k) - \eta(k) \frac{\partial J}{\partial \alpha_m(k)} \\
\beta_o(k+1) &= \beta_o(k) - \eta(k) \frac{\partial J}{\partial \beta_o(k)} \\
\beta_{ml}(k+1) &= \beta_{ml}(k) - \eta(k) \frac{\partial J}{\partial \beta_{ml}(k)}
\end{aligned}\right\}$$

Where η is the network learning rate

α and β are the network parameters

At this point, a system of differential equations needs to be solved. The solution could be determined by using numerical analysis. In addition, the solution can be obtained by linearizing the function, as presented in Method 2.

Method 2: Taylor linearization

The non-linear model is written using a Taylor series linearization

$$y_t = f(x_t;\theta) \cong f(x_t;\theta_o) + (\theta_1 - \theta_{1,o})\left[\frac{\partial f(x_t,\theta)}{\partial \theta_1}\right]_{\theta=\theta_o} + (\theta_2 - \theta_{2,o})\left[\frac{\partial f(x_t,\theta)}{\partial \theta_2}\right]_{\theta=\theta_o}$$

$$+\ldots+(\theta_p - \theta_{p,o})\left[\frac{\partial f(x_t,\theta)}{\partial \theta_p}\right]_{\theta=\theta_o}$$

where $\theta = (\alpha_o, \alpha_1, \ldots, \alpha_q, \beta_{oo}, \beta_{o1}, \ldots, \beta_{or}, \ldots, \beta_{qr})$

The estimated network output, $\hat{y}$, is derived by hand to be

$$\hat{y} \cong \left[\alpha_o^o + \sum_{h=1}^{q} \frac{\alpha_h^o}{1 - e^{-\beta_{ho}^o - \sum_{i=1}^{r}\beta_{hi}^o x_i}}\right] + (\alpha_o - \alpha_o^o) + (\alpha_1 - \alpha_1^o)\left[\frac{1}{1 - e^{-\beta_{oo}^o - \sum_{i=1}^{r}\beta_{oi}^o x_i}}\right]$$

$$+\ldots+(\alpha_q - \alpha_q^o)\left[\frac{1}{1 - e^{-\beta_{qo}^o - \sum_{i=1}^{r}\beta_{qi}^o x_i}}\right] + (\beta_{oo} - \beta_{oo}^o)\left[\frac{\alpha_o e^{-\beta_{oo}^o - \sum_{i=1}^{r}\beta_{oi}^o x_i}}{\left(1 - e^{-\beta_{oo}^o - \sum_{i=1}^{r}\beta_{oi}^o x_i}\right)^2}\right] + \ldots$$

$$+(\beta_{1o} - \beta_{1o}^o)\left[\frac{-\left(\sum_i x_i\right)\alpha_o e^{-\beta_{1o}^o - \sum_{i=1}^{r}\beta_{1i}^o x_i}}{\left(1 - e^{-\beta_{1o}^o - \sum_{i=1}^{r}\beta_{1i}^o x_i}\right)^2}\right] + \ldots + (\gamma_{ro} - \gamma_{ro}^o)\left[\frac{-\left(\sum_i x_i\right)\alpha_o e^{-\beta_{ro}^o - \sum_{i=1}^{r}\beta_{1i}^o x_i}}{\left(1 - e^{-\beta_{ro}^o - \sum_{i=1}^{r}\beta_{1i}^o x_i}\right)^2}\right] + \ldots$$

$$+(\beta_{o1} - \beta_{o1}^o)\left[\frac{-x_1\alpha_o e^{-\beta_{oo}^o - \sum_{i=1}^{r}\beta_{oi}^o x_i}}{\left(1 - e^{-\beta_{oo}^o - \sum_{i=1}^{r}\beta_{oi}^o x_i}\right)^2}\right] + \ldots + (\beta_{qr} - \beta_{qr}^o)\left[\frac{-x_r\alpha_o e^{-\beta_{qo}^o - \sum_{i=1}^{r}\beta_{qi}^o x_i}}{\left(1 - e^{-\beta_{qo}^o - \sum_{i=1}^{r}\beta_{qi}^o x_i}\right)^2}\right]$$

This is repeated for the parameter estimation procedure at each of the iterations.

3.8. Results

This section contains simulation results and results using data from the NHIS. This section compares the applied performance of non-linear ANN in imputation to EM single imputation, MCMC and propensity score multiple imputation techniques. In ANN analysis, the data is usually divided into several parts. While, some observations are used for training and estimating the network parameters, other observations are used to check the performance of the network. A third set, or test set, of cases is used to evaluate the imputation error.

3.8.1. *Simulation*

A simulation study was performed to compare the results of imputation using non-linear ANN to traditional imputation techniques. Data for this simulation was generated using a stratified simple random design with two strata having equal allocation. For this simulation study, each of the two strata (Z=1,2) had three variables X_1, X_2 and Y, and 1000 observations. Using Matlab software, the X's were generated separately with a normal distribution in each stratum. The Y was generated as a function of the X's and with normal random error. The relationship between the X's and Y was simulated as both a linear and a non-linear relationship for the purpose of generating simulation data. The linear model was simulated using a linear combination of the X's and the error term using the following equation: $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \xi$.

In the non-linear case, three different models were used to generate the data covering different data patterns:

Model 1: $y = \alpha_o + \dfrac{\beta_3}{\beta_4 + \beta_5 x_1} + \gamma_o e^{\gamma_1 - \gamma_2 x_2} + \xi$ (exponential)

Model 2: $y = \beta_6 x_1^3 - \beta_7 x_2^3 + \beta_8 x_1 x_2 + \xi$ (Cubic)

Model 3: $y = \beta_9 + \beta_{10} x_1 x_2 + \beta_{11} \sin(x_1 - x_2) + \xi$ (Cyclic)

The parameters in the chosen models for data simulation (α, β, and γ) were set arbitrarily and separately for each stratum. A random number was generated and used to select certain observations to have missing Y values. This random number was generated from three different distributions: a normal distribution, a Poisson distribution and a hypergeometric distribution. The reason behind generating the missing data with different distributions was to investigate the effect of the distribution of the missing data on the imputation techniques. The distribution of the missing data is customarily used to be normal distribution in the statistical literature. The normality assumption is unrealistic in many cases; therefore we decided to investigate the effect of different distributions on the analysis results. The number of missing observations in each simulation run represented 10 percent of the sample size in each stratum. In addition, the missing observations were used for evaluating the performance of the imputation techniques.

Several measures of accuracy were used for comparing the results from the different imputation techniques. These included mean absolute percentage error (MAPE), mean absolute deviation (MAD), and mean squared deviation (MSD). Let y be the actual value, $\hat{y}$ the imputed value, and m the number of missing/imputed values. The mean absolute percentage error measures the accuracy of imputed values as a percentage:

$$MAPE = \frac{\sum_{t=1}^{m} \left| \left(y_t - \hat{y}_t \right) / y_t \right|}{m} \times 100 \quad \left(y_t \neq 0 \right).$$

The mean absolute deviation is expressed in the same units as the data to help conceptualize the amount of error.

$$MAD = \frac{\sum_{t=1}^{m} \left| y_t - \hat{y}_t \right|}{m}.$$

The mean squared deviation (MSD) is similar to mean squared error deviations (MSE) a commonly used measure of accuracy.

$$MSD = \frac{\sum_{t=1}^{m} (y_t - \hat{y}_t)^2}{m}$$

Note that the mean squared deviation is computed using the same denominator, m, regardless of the model, in order to compare MSD values across all models. Mean squared error deviations are computed with different degrees of freedom for different models, so MSE values are not easily compared across models.

A variety of software packages were combined in the analysis of the data to make use of several imputation techniques. SAS was used for MI with MCMC, SPSS was used for EM imputation, and Solas was used for MI using propensity score. In addition, the Matlab Neural Network toolbox was used for ANN imputation with a feed-forward network with BP and two layers in each stratum. Imputation was performed separately in each stratum. In ANN imputation, online training was used to provide the network with one observation at each pass. Initial parameter values were randomly assigned. Several neural networks were observed based on increasing the number of nodes in the hidden layer. Increasing the number of nodes in the hidden layer above three nodes did not improve the results. For consistency and comparability, the hidden layer was set to have three nodes for all neural networks used in imputation. With one response variable, y, the output layer consisted of one node.

In case of the data generated based on the linear model with a normally distributed data, EM algorithm for imputation and ANN imputation consistently gave the smallest error for all three measures of accuracy. Multiple imputation using propensity score gave the highest error in this case. However with a linear model but with Poisson and hypergeometric distributions, multiple imputation with MCMC resulted in the smallest error for all three accuracy measures. ANN came in second

place after multiple imputation using MCMC. However, multiple imputation with propensity score resulted in the largest error under Poisson distribution.

Table 3. Non-linear imputation simulation

a. Linear model

	Normal	Poisson	Hypergeometric
MCMC			
MAPE	2.84	2.54*	1.86*
MAD	0.85	0.88*	0.86*
MSD	1.21	1.27*	1.23*
EM			
MAPE	0.0009*	15.581	9.45
MAD	0.0003*	5.09	4.37
MSD	1^{E-7}*	38.61	33.15
Propensity Score			
MAPE	25.25	20.02	7.18
MAD	5.41	6.76	3.46
MSD	51.65	76.151	24.72
Non-linear ANN			
MAPE	0.15	4.05	7.66
MAD	0.02	1.65	3.49
MSD	0.0007	9.52	35.78

b. Non-linear model 1

	Normal	Poisson	Hypergeometric
MCMC			
MAPE	64.88	25.83	25.76
MAD	1.26	1.37	1.33
MSD	2.47	2.95	2.76
EM			
MAPE	280.73	87.79	94.85
MAD	5.52	4.46	4.54
MSD	36.39	25.93	26.88
Propensity Score			
MAPE	19.96*	24.79*	26.2
MAD	1.27	1.28*	1.29
MSD	3.49	2.58*	2.49
Non-linear ANN			
MAPE	55.4	34.16	20.14*
MAD	0.90*	2.41	0.88*
MSD	1.27*	7.47	1.36*

Table 3. Non-linear imputation simulation (Continued)

c. Non-linear model 2

	Normal	Poisson	Hypergeometric
MCMC			
MAPE	24946.37	74.67	63771.28
MAD	45987.29	23730.75	25439.04
MSD	4.83^{E+9}	9.84^{E+8}	1.26^{E+9}
EM			
MAPE	16728.15	80.39	36454.33
MAD	19724.15*	19639.4	12202.66
MSD	9.6^{E+8*}	3.89^{E+8*}	2^{E+8}
Propensity Score			
MAPE	8391.01*	42.72*	3148.42
MAD	33713.19	14885.09*	2209.44
MSD	3.1^{E+9}	4.19^{E+8}	7.8^{E+6}
Non-linear ANN			
MAPE	21011.26	100	100.08*
MAD	36166.38	34413.12	1000.06*
MSD	3.1^{E+9}	1.5^{E+9}	1.6^{E+6*}

d. Non-linear model 3

	Normal	Poisson	Hypergeometric
MCMC			
MAPE	757.06	29.11	36.49
MAD	38.58*	34.04	36.56
MSD	2637.71*	1871.39	2205.34
EM			
MAPE	351.57	33.38	39.61
MAD	52.45	39.09	39.88
MSD	5198.47	2554.83	2880.36
Propensity Score			
MAPE	370.59	66.17	42.67
MAD	52.53	77.56	42.61
MSD	5729.33	9817.84	3638.51
Non-linear ANN			
MAPE	190.35*	10.96*	14.24*
MAD	42.57	12.67*	14.35*
MSD	4781.13	285.1*	351.75*

In case of non-linear models, the results varied from one model to another. In case of non-linear model 1, ANN resulted in the smallest error in case of normal and

hypergeometric data followed by multiple imputation using MCMC and propensity score. However, multiple imputation with propensity score resulted in lowest error under Poisson distribution under non-linear model 2. The EM algorithm gave consistent high error under all distributions with non-linear model 1 but resulted in the lowest error across all three accuracy measures in non-linear model 2 with normally distributed data. In non-linear model 2 with Poisson distribution, multiple imputation with propensity score resulted in the smallest error while ANN had consistently higher error. In the hypergeometric case of non-linear model 2 ANN had the lowest error followed by multiple imputation with propensity score. Multiple imputation with MCMC had the largest errors under non-linear model 2 with both normal and hypergeometric data. Finally in case of non-linear model 3, results were consistent, where ANN had lowest error under Poisson and hypergeometric data followed by multiple imputation using MCMC. However, in case of normal data under non-linear model 3 multiple imputation using MCMC had the lowest error. Multiple imputation using propensity scores resulted in the highest error across all accuracy measures under Poisson and hypergeomeric distributions.

The MSD and MAD show that MCMC imputation had smaller error under the normal distribution compared to Poisson and hypergeometric cases under linear model and non-linear-model 1. As for non-linear models 2 and 3, MCMC had smaller error under Poisson distribution compared to normal and hypergeometric distributions. MCMC gave lowest imputation errors under non-linear model 3 when compared to the other imputation techniques. These results show that the different methods used with multiple imputation have a strong effect of the performance of multiple imputation.

As for the performance of the different methods compared across the different distributions, the results could be interpreted in a different way. The EM algorithm gave small errors in case of linear model with normal distribution compared to Poisson and hypergeometric. However, in case of non-linear models 1 and 3, the Poisson distribution had smaller error when compared to normal and hypergeometric

distributions. When compared to the other imputation techniques, EM gave the smallest error under the linear model and non-linear model 2 with normal distribution. The propensity score imputation performed better under hypergeometric distribution compared to normal and Poisson distributions. Non-linear ANN imputation yielded the smallest errors under the linear model with normal distribution. However, in case of non-linear models, non-linear ANN imputation performed better under hypergeometric and Poisson distributions.

As a result, we notice that the distribution of the missing data affected the efficiency of the imputation results. The EM algorithm compared favorably to MI with propensity score. When compared to the different imputation techniques, non-linear ANN had smaller error more frequently and consistently. The results, presented in Table 6, indicate that MI with MCMC performed as well as ANN in several of the simulation cases. This may be due to the similarity of ANN to a Markov chain at steady-state (Burton, 1997; De Wilde, 1995).

3.8.2. *Application*

To illustrate the difference between imputation techniques, data from the National Health Interview Survey (NHIS) was used. The National Health Interview Survey is a survey conducted by the National Center for Health Statistics, Centers for Disease Control and Prevention (NCHS-CDC). The NHIS is a cross-sectional, household interview survey, which represents the principal source of information on the health of the civilian non-institutionalized population of the United States. The data under investigation is from the last quarter (October – December) of 2001. The NHIS data is obtained through a complex sample design involving stratification, clustering, and multistage sampling designed to represent the civilian, non-institutionalized population of the United States (National Center for Health Statistics, 2002).

