# QUANTIFYING THE VALUE OF MODELS AND DATA: A COMPARISON OF THE PERFORMANCE OF REGRESSION AND NEURAL NETS WHEN DATA QUALITY VARIES

Arun Bansal

Robert J. Kauffman

Rob R. Weitz

Department of Information, Operations, and Management Sciences

Leonard N. Stern School of Business, New York University

44 West 4th Street, New York, NY 10012

# QUANTIFYING THE VALUE OF DATA AND MODELS:

# A COMPARISON OF THE PERFORMANCE OF REGRESSION AND NEURAL NETS

# WHEN DATA QUALITY VARIES

**October 5, 1992**


**ARUN BANSAL**
Doctoral Program in Information Systems
Stern School of Business
New York University

**ROBERT J. KAUFFMAN**
Associate Professor of Information Systems
Stern School of Business
New York University

**ROB R. WEITZ**
Associate Professor
Department of Computing and Decision Sciences
Stillman School of Business
Seton Hall University



The authors wish to express their thanks to Robert M. Mark and Edward Peters for providing funding, and access to data and people at the Manufacturer Hanover Trust Corporation. All errors are the responsibility of the authors.



**Working Paper Series**
STERN IS-92-33

# QUANTIFYING THE VALUE OF DATA AND MODELS: A COMPARISON OF THE PERFORMANCE OF REGRESSION AND NEURAL NETS WHEN DATA QUALITY VARIES

---


## ABSTRACT

Under circumstances where data quality may vary, knowledge about the potential performance of alternate predictive models can enable a decision maker to design an information system whose value is optimized in two ways. The decision maker can select a model which is least sensitive to predictive degradation in the range of observed data quality variation. And, once the "right" model has been selected, the decision maker can select the appropriate level of data quality in view of the costs of acquiring it. This paper examines a real-world example from the field of finance -- prepayments in *mortgage-backed securities* (**MBS**) portfolio management -- to illustrate a methodology that enables such evaluations to be made for two modeling alternative: regression analysis and neural network analysis. The methodology indicates that with "perfect data," the neural network approach outperforms regression in terms of predictive accuracy and utility in a prepayment *risk management forecasting system* (**RMFS**). Further, the performance of the neural network model is more robust under conditions of data quality degradation.



*KEY WORDS AND PHRASES: business value of information technology, data quality, decision support systems, information economics, neural networks, risk management, risk management forecasting systems, systems design.*

# 1. INTRODUCTION

## 1.1. Motivation

When data quality may vary, knowledge about the potential performance of alternate predictive models can enable a decision maker to design an information system whose value is optimized. Model performance comparisons often presume perfect data -- a presumption arguably more appropriate to textbook examples than to real-world problems. This research empirically examines the effects of data accuracy on the performance of two alternate forecasting frameworks: regression analysis and neural network analysis. This can enable a decision maker to select the model which is least sensitive to predictive degradation in the range of observed data quality variation. And, once the "right" model has been selected, the decision maker can also select the appropriate level of data quality in view of the costs of acquiring it. Such an analysis, in outlining the relative strengths of two alternate approaches, can provide especially useful guidance for maximizing the business value of information systems in circumstances where data quality is known, perhaps historically, to be less than perfect.

## 1.2. Application

The application selected is a risk management problem associated with the forecasting of prepayment rates in *mortgage-backed securities* (**MBS**) portfolio management. This area is well-known in financial circles and has shown itself to be fertile ground for predictive models. Typically forecasting prepayments requires large data sets, which are available via commercial sources. In this work, model performance is based on a traditional statistic, $R^2$, and on a utility measure called *deviation around mean utility* (**DAMU**), developed especially for this research application. The utility measure allows for evaluating the trade-offs (if any) between the business value of improved decisions resulting from the use of more accurate data, and the cost of obtaining such data.

### 1.3. Organization

The paper is organized as follows. Section 2 examines the literature in information economics, management science, accounting and information systems that we surveyed to obtain guidance about how to explore the problem of data quality variations and its impacts on model performance. Section 3 reviews prior research that compares the performance of regression analysis and neural networks, when data quality is a constant. It also presents the basics of the neural network analysis approach. Section 4 introduces the reader to the application: data quality variations in mortgage-backed securities portfolio management.

Section 5 presents details of the mortgage prepayment rate forecasting model to which we applied regression and neural net analysis. Section 6 lays out the method used to simulate quality variations, in data that are commonly used by mortgage-backed security portfolio managers to predict prepayment rates. Section 7 discusses two issues -- accuracy and utility -- that are crucial to the mechanism that enables us to evaluate the business value associated with the forecasting performance of a *risk management forecasting system* (**RMFS**) for mortgage-backed securities. It examines the measures for forecasting model fit and utility, in a way that shows how inaccurate information can bias the creation of hedged positions and introduce excess risk. Section 8 presents the details of two hypothesis tests that focus on the business value of data quality and model choice. Section 9 describes the results of the paper, and Section 10 concludes with a reckoning of the major results, a consideration of some potential threats to their validity and directions that we are pursuing