A Preliminary Analysis of Data Quality in Neural Networks

by

Barbara D. Klein
bdklein@fob-f1.umd.umich.edu

Donald F. Rossin
drossin@fob-f1.umd.umich.edu

School of Management
University of Michigan-Dearborn
Dearborn, MI 48128

Abstract

This paper reports the results of a study investigating the effect of data quality on neural network models. Neural networks have recently been applied in a wide variety of business domains. Although databases used in many organizations have been found to contain errors, little is known about the effect of these errors on predictions made by neural network models. The paper uses a real-world example, the prediction of the net asset values of mutual funds, to investigate this topic. The results of a three-factor experiment in which the fraction of data containing errors, the amount of the data errors, and mutual fund type are found to affect the predictive accuracy of neural networks are reported. The findings have implications for users of neural networks working with databases containing errors.

1. Introduction

There is strong evidence (e.g., Laudon, 1986; Morey, 1982; Redman, 1992; Redman, 1995) that data stored in organizational databases have a significant rate of errors. The effect of data errors on the outputs of computer-based models has been investigated by a number of researchers (e.g., Ballou and Pazer, 1985; Ballou, Pazer, Belardo, and Klein, 1987; Bansal, Kauffman, and Weitz, 1993). This investigation builds on this prior research by examining the effect of data quality on neural network models. The study uses a financial application of a neural network to examine this question.

A neural network is a type of model that can be used to predict continuously-valued outputs or to classify observations. Neural network models have been applied to a variety of problem domains such as the prediction of graduate student success (Hardgrave, Wilson, and Walstrom, 1994), the prediction of bank failure (Tam and Kiang, 1992), the detection of fraud in insurance claims (Clayton, 1997), the analysis of product quality in refineries (Wadi, 1996), and the forecasting of extended warranty claims (Wasserman and Sudijianto, 1996). Typically one has a neural network learn about a problem by training it with examples. Training algorithms search for a set of weights that offer the best fit with the given examples. Once trained, a network can be used to make predictions. Although several architectures for neural networks have been developed, the scope of this study is limited to the back propagation feedforward neural network architecture.

In general, when claims about the predictive accuracy of neural networks are made, it is assumed that data used to test the models are free of errors. One notable exception to the assumption of accurate test data is the work of Bansal, Kauffman, and Weitz (1993) which examined the predictive accuracy of neural network models designed to predict the prepayment rate of mortgage-backed security portfolios.

An understanding of the effect of data errors on neural network models is particularly important because the availability of inexpensive software packages for personal computers makes the development and use of neural networks by end users feasible. Researchers have argued that end-user computing has increased the potential for data errors in computer applications (Boockholdt, 1989). As end users develop applications, it is possible that fewer data validation methods such as logic tests and control totals will be in place and it is likely that less rigorous testing will occur before applications are used in production (Corman, 1988; Davis, 1984; Davis, Adams, and Schaller, 1983).

The remaining sections of this paper present (1) a review of relevant prior research on data quality and on financial applications of neural networks, (2) a brief explanation of back propagation neural networks, (3) the methodology of the experiment investigating the effect of errors in test data, (4) the results of the experiment, and (5) conclusions and suggestions for future research. Preliminary results from an experiment investigating the effect of errors in training data are also presented in the last section of the paper.

2. Background

This study builds on prior research examining the effect of data errors on computer-based models and on studies investigating the application of neural network models to the analysis of financial instruments. Data errors are discussed in section 2.1 and financial applications of neural networks are discussed in section 2.2.

2.1 The Effect of Data Errors on Computer-Based Models

Data quality is generally recognized as a multidimensional concept (Wand and Wang, 1996; Wang and Strong, 1996). While no single definition of data quality has been accepted by researchers working in this area, there is agreement that data accuracy, currency, completeness, and consistency are important areas of concern (Agmon and Ahituv, 1987; Ballou and Pazer, 1985; Davis and Olson, 1985; Fox, Levitin, and Redman, 1993; Huh, Keller, Redman, and Watkins, 1990; Madnick and Wang, 1992; Wand and Wang, 1996; Wang and Strong, 1996; Zmud, 1978). This investigation adopts the conceptualization of data quality proposed by Ballou and Pazer (1985) which includes four dimensions: accuracy, timeliness, completeness, and consistency. This study is primarily concerned with data accuracy, defined as conformity between a recorded data value and the corresponding actual data value.

Prior research has found that organizational databases are not in general free of errors (e.g., Laudon, 1986; Morey, 1982; Redman, 1992; Redman, 1995). Between one and ten percent of data items in critical organizational databases are estimated to be inaccurate (Laudon, 1986; Madnick and Wang, 1992; Morey, 1982; Redman, 1992). Inaccurate data have been reported in a student loan database maintained by the U.S. Department of Education (Knight, 1992), in records maintained by the U.S. Department of Agriculture ("Dead farmer," 1992), in records maintained by credit reporting bureaus ("Consumer enemy," 1991), and in databases containing information about stock prices (Bennin, 1980; Rosenberg and Houglet, 1974).