Researchers found that the prevalence of obesity has increased rapidly in the United States. Body mass index (BMI = weight/height^2) was considered as the response variable in this analysis. The other variables in the data were gender, age, income, family size and respondents' self-conscience (e.g. desired weight, frequency of happiness, sadness, etc). The data was filtered to create a set of 7543 records with no missing information in the 27 variables under investigation. Data was assumed to be a simple random sample for simplicity and demonstration purposes, further details about the sampling design will be discussed in Chapter 4. Missing data, representing 10, 20, and 30 percent of the records, were generated completely at random. Imputation was performed using the same imputation techniques discussed earlier in section 3.4. The objective was to compare the performance of the different imputation techniques as well the effect of the increasing percent of missing data on the imputation results. The results of the imputation are shown in Table 4 with the lowest error marked with an *.

Table 4. Non-linear imputation application

Missing data		MAPE	MAD	MSD
10 %	Nonlinear ANN	4.75*	1.21*	5.17*
	MCMC	9.53	2.49	11.06
	EM	16.04	4.27	31.09
	Propensity Score	8.81	2.51	17.20
20 %	Nonlinear ANN	4.70*	1.22*	5.21*
	MCMC	9.28	2.50	11.13
	EM	15.95	4.26	31.22
	Propensity Score	8.55	2.45	16.57
30 %	Nonlinear ANN	5.80*	1.26*	5.25*
	MCMC	9.46	2.50	11.14
	EM	15.93	4.20	30.71
	Propensity Score	9.22	3.99	17.05

A nonlinear neural network with as few as five nodes in the hidden layer consistently resulted in a lower error than the other imputation techniques tested. This could be explained by the adaptive nature of ANN, where ANN adapt to the empirical distribution of the sampled data and attempts to impute the missing data based on

information acquired from the sample. Multiple imputation using MCMC and propensity score performed similarly, while EM algorithm had the highest error.

The percent of missing data, also referred to as the non-response rate, was expected to affect the performance of the different imputation techniques. The percent of missing data increase the denominator of the accuracy measures. However, the numerator of the accuracy measures increases also due to the larger number of imputed data. Generally, in cases where the non-response rate is small, the bias may be ignored, but with high non-response rates substantial bias may occur. Analysis results showed that the imputation error did not increase as the percent of missing values increased.

3.9. Conclusion

Model-based imputation techniques are flexible and practical approaches to missing data problems but the challenge is to find the correct imputation model. In cases where different models are used for imputation and for analysis, or the imputer and the analyst make different assumptions about the models used, the results could be very different and lead to incorrect conclusions. Neural networks offer an alternative to model specification, by adapting to the empirical distribution of the complete case and using this distribution in the imputation. This may be acceptable in the case of missing data generated through an ignorable mechanism. Without the need to specify an absolute model, NN help avoid model misspecification, which could be a source of concern for some statisticians.

MCMC and ANN are useful techniques that require considerable care in their application and interpretation of the results. These techniques require training by the analyst. If difficulties are encountered in obtaining convergence, the analyst should change the starting values. Multiple imputation requires more effort, longer processing time and more computer memory compared to single imputation. Multiple imputation has the advantage of accounting for imputation uncertainty. Incorporating features of a

complex sampling design are not easy using multiple imputation. ANN may have the ability to manage any complex sample design as will be discussed in chapter 4 but requires a trained user for its application. Finally, the distribution of the missingness pattern appears to affect the accuracy of the imputations. The distribution of the missingness pattern needs to be considered and followed up with further research.

3.10. References

Allison, Paul D. (1999) Multiple Imputation for Missing Data: A Cautionary Tale. (http://www.ssc.upenn.edu/~allison/MultInt99.pdf)

Amari, S. (993). Universal Theorem on Learning Curves. Neural Networks, 6, 161-166.

Bard, Y. (1974). Nonlinear Parameter Estimation. New York: Academic Press.

Bates, Douglas M. and Watts, Donald G. (1988). Nonlinear Regression Analysis and its Applications. New York: Wiley.

Bishop, C.M. (1995). Neural Networks for Pattern Recognition. Oxford: Clarendon Press.

Box, G. E. P. and Lucas, H. L. (1959). Design of Experiments in Non–linear Situations. Biometrika, 46, 77–90.

Box, G.E.P. (1960). Fitting Experimental Data. Annals of the New York Academy of Sciences, 86, 792-816.

Box, G.E.P. and Hunter, W.G. (1962). A Useful Method for Model-building. Technometrics, 4, 301-318.

Box, G.E.P. and Hunter, W.G. (1965). The Experimental Study of Physical Mechanisms. Technometrics, 7(1), 23-42.

Box, G.E.P. and Hill, W.J. (1967). Discrimination among Mechanistic Models. Technometrics, 9(1), 57-71.

Box, G.E.P. and Hill, W.J. (1974). Correcting Inhomogeneity of Variance with Power Transformation Weighting. Technometrics, 13(3), 385-389.

Box, G.E.P., Hunter, W.G. and Hunter, J.S. (1978). Statistics for Experimenters. New York: Wiley.

Burton, R.M. (1997). Lecture Notes, Department of Mathematics, Oregon State University.

Chambers, J.R. (1977). Computational Methods for Data Analysis. New York: Wiley.

Cochran, W.G. (1977). Sampling Techniques, 3rd Edition, New York: Wiley.

Draper, N.R. and H. Smith (1981). Applied Regression Analysis. New York: John Wiley and Sons.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, Series B, 39, 1-38.

De Wilde, P. (1995). Neural Network Models: Analysis, Lecture Notes in Control and Information Sciences, Vol. 210. New York: Springer-Verlag.

Fellegi, I. P. and Holt, D. (1976). A Systematic Approach to Automatic Edit and Imputation, Journal of the American Statistical Association, 71, 17-35.

Fisher, R.A. (1922). On the Mathematical Foundations of Theoretical Statistics. Philosophical Transactions of the Royal Society of London, Series A, 222, 309-368.

Fisher, R.A. and Mackenzie, W.A. (1923). The Manorial Response of Different Potato Varieties. Journal of Agricultural Science, 13, 311-320.

Gallant, A.R. (1975). Nonlinear Regression. American Statistician, 29(2), 73-81.

Gelman, A.E., Carlin, J.B., Stern, H.S., and Rubin, D.B. (1995). Bayesian Data Analysis. London: Chapman and Hall.

Gelman, A.E and Rubin, D.B. (1992). Inference from Iterative Simulation Using Multiple Sequences. Statistical Science, 7, 457-472.

Geyer, C.J. (1992). Practical Markov Chain Monte Carlo. Statistical Science, 7(4).

Grossberg, S. (1974). Classical and Instrumental Learning by Neural Networks. Progress in Theoretical Biology, 3, 51-141.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer.

Hebb, D. (1949). The Organization of Behavior: A Neuropsychological Theory. New York: Wiley.

Hocking, R. R. and Smith, W. B. (1972). Optimum Incomplete Multinormal Samples, Technometrics, 4, 299-307.

Horton, N.J. and Lipsitz, S.R. (2001). Multiple Imputation in Practice: Comparisons of Software Packages for Regression Models with Missing Variables, The American Statistician, 5(3).

Kennedy, William J. and Gentle, James E. (1980). Statistical Computing. New York: Marcel Dekker.

Kotz Samuel, Read Campbell B., and Banks David L. (Eds.) (1998). Encyclopedia of Statistical Sciences. Wiley-Interscience.

Lawton, W.H., Sylvester, E.A. and Maggio, M.S. (1972). Self Modeling Non-linear Regression. Technometrics, 14, 513-532.

Levenberg, K. (1944). A Method for the Solution of Certain Non-linear Problems in Least Squares. Quarterly of Applied Mathematics, 2, 164-168.

Lewicki, P., Hill, T., and Czyzewska, M. (1992). Nonconscious Acquisition of Information. American Psychologist, 47, 796-801.

Little, R.J.A. (1986). Survey Nonresponse Adjustments for Estimates of Means. International Statistical Review, 54, 139-157.

Little, R.J.A and Rubin, D.B. (2002). Statistical Analysis with Missing Data. John Wiley and Sons Inc.

Mandic, D.P. and Chambers, J.A. (2001). Recurrent Neural Networks for Prediction: Learning Algorithms, Architectures and Stability. England: John Wiley and Sons.

Marquardt, D.W. (1963). An Algorithm for Least-squares Estimation of Nonlinear Parameters. Journal of the Society of Industrial and Applied Mathematics, 11, 431-441.

McCulloch, W.S. and Pitts, W.H. (1943). A Logical Calculus of Ideas Immanment in Nervous Activity. Bulletin of Mathematical Biophysics, 5, 115-133.

Meng, X.-L. (1994). Multiple-imputation Inferences with Uncongenial Sources of Input. Statistical Science, 9, 538-558.

Minsky, M. and Papert, S. (1969). Perceptrons. Cambridge: MIT Press.

National Center for Health Statistics (2002). Data file Documentation, National Health Interview Survey, 2001 (machine readable file and documentation). National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, Maryland.

Neter, J., Wasserman, W., and Kutner, M. H. (1985). Applied Linear Statistical Models: Regression, Analysis of Variance, and Experimental Designs. Homewood, IL: Irwin.

Press, W.H., Flannery, B.P., Teukolsky, S.A., and Vetterling, W. T. (1992). Numerical Recipes in FORTRAN: The Art of Scientific Computing, 352-355 (2nd edition). England: Cambridge University Press.

Principe, J.C., Euliano, N.R. and Lefebvre, W.C. (2000). Neural and Adaptive Systems: Fundamentals Through Simulation. New York: John Wiley and Sons.

Ripley, B.D. (1977). Modeling Spatial Patterns. Journal of the Royal Statistical Society, Series B, 39, 172-212.

Rosenbaltt, F. (1958). The Perceptron: a Probabilistic Model for Information Storage and Organization in the Brain. Psychological review, 65, 386-408.

Rubin D.B. (1976a). Inference and Missing Data, Biometrika, 63, 581-592.

Rubin D.B. (1977). Formalizing Subjective Notions About The Effect of Non-Respondents in Sample Surveys. Journal of the American Statistical Association, 77, 538-543.

Rubin, D. B. (1978b). Multiple Imputations in Sample Surveys- a Phenomenological Bayesian Approach to Nonresponse, Proceedings of the Survey Research Methods Section, American Statistical Association, 20-34.

Rubin, D.B. (1985a). The Use of Propensity Scores in Applied Bayesian Inference, in Bayesian Statistics 2 (J.M. Bernardo, M.H. De Groot, D.V. Lindley, and A.F.M. Smith, eds.), Amsterdam: North Holland, 463-472.

Rubin, D.B. (1996). Multiple Imputation After 18+ Years, Journal of American Statistical Association, 91, 473-489.

Rumelhart, D.E and McClelland J.L. (1986). Parallel Distributed Processing, Explorations in the Microstructure of Cognition. Cambridge, MA: MIT Press.

Schafer, J.L. and Schenker, N. (2000). Inference with Imputed Conditional Means. Journal of the American Statistical Association, 95, 144–154.

Schafer, J.L. (1997). Analysis of Incomplete Multivariate Data. London: Chapman and Hall.

Schimert, J., Schafer, J.L.; Hesterberg, T.M., Fraley, C., and Clarkson, D.B. (2000). Analyzing Data with Missing Values in S-Plus. Seattle: Insightful Corp.

Smith, A.F.M. and Roberts, G.O. (1992). Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo methods. Journal of the Royal Statistical Society, Series B, 5(1).

StatSoft, Inc. (2004). Electronic Statistics Textbook. Tulsa, OK: StatSoft. (http://www.statsoft.com/textbook/stathome.html)

Tanner, M.A. and Wong, W.H. (1987). The Calculation of Posterior Distributions by Data Augmentation. Journal of the American Statistical Association, 82(398), 528-550.

Trawinsky, I.M. and Bargmann, R.E. (1964). Maximum Likelihood Estimation with Incomplete Multivariate Data. Annals of Mathematical Statistics, 35, 647-657.

Treiman, D.J., Bielby, W., and Cheng, M. (1993). Multiple Imputation by Splines. Bulletin of the International Statistical Institute, Contributed Papers II, 503-504.

Wachter, K.W. and Trussell, J. (1982). Estimating Historical Heights, Journal of the American Statistical Association, 77, 279-301.

White, H. (1989). Learning in Artificial Neural Networks: A Statistical Perspective. Neural Computation, 1(4), 425-464.

White, H. (1992). Artificial Neural Networks: Approximation and Learning Theory. UK: Oxford and USA: Cambridge.

4. IMPUTATION, COMPLEX SURVEY DESIGN AND INFERENCE

4.1. Abstract

Missing data yields many analysis challenges. In addition to dealing with missing data, researchers need to account for the sampling design in order to achieve useful inferences. Methods for incorporating sampling weights in neural network imputation were investigated to account for complex survey designs. An estimate of variance to account for the imputation uncertainty as well as the sampling design using neural networks will be provided. A simulation study was conducted to compare estimation results based on complete case analysis, multiple imputation using a Markov Chain Monte Carlo method, and a neural network imputation. In addition, a public-use dataset was used as an example to illustrate neural networks imputation under a complex survey design.

Keywords: complex survey, imputation, neural networks, estimate, variance

4.2. Introduction

Traditional methods presented in the statistical literature, outside of survey sampling, have been based on a simple random sample (Casella and Berger, 1990). This assumption is not appropriate when the data were generated using a complex sampling design (Lohr, 1999). As an alternative to standard formulas and techniques used in case of simple random sample, design-based procedures were developed to handle probability sampling. Design-based procedures provide accurate inference for complex surveys. Design-based procedures date back to the 1950's, and they have been evolving to account for complex sampling designs (Hansen, Hurwitz and Madow, 1953; Kish, 1965; Cochran, 1977; Kish, 1995).

In simple random sampling, the sampling design does not provide any information and has no effect on the estimation of the population parameters. In a complex survey design, characteristics of the population may affect the sample and are

used as design variables (e.g. strata). Sample design involves the concepts of stratification, clustering, etc. These concepts usually reflect a complex population structure and should be accounted for during the analysis. In design-based inference, the main source of random variation is induced by the sampling mechanism (Chambers and Skinner, 2003). In case of a complex sample design, the population average is considered as a fixed unknown quantity and the sample indicators are the random variables. Furthermore, in complex survey design, the variance is the average squared deviation of the estimate from its expected value, averaged over all possible samples which could be obtained using a given design. This is opposed to the traditional variance definition where the variance is the average squared deviation of the estimate from its expected value averaged over all possible samples (Lohr, 1999). Design-based approaches make use of sampling weights as part of the estimation and inference procedures.

In survey sampling, two types of weights are of interest. These weights are sampling weights and nonresponse weights. If a unit is sampled with a specific selection probability then the sampling weight is the inverse of the probability of sample selection. For example, stratified sampling occurs when the population is divided into distinct subpopulations (strata) and a separate sample is selected within each stratum. From the sample in each stratum, a separate stratum mean is computed. The stratum means are weighted to calculate a combined estimate for the entire population. Weighting is used as a nonresponse adjustment for unit nonresponse as well. Nonresponse weight is the reciprocal of the probability that a unit is selected in the sample and responds. A combined weight results from multiplying the response weight times the sampling weight.