Errors in data are acknowledged as a significant problem by at least some information system managers. In a survey of fifty Chief Information Officers of large organizations, half were found to believe that the usefulness of their organization's data is limited because of data accuracy problems (Nayar, 1993). Knight (1992) reports the findings of a study in which two-thirds of surveyed organizations acknowledged problems stemming from inaccurate or incomplete data.

Several studies have investigated the effect of data errors on the outputs of computer-based models. Bansal, Kauffman, and Weitz (1993) studied the effect of data errors on predictions made by neural network and linear regression models. Ballou and his colleagues have conducted a stream of research on the effect of data errors on information system outputs (Ballou and Pazer, 1985; Ballou, Pazer, Belardo, and Klein, 1987; Ballou and Tayi, 1989; Ballou and Pazer, 1995; Ballou, Wang, Pazer, and Tayi, in press). O'Leary (1993) investigated the effect of data errors in the context of a rule-based artificial intelligence system. Each of these studies is discussed in turn below.

Bansal, Kauffman, and Weitz (1993) compare the effect of errors in data on linear regression and neural network models. Models to predict the prepayment rate of mortgage-backed security portfolios were built using a training data set that was free of errors. Test data sets containing data errors were then constructed to evaluate the sensitivity of these models to data errors. The size of the data errors (5%, 10%, 15%, and 20%) and the fraction of the data set containing errors (4%, 8%, and 12%) were manipulated. The linear regression and neural network forecasts were evaluated using two metrics: (1) R^2 as a measure of predictive accuracy and (2) a payoff measure designed to capture the value of an accurate forecast to a portfolio manager. Error size had a statistically significant effect on predictive accuracy for both the linear regression and the neural network models and on the measure of payoff for linear regression. The fraction of the data set containing errors had a statistically significant effect on predictive accuracy and the payoff measure for linear regression but had no effect on either metric for the neural network model. They concluded that the neural network model is more robust than the linear regression model as data quality decreases.

Ballou and Pazer (1985) present a model for analyzing the effect of errors in data on the outputs of information systems. The objective of this model is to understand the way in which data errors are magnified or dampened as data are manipulated in an information system. Ballou, Pazer, Belardo, and Klein (1987) apply this model to an analysis of the impact of data errors in a spreadsheet model. The problem of the selection of an appropriate forecasting model is examined. Four variables are forecasted using ten different historical data sets containing errors.

Across the four variables, the data errors were found to affect the selection of a forecasting model for at least six and for as many as nine of the ten historical data sets.

Other studies in the research stream conducted by Ballou and his colleagues have examined the allocation of resources to data quality improvement projects (Ballou and Tayi, 1989), developed a framework for analyzing tradeoffs between the accuracy and timeliness dimensions of data quality (Ballou and Pazer, 1995), and developed a framework applying total quality management concepts to the measurement of data quality (Ballou, Wang, Pazer, and Tayi, in press).

O'Leary (1993) presents a general methodology for analyzing the impact of data accuracy on the performance of an artificial intelligence system designed to generate rules from data stored in a database. The methodology is applicable to artificial intelligence systems that analyze data and generate a set of rules of the form "if x then y." It is assumed that a subset of the generated rules are added to the system's rule base on the basis of a measure of the goodness of each rule. O'Leary shows that errors in data can affect the subset of rules that are added to the rule base and that inappropriate rules may be retained while useful rules are discarded if data accuracy is ignored.

2.2 Financial Applications of Neural Network Models

Neural networks are used by both academics and practitioners working in the area of financial analysis. Much of the research in this area focuses on the predictive accuracy of neural network models and on comparing the predictions of neural networks to those of more traditional models such as linear regression. Chiang, Urban, and Baldridge (1996) developed neural network, linear regression, and nonlinear regression models to forecast the net asset value of mutual funds and found that neural network models perform better than both linear and nonlinear regression models. Yoon, Swales, and Margavio (1993) compared the performance of back propagation neural network models and discriminant analysis for predicting the performance of stocks and found that neural network models made more accurate classifications than discriminant analysis. Schoneburg (1990) developed neural network models to predict daily stock prices for three German stocks. Jain and Nag (1995) applied a neural network to the problem of pricing initial public offerings. Neural network based software to make predictions about financial instruments is also commercially available (e.g., Neural Applications Corporation, 1996; Trendy Systems, 1997).

With the exception of the Bansal, Kauffman, and Weitz (1993) study, both academic and commercial financial applications of neural network models assume that all data used to construct the model and all data input to the model in production are accurate. The remaining sections of this paper present the design and results of an investigation into the performance of neural network models when this assumption is relaxed.

3. Back Propagation Neural Networks

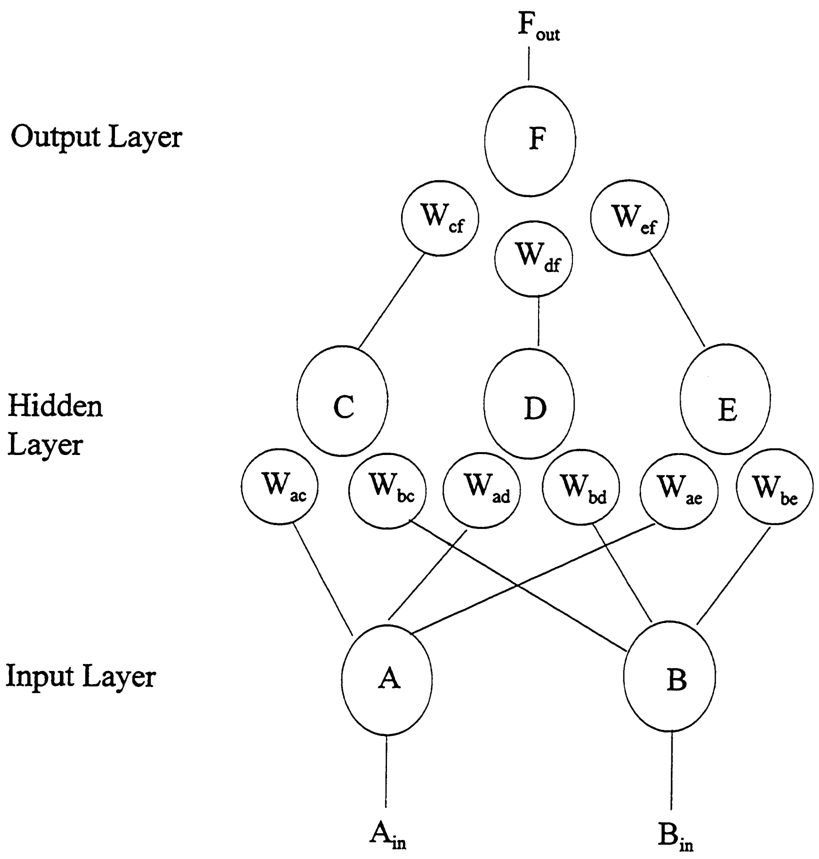
Figure 1 shows a visual representation of a typical back propagation neural network. It has three layers; an input layer which receives information from the environment, a hidden layer, and an output layer which transmits a response back to the environment. Connections denote whether information flows between processing elements occur. In this network inputs are processed through a hidden layer to an output layer. The basic objective of back propagation is to minimize the mean squared error between the actual output and the desired output as specified in the training set.

Updating of a back propagation neural network consists of two phases, a forward phase and a reverse phase. During the forward phase, input in the sense of paired values for A_{in} and B_{in} is presented, and propagated forward through the network to compute an output value for each processing element (PE). This is accomplished by summing the results of multiplying the weights associated with the connections to a particular PE and outputs associated with those weights.

Use of a linear output or activation function for F_{out} allows the output value to take on any value. Sometimes it is desired that the output range be between 0 and 1. In this case, typically a sigmoid or possibly a sine function is used.

The backward phase adjusts the weights associated with the nodes. Starting at the output node where the error measure of desired output minus actual output is readily available, the error measure is propagated back through the layers toward the input node. More detailed information can be found in summary papers such as Masson and Yang (1990), Wang and Malakooti (1992), and Zahedi (1991).

Figure 1: Typical Back Propagation Neural Network



4. Model Construction and Experimental Design

The application for study in this paper is the prediction of prices or net asset values (NAV) of mutual funds. Mutual funds consist of diversified portfolios of stocks that are managed by professionally trained individuals. They have become the major investment vehicle of choice.

Prices or net asset values (NAV) of mutual funds should reflect known economic information. The relationships are often unclear and ill-defined, making the prediction of the NAV very difficult and complex. Neural networks may provide a mechanism by which these economic relationships can be exploited.

To start the construction of a model neural network for predicting the net asset value for a mutual fund, 14 economic variables were identified as input. They are specified and defined in Figure 2. A 10-year economic data set (1986-1995) was constructed from (*Statistical Abstract*, 1996). In addition, end-of-year net asset values for 213 U.S. mutual funds were obtained from (*The Individual Investor's Guide*, 1997). The criteria for inclusion was having historical net asset value figures back to 1987.

As the purpose of this study is the effect of data quality on neural network forecasting, it was decided to limit the number of input variables to a more manageable amount. Stepwise linear regression was conducted for the 213 mutual funds. A 5% significance level (the SPSS default)

Figure 2: Potential Independent Variables

Name	Description
GDP	Gross Domestic Product (in billions of dollars). Output attributable to all labor and property supplied by United States residents.
CD*	Consumption Demand (in billions of dollars). Personal consumption expenditures.
ID	Investment Demand (in billions of dollars). Investment spending by firms. Excludes residential investments.
GD*	Government Demand (in billions of dollars). U.S. government spending. Includes consumption expenditures and gross investment.
NEX	Net Exports (in billions of dollars). Net exports of goods and services.
CPI*	Consumer Price Index. Measure of the average change is prices over time in a fixed market basket of goods and services. 1982-84 = 100.
M1*	Money, M1 (in billions of dollars). Includes currency in the hands of the nonbank public, travelers checks, demand deposits, and other checkable deposits.
M2	Money, M2 (in billions of dollars). Includes M1 plus money market funds, savings deposits, and small time deposits.
UR	Unemployment Rate. Percent of the labor force unemployed
TBR	Treasury Bill Rate. Interest rate for 3-month Treasury bill.
FFR	Federal Funds Rate.
CILEAD	Composite Index - Leading Indicators. 1987 = 100.
CICOIN	Composite Index - Coincident Indicators. 1987 = 100.
CILAG	Composite Index - Lagging Indicators. 1987 = 100.