Several estimators and their corresponding variances have been introduced in the literature for different sampling designs (e.g. Horvitz-Thompson estimator). Point estimators are usually calculated using survey weights, which may involve auxiliary population information. However, sampling variance estimation is more complicated than parameter estimation (Lohr, 1999). Alternatives to conventional variances, in

case of complex survey designs, were proposed to facilitate the variance calculation. Methods like the random group method (Mahalanobis, 1946) are based on the replication of the survey design. These methods are simple to apply to nonparametric problems, but lead to imprecise estimates of the variances (Lohr, 1999). Woodruff (1971) illustrated the Taylor series linearization method to approximate the variance in complex surveys. In case of Taylor series linearization, in spite of a complex calculation, the linearization theory may be applied if the partial derivatives are known.

4.3. Neural network and complex survey design

One major advantage of artificial neural networks (ANN) is their flexibility in modeling many types of non-linear relationships. Artificial neural networks can be structured to account for complex survey designs and for unit nonresponse as well. Sampling weights have been used to adjust for the complex sampling design using unequal sampling (Lohr, 1999). We suggest two different methods to include sampling weights into ANN. The first method is to include the sampling weights in the ANN similar to weighted least squares (WLS). The second method is based on accounting for the sampling design structure in ANN.

4.3.1. Method based on weighted least squares

Weighted least squares are used in regression in several situations to account for variance. When the deviations from the responses are available; the weights are the reciprocal of the response variance to give observations with smaller error more weight in the estimation procedure. In addition, weights are used when the responses are averaged from samples with different sizes; the weights are equal to the sample sizes. Additionally, when the variance is proportional to a covariate, the weight is the inverse of the covariate. In survey sampling, statisticians debated about the relevance of the sampling weights for inference in regression (Brewer and Mellor, 1973). Part of this debate is based on the idea that weights are needed for estimating population

parameters in complex survey sampling and by analogy should be used in regression. Chapter 2 illustrated the similarities between ANN and linear regression. Based on these similarities, we propose including the sampling weights in the network in the same manner it would be incorporated in case of regression. If sampling weights are used in the weighted least squares estimation, point estimates will be similar to design based estimates. However, the standard errors also need to be developed based on the survey design.

Example of applying method 1: Weighted least squares (WLS) in a linear neural network

Assume Y is a variable with missing values and X is the auxiliary variable. Using the formula for weighted least squares, we have: $w_i y_i = \beta w_i x_i + w_i \alpha + w_i \xi_i$ such that w_i is the sampling weight for observation i, and α and β the model parameters (Neter, Wasserman and Kutner, 1985). By setting $y^* = wy$, $x^* = wx$, and $\xi^* = w\xi$, we get $y_i^* = \beta x_i^* + \alpha^* + \xi_i^*$ where $\hat{y}_i^* = \alpha^* + \beta x_i^*$ such that $\xi_i^* = y_i^* - \hat{y}_i^*$.

Following the same steps as in chapter 2, we want to minimize the objective function J, where:

$$
\begin{aligned}
J &= \frac{1}{2n}\sum_{i=1}^{n}\xi_i^{*2} \\
&= \frac{1}{2n}\sum_{i=1}^{n}\left(y_i^* - \hat{y}_i^*\right)^2 \\
&= \frac{1}{2n}\sum_{i=1}^{n}\left(y_i^* - \beta x_i^*\right)^2 \\
&= \frac{1}{2n}\left[\sum_{i=1}^{n} y_i^{*2} - 2\beta\sum_{i=1}^{n} y_i^* x_i^* + \beta^2\sum_{i=1}^{n} x_i^{*2}\right]
\end{aligned}
$$

To find the minimum, we need to take the derivative with respect to the parameter:

$$\frac{\partial J}{\partial \beta}=\frac{1}{2n}\left[-2\sum_i y_i^* x_i^* + 2\beta\sum_i x_i^{*2}\right]=\frac{1}{n}\left[-\sum_i y_i^* x_i^* + \beta\sum_i x_i^{*2}\right]^{set}=0$$

Therefore, the estimate is $\beta^* = \dfrac{\sum y_i^* x_i^*}{\sum_i x_i^{*2}}$. Note that this estimate is identical to the estimate obtained in chapter 2. This estimate yields a minimum value for the objective function. Where

$$J_{\min}=\frac{1}{2n}\left[\sum_i y_i^{*2} - \frac{\left(\sum_i x_i^* y_i^*\right)^2}{\left(\sum x_i^{*2}\right)}\right]$$

Then the relationship between J and $J_{\min}$ can be written as

$$J = J_{\min} + \frac{1}{2n}\left(\beta-\beta^*\right)\left(\sum x_i^2 + \lambda^*\right)\left(\beta-\beta^*\right)$$

Subtract $J_{\min}$ from both sides, we get:

$$J - J_{\min} = \frac{1}{2n}\left(\beta-\beta^*\right)\left(\sum x_i^2 + \lambda^*\right)\left(\beta-\beta^*\right)$$

$$\nabla J = \frac{\partial\left(J - J_{\min}\right)}{\partial\beta} = \upsilon\left(\beta-\beta^*\right) \quad \text{where} \quad \upsilon = \frac{1}{n}\left(\sum x_i^2 + \lambda^*\right)$$

In this case $\beta(k+1) = \beta^* + (1-\eta\upsilon)^{k+1}\left(\beta(o)-\beta^*\right)$

As a result $J = J_{\min} + \lambda^*(1-\eta\upsilon)^{2k}\left(\beta(o)-\beta^*\right)^2$

In case of complex survey design, using the approach of weighted least squares proves to be useful specially when the analyst is not involved in the design stage but is presented with the final weights.

4.3.2. Method based on ANN structure

When the analyst has access to the design variables, an alternative method to WLS is to construct the neural network using the sampling design features. For example, in a stratified sampling design, a separate network could be built and trained using data from a specific stratum in the imputation procedure. We suggest using a separate network for each of the strata. These networks are then connected with a binary activation function at the input layer. This binary activation function directs each observation to the corresponding stratum, taking the value 0 when the observation is not in the stratum and 1 when the observation belongs to that stratum. This leads to a different network parameter estimates for each stratum. The disadvantage of using a separate network for each stratum (without connecting the networks or assigning a probability for each stratum) is that it does not provide estimates for the entire sample. Therefore, a suggested solution is to train separate networks for each output and then to combine all strata to account for the full sample.

Using a mixture of network models is common in ANN and can be considered a technique to account for the sampling design (Bishop, 1995). Mixture of expert networks are mixture models used to solve complex problems. The solution of these complex problems is achieved by dividing the problem into smaller sub-problems, where each network is considered an expert for a subgroup of the observations. These expert networks are connected together through a gating network, which divides the input space into different subgroups of conditional densities as shown in Figure 6 (Jacobs, Jordan, Nolman and Hinton, 1991). The use of a mixture of expert networks allows the introduction of sampling probabilities and construction of a separate model for each stratum in a stratified sample design. In mixtures of expert network models, we have: $y_j = f_j\left(x_j;\theta\right)$ where strata j=1,...,M .

Figure 6. A mixture of Expert Networks for a stratified sampling design

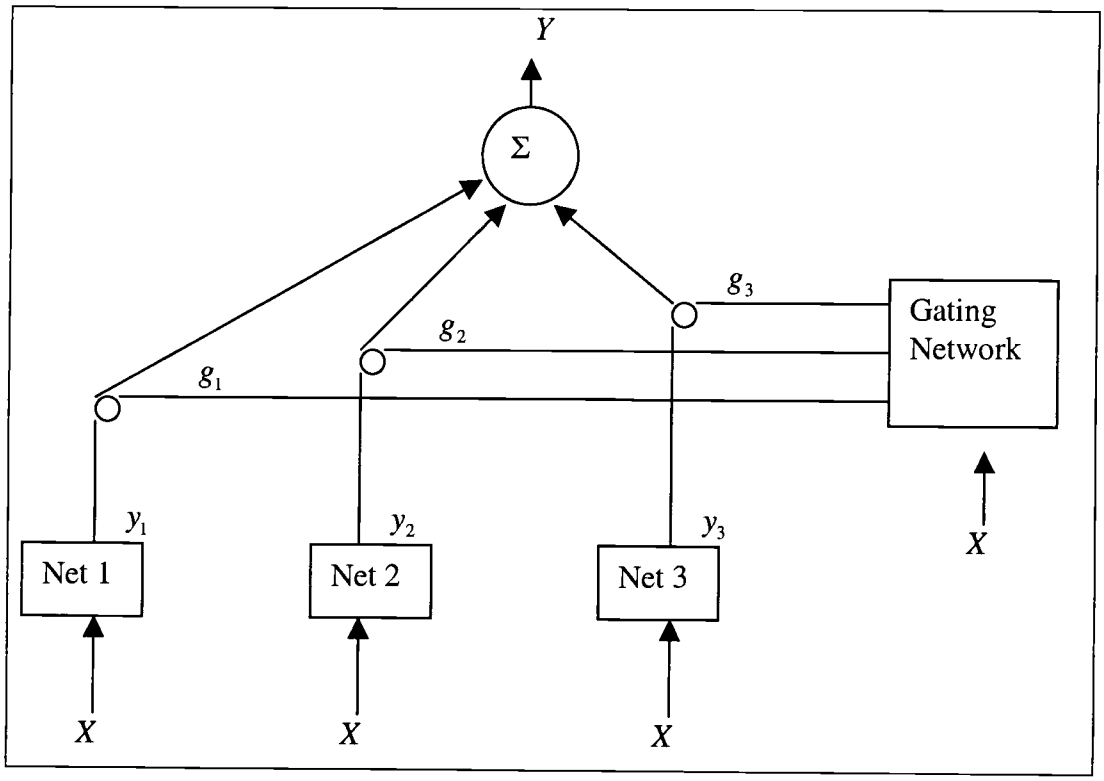


Figure 6 above corresponds to a model representing a stratified sampling design with three strata and a gating network. Each of the networks Net 1, Net 2 and Net 3 serves as a unique network for imputation in each stratum separately. The covariates are represented by the matrix X. The matrix is fed into each of the networks. The gating network serves as a portal to synchronize between the different strata. The gating network acts as dummy variables that differentiates between the different strata and assigns a sampling weight to each network. The output node is the sum of the results from the different strata. The neural network model corresponding to such a design can be formulated as:

$$\hat{y} = \sum_{j=1}^{M} \frac{1}{\rho_j(x)} f_j(y \mid x)$$

where $\rho_j(x)$ = Probability of each stratum

$$= \frac{1}{w_j(x)}$$

$w_j(x)$ = Sampling weight $= \frac{N_j}{n_j}$

such that n_j is the sample size from stratum j

N_j is the population size from stratum j

$$\sum_{j=1}^{M} \rho_j(x) = 1 \quad \text{and} \quad 0 \le \rho_j(x) \le 1$$

M = Number of Strata

$f_j(y \mid x)$ is the function representative of network j

Using a mixture of experts is a convenient way to adjust for complex designs. After the design is taken into account, the network may then be used for imputation.

4.4. Bias/Variance trade-off in ANN and inference

To calculate the mean squared error (MSE) in ANN, the expected error rate can be broken down into three main components: bias, variance and noise (Hastie, Tibshirani and Friedman, 2001). Partitioning of the MSE into bias and variance is helpful to understand the variability in ANN.

Let $y = f(x) + \xi$ where ξ is $N(0, \sigma^2)$, $\hat{y} = \hat{f}(x)$, and $E(\hat{y}) = \bar{\hat{y}}$. We have:

$$\begin{aligned} E\left[(\hat{y} - y)^2\right] &= E\left[\left(\{\hat{y} - E(\hat{y})\} + E(\hat{y}) - f(x) - y + f(x)\right)^2\right] \\ &= E\left[\left(\hat{y} - \bar{\hat{y}}\right)^2\right] + \left[E(\hat{y}) - f(x)\right]^2 + E\left[\left(y - f(x)\right)^2\right] \\ &= E\left[\left(f(x) - \bar{\hat{y}}\right)^2\right] + \left(\bar{\hat{y}} - f(x)\right)^2 + E\left[\left(y - f(x)\right)^2\right] \end{aligned}$$

$$E\left[(\hat{y}-y)^2\right] = Var[\hat{y}] + Bias(\hat{y})^2 + E\left[\xi^2\right]$$
$$= Var[\hat{y}] + Bias(\hat{y})^2 + \sigma^2$$

Therefore, $MSE = Bias^2 + Var$ where $Var = Var[\hat{y}] + \sigma^2$. The performance of the network in prediction is based on the MSE. Similar to regression, neural networks attempt to minimize the MSE. In ANN, bias arises when the true function cannot be represented correctly, i.e., under-fitted. However, variance in ANN comes from over-fitting, where the network adapts to the specific data used and cannot be generalized to new observations. When a neural network is used in imputation, the values of $\hat{y}$ are the predictions from the network, therefore $Var(\hat{y}) = Var\left(y_{imputed}\right)$, and $\sigma^2 = Var(\xi)$.

There are several ways to calculate variance and to construct confidence intervals for neural networks. Rivals and Personnaz (2000) suggest using Taylor series approximation as a variance estimation procedure. Consider a non-linear function $f(\theta)$; the estimate is:

$$f\left(\hat{\theta}\right) \overset{approximately}{\sim} N\left(f(\theta), \sigma^2 u'(X'X)^{-1} u\right)$$

where $u' = \left[\frac{\partial f(\theta)}{\partial \theta_1}, \frac{\partial f(\theta)}{\partial \theta_2}, \ldots, \frac{\partial f(\theta)}{\partial \theta_p}\right]$ is the vector of derivatives

At $\hat{\theta}$, the vector of derivatives u would become:

$$z_o' = \left[\frac{\partial f(X_o,\theta)}{\partial \theta_1}, \frac{\partial f(X_o,\theta)}{\partial \theta_2}, \ldots, \frac{\partial f(X_o,\theta)}{\partial \theta_p}\right]_{\theta=\hat{\theta}}$$

As a result, the standard error of a predicted response $f\left(X_o, \hat{\theta}\right)$ is given by $s_{f(X_o,\hat{\theta})} = \sqrt{s z_o' (X'X)^{-1} z_o}$. Therefore, approximate 100(1-α) % confidence interval on

the mean response at an arbitrary location X_0 is $f\left(X_o,\hat{\theta}\right)\pm t_{\alpha/2,m-p-1}s\sqrt{z_o^{'}\left(X'X\right)^{-1}z_o}$ and an approximate 100(1-α) % prediction limits on a new observation at x_0 is $f\left(X_o,\hat{\theta}\right)\pm t_{\alpha/2,m-p-1}s\sqrt{1+z_o^{'}\left(X'X\right)^{-1}z_o}$.