Note: Asterisk indicates selection for model development.

was used to bring variables into the models. Four input variables were chosen based on the number of times each had been selected in the regression step. These variables are identified by an asterisk in Figure 2. In addition, it was decided to limit the number of mutual funds to 10 per fund type. Fund type definitions are per (*The Individual Investor's Guide*, 1997). The randomly chosen 40 funds are indicated in Figure 3.

Figure 3: Randomly Chosen Mutual Funds

Aggressive Growth (out of 64 possible)

Fairmont
Fidelity Sel Air Transportation
Fidelity Sel Automotive
Fidelity Sel Brokerage & Investment
Fidelity Sel Computers
Fidelity Sel Leisure
Fidelity Sel Software & Computer
INVESCO Dynamics
Kaufmann
USAA Aggressive Growth

Growth (out of 80 possible)

Fidelity Capital Appreciation
Fiduciary Capital Growth
Founders Growth
Janus Fund
Mathers
Meridian
Schwartz Value
Scudder Equity Trust: Capital Growth
Sound Shore
Vanguard/Morgan Growth

Balanced (out of 24 possible)

Dodge & Cox Balanced
Fidelity Puritan
Founders Balanced
Greenspring
INVESCO Industrial Income
Northeast Investors Trust
SAFECO Income
Strong Asset Allocation
USAA Income
Value Line Income

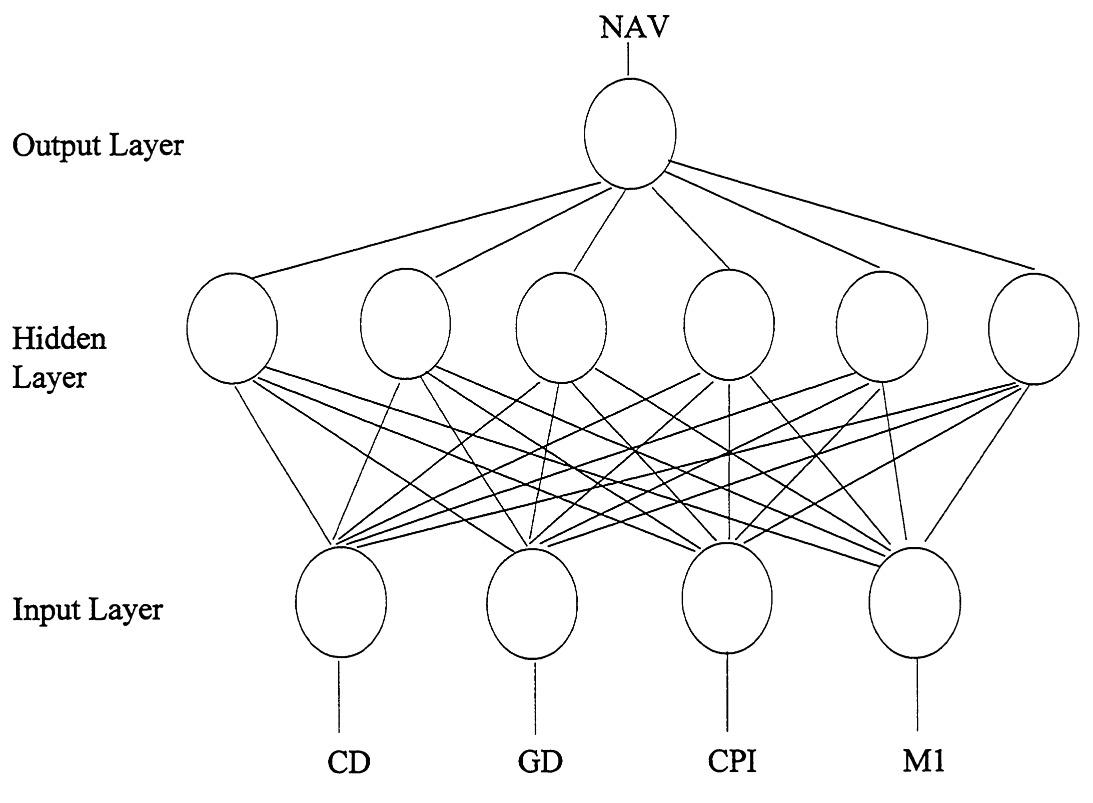
Growth & Income (out of 45 possible)

AARP Growth & Income
Berger Growth & Income
Dreyfus Third Century
Fidelity Sel Utilities Growth
IAI Growth & Income
INVESCO Value: Value Equity
SAFECO Equity
Stratton Monthly Dividend Shares
Strong Total Return
T. Rowe Price Growth & Income

For construction of the neural network models, the first nine years of data (the *training* set) are used. Data from the tenth year (the *testing* set) are used to develop the NAV forecast for a specific mutual fund.

Forty neural networks were constructed (one for each mutual fund). Various additional parameter value decisions were made by a combination of trial and error and experience. For example, it was decided to have one hidden layer with six nodes and one output node for NAV prediction for a particular mutual fund for 1996. This is in addition to the four input nodes which were decided earlier. Figure 4 gives a simplified schematic representation of the neural network.

Figure 4: Simplified Neural Network for NAV Example



Also, a learning rate of 0.10, a momentum rate of 0.10, and 0.30 for initial weights were chosen. An activation function of hyperbolic tangent for the hidden layer and linear for the output node were chosen based on software advice. Stopping rules were set to allow the neural network enough time to make significant adjustments. The maximum number of learning epochs was set to 10,000. A learning epoch indicates the network going through the 9 years of training data once. Finally, better solutions are indicated by a decrease in the minimum average error. Decreases occur frequently early in training. It was decided that the procedure should be stopped if 1,000 epochs occurred without a change in minimum average error. All runs were conducted using NeuroShell 2 software (*NeuroShell,* 1996).

A neural network was constructed for each of the 40 mutual funds using the 9 oldest years of the data for training. The 1995 testing data was then used to predict a NAV value for each of the 40 mutual funds for the end-of-year 1996. Actual end-of-year 1996 NAV values and predicted end-of-year NAV values by fund type were compared using both R^2 and mean absolute percent error (MAPE) measures of accuracy. This formed the base case.

The experimental design included three factors; (1) fraction-error or percent of the four variables in the testing set that would be changed with levels of 25 percent, 50 percent, 75 percent, and 100 percent, (2) amount-error or the percent amount by which the variables identified in the fraction-error factor would be changed with levels of plus or minus 5 percent, and plus or minus 10 percent, and (3) fund type with levels of aggressive growth, balanced, growth, growth & income.

As there are four variables used as input, the fraction-error levels of 25 percent, 50 percent, 75 percent, and 100 percent indicate whether 1, 2, 3, or 4 variable values would be altered to simulate an inaccuracy. The two levels of amount-error, 5 percent and 10 percent, were chosen as being representative of many situations. Random numbers were utilized to determine which variables were to be altered and by how much for a particular combination of levels. Amount-error was equally likely to be positive as well as negative.

In order to reduce variability, a second stream of random numbers was constructed along with the first. This second stream contained antithetic random numbers which are simply the difference found by subtracting the random number in the first stream from the number one. It was decided that four estimations would be appropriate for each amount-error fraction-error combination. Therefore, a total of 8 runs (four estimations with two random number streams) was accomplished for each combination.

5. Results Discussion

Predictive accuracy results, using the simulated inaccuracies for amount-error and fraction-error, in terms of R^2 and MAPE for the NAV forecasts for 1996 are given in Table 1 and Table 2. Results (R^2 in Table 1 and MAPE in Table 2) reflect the use of the appropriately perturbed portion of the test data except for the 0% fraction-error and 0% amount-error cell. This cell reflects results using the unperturbed test data. All other cells reflect average values for four estimations; each estimation utilized an independently drawn random number stream for one run and the antithetic stream for a second run. Therefore, each cell reflects the average of eight runs, four independent pairs each.

5.1 ANOVA Analysis

Two three-factor analysis of variance (ANOVA) tests were conducted, using the data in Tables 1 and 2. One ANOVA run was performed for each performance measure (R^2 and MAPE). For each run, the factors are fraction-error (25 percent, 50 percent, 75 percent, and 100 percent), amount-error (plus or minus 5 percent, and plus or minus 10 percent), and fund type (aggressive growth, balanced, growth, growth & income).

Tables 3 and 4 gives the calculated F values in each instance; critical values are given in the left-hand column; Table 3 for R^2 and Table 4 for MAPE. Significant results are indicated with an asterisk. Significant results indicate those factors which have a significant effect on the predictive measure. For example, Table 3 results indicate that as the amount-error increases from 5 percent to 10 percent, the decrease in R^2 is significant. In other words, the mean value for R^2 is not equal for both values of amount-error.

When there are more than two factors, ANOVA results do not indicate where the significant differences occur. For example, while fraction-error is a significant factor for both R^2 and MAPE, this difference may come as fraction-error changed from 25 percent to 50 percent, 50 to 75 percent, or 75 to 100 percent. It could also have come from a larger jump, such as 25 percent to 75 percent or 25 percent to 100 percent. Independent Samples T-Tests were performed in order to determine exactly where significant differences occurred.

Table 1: Variations in Predictive Accuracy (R^2) When Accuracy of Test Data Varies

Amount Error	Fund Type	0%	Fraction Error 25%	50%	75%	100%
0%	Aggressive Growth	91.0				
	Balanced	98.9				
	Growth	76.1				
	Growth & Income	88.7				
5%	Aggressive Growth		91.0	91.7	90.7	89.8
	Balanced		98.8	98.9	98.8	98.8
	Growth		77.4	76.2	77.1	75.5
	Growth & Income		88.7	89.8	88.7	89.0
10%	Aggressive Growth		90.7	89.9	89.1	83.7
	Balanced		98.8	98.7	98.0	97.3
	Growth		76.8	68.9	67.3	60.9
	Growth & Income		88.7	86.6	84.7	75.5

Note: Data used to obtain these results were the test data. The 0% fraction error and 0% amount error cell reflects the accuracy of the unmodified test data used in conjunction with the unmodified neural network. All other cells reflect average accuracy results for 4 simulated estimations involving appropriately simulated data inaccuracies. Each estimation is the result of two runs, one using the drawn random numbers, the other the antithetic random number. Therefore, each cell represents the average of 8 runs.