4.5. Imputation

When survey nonresponse is encountered, either nonresponse weighting or imputation may be used to handle the missing data. Imputation is the procedure of filling in the missing values. Imputation can be performed as single imputation, or repeated several times resulting in multiple imputations (Fellegi and Holt, 1976). One drawback to single imputation is the unaccounted uncertainty attributed to the imputation from the filled-in data. Multiple imputation (MI), as proposed by Rubin (1977), replaces the missing value by a vector of imputed values to obtain a number of complete data sets. Regular analysis run on the multiply imputed data sets yield estimates that are subsequently combined to get the final results. The combined estimate from a multiply-imputed data set is the average of the estimates resulting from the analysis of each completed data set separately. However, the variance of this estimate is divided into two components, the average within imputation variance and the between imputation component. The total variance is then a weighted sum of these two variance components (Little and Rubin, 2002). Inferences resulting from combining the imputations reflect the uncertainty due to nonresponse. In real data analyses, MI may not result in good performance if it is not applied properly or if the mechanisms generating either the data or the missing values depart substantially from the underlying statistical assumptions (Collins, Schafer and Kam, 2001).

Many single imputation techniques can be repeated several times resulting in multiple imputation. Allison (1999), and Horton and Lipsitz (2001) offer a review of MI. Schafer (1997) offers an extended review of techniques used for MI. In this chapter, the MCMC data augmentation technique will be used as an example. The MCMC procedures are a collection of methods for simulating random draws from the

joint distribution of $(Y_{mis}, \theta | Y_{obs})$ where Y_{mis} are the missing values of Y, Y_{obs} are the observed values of Y, and θ is the distribution parameter. This conditional distribution is assumed to be a multivariate normal distribution (Geman and Geman, 1984; Ripley, 1977). The simulated random draws result in a sequence of values that form a Markov chain (Gelman, Carlin, Stern and Rubin, 1995; Geyer, 1992; Smith and Roberts, 1992). A Markov chain is a sequence of random variables where the distribution of each element depends only on the value of the previous one and the iterative procedure consists of two steps. The first step is an imputation step (I-step), which is a draw Y_{mis} from the conditional predictive distribution $P(Y_{mis} | Y_{obs}, \theta)$ given a value for θ. The second step is a posterior step (P-step), given Y_{mis}, draw θ from its complete data posterior $P(\theta | Y_{obs}, Y_{mis})$. The goal of MCMC procedure is to sample values from a convergent Markov chain in which the limiting distribution is the joint posterior of the quantities of interest (Schimert, Schafer, Hesterberg, Fraley and Clarkson, 2000). In practice, the major challenge in using MCMC is the difficulty, for the user, to assess convergence (Gelman and Rubin, 1992). Overall, multiple imputation is difficult to implement in large data sets, due to the amount of computer memory needed to store the different, multiply-imputed data sets and the time required to run the analysis.

Increased computer power and decreased cost have encouraged more research into the automated edit and imputation techniques such as ANN. The type of ANN used in this chapter for imputation in each stratum are called feed-forward, where input terminals receive values of explanatory variables X, while the output provides the imputed variable Y. Multilayer feed-forward networks consist of one or more hidden layers. The role of the hidden layer of neurons is to intervene between the external input and the network output. Inputs and outputs are connected through neurons that transform the sum of all received input values to an output value, according to connection weights and an activation function. The connection weights represent the strength of the connection between the neurons. The network weights

(parameters) are randomly initialized and are then changed in an iterative process to reflect the relationships between the inputs and outputs. Many linear or non-linear functions are suitable candidates for an activation function. ANN do not require a model, which is advantageous in large data set imputation. Advances in computer software and increased memory have made the use of both MI and ANN more practical.

Most traditional imputation techniques do not account for sampling design during the imputation procedure (Burns, 1989). For example, multiple imputation (MI) is considered imperfect because it does not account for survey design. One solution is to run a separate imputation within each sampling subgroup and run a weighted analysis for each imputed data set. Another solution is to base MI on models that specifically include design characteristics. Binder and Sun (1996) suggest that finding accurate methods for imputation may be very difficult under complex survey design and requires a correct model for imputation. Marker, Judkins and Winglee (2002) state that variance estimation of imputed values under complex survey design has not been solved and needs further research. Remedies such as imputing the non-respondents with the sample weighted mean have been suggested (Vartivarian and Little, 2003). In this case, the weighted mean from complete cases is calculated and used for imputation.

The imputation methods discussed in the previous chapters did not account for the sampling design. These methods are appropriate when the data is selected using a simple random sample. In this chapter, imputation methods are extended to complex surveys. We will focus on computing weighted estimates for large public use data files. With large sample sizes, we assume the central limit theorem applies and the sampling distribution of the parameter estimator is approximately normal. Using the central limit theorem, different methods for including the survey design into a neural network used for imputation will be investigated in the next section.

4.6. Imputation and inference under ANN with a complex survey design

In case of nonresponse, bias needs to be quantified and both estimation and inference procedures are harder to handle (Särndal, Swensson, and Wretman, 1991). With an increasing rate of nonresponse, when the mean of the nonrespondents differs from respondents, bias increases. Therefore, the mean square error (MSE) is customarily used when comparing different estimates. According to Lee, Rancourt and Särndal (2002), there are two reasons why MSE should be considered instead of variance. First, the assumption of obtaining an unbiased estimate after imputation is not usually guaranteed. Secondly, the MSE is a measure of accuracy. In general, Total error = Sampling error + Non-sampling error. Sampling error accounts for most of the variable errors in a survey. Non-sampling error is mostly bias, caused by measurement, editing and/or imputation errors (Kish, 1965).

The estimate of the population parameters from the imputed data set includes several sources of bias. The first source of bias is from the estimate provided using traditional statistical techniques in complete case analysis. The second source of bias is due to imputation using ANN. The total MSE is defined as:

$$MSE = \left(\sum_{g} B_g \right)^2 + \sum_{v} \frac{S_v^2}{m_v}$$

In this case, the bias is the sum of the bias expected from a sample survey and the bias from the neural network estimate based on imputed values. However, analytically, the bias cannot be estimated. Therefore, most analysts estimate the variance only. The variance can be divided into several parts, as follows:

Total variance $= S^2_{obs} + S^2_{imp} + 2S_{\text{joint}}$ (Sarndal, 1992)

where

S^2_{obs} = Observed sample variance under complex survey design

S^2_{imp} = Imputation variance

$S_{\text{joint}} \overset{Asymptotically}{\approx} 0$ (Rancourt, Sarndal and Lee, 1994)

It is necessary to identify the observed and imputed values using ANN in the data file before the analysis. Assume there is a stratified random sample of size n with X observed for all sampled units. Let $\{x_{thi} : i = 1,\ldots,t\}$ and $\{x_{mhj} : j = t+1,\ldots,n\}$ denote the observed X values which correspond to the t observed Y values and m missing Y values in strata h for $h = 1,\ldots,H$, respectively. The weighted sample mean for the completed data $\bar{y}_{cw}$ can be calculated to estimate the population mean $\bar{Y}$. The standard error of $\bar{y}_{cw}$ can also be calculated to estimate the variability associated with this estimate. The weighted mean can be expressed as

$$\bar{y}_w = \sum_h W_h \bar{y}_h \quad \text{where } W_h = \frac{N_h}{N} \ \ni \ \sum_h W_h = 1$$

At each stratum we have $\bar{y}_h = \dfrac{\sum_{k=1}^{n_h} y_{hk}}{n_h} = \dfrac{1}{n_h}\left(\sum_{i=1}^{t_h} y_{thi} + \sum_{j=t_h+1}^{n_h} y_{mhj} \right)$ where the observed values are presented by y_{thi} and the imputed values presented by y_{mhj}.

Let $\bar{y}_t = \dfrac{\sum_h t_h \bar{y}_{th}}{t}$ = mean of the observed data and $\bar{y}_m = \dfrac{\sum_h m_h \bar{y}_{mh}}{m}$ = mean of the imputed data, then $\bar{y}_{cw} = (1-\pi)\bar{y}_t + \pi \bar{y}_m$ where π is the percent of missing data. The sample variance is expressed as:

$$Var(\bar{y}_w) = \sum_h W_h^2 Var(\bar{y}_h)$$

$$Var(\bar{y}_h) = \frac{1}{n-1}\left[\sum_i (y_{thi} - \bar{y}_h)^2 + \sum_j (y_{mhj} - \bar{y}_h)^2\right]$$

$$Var(\bar{y}_{cw}) = (1-\pi)^2 Var(\bar{y}_t) + \pi^2 Var(\bar{y}_m)$$

$$Var(\bar{y}_t) = \sum_h \left(\frac{t_h}{t}\right)^2 Var(\bar{y}_{th})$$

$$Var(\bar{y}_m) = \sum_h \left(\frac{m_h}{m}\right)^2 Var(\bar{y}_{mh})$$

A total weighted variance is derived by combining the variances from the complete cases and from the imputation procedure with ANN. The imputed values are given their relative importance depending on the percentage of missing data.

4.7. Results

This section contains simulation results as well as results using data from the NHIS under a complex survey design.

4.7.1 Simulation

A simulation study was performed to compare the results of imputation using non-linear ANN to multiple imputation (MI) using Markov chain Monte Carlo (MCMC) method under a complex survey design. Data for this simulation was generated using a stratified simple random design with two strata having equal allocation. The sampling weights of the two strata were set respectively to 1.33 and 4. For this simulation study, each of the two strata (Z = 1, 2) had three variables X_1, X_2 and Y, and 1000 observations. Using the Matlab software, the X's were generated separately with a normal distribution in each stratum. The Y was generated as a function of the X's with normal random error. The relationship between the X's and Y was simulated to be linear. The linear model was simulated using a linear combination of the X's and the error term using the following equation: $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \xi$. The

parameters in this model used for data simulation (α, β_1, and β_2) were set arbitrarily and separately for each stratum. A random number was generated and used to select certain observations to have missing Y values. The number of missing observations represented 10 percent of the sample size in each stratum. In addition, the missing observations were used for evaluating the performance of the imputation techniques.

Two software packages were combined in the analysis of the data to make use of several imputation techniques. SAS was used for MI with MCMC and the Matlab Neural Network toolbox was used for ANN imputation. In case of MI with MCMC, imputation was performed separately in each stratum. Weighted estimates were calculated from each stratum. The weighted estimates from the imputed data were combined using the formulas presented by Little and Rubin (2002). In ANN imputation, a separate network was used for imputation in each stratum. Online training was used to provide the network with one observation at each pass. The activation function at the hidden layer was chosen as a logistic function while the activation function at the output layer is set to be a linear function. Initial parameter values for each network were randomly assigned. Using a gating network, the results were combined based on the weights to yield the final estimates.

Table 5.
Comparison between ANN and MI using MCMC in complex survey design

	Sample size	Weighted Mean	SE
Complete cases	1800	39.9713	0.3575
ANN	2000	39.7680	0.3433
MCMC 1	2000	39.7537	0.3432
MCMC 2	2000	39.7814	0.3434
MCMC 3	2000	39.7763	0.3436
MCMC 4	2000	39.7673	0.3434
MCMC 5	2000	39.7680	0.3433
MI using MCMC combined		39.7693	0.3436

Table 5 shows the weighted results using multiple imputation with MCMC and non-linear ANN. Results show that both the weighted mean and SE resulting from MI

using MCMC and ANN are approximately equal. However, the complete case analysis provides a slightly higher weighted mean and SE.

4.7.2 Application

The National Health Interview Survey (NHIS), a health survey conducted by the National Center for Health Statistics (NCHS), Centers for Disease Control (CDC), is the principal source of information on the health of the civilian, non-institutionalized, household population of the United States. NCHS-CDC has been releasing microdata files for public use on an annual basis since 1957 (NCHS, 2002). The focus of this application is on the 2001 sample adult core survey, where one adult from each household is randomly sub-sampled to receive a questionnaire. This questionnaire collects basic information on health status, health care services and behavior of adults in the population. The U.S. Census Bureau collects the data for the NHIS by personal interviews. The sample for the last quarter (September-December) of 2001 survey consisted of 8673 adults for the sample adult component. The response rate for the sample adult component was 73.8%.

The NHIS data is obtained through a complex sample design involving stratification, clustering, and multistage sampling designed to represent the civilian, non-institutionalized population of the United States. The respondent weights are further modified by adjusting them to Census control totals for sex, age, and race/ethnicity population using post-stratification. The probability of selection for each person and adjustments for nonresponse and post-stratification are reflected in the sample weights. These weights are necessary for the analysis to yield correct estimates and variance estimation. If the data is not weighted, and standard statistical methods are used, then the estimators are overly biased and the results will be misleading. Variance estimation is suggested to be calculated using the Taylor series linearization method. For more information about the sampling design, the reader may refer to NCHS 2002.

The NHIS contains demographic information in addition to information about whether the respondent had cardiovascular disease, emphysema, asthma, ulcers, cancer, diabetes, respiratory conditions, liver conditions, joint symptoms, pain. Information is also available on the mental health of respondent (sadness, nervousness, etc.), daily activities, social activities, smoking, and the ability to perform physical tasks. Information on body mass index (BMI = weight/height^2) was also provided. In addition, sampling weights were included. A total of 73 variables were maintained in the dataset used in the imputation.

The BMI was the variable of interest for the imputation procedure approximately 4.5% of the respondents had BMI missing in this data set. Artificial neural network was used for imputation of the BMI missing values. A feed-forward network with 38 input nodes corresponding to the auxiliary variables in the data set, a hidden layer with three nodes, and one node at the output layer corresponding to the output (imputed) variable was used for imputation. The number of nodes at the hidden layer was based on multiple trials to minimize the total network error. Results of the weighted BMI mean and standard error after ANN imputation were compared to the results from the weighted analysis using the complete cases only. The results are reported in Table 6.

Table 6. Imputation results

	Sample size	Weighted Mean	SE
Complete cases	8282	26.92	0.071
ANN Imputed cases	391	27.18	0.078
Overall after ANN imputation	8673	26.93	0.065
Overall after weighted mean imputation	8673	26.92	0.068

Table 6 shows a comparison between the results from running a weighted analysis using the complex survey weights on each of the following: complete cases, imputed cases using ANN, and the full data set after imputation. Imputation was performed using ANN and using a weighted mean. The weighted mean imputation was chosen for its simplicity. Multiple imputation using MCMC was not applied to this example due to the difficulties of its application and due to the need for a model

based approach which is beyond the scope of this thesis. The comparison results in table 6 show that ANN yield an estimate with higher precision than the complete case analysis where the difference detected in the variance is estimated to be approximately eight percent. This difference in the variance is not trivial and requires further investigation in future research.

4.8. Conclusion

Design-based inference accounts for the survey design and provides reliable inferences in large samples without requiring any modeling assumptions. Variance, standard error, and tests of significance based on the assumption of independent selections are misleading and not valid for complex samples. The mean may be an acceptable estimate, but the standard error is underestimated. Measures of variability depend on the sample design and are subject to design effects. It is important to incorporate the complex survey design during the imputation procedure and in the inference after imputation.

Multiple imputation (MI) has the capability of providing a variance estimate. However, MI lacks the ability to account for the survey design in case of more complex survey designs such as NHIS. In the simulation study, the design was very simple which provided a design based analysis within each stratum. However, in the real-world data application, the design was more complex and in order to account for the design weights more research needs to be pursued. Artificial neural network represents an alternative imputation technique that requires fewer resources and offers a variance that accounts for the imputation as well as the survey design.