Table 2: Variations in Predictive Accuracy (MAPE) When Accuracy of Test Data Varies

Amount			Fraction Error			
Error	Fund Type	0%	25%	50%	75%	100%
0%	Aggressive Growth	9.95%				
	Balanced	7.70%				
	Growth	10.59%				
	Growth & Income	9.73%				
5%	Aggressive Growth		10.58%	10.66%	10.87%	12.27%
	Balanced		7.66%	7.81%	8.61%	9.62%
	Growth		10.31%	10.94%	11.21%	12.10%
	Growth & Income		9.70%	9.75%	10.31%	11.39%
10%	Aggressive Growth		11.21%	12.60%	13.15%	15.72%
	Balanced		8.23%	11.21%	12.77%	14.75%
	Growth		10.55%	15.43%	15.58%	17.51%
	Growth & Income		11.45%	13.77%	14.76%	17.89%

Note: Data used to obtain these results were the test data. The 0% fraction error and 0% amount error cell reflects the accuracy of the unmodified test data used in conjunction with the unmodified neural network. All other cells reflect average accuracy results for 4 simulated estimations involving appropriately simulated data inaccuracies. Each estimation is the result of two runs, one using the drawn random numbers, the other the antithetic random number. Therefore, each cell represents the average of 8 runs.

Table 3: Significance of Varying Amount Error, Fraction Error, and Fund Type on Predictive Performance (R^2) - ANOVA Results

Factor/significance criterion	Predictive Accuracy
Fraction error F(0.01;3;96) = 3.992	12.190*
Amount error F(0.01;1;96) = 6.906	43.551*
Fund type F(0.01;3:96) = 3.992	293.371*
Fraction error-amount error interaction F(0.01;3;96) = 3.992	8.275*
Fraction error-fund type interaction F(0.01;9;96) = 2.598	1.589
Amount error-fund type interaction F(0.01;3;96) = 3.992	6.803*
Fraction error-amount error-fund type interaction F(0.01;9;96) = 2.598	1.134

Significant results are marked with an asterisk.

Table 4: Significance of Varying Amount Error, Fraction Error, and Fund Type on Predictive Performance (MAPE) - ANOVA Results

Factor/significance criterion	Predictive Accuracy
Fraction error	7.028*
F(0.01;3;96) = 3.992	
Amount error	28.671*
F(0.01;1;96) = 6.906	
Fund type	4.247*
F(0.01;3:96) = 3.992	
Fraction error-amount error interaction	2.150
F(0.01;3;96) = 3.992	
Fraction error-fund type interaction	0.107
F(0.01;9;96) = 2.598	
Amount error-fund type interaction	0.492
F(0.01;3;96) = 3.992	
Fraction error-amount error-fund type interaction	0.052
F(0.01;9;96) = 2.598	

Significant results are marked with an asterisk.

From the ANOVA, and Independent Samples tests, the following conclusions are drawn:

Performance Measure: R^2

1. No significant difference among and between the various levels of fraction-error.
2. No significant difference among and between the various levels of amount-error.
3. Balanced fund had the significantly highest R^2.
4. Aggressive growth and growth & income were indistinguishable from each other and had the next significantly highest R^2.
5. Growth had the significantly lowest R^2.

Performance Measure: MAPE

1. The 100% fraction-error had significantly higher MAPE than the 25% fraction-error.
2. The 75% fraction-error had significantly higher MAPE than the 25% fraction-error.
3. The 50%, 75% and 100% fraction-error MAPE means were not significantly different.
4. The 5% and 10% amount-error MAPE means were significantly different.
5. Growth fund had significantly higher MAPE than aggressive growth fund.
6. MAPE means were not significantly different for aggressive growth, balanced and growth & income.
7. MAPE means were not significantly different for balanced, growth & income and growth.

6. Conclusions

This paper contributes to the literature on data quality by demonstrating that the predictive accuracy of neural networks is affected by two factors: fraction-error and amount-error. The findings of this study have implications for practitioners working in a variety of settings characterized by imperfect data. They suggest that an understanding of the error rate and the magnitude of errors in a dataset should be important considerations for users of neural network models. The fact that our experimental manipulations of the amount-error factor were limited to 5 and 10 percent is of particular practical importance. Data errors larger than 10 percent have been documented in practice, and our results suggest that larger errors would have a more detrimental effect.

The results presented in the prior section assume that data used to train the network are free of errors. One area for future research is the investigation of the effect of errors in data used to train neural networks. We are currently investigating this problem using the task of mutual fund net asset value prediction described in section 4. A three factor experiment is being conducted in which fund type, fraction of the training data containing errors (5% and 10%), and amount of the data errors (5% and 10%) are varied for the training set.