4.9. References

Akaike, H. (1973). Information Theory and an Extension of The Maximum Likelihood Principle. In Petrov, B.N. and Csake, F. (eds.), Second International Symposium on Information Theory. Budapest: Akademiai Kiado, 267-281.

Allison, Paul D. (1999). Multiple Imputation for Missing Data: A Cautionary Tale. (http://www.ssc.upenn.edu/~allison/MultInt99.pdf)

Binder, D.A., SUN, W. (1996). Frequency Valid Multiple Imputation for Surveys With a Complex Design. Proceedings of the Section on Survey Research Methods, American Statistical Association, 281-286.

Bishop, C.M. (1995). Neural Networks for Pattern Recognition. Oxford: Clarendon Press.

Brent, R.P. (1973). Algorithms for Minimization Without Derivatives. New Jersey: Prentice-Hall.

Brewer, K.R.W. and Mellor, R.W. (1973). The Effect of Sample Structure on Analytical Surveys. Australian Journal of Statistics, 15, 145-152.

Burton, R.M (1997). Lecture Notes, Department of Mathematics, Oregon State University.

Burton, R.M. and Dehling, H. G. (1997). Mathematical Aspects of Neural Computing, Department of Mathematics, Oregon State University.

Burns, E.M. (1989). Multiple Imputation in a Complex Sample Survey. Proceedings of the Survey Research Methods Section of the American Statistical Association, 233-238.

Casella, G. and Burger, R.L. (1990). Statistical Inference. California: Duxbury press.

Chambers, R.L. and Skinner, C.J. (eds.) (2003). Analysis of Survey Data. Chester: Wiley.

Chatfield, C. (1995). Model Uncertainty, Data Mining and Statistical Inference. Journal of Royal Statistical Society, Series A, 158 (3), 419-466.

Chen, J., Rao, J.N.K. and Sitter, R.R. (2000). Adjusted Imputation For Missing Data in Complex Surveys. Statistica Sinica, 10, 1153-1169.

Cochran, W.G. (1977). Sampling Techniques, (3rd Edition). New York: Wiley.

Collins, L.M., Schafer, J. L. and Kam, C-M. (2001) A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures, Psychological Methods, 6 (4), 330-351.

Dennis, J.E. and Schnabel, R.B. (1983). Numerical Methods for Unconstrained Optimization and Nonlinear Equations. New Jersey: Prentice-Hall.

Eason, E. D. and Fenton, R. G. (1974). A Comparison of Numerical Optimization Methods For Engineering Design. ASME Paper 73-DET-17.
Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. Annals of Statistics, 7, 1-26.

Fay, R. E. (1996). Alternative Paradigms for the Analysis of Imputed Survey Data. Journal of the American Statistical Association, 91(426), 490-498.

Fellegi, I. P. and Holt, D. (1976). A Systematic Approach to Automatic Edit and Imputation. Journal of the American Statistical Association, 71, 17-35.

Fletcher, R. and Powell, M. J. D. (1963). A Rapidly Convergent Descent Method for Minimization. Computer Journal, 6, 163-168.

Fletcher, R. and Reeves, C. M. (1964). Function Minimization by Conjugate Gradients. Computer Journal, 7, 149-154.

Fletcher, R. (1969). Optimization. New York: Academic Press.

Gelman, A.E., Carlin, J.B., Stern, H.S., and Rubin, D.B. (1995). Bayesian Data Analysis. London: Chapman & Hall.

Gelman, A.E.; and Rubin, D.B. (1992). Inference from Iterative Simulation Using Multiple Sequences. Statistical Science, 7, 457-472.

Geman, S. and Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 6, 721-741.

Geyer, C.J. (1992). Practical Markov Chain Monte Carlo, Statistical Science, 7(4).

Gill, P. E. and Murray, W. (1974). Numerical Methods for Constrained Optimization. New York: Academic Press.

Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953). Sampling Survey Methods and Theory, Vols. I and II. New York: Wiley.

Hastie, T., Tibshirani, R. and Friedman, J. (2001) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York.

Hooke, R. and Jeeves, T.A. (1961). Direct Search Solution of Numerical and Statistical Problems. Journal of the Association for Computing Machinery, 8, 212-229.

Horton, N.J. and Lipsitz, S.R. (2001). Multiple Imputation in Practice: Comparisons of Software Packages for Regression Models with Missing Variables. The American Statistician, 5(3).

Jacobs, R.A., Jordan, M.I., Nolman, S.J. and Hinton, G.E. (1991). Adaptive Mixtures of Local Experts. Neural Computation, 3, 79-87.

Jacoby, S.L.S., Kowalik, J.S., and Pizzo, J.T. (1972). Iterative Methods for Nonlinear Optimization Problems. NJ: Prentice-Hall.

Kim, J.-K. and Fuller, W. A. (1999). Jackknife Variance Estimation after Hot Deck Imputation. American Statistical Association Proceedings of the Section on Survey Research Methods, 825–830.

Kish, L. (1965). Survey Sampling. New York: Wiley.

Kish, L. (1995). The Hundred Years' Wars of Survey Sampling. Statistics in Transition, 2, 813-830.

Krewski, D. and Rao, J.N.K. (1981). Inference From Stratified Samples: Properties of the Linearization, Jackknife, and balanced Repeated Replication Methods. Annals of Statistics, 9, 1010-1019.

Lee, H., Rancourt, E. and Särndal, C.E. (2002). Variance Estimation from Survey Dara under Single Imputation. Survey Nonresponse, Groves, R.M., Dillman, D.A., Eltinge, J.L. and Little, R.J.A. (Eds). New York: John Wiley and Sons.

Little, R. (2003). To Model or Not to Model? Competing Modes of Inference for Finite Population Sampling, The University of Michigan Department of Biostatistics Working Paper Series. (http://www.bepress.com/umichbiostat/paper4)

Little, Roderick J.A. and Rubin, Donald B. (2002). Statistical Analysis with Missing Data. New Jersey: John Wiley & Sons.

Lohr, S. L. (1999). Sampling: Design and Analysis. Duxbury Press.

Mahalanobis, P.C. (1946). Recent Experiments in Statistical Sampling in The Indian Statistical Institute. Journal of the Royal Statistical Society, 109, 325-370.

Marker, D.A., Judkins, D.R. and Winglee, M. (2002). Large-Scale Imputation for Complex Surveys. Groves, R.M., Dillman, D.A., Eltinge, J.L. and Little, R.J.A (Eds.) Survey Nonresponse, New York: John Wiley and Sons.

National Center for Health Statistics (2002). Data file Documentation, National Health Interview Survey, 2001 (machine readable file and documentation). National Center

for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, Maryland.

Nelder J.A. and Mead R. (1964). A Simplex Method for Function Minimization. The Computer Journal, 7, 308-313.

Neter, J., Wasserman, W., and Kutner, M. H. (1985). Applied Linear Statistical Models: Regression, Analysis of Variance, and Experimental Designs. Homewood, IL: Irwin.

Peressini, A. L., Sullivan, F. E., and Uhl, J. J., Jr. (1988). The mathematics of nonlinear programming. New York: Springer.

Quenouille, M.H. (1949). Problems in Plane Sampling. Annals of Mathematical Statistics, 20, 355-375.

Ramsey, F.L. and Schafer, D.W. (1996). The Statistical Sleuth: A Course in Methods of Data Analysis. USA: Duxbury Press

Rancourt, E. , Särndal, C.-E., and Lee, H. (1994). Estimation of the Variance in Presence of Nearest Neighbor Imputation. Proceedings of the Section on Survey Research Methods, American Statistical Association, 888-893.

Rao, J. N. K. and Shao, J. (1992). Jackknife Variance Estimation with Survey Data Under Hot Deck Imputation. Biometrika, 79, 811-822.

Rao, J.N.K. and Wu, C.F.J. (1985). Inference From Stratified Samples: Second-order Analysis of Three Methods for Nonlinear Statistics. Journal of The American Statistical Association, 80, 620-630.

Ripley, B.D. (1977). Modeling Spatial Patterns, Journal of the Royal Statistical Society, Series B, 39, 172-212.

Rivals, I. and Personnaz, L. (2000). Construction of Confidence Intervals for Neural Networks Based on Least Squares Estimation. Neural Networks, 13, 463-484

Rubin D.B. (1977). Formalizing Subjective Notions About the Effect of Non-respondents in Sample Surveys. Journal of the American Statistical Association, 77, 538-543.

Rubin, D. B. (1978). Bayesian Inference for Causal Effects: The Role of Randomization. Annals of Statistics, 6, 34-58.

Rubin, D. B. (1987). Multiple Imputation for Non-response in Surveys. New York: Wiley.

Särndal, C.-E., Swensson, B. and Wretman, J. (1991). .Model Assisted Survey Sampling. Springer-Verlag.
Särndal, C.-E. (1992). Methods for Estimating the Precision of Survey Estimates When Imputation Has Been Used. Survey Methodology, 18, 241-265.

Schafer, J.L. (1997). Analysis of Incomplete Multivariate Data. London: Chapman and Hall.

Schimert, J., Schafer, J.L., Hesterberg, T.M., Fraley, C., Clarkson, D.B. (2000). Analyzing Data with Missing Values in S-Plus. Seattle: Insightful Corp.

Shao, J. and Tu, D. (1995). The Jackknife and Bootstrap. New York: Springer-Verlag.

Smith, A.F.M. and Roberts, G.O. (1992). Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods. Journal of the Royal Statistical Society, Series B, 5(1).

StatSoft, Inc. (2004). Electronic Statistics Textbook. Tulsa, OK: StatSoft. (http://www.statsoft.com/textbook/stathome.html)

Tukey, J.W. (1958). Bias and Confidence in Not-quite Large Samples. Annals of Mathematical Statistics, 29, 614.

Vartivarian, S.L. and Little, R.J. (2003). Weighting Adjustments for Unit Nonresponse with Multiple Outcome Variables. The University of Michigan Department of Biostatistics Working Paper Series. Working Paper 21.
(http://www.bepress.com/umichbiostat/paper21)

Wilde, D. J. and Beightler, C. S. (1967). Foundations of optimization. Englewood Cliffs. NJ: Prentice-Hall.

Wolter, K. M. (1985). Introduction to Variance Estimation. Springer-Verlag.

Woodruff, R.S. (1971). A Simple Method For Approximating The Variance of a Complicated Estimate. Journal of the American Statistical Association, 66, 411-414.

5. CONCLUSIONS

5.1. Summary

Missing data complicates any statistical analysis. The simplest way is to discard the entire unit, and use a different number of cases in different tables, or provide the percentage of missing observations in each table of the analysis. However, editing and imputation of the missing data may be more effective than discarding observations with missing variables (Kish, 1965). Imputation is a laborious process that should be managed carefully. An effective imputation method has several important properties, including the ability to condition on the observed variables to make use of auxiliary information, and the ability to preserve marginal distributions, joint distributions, and the association between variables. In large public-use data sets, employing a complex sampling design (e.g. NHIS), the imputer needs to account for the sample design as well as the uncertainty from the imputation process, while taking into consideration the general patterns of missing data. The challenge of the imputation reside in the need to maximize the use of available data while reducing the mean square error and preserving the covariance structure of the data (Marker, Judkins and Winglee, 2002). All imputation methods require a model for the missing data mechanism (either implicitly or explicitly), assumptions about the pattern of missing data and the mechanism that generated such pattern. Most imputation methods assume a multivariate normal distribution of the variables, no interactions between the auxiliary variables, linear relationships between response and auxiliary variables, and a simple random sample.

In choosing an imputation method to account for missing data, the researcher should consider the size of the sample (i.e. number of observations in the data file), the percentage of missing values, the reason the data are missing, and the data user. All imputation methods have disadvantages; however, ideally the researcher chooses the method with the fewest disadvantages. Details about the imputation should be reported.

5.1.1. Summary of the results

Overall, nine imputation methods were compared in this study to analyze data with an ignorable missing mechanism. Deterministic methods such as hot deck, mean, and regression imputations were used. These methods are considered to be biased and inefficient under the assumption of data missing at random, generally underestimate variance and overstate sample size although they are relatively simple. Maximum likelihood imputation procedures, such as EM, are efficient and consistent when used with the correct model, are able to specify a model for joint distribution, and estimate parameters to maximize likelihood. However, when used in the simulations of this research, the EM algorithm resulted in higher error compared to other techniques. Multiple imputation (MI) produces reasonable point estimates and measures of uncertainty. Multiple imputation has attracted attention because of its ability to account for imputation uncertainty, but it has not been used routinely because of the required effort to apply it with large datasets. Multiple imputation is a model-based procedure and is susceptible to model misspecification. Multiple imputation is computationally intensive, and the results depend on the imputation technique used to achieve the multiple imputations. For instance, in this thesis, when MI was used with propensity scores, the results of the imputations were not as efficient as when MI was used with MCMC. In addition, MI does not easily incorporate survey design features. Therefore, MI and MI-inference are too demanding for practical imputations. Linear and non-linear ANN were also used for imputation. Linear ANN gave similar results to linear regression, and non-linear ANN had similar results to MI with MCMC.

One public-use data example, the NHIS data (2002), was followed in the different chapters to compare the results in an actual setting. For simplicity, chapters 2 and 3 investigated the performance of the nine imputation techniques assuming a simple random sample. Table 7 below summarizes the results obtained from the different imputation techniques in case of the NHIS data set. These results show that non-linear ANN resulted in the smallest error with differences ranging from one percent and up to 16 percent in error compared to other imputation techniques.

Table 7. NHIS results

	MAPE	*MAD*	*MSD*
Hot deck	20.6	5.8	57.8
Mean	16.2	4.3	31.1
EM	16.04	4.27	31.09
MI			
With Propensity Score	8.81	2.51	17.20
With MCMC	9.52	2.49	11.06
Stochastic Regression	9.4	2.6	17.2
Regression [a]	5.7	1.6	5.4
Linear ANN	5.6	1.6	5.9
Nonlinear ANN	4.75*	1.21*	5.17*

a. The R^2 from the regression model was 0.82

5.1.2. Artificial neural networks

Artificial neural networks (ANN) offer an easier alternative that might attract survey samplers. Artificial neural network is a powerful, automatic, easy to use technique. It combines deterministic and stochastic methods for imputation while allowing a broad range of relationships between the variables. There is no specific model is imposed on the data and the imputer can use an unlimited number of input variables. Artificial neural network provides better convergence results with larger data sets. It can also provide a variance estimate which takes into consideration the relationships between the variables in the network. The flexibility of the network structure provides the possibility of accounting for complex survey design as well. Literature demonstrates that neural networks have been used effectively as data analysis techniques. The main objective in ANN is generalization/prediction.

Since artificial neural networks are semi-parametric procedures based on an iterative process, ANN require the analyst to determine initial parameter values and the number of nodes to use in the network. Data analysts should be familiar with ANN and a data mining software package that includes neural networks. Like other imputation techniques, the data analyst needs to have experience working with ANN. For example, Curdas and Chambers (1997) had to commission a professional company to run ANN for their research. Wilmot and Shivananjappa (2001) state that more experience in the development and use of ANN could lead to an improvement in

the performance of ANN in imputation. This dissertation reviewed the basic theoretical background of ANN and linked it to statistical terminology and techniques in imputation problems.

The primary difference between ANN and other statistical techniques when dealing with missing values is the method used for data processing. In general, statistical analysis procedures handle all the observations in the data set concurrently. With statistical analysis, processing is mostly done by batch and the data are used only once. For example, regression incorporates all observations at the same time to estimate the regression line. In neural networks, processing is sequential. ANN use an online training procedure where observations are presented to the system one at a time. Each datum in the sample is fed into the network repeatedly until the network learns the association of input to output. This online procedure is a non-probabilistic framework which avoids unnecessary assumptions about the data distribution.

Traditional statistical techniques depend on structured models for interpretability which make them more computationally complex. While providing asymptotically good estimates, statistical techniques are less robust to multicollinearity and outliers. Artificial neural networks make few assumptions about the data, are asymptotically good and more robust to multicollinearity and outliers. Artificial neural networks could be considered as implicit models that take into account the sample design without making strong parametric assumptions. Avoiding strong parametric assumptions can produce reliable and efficient inferences in survey settings (Little, 2003). Therefore, ANN represent a favorable alternative to regression and to MI (Little, 2003). This was illustrated with a simple design, stratified random sample, in this dissertation. Overall, ANN could be time and resources efficient for an experienced user.

5.1.3. Sources of error

Surveys are susceptible to several sources of error classified as sampling or non-sampling errors (Särndal, Swensson and Wretman, 1991). Sampling error is

caused by selecting part of the population, rather than the whole population. Under probability sampling and by incorporating inclusion weights, the sampling error in a survey can be estimated. Non-sampling error is attributed to bias from frame error, measurement error, coding, editing, imputation, or non-response errors. When ANN are used for imputation, network errors are part of the imputation procedure. Therefore, ANN error can be classified as non-sampling error similar to measurement error. A neural network has several sources of error that contribute to the final output error from the network. Although we cannot divide the final network output error into its different components, it seemed beneficial to briefly explain these errors. Overall, the network errors contain three components: generalization, training, and induced errors.

The generalization error is the probability that the trained neural network, using the observed training set makes a wrong decision for a new input (Karayiannis and Venetsanopoulos, 1993). The estimated generalization error for an application depends on the selected random training set. However, the generalization error is difficult to calculate and is substituted by the training error. During the ANN training phase, the network learns the empirical distribution of the data by example using a collection of input vectors and corresponding outputs from a training data set. During the learning process, the neural network attempts to minimize the training error which constitutes of local and global parts (Karayiannis and Venetsanopoulos, 1993). The first part of the training error is the local error, which is generated during each training example. The second part is the global error, which is the average of the local errors. Both local and global errors measure the deviation between the desired output and the actual output (Burton and Dehling, 1997). The objective is to minimize the global error.

The induced error accounts for the imputation error and avoids converging to local minima. In order to avoid converging to a local minima during the ANN optimization process, Burton (1997) proposed a noise algorithm to enhance the descent. This noise is not found in every neural network, it is only found if the

algorithm proposed by Burton (1997) is used. The procedure consists of injecting noise into the algorithm in such a way that the variance of the noise decreases to zero exponentially. This noise is the induced error. This algorithm is adequate for an error surface with few local minima. An alternative algorithm, the time invariant noise algorithm (TINA) proposed by Burton and Mpitsos (1992) is preferred if we suspect that the error surface has more local minima.

5.1.4. Software overview and comments

Running the data analysis through software without understanding what the software is estimating and how it is estimated can lead to unreliable conclusions. When dealing with missing data imputation in large datasets, the researcher needs to make a choice between different software packages for imputation and analysis. In this thesis, we had access to SAS version 9 proc impute, SPSS 11 missing value analysis, SOLAS 1.1 Software for missing data, and MATLAB 13 student version with the Neural Network toolbox. The time commitment varied based on the software used. However, if the user is familiar with the different software packages, the run time of the imputation using ANN was less than half the run time using other statistical procedures.

Part of the poor performance of some of the imputation techniques explored in this thesis compared to ANN could be explained by the software packages used. Allison (1999) show that SOLAS yields biased estimates from the multiple imputation procedure using propensity scores with data missing at random. A recent paper by Von Hippel (2004) shows that the stochastic regression imputation and the EM algorithm presented in SPSS software yield biased estimates when values are missing at random.

We should also acknowledge that the performance of ANN depends on the software used as well as the experience of the user. Artificial neural network was investigated using MATLAB neural network toolbox. Other software packages are

available for ANN (e.g. WinNN, STATISTICA, NeuroSolutions) and may be more user friendly but we were restricted by software availability.

5.2. Further research ideas

While allowing for several input variables, we focused only on univariate imputation in this thesis. However, multivariate imputation is accomplished as a set of sequential univariate imputations. Therefore, in order to impute more than one variable in the data set, the researcher can impute one variable after the other. We used all variables in the data set, except the response variable, as the inputs to the network. However, selecting good inputs is a hard task that needs to be considered (Bishop, 1995). Each input to the network adds a new dimension. To avoid an over-parameterized network with many nodes, we need to select a smaller number of input variables. Some correlated input variables can explain the same information; a choice of a representative subset would be appealing. Variable selection is an important part of neural network design. Statistical test, combinations of inputs, and sensitivity analysis are possible solutions to the variable selection procedure that needs to be taken into account. Another important research point is the comparison between the notions of degrees of freedom and the Vapnik-Chervonenkis dimension (Vapnik and Chervonenkis, 1971) as indicators of the complication level of the network versus the statistical modeling technique. To allow a wider scope of research, a comparison between ANN, statistical modeling and compartmental modeling seems to be necessary. In this thesis, the output variable assumed was continuous. If the output variable is categorical or nominal, either ANN or classification trees could be used as alternatives to discriminant analysis and logistic regression.

5.3. Conclusions

No one imputation method works best in all data sets for filling in missing values. The imputation results depend on the dataset and variables (e.g. scale of measurement, dependencies between variables, type of missingness). A good training dataset is critical for calibration and developing a strategy in NN analysis. Data should

always be explored using some descriptive statistics and graphs prior to editing and imputation to learn about relationships. An editing and imputation approach is a mixture of methods tuned to each particular dataset. Naïve users will not get maximum benefit from complex imputation methods and may do better with simpler less efficient imputations. In some cases, a mixed imputation procedure is used to handle different variables in the data set. For example, the imputer could use a mean imputation to estimate some values. This could be used jointly with hot-deck imputation estimating other values, followed by a model-based imputation. Experience has shown that combining the predictions from multiple methods often yields more accurate predictions than can be derived from any one method (Witten and Frank, 2000).

The objective in all imputation analyses is to better describe the data and not to create it. Imputers are not seeking optimal point prediction, but valid statistical inference while properly reflecting uncertainty, preserving important aspects of data distributions, and preserving important relationships (Rubin, 1996). Imputation methods are not all equally good, and none are universally good. Finally, skepticism surrounds any new method (e.g. ANN imputation) which has not been well established in a specific application. This skepticism does not diminish the importance of use of these new methods and their evaluation.

5.4. References

Allison, Paul D. (1999). Multiple Imputation for Missing Data: A Cautionary Tale. (http://www.ssc.upenn.edu/~allison/MultInt99.pdf)

Bishop, C. M. (1995). Neural Networks for Pattern Recognition. Oxford: University Press.

Burton, R.M (1997). Lecture Notes, Department of Mathematics, Oregon State University.

Burton, R.M. and Dehling, H. G. (1997). Mathematical Aspects of Neural Computing, Department of Mathematics, Oregon State University.

Burton, R.M. and Mpitsos, G.J. (1992). Event-dependent Control of Noise Enhances Learning in Neural Networks. Neural Networks, 5, 627-637.

Chen, J., Rao, J.N.K. and Sitter, R.R. (2000). Adjusted Imputation for Missing Data in Complex Surveys. Statistica Sinica, .10, 1153-1169.

Coppola, L., Zio, M.D., Luzi, O., Ponti, A., and Scanu, M. (2001). Bayesian Networks for Imputation, ISTAT, via Cesare Bablo 14, 00184 Roma.

Curdas, Marie and Chambers, Ray (1997). Neural Network Imputation: Statistical Evaluation. Conference of European Statisticians.

Fausett, L. (1994). Fundamentals of Neural Networks: Architectures, Algorithms and Applications. Prentice Hall.

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

Karayiannis, N.B. and Venetsanopoulos, A. N. (1993). Artificial Neural Networks: Learning Algoriyhms, Performance Evaluation, and Applications. Massachusetts: Kluwer Academic Publishers.

Kish, L. (1965) Survey Sampling, New York: Wiley.

Kohonen, T. (1982). Self-organized Formation of Topologically Correct Feature Maps. Biological Cybernetics, 43, 59-69.

Little, R. (2003). To Model or Not to Model? Competing Modes of Inference for Finite Population Sampling. (www.bepress.com/umichbiostat/paper4).

Little, R. (2003). Discussion at FCSM Conference.
(http://sitemaker.umich.edu/rlittle/files/fcsmdisc.pdf).

Marker, D.A., Judkins, D.R. and Winglee, M. (2002). Large-Scale Imputation for Complex Surveys. Groves, R.M., Dillman, D.A., Eltinge, J.L. and Little, R.J.A (Eds.) Survey Nonresponse, New York: John Wiley and Sons.

National Center for Health Statistics (2002). Data file Documentation, National Health Interview Survey, 2001 (machine readable file and documentation). National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, Maryland.

Patterson, D.W. (1996). Artificial Neural Networks: Theory and Applications. Singapure: Prentice Hall.

Piela, P. (2001). Tree-Structured Self-Organizing Maps for Imputation, Statistics Finland (Euredit project).

Rubin, D.B. (1996). Multiple Iimputation After 18+ Years. Journal of American Statistical Association, 91, 473-489.

Särndal, C.-E., Swensson, B. and Wretman, J. (1991). Model Assisted Survey Sampling. Springer-Verlag.

Schafer, J.L. and Graham, J.W. (2002). Missing Data: Our View of the State of the Art. Psychological Methods, 7(2), 147-177.

Southcott, M.L. and Bogner, R.E. (1993). Classification of Incomplete Data Using Neural Networks. ACNN'93, 220-223.

Titterington, D.M. and Cheng, B. (1994). Neural Networks: A Review from a Statistical Perspective. Statistical Sciences, 9 (1), 2-54.

Vapnik, V. and Chervonenkis, A. (1971). On the uniform convergence of relative frequencies of events to their probabilities. Theory of Probability and its Applications, 16(2),264-280.

Von Hippel, P.T. (2004).Biases in SPSS 12.0 Missing Value Analysis. The American Statistician, 58 (2), 160-164.

Wayman, J.C. (2003). Multiple Imputation For Missing Data: What Is It And How Can I use It?. The 2003 Annual Meeting of the American Educational Research Association. Chicago, IL.

Wilmot, C.G. and Shivananjappa, S. (2001). Comparison of Hot Deck and Neural-Network Imputation.

Witten, I.H. and Frank, E. (2000). Data mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco, CA.

BIBLIOGRAPHY

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

Afifi, A.A. and Elashoff, R.M. (1967). Missing Observations in Multivariate Statistics II, Point Estimation in Simple Linear Regression. Journal of American Statistical Association, 62, 10-29.

Akaike, H. (1973). Information Theory And An Extension of The Maximum Likelihood Principle. In Petrov, B.N. and Csake, F. (eds.), Second International Symposium on Information Theory. Budapest: Akademiai Kiado, 267-281.

Allison, Paul D. (1999) Multiple Imputation for Missing Data: A Cautionary Tale. (http://www.ssc.upenn.edu/~allison/MultInt99.pdf).

Allison, Paul D. (2000). Multiple Imputation for Missing Data: A Cautionary Tale. Sociological Methods and Research, 28, 301-309.

Amari, S. (993). Universal Theorem on Learning Curves. Neural Networks, 6, 161-166.

Anderson, D. and McNeill, G. (1992). Artificial Neural Network Technology, A DACS State-of-the-Art Report. Kaman Sciences Corporation.

Anderson, J.A., Pellionisz, A. and Rosenfeld, E. (Eds.). (1990). Neurocomputing 2: Directions for Research. Cambridge, Massachusetts: MIT Press.

Anderson, T.W. (1957). Maximum Likelihood Estimates for a Multivariate Normal Distribution When Some Observations are Missing. Journal of the American Statistical Association, 52, 200-203.

Bard, Y. (1974). Nonlinear Parameter Estimation. New York: Academic Press.
Bates, Douglas M. and Watts, Donald G. (1988). Nonlinear Regression Analysis and its Applications. New York: Wiley.

Beale, E.M.L. and Little, R.J.A. (1975). Missing Values in Multivariate Analysis. Journal of Royal Statistical Society, Series B, 37, 129-146.

Binder, D.A., SUN, W. (1996). Frequency Valid Multiple Imputation forSsurveys With a Complex Design. Proceedings of the Section on Survey Research Methods, American Statistical Association, 281-286.

Bishop, C.M. (1995). Neural Networks for Pattern Recognition. New York: Oxford University Press.

Box, G. E. P. and Lucas, H. L. (1959). Design of Experiments in Non–linear Situations. Biometrika, 46, 77–90.

Box, G.E.P. (1960). Fitting Experimental Data. Annals of the New York Academy of Sciences, 86, 792-816.

Box, G.E.P. and Hill, W.J. (1967). Discrimination Among Mechanistic Models. Technometrics, 9(1), 57-71.

Box, G.E.P. and Hill, W.J. (1974). Correcting Inhomogeneity of Variance with Power Transformation Weighting. Technometrics, 13(3), 385-389.

Box, G.E.P. and Hunter, W.G. (1962). A Useful Method for Model-building. Technometrics, 4, 301-318.

Box, G.E.P. and Hunter, W.G. (1965). The Experimental Study of Physical Mechanisms. Technometrics, 7(1), 23-42.

Box, G.E.P., Hunter, W.G. and Hunter, J.S. (1978). Statistics for Experimenters. New York: Wiley.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). Classification and Regression Trees. Wadsworth.

Brent, R.P. (1973). Algorithms for Minimization Without Derivatives. New Jersey: Prentice-Hall.

Brewer, K.R.W. and Mellor, R.W. (1973). The Effect of Sample Structure on Analytical Surveys. Australian Journal of Statistics, 15, 145-152.

Buck, S.F. (1960). A Method of Estimation of Missing Values in Multivariate Data Suitable for Use with an Electronic Computer. Journal of Royal Statistical Society, Series B, 22, 302-306.

Burns, E.M. (1989). Multiple Imputation in a Complex Sample Survey. Proceedings of the Survey Research Methods Section of the American Statistical Association, 233-238.

Burton, R.M (1997). Lecture notes, Department of Mathematics, Oregon State University.

Burton, R.M. and Dehling H.G. (1997). Mathematical Aspects of Neural Computing, Oregon State University.

Burton, R.M. and Mpitsos, G.J. (1992). Event-dependent Control of Noise Enhances Learning in Neural Networks. Neural Networks, 5, 627-637.