While we plan to perform 8 runs in each cell of the experimental design currently only one run per cell has been completed. Results from the completed runs are presented in Table 5 and Table 6. The number of completed runs in each cell of the experimental design is too small to perform a statistical analysis of the effect of fraction error, amount error, and fund type on predictive accuracy (R^2 and MAPE). However, we note that for 15 of the 16 runs completed, R^2 is higher for the runs with errors in the training data than for the base case runs without data errors. MAPE is higher for all of the runs with 5 percent fraction error than for the base case runs. For 10 percent fraction error, two runs have higher MAPE than the base case, and six runs have lower MAPE than the base case.

The scope of this study is limited to the back propagation neural network architecture with hyperbolic tangent and linear activation functions. Future research could consider other neural network architectures such as probabilistic neural networks and general regression neural networks as well as other activation functions such as the logistic activation function.

Table 5: Variations in Predictive Accuracy (R^2) When Accuracy of Training Data Varies

Amount			Fraction Error	
Error	Fund Type	0%	5%	10%
0%	Aggressive Growth	91.0		
	Balanced	98.9		
	Growth	76.1		
	Growth & Income	88.7		
5%	Aggressive Growth		95.9	97.7
	Balanced		99.3	99.4
	Growth		89.3	92.8
	Growth & Income		96.3	97.9
10%	Aggressive Growth		95.2	97.1
	Balanced		99.6	98.6
	Growth		94.3	96.4
	Growth & Income		95.2	95.3

Note: Data used to obtain these results were the training data. The 0% fraction error and 0% amount error cell reflects the accuracy of the unmodified test data used in conjunction with the unmodified neural network. All other cells reflect accuracy results for 1 simulated run involving appropriately simulated data inaccuracies.

Table 6: Variations in Predictive Accuracy (MAPE) When Accuracy of Training Data Varies

Amount			Fraction Error	
Error	Fund Type	0%	5%	10%
0%	Aggressive Growth	9.95%		
	Balanced	7.70%		
	Growth	10.59%		
	Growth & Income	9.73%		
5%	Aggressive Growth		18.94%	9.30%
	Balanced		13.08%	7.56%
	Growth		10.82%	6.55%
	Growth & Income		12.05%	7.25%
10%	Aggressive Growth		16.77%	11.62%
	Balanced		10.68%	8.77%
	Growth		10.70%	5.49%
	Growth & Income		15.62%	7.59%

Note: Data used to obtain these results were the training data. The 0% fraction error and 0% amount error cell reflects the accuracy of the unmodified test data used in conjunction with the unmodified neural network. All other cells reflect accuracy results for 1 simulated run involving appropriately simulated data inaccuracies.

References

Agmon, N. and N. Ahituv, "Assessing Data Reliability in an Information System," *Journal of Management Information Systems*, 4, 2 (1987), pp. 34-44.

Ballou, D. P. and H. L. Pazer, "Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems," *Management Science*, 31, 2 (1985), pp. 150-162.

Ballou, D. P. and H. L. Pazer, "Designing Information Systems to Optimize the Accuracy-timeliness Tradeoff," *Information Systems Research*, 6, 1 (1995), pp. 51-72.

Ballou, D. P., H. L. Pazer, S. Belardo, and B. Klein, "Implications of Data Quality for Spreadsheet Analysis," *Data Base*, 18, 3 (1987), pp. 13-19.

Ballou, D. P. and G. K. Tayi, "Methodology for Allocating Resources for Data Quality Enhancement," *Communications of the ACM*, 32, 3 (1989), pp. 320-329.

Ballou, D. P., R. Wang, H. Pazer, and G. K. Tayi, "Modeling Information Manufacturing Systems to Determine Information Product Quality," *Management Science*, (in press).

Bansal, A., R. J. Kauffman, and R. R. Weitz, "Comparing the Modeling Performance of Regression and Neural Networks as Data Quality Varies: A Business Value Approach," *Journal of Management Information Systems*, 10, 1 (1993), pp. 11-32.

Bennin, R. (1980). "Error Rates in CRSP and COMPUSTAT: A Second Look," *The Journal of Finance*, 35, 5 (1980), pp. 1267-1271.

Boockholdt, J. L. "Implementing Security and Integrity in Micro-Mainframe Networks," *MIS Quarterly*, 13 (1989), pp. 135-144.

Chiang, W., T. L. Urban, and G. W. Baldridge, "A Neural Network Approach to Mutual Fund Net Asset Value Forecasting," *Omega*, 24, 2 (1996), pp. 205-215.

Clayton, M. "Risk Data to Provide Comp Fraud Detection System," *The Insurance Accountant*, 10, 4, p. 1.

"Consumer Enemy No. 1," *Newsweek*, (October 28, 1991), pp. 42, 47.

Corman, L. S. "Data Integrity and Security of the Corporate Data Base: The Dilemma of End User Computing," *Data Base*, 19 (1988), pp. 1-5.

Davis, G. B. "Caution: User Developed Systems Can Be Dangerous to Your Organization," MISRC Working Paper 82-04, MIS Research Center, University of Minnesota, (1984).

Davis, G. B., D. L. Adams, and C. A. Schaller, *Auditing & EDP*, American Institute of Certified Public Accountants, Inc., New York, (1983).