Casella, G. and Burger, R.L. (1990). Statistical Inference. California: Duxbury press.

Chambers, J.R. (1977). Computational Methods for Data Analysis. New York: Wiley.

Chambers, R.L. and Skinner, C.J. (eds.) (2003). Analysis of Survey Data. Chester: Wiley.

Chatfield, C. (1995). Model Uncertainty, Data Mining and Statistical Inference. Journal of Royal Statistical Society, Series A, 158 (3), 419-466.

Chatfield, C. (2002). Confessions of a Pragmatic Statistician. The Statistician, 51(1), 1-20.

Chen, J., Rao, J.N.K. and Sitter, R.R. (2000). Adjusted Imputation for Missing Data in Complex Surveys. Statistica Sinica, 10, 1153-1169.

Cheng B. and Titterington D.M. (1994). Neural Networks: A review from a statistical perspective.

Cherkassky, V. and Mulier, F. (1994). Statistical and Neural Network Techniques for Non-Parametric Regression. In: Cheeseman, P. and Oldford, R.W., Selecting models from data: Artificial Intelligence and Statistics IV. New York: Springer.

Cherkassky, V. and Mulier, F. (1998). Learning from data. New York: Wiley.

Citro, C.F., Cork, D.L. and Norwood, J.L. (Eds.) (2002). The 2000 Census: Interim Assessment. Washington, D.C.: National Academy Press.

Clarck, Alex (1996). Planning for the 2001 Census of the United Kingdom. (http://www.census.gov/prod/2/gen/96arc/xastreet.pdf)

Cochran, W. (1963). Sampling Techniques. (2nd edition) New York: Wiley and Sons.

Cochran, W.G. (1977). Sampling Techniques, (3rd Edition). New York: Wiley.

Collins, L.M., Schafer, J. L. and Kam, C-M. (2001) A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures, Psychological Methods, 6 (4), 330-351.

Coppola, L., Zio, M.D., Luzi, O., Ponti, A., and Scanu, M. (2001). Bayesian Networks for Imputation, ISTAT, via Cesare Bablo 14, 00184 Roma.

Curdas, Marie and Chambers, Ray (1997). Neural Network Imputation: Statistical Evaluation. Conference of European Statisticians.

De Wilde, P. (1995). Neural Network Models: Analysis, Lecture Notes in Control and Information Sciences, Vol. 210. New York: Springer-Verlag.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, Series B, 39, 1-38.

Dempster, A.P. and Rubin, D.B. (1983). Incomplete Data in Sample Surveys: Theory and Bibliography. New York: Academic Press.

Dennis, J.E. and Schnabel, R.B. (1983). Numerical Methods for Unconstrained Optimization and Nonlinear Equations. New Jersey: Prentice-Hall.

Dillman, D.A., Eltinge, J.L., Groves, R.M. and Little, R.J.A. (2002). Survey Nonresponse in Design, Data Collection, and Analysis. In Survey Nonresponse by Groves, R.M., Dillman, D.A., Eltinge, J.L. and Little, R.J.A. (Eds.). New York: John Wiley and Sons.

Draper, N.R. and H. Smith (1981). Applied Regression Analysis. New York: John Wiley and Sons.

Eason, E. D. and Fenton, R. G. (1974). A Comparison of Numerical Optimization Methods For Engineering Design. ASME Paper 73-DET-17.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. Annals of Statistics, 7, 1-26.

Efron, B. (1994). Missing Data, Imputation, and the Bootstrap. Journal of the American Statistical Association, 89, 463-474.

Fausett, L. (1994). Fundamentals of Neural Networks: Architectures, Algorithms and Applications. Prentice Hall.

Fay, R. E. (1996). Alternative Paradigms for the Analysis of Imputed Survey Data. Journal of the American Statistical Association, 91(426), 490-498.

Fellegi, I. P. and Holt, D. (1976). A Systematic Approach to Automatic Edit and Imputation. Journal of the American Statistical Association, 71, 17-35.

Fetter, M. (2001). Mass Imputation of Agricultural Economic Data Missing by Design: A Simulation Study of Two Regression Based Techniques. Federal Committee on Statistical Methodology Research Conference.

Fichman, M., and Cummings, J. (2003). Multiple Imputation for Missing Data: Making the Most of What You Know. Organizational Research Methods, 6(3), 282-308.

Fisher, R.A. (1922). On the Mathematical Foundations of Theoretical Statistics. Philosophical Transactions of the Royal Society of London, Series A, 222, 309-368.

Fisher, R.A. and Mackenzie, W.A. (1923). The Manorial Response of Different Potato Varieties. Journal of Agricultural Science, 13, 311-320.

Fitzmaurice, G.M., Heath, A.F. and Clifford, P. (1996). Logistic Regression Models for Binary Panel Data with Attrition. Journal of the Royal Statistical Society, Series A (Statistics in Society), 159 (2), 249-263.

Fletcher, R. (1969). Optimization. New York: Academic Press.

Fletcher, R. and Powell, M. J. D. (1963). A Rapidly Convergent Descent Method for Minimization. Computer Journal, 6, 163-168.

Fletcher, R. and Reeves, C. M. (1964). Function Minimization by Conjugate Gradients. Computer Journal, 7, 149-154.

Furlong, J., Dupuy, M. and Heinsimer, J. (1991). Neural Network Analysis of Serial Cardiac Enzyme Data. American Journal of Clinical Pathology, 96, 134-141.

Gallant, A.R. (1975). Nonlinear Regression. American Statistician, 29(2), 73-81.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (1995). Bayesian Data Analysis. Chapman.

Gelman, A.E and Rubin, D.B. (1992). Inference from Iterative Simulation Using Multiple Sequences. Statistical Science, 7, 457-472.

Gelman, A.E., Carlin, J.B., Stern, H.S., and Rubin, D.B. (1995). Bayesian Data Analysis. London: Chapman and Hall.

Geman, S. and Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 6, 721-741.

Geman, S., Bienenstock, E. and Doursat, R. (1992). Neural Networks and the Bias/Variance Dilemma. Neural computation, 4, 1-8.

Geyer, C.J. (1992). Practical Markov Chain Monte Carlo, Statistical Science, 7(4).

Gill, P. E. and Murray, W. (1974). Numerical Methods for Constrained Optimization. New York: Academic Press.

Graham, J.W., Hofer, S.M., and MacKinnon, D.P. (1996). Maximizing the Usefulness of Data Obtained with Planned Missing Value Patterns: An Application of Maximum Likelihood Procedures. Multivariate Behavioral Research, 31, 197-218.

Grossberg, S. (1974). Classical and Instrumental Learning by Neural Networks. Progress in Theoretical Biology, 3, 51-141.

Hand, D.J. (1984). Statistical Expert Systems: Design. The Statistician, 33, 351-369.

Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953). Sampling Survey Methods and Theory, Vols. I and II. New York: Wiley.

Hartley, H.O. (1958). Maximum Likelihood Estimation from Incomplete Data. Biometrics, 14, 174-194.

Hartley, H.O. and Hocking, R.R. (1971). The Analysis of Incomplete Data. Biometrics, 27, 783-823.

Hastie, T., Tibshirani, R. and Friedman, J (2001). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer.

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

Healey C. (1999). Semi-Continuous Cardiac Output Monitoring using a Neural Network. Critical Care Medicine, 27(8), 1505-1510.

Hebb, D. (1949). The Organization of Behavior: A Neuropsychological Theory. New York: Wiley.

Heitjan, D.F. and Rubin, D.B. (1991). Ignorability and Coarse Data. Annals of Statistics, 19, 2244-2253.

Herzog, T.N. and D.B. Rubin (1983). Using Multiple Imputations to Handle in Nonresponse in Sample Surveys. Incomplete Data in Sample Surveys, Vol.II: Theory and Annotated Bibliography (W.G.Madow, I.Olkin, and D.B.Rubin, Eds.). Academic Press.

Hocking, R. and Smith, W.R. (1968). Estimation of Parameters in the Multivariate Normal Distribution with Missing Observations. Journal of the American Statistical Association, 63, 159-173.

Hocking, R. R. and Smith, W. B. (1972). Optimum Incomplete Multinormal Samples, Technometrics, 4, 299-307.

Hooke, R. and Jeeves, T.A. (1961). Direct Search Solution of Numerical and Statistical Problems. Journal of the Association for Computing Machinery, 8, 212-229.

Hopke, P. K., Liu, C., and Rubin, D.B. (2001). Multiple Imputation for Multivariate Data with Missing and Below-Threshold Measurements: Time Series Concentrations of Pollutants in the Arctic. Biometrics, 57, 22-33.

Horton, N.J. and Lipsitz, S.R. (2001). Multiple Imputation in Practice: Comparisons of Software Packages for Regression Models with Missing Variables. The American Statistician, 5(3).

Jacobs, R.A., Jordan, M.I., Nolman, S.J. and Hinton, G.E. (1991). Adaptive Mixtures of Local Experts. Neural Computation, 3(1), 79-87.

Jacoby, S.L.S., Kowalik, J.S., and Pizzo, J.T. (1972). Iterative Methods for Nonlinear Optimization Problems. NJ: Prentice-Hall.

Johnson, N.L. and Kotz, S. (1970). Distributions in Statistics: Continuous Univariate Distributions I, New York: Wiley.

Jones, R. H. (1980). Maximum Likelihood Fitting of ARMA Models to Time Series with Missing Observations. Technometrics, 22 (3), 389-395.

Karayiannis, N.B. and Venetsanopoulos, A. N. (1993). Artificial Neural Networks: Learning Algoriyhms, Performance Evaluation, and Applications. Massachusetts: Kluwer Academic Publishers.

Kennedy, William J. and Gentle, James E. (1980). Statistical Computing. New York: Marcel Dekker.

Kim, J.-K. and Fuller, W. A. (1999). Jackknife Variance Estimation after Hot Deck Imputation. American Statistical Association Proceedings of the Section on Survey Research Methods, 825–830.

Kim, J.O. and Curry, J. (1977). Treatment of Missing Data in Multivariate Analysis. Sociological Methods and Research, 6, 215-240.

King, G., Honaker, J., Joseph, A., and Scheve, K. (2001). Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation. American Political Science Review, 95 (1), 49-69.

Kish, L. (1965) Survey Sampling, New York: Wiley.

Kish, L. (1995). The Hundred Years' Wars of Survey Sampling. Statistics in Transition, 2, 813-830.

Kohonen, T. (1982). Self-Organized Formation of Topologically Correct Feature Maps. Biological Cybernetics, 43, 59-69.

Kotz, S., Read, C. B. and Banks, D. L. (Eds.) (1998). Encyclopedia of Statistical Sciences. Wiley-Interscience.

Krewski, D. and Rao, J.N.K. (1981). Inference From Stratified Samples: Properties of the Linearization, Jackknife, and balanced Repeated Replication Methods. Annals of Statistics, 9, 1010-1019.

Lawton, W.H., Sylvester, E.A. and Maggio, M.S. (1972). Self Modeling Non-linear Regression. Technometrics, 14, 513-532.

Lee, H., Rancourt, E. and Särndal, C.E. (2002). Variance Estimation from Survey Dara under Single Imputation. Survey Nonresponse, Groves, R.M., Dillman, D.A., Eltinge, J.L. and Little, R.J.A. (Eds). New York: John Wiley and Sons.

Lessler, J.T. and Kalsbeek, W.D. (1992). Nonsampling Error in Surveys. John Wiley and Sons.

Levenberg, K. (1944). A Method for the Solution of Certain Non-linear Problems in Least Squares. Quarterly of Applied Mathematics, 2, 164-168.

Lewicki, P., Hill, T., and Czyzewska, M. (1992). Nonconscious Acquisition of Information. American Psychologist, 47, 796-801.

Little, R. (2003). Discussion at FCSM Conference. (http://sitemaker.umich.edu/rlittle/files/fcsmdisc.pdf).

Little, R. (2003). To Model or Not to Model? Competing Modes of Inference for Finite Population Sampling, The University of Michigan Department of Biostatistics Working Paper Series. (http://www.bepress.com/umichbiostat/paper4).

Little, R.J. and Schenker, N. (1995). Missing data. In: G. Arminger, C.C. Clogg and M.E. Sobel (Eds.) Handbook of Statistical Modeling for the Social and Behavioral Sciences. New York: Plenum Press.

Little, R.J.A and Rubin, D.B. (2002). Statistical Analysis with Missing Data. John Wiley and Sons Inc.

Little, R.J.A. (1976). Inferences about Means from Incomplete Multivariate Data. Biometrika, 63, 593-604.

Little, R.J.A. (1982). Models for Nonresponse in Sample Surveys. Journal of the American Statistical Association, 77, 237-250.

Little, R.J.A. (1986). Survey Nonresponse Adjustments for Estimates of Means. International Statistical Review, 54, 139-157.

Little, R.J.A. (1992). Regression With Missing X's: A Review. Journal of the American Statistical Association, 87(420), 1227-1237.

Little, R.J.A. and Rubin, D.B. (1987). Statistical Analysis with Missing Data. (1st edition). New York: John Wiley and Sons.

Little, R.J.A. and Rubin, D.B. (2002). Statistical Analysis with Missing Data. (2nd edition). New York: John Wiley and Sons.

Liu, H.-M., Tseng, C.-H., and Tsao, F.-M. (2000). Perceptual and Acoustic Analyses of Speech Intelligibility in Mandarin-Speaking Young Adults with Cerebral Palsy. Clinical Linguistics and Phonetics, 14, 447-464.

Lohr, S. L. (1999). Sampling: Design and Analysis. Duxbury Press.

Mahalanobis, P.C. (1946). Recent Experiments in Statistical Sampling in The Indian Statistical Institute. Journal of the Royal Statistical Society, 109,. 325-370.

Mandic, D.P. and Chambers, J.A. (2001). Recurrent Neural Networks for Prediction: Learning Algorithms, Architectures and Stability. England: John Wiley and Sons.

Mann, N.R., Schafer, R.E., and Singpurwalla, N.D. (1974). Methods for Statistical Analysis of Reliability and Life Data. New York: Wiley.

Marker, D.A., Judkins, D.R. and Winglee, M. (2002). Large-Scale Imputation for Complex Surveys. Groves, R.M., Dillman, D.A., Eltinge, J.L. and Little, R.J.A (Eds.) Survey Nonresponse, New York: John Wiley and Sons.

Marquardt, D.W. (1963). An Algorithm for Least-squares Estimation of Nonlinear Parameters. Journal of the Society of Industrial and Applied Mathematics, 11, 431-441.

Masters, T. (1995). Advanced Algorithms for Neural Networks: A C++ Sourcebook. New York: John Wiley and Sons.

McCulloch, W.S. and Pitts, W. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. Bulletin of Mathematical Biophysics, 5, 115-133.

Meng, X.-L (1994). Multiple-Imputation Inferences with Uncongenial Sources of Input. Statistical Science, 9, 538-558.

Michie, D.J., Siegelhalter, D.J. and Taylor, C.C. (1994). Machine Learning, Neural and Statistical Classification. New York: Ellis Horwood.

Minsky, M. and Papert, S. (1969). Perceptrons. Cambridge: MIT Press.