Davis, G. B. and M. H. Olson, *Management Information Systems: Conceptual Foundations, Structure, and Development*, McGraw-Hill Book Company, New York, (1985).

"Dead farmer syndrome haunts efforts to trim USDA offices," *Minneapolis Star Tribune*, (April 19, 1992), p. 5A.

Fox, C., A. Levitin, and T. Redman, "The Notion of Data and Its Quality Dimensions, " *Information Processing & Management*, 30, 1 (1993), pp. 9-19.

Hardgrave, B. C., R. L. Wilson, and K. A. Walstrom, "Predicting Graduate Student Success: A Comparison of Neural Networks and Traditional Techniques," *Computers and Operations Research*, 21, 3 (1994), pp. 249-263.

Huh, Y. U., F. R. Keller, T. C. Redman, and A. R. Watkins, "Data Quality," *Information and Software Technology*, 32, 8 (1990), pp. 559-565.

The Individual Investor's Guide to Low-Load Mutual Funds, 16th ed., American Association of Individual Investors, Chicago, IL, (1997).

Jain, B. A. and B. N. Nag, "Artificial Neural Network Models for Pricing Initial Public Offerings," *Decision Sciences*, 26, 3 (1995), pp. 283-302.

Knight, B. "The Data Pollution Problem," *Computerworld*, 26, 39 (1992), pp. 81-83.

Laudon, K. C. "Data Quality and Due Process in Large Interorganizational Record Systems," *Communications of the ACM*, 29, 1 (1986), pp. 4-11.

Madnick, S. E. and R. Y. Wang, "Introduction to the TDQM Research Program," Total Data Quality Management Research Program Working Paper #92-01, (1992).

Masson, E. and Y. Wang, "Introduction to Computation and Learning in Artificial Neural Networks," *European Journal of Operational Research*, 47, (1990), pp. 1-28.

Morey, R. C. "Estimating and Improving the Quality of Information in a MIS," *Communications of the ACM*, 25, 5 (1982), pp. 337-342.

Nayar, M. K. "Achieving Information Integrity: A Strategic Imperative," *Information Systems Management*, 10, 2 (1993), pp. 51-61.

Neter, J., W. Wasserman, and M. Kutner, *Applied Linear Statistical Models*, 3rd ed., Irwin, Homewood, IL, (1990).

NeuroShell 2, 4th ed., Ward Systems Group, Frederick, MD, (1996).

"Neural Applications Corporation: Intelligent Process Optimization." (1996). <http://www.neural.com/> (27 June 1997).

O'Leary, D. E. "The Impact of Data Accuracy on System Learning," *Journal of Management Information Systems*, 9, 4 (1993), pp. 83-98.

Redman, T. C. *Data Quality: Management and Technology*, Bantam Books, New York, (1992).

Redman, T. C. "Improve Data Quality for Competitive Advantage," *Sloan Management Review*, 36, 2 (1995), pp. 99-107.

Rosenberg, B. and M. Houglet, "Error Rates in CRSP and COMPUSTAT Data Bases and Their Implications," *The Journal of Finance*, 29, (1974), pp. 1303-1310.

Schoneburg, E. "Stock Price Prediction Using Neural Networks: A Project Report," *Neurocomputing* 2, (1990), pp. 17-27.

Statistical Abstract of the United States, U.S. Bureau of the Census, Government Printing Office, Washington, D.C., (1996).

Tam, K. Y. and M. Y. Kiang, "Managerial Applications of Neural Networks: The Case of Bank Failure Predictions," *Management Science*, 38, 7 (1992), pp. 926-947.

"Trendy Systems, LLC...End of Day S&P Futures Trading." (1997). <http://www.trendysystems.com/> (27 June 1997).

Wadi, I. "Neural Network Model Predicts Naphtha Cut Point," *Oil & Gas Journal*, 94, 48 (Nov. 25, 1996), pp. 67-70.

Wand, Y. and R. Y. Wang, "Anchoring Data Quality Dimensions in Ontological Foundations," *Communications of the ACM*, 39, 11 (1996), pp. 86-95.

Wang, J. and B. Malakooti, "A Feedforward Neural Network for Multiple Criteria Decision Making," *Computers and Operations Research*, 19, (1992), pp. 151-167.

Wang, R. Y. and D. Strong, "Beyond Accuracy: What Data Quality Means to Data Consumers," *Journal of Management Information Systems*, 12, 4 (1996), pp. 5-34.

Wasserman, G. S. and A. Sudjianto, "A Comparison of Three Strategies for Forecasting Warranty Claims," *IEEE Transactions*, 28, 12 (1996), 967-977.

Yoon, Y., G. Swales, and T. M. Margavio, "A Comparison of Discriminant Analysis Versus Artificial Neural Networks," *Journal of Operational Research*, 44, 1 (1993), pp. 51-60.

Zahedi, F. "An Introduction to Neural Networks and a Comparison with Artificial Intelligence and Expert Systems," *Interfaces*, 21 (1991), pp. 25-38.

Zmud, R. W. "An Empirical Investigation of the Dimensionality of the Concept of Information," *Decision Sciences*, 9, 2 (1978), pp. 187-195.