Molenberghs, G. and Goetghebeur, E. (1997). Simple Fitting Algorithms for Incomplete Categorical Data. Journal of the Royal Statistical Society, Series B, 59, 401-414.

Murtagh, F. (1994). Neural Networks and Related Massively Parallel Methods for Statistics: a Short Overview. International Statistical Review, 62, 275-288.

National Center for Health Statistics (2002). Data file Documentation, National Health Interview Survey, 2001 (machine readable file and documentation). National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, Maryland.

Nelder J.A. and Mead R. (1964). A Simplex Method for Function Minimization. The Computer Journal, 7, 308-313.

Neter, J., Wasserman, W., and Kutner, M. H. (1985). Applied Linear Statistical Models: Regression, Analysis of Variance, and Experimental Designs. Homewood, IL: Irwin.

Nordbotten S. (1963). Automatic Editing of Individual Statistical Observations Statistical Standards and Studies, Handbook No. 2. New York.

Nordbotten, S. (1963). Automatic Editing of Individual Observations. Conference of European Statisticians. U.N. Statistical and Economic Commission of Europe.

Nordbotten, S. (1995). Editing and Imputation by Means of Neural Networks. Statistical Journal of the UN/ECE,12.

Nordbotten, S. (1995). Editing Statistical Records by Neural Networks. Journal of Official Statistics, 11 (4), 391-411.

Nordbotten, S. (1996). Neural Network Imputation Applied to the Norwegian 1990 Population Census Data. Journal of Official Statistics, 12 (4), 385-401.

Nordbotten, S. (1997). New Methods of Editing and Imputation. (http://www.unece.org/stats/)

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

O'Sullivan, J. W. (1994) Neural Nets: A Practical Primer, AI In Finance, Spring.

Parker, D.B. (1985). Learning Logic, Technical Report TR-47, Center for Computational Research in Economics and Management Science. MA: MIT, Cambridge.

Patterson, D.W. (1996). Artificial Neural Networks: Theory and Applications. Singapore: Prentice Hall.

Peressini, A. L., Sullivan, F. E., and Uhl, J. J., Jr. (1988). The mathematics of nonlinear programming. New York: Springer.

Piela, P. (2001). Tree-Structured Self-Organizing Maps for Imputation, Statistics Finland (Euredit project).

Pregibon, D. (1997). Data Mining. Statistical Computing and Graphics, 7(3), 8.

Press, W.H., Flannery, B.P., Teukolsky, S.A., and Vetterling, W. T. (1992). Numerical Recipes in FORTRAN: The Art of Scientific Computing, 352-355 (2nd edition). England: Cambridge University Press.

Principe, J.C., Euliano, N.R. and Lefebvre, W.C. (2000). Neural and Adaptive Systems: Fundamentals Through Simulation. New York: John Wiley and Sons.

Quenouille, M.H. (1949). Problems in Plane Sampling. Annals of Mathematical Statistics, 20, 355-375.

Ramsey, F.L. and Schafer, D.W. (1996). The Statistical Sleuth: A Course in Methods of Data Analysis. USA: Duxbury Press.

Rancourt, E., Särndal, C.-E., and Lee, H. (1994). Estimation of the Variance in Presence of Nearest Neighbor Imputation. Proceedings of the Section on Survey Research Methods, American Statistical Association, 888-893.

Rao, J. N. K. and Shao, J. (1992). Jackknife Variance Estimation with Survey Data Under Hot Deck Imputation. Biometrika, 79, 811-822.

Rao, J.N.K. and Wu, C.F.J. (1985). Inference From Stratified Samples: Second-order Analysis of Three Methods for Nonlinear Statistics. Journal of The American Statistical Association, 80, 620-630.

Ripley, B.D. (1977). Modeling Spatial Patterns, Journal of the Royal Statistical Society, Series B, 39, 172-212.

Ripley, B.D. (1993). Statistical Aspects of Neural Networks. Networks and Chaos: Statistical and probabilistic Aspects (U. Bornndor-Nielsen, J Jensen, and W. Kendal, (Eds.). Chapman and Hall.

Ripley, B.D. (1994). Neural Networks and Related Methods for Classification, Journal of Royal Statistical Society B, 6(3), 409-456.

Ripley, B.D. (1996). Pattern Recognition and Neural Networks. Cambridge: University Press.

Rivals, I. and Personnaz, L. (2000). Construction of Confidence Intervals for Neural Networks Based on Least Squares Estimation. Neural Networks, 13, 463-484

Robins, J.M. and Wang, N. (2000). Inference for Imputation Estimators. Biometrika, 87, 113-124.

Rosenbaltt, F. (1958). The Perceptron: a Probabilistic Model for Information Storage and Organization in the Brain. Psychological review, 65, 386-408.

Rosenbaltt, F. (1962). Principles of Neurodynamics. Spartan Books, New York.

Rubin D.B. (1976a). Inference and missing data, Biometrika, 63, 581-592.

Rubin D.B. (1977). Formalizing Subjective Notions about The Effect of Non-Respondents in Sample Surveys. Journal of the American Statistical Association, 77, 538-543.

Rubin, D. B. (1978). Bayesian Inference for Causal Effects: The Role of Randomization. Annals of Statistics, 6, 34-58.

Rubin, D. B. (1978b). Multiple Imputations in Sample Surveys- a Phenomenological Bayesian Approach to Nonresponse, Proceedings of the Survey Research Methods Section, American Statistical Association, 20-34.

Rubin, D.B. (1972). A Non-iterative Algorithm for Least Squares Estimation of Missing Values in any Analysis of Variance Design. Applied Statistics, 21, 136-141.

Rubin, D.B. (1976). Noniterative Least Squares Estimates, Standard Errors and F-tests for Analyses of Variance with Missing Data. Journal of Royal Statistical Society, Series B, 38, 270-274.

Rubin, D.B. (1985a) The Use of Propensity Scores in Applied Bayesian Inference, in Bayesian Statistics 2 (J.M. Bernardo, M.H. De Groot, D.V. Lindley, and A.F.M. Smith, eds.), Amsterdam: North Holland, 463-472.

Rubin, D.B. (1986). Basic Ideas of Multiple Imputation for Nonresponse. Survey Methodology, 12, 37-47.

Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: Wiley.

Rubin, D.B. (1996). Multiple Iimputation After 18+ Years. Journal of American Statistical Association, 91, 473-489.

Rubin, D.B. and Schenker, N. (1986). Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse. Journal of the American Statistical Association, 81, 366-374.

Rumelhart, D.E., Hinton, G.E. and Williams, R. J. (1986). Learning Internal Representations by Error Propagation, in D. E. Rumelhart and J. L. McClelland (Eds.). Parallel Distributed Processing: Explorations in the Microstructures of Cognition, 1, 318-362.Cambridge, MA: MIT Press/

Rumelhart, D.E., McClelland, J.L. (1986). Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vols. 1 and 2. Cambridge: MIT Press.

Särndal, C.-E. (1992). Methods for Estimating the Precision of Survey Estimates When Imputation Has Been Used. Survey Methodology, 18, 241-265.

Särndal, C.-E., Swensson, B. and Wretman, J. (1991) .Model Assisted Survey Sampling. Springer-Verlag.

Schafer, J.L. (1997). Analysis of Incomplete Multivariate Data. London: Chapman and Hall.

Schafer, J.L. and Graham, J.W. (2002). Missing Data: Our View of the State of the Art. Psychological Methods, 7(2), 147-177.

Schafer, J.L. and Schenker, N. (2000). Inference with Imputed Conditional Means. Journal of the American Statistical Association, 95, 144–154.

Schimert, J., Schafer, J.L., Hesterberg, T.M., Fraley, C. and Clarkson, D.B. (2000). Analyzing Data with Missing Values in S-Plus. Seattle: Insightful Corp.

Shao, J. and Tu, D. (1995). The Jackknife and Bootstrap. New York: Springer-Verlag.

Shepherd, A.J. (1997). Second-Order Methods for Neural Networks. New York: Springer.

Shumway, R.H. (1984). Proceedings of the Symposium on Time Series Analysis of Irregularly Observed data, E. Parzen, (Ed.) Lecture Notes in Statistics. New York: Springer-Verlag.

Smith, A.F.M. and Roberts, G.O. (1992). Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo methods. Journal of the Royal Statistical Society, Series B, 5(1).

Southcott, M.L. and Bogner, R.E. (1993). Classification of Incomplete Data Using Neural Networks. ACNN'93, 220-223.

StatSoft, Inc. (2004). Electronic Statistics Textbook. Tulsa, OK: StatSoft. (http://www.statsoft.com/textbook/stathome.html)

Tanner, M.A. and Wong, W.H. (1987). The Calculation of Posterior Distributions by Data Augmentation. Journal of the American Statistical Association, 82(398), 528-550.

Teague, A. and Thomas, J. (1996). Neural Networks as a Possible Means for Imputing Missing Census Data in the 2001 British Census of Population. Survey and Statistical Computing, 199-203.

Titterington, D.M. and Cheng, B. (1994). Neural Networks: A Review from a Statistical Perspective. Statistical Sciences, 9 (1), 2-54.

Trawinski, I.M. and Bargmann, R.E. (1964). Maximum Likelihood Estimation with Incomplete Multivariate Data. Annals of Mathematical Statistics, 35, 647-657.

Treiman, D.J., Bielby, W., and Cheng, M. (1993). Multiple Imputation by Splines. Bulletin of the International Statistical Institute, Contributed Papers II, 503-504.

Tukey, J.W. (1958). Bias and Confidence in Not-quite Large Samples. Annals of Mathematical Statistics, 29, 614.

Vartivarian, S. and Little, R. (2003). On the Formation of Weighting Adjustment Cells for Unit Nonresponse, The Berkeley Electronic Press, working paper 10.(http://www.bepress.com/umichbiostat/paper10)

Vartivarian, S.L. and Little, R.J. (2003). Weighting Adjustments for Unit Nonresponse with Multiple Outcome Variables. The University of Michigan Department of Biostatistics Working Paper Series. Working Paper 21. (http://www.bepress.com/umichbiostat/paper21)

Von Hippel, P.T. (2004).Biases in SPSS 12.0 Missing Value Analysis. The American Statistician, 58 (2), 160-164.

Wachter, K.W. and Trussell, J. (1982). Estimating Historical Heights, Journal of the American Statistical Association, 77, 279-301.

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

Wayman, J.C. (2003). Multiple Imputation For Missing Data: What Is It And How Can I use It?. The 2003 Annual Meeting of the American Educational Research Association. Chicago, IL.

Welstead, S.T. (1994). Neural Network and Fuzzy Logic Applications in C/C++. New York: Wiley.

Werbos, P.J. (1974). Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences, PhD thesis, Harvard University.

White, H. (1989). Learning in Artificial Neural Networks: A Statistical Perspective. Neural Computation.

White, H. (1992). Artificial Neural Networks Approximation and Learning Theory. Cambridge, MA: Blackwell Publishers.

Wilde, D. J. and Beightler, C. S. (1967). Foundations of optimization. Englewood Cliffs. NJ: Prentice-Hall.

Wilkinson, J. (1958). The Calculation of Eigenvectors of Co-diagonal Matrices. The Computer Journal, 1, 90-96.

Wilmot, C.G. and Shivananjappa, S. (2001). Comparison of Hot Deck and Neural-Network Imputation.

Witten, I.H. and Frank, E. (2000). Data mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco, CA.

Wolter, K. M. (1985). Introduction to Variance Estimation. Springer-Verlag.

Woodruff, R.S. (1971). A Simple Method For Approximating The Variance of a Complicated Estimate. Journal of the American Statistical Association, 66, 411-414.

Woodruff, Stephen M. (1988). Estimation in the Presence of Non-Ignorable Missing Data and a Markov Super-population Model. Washington: D.C. Bureau of Labor Statistics. (http://www.amstat.org/sections/srms/Proceedings/papers/1988_114.pdf).

APPENDIX

GLOSSARY OF TERMS

Activation function: A function used by a node in a neural network to transform input data from any domain of values into a finite range of values.

Affine function: A function $A:\mathbb{R}^m \to \mathbb{R}^n$ is affine if there is a linear function $L:\mathbb{R}^m \to \mathbb{R}^n$ and a vector b in $\mathbb{R}^n$ such that $A(x)=L(x)+b$ for all x in $\mathbb{R}^m$. An affine function is a linear function plus a translation.

Backpropagation: A training method used to calculate the weights in a neural network from the data.

Data mining: An information extraction activity whose goal is to discover hidden facts contained in databases.

Ensembles: Ensembles are collections of neural networks that cooperate in performing a prediction.

Epoch: During iterative training of a neural network, an Epoch is a single pass through the entire training set, followed by testing of the verification set.

Feed-forward: A neural network in which the signals only flow in one direction, from the inputs to the outputs.

Hidden Layers: Intermediate layers of a neural network that fall between the input and output layers.

Hidden nodes: The nodes in the hidden layers in a neural net.

Imputation: The act of imputing or ascribing; attribution. In missing data, imputation is filling in missing data.

Layer: Nodes in a neural net are usually grouped into layers, with each layer described as input, output or hidden.

Learning: Training models (estimating their parameters) based on existing data.

Learning rate: A control parameter of some training algorithms, which controls the step size when parameters are iteratively adjusted.

Loss function: The function that is minimized in the process of fitting a model.

Machine learning: Application of generic model-fitting or classification algorithms for predictive data mining.

Missing data: Data values can be missing because they were not measured, not answered, corrupted, were unknown or were lost.

Multilayer perceptrons: Feed-forward neural networks.

Multiple imputation: Multiple imputation is based on imputing the missing value several times, e.g. r, resulting in r complete data sets.

Network bias: In a neural network, bias refers to the constant terms in the model.

Neural network: A complex nonlinear modeling technique based on a model of a human neuron. Analytical techniques modeled after the processes of learning in the cognitive system and the neurological functions of the brain.

Neuron: A unit in a neural network.

Node: A point in a neural network that combines input from other nodes and produces an output through application of an activation function.

Noise: The difference between a model and its predictions.

Over-fitting: When attempting to fit a curve to a set of data points, it is the act of producing a curve with high curvature which fits the data points well, but does not model the underlying function well. In this case, the shape of the fitted curve is distorted by the noise inherent in the data.

Parallel processing: Several computers or CPUs linked together so that each can be computing simultaneously.

Perceptrons: Perceptrons are a simple form of neural networks. They have no hidden layers, and can only perform linear classification tasks.

Supervised learning: The collection of techniques where analysis uses a well-defined dependent variable.

Test data: A data set independent of the training data set, used to fine-tune the estimates of the model parameters.

Topology: For a neural net, topology refers to the number of layers and the number of nodes in each layer.

Training: Term for estimating a model's parameters based on the data set at hand.

Training data: A data set used to estimate or train a model (In the results of chapters 2 and 3, 60% of the original data was used for training).

Unsupervised learning: A collection of techniques where groupings of the data are defined without the use of a dependent variable.

Validation: The process of testing the models with a data set different from the training data set (In the results of chapters 2 and 3, 30% of the original data was used for validation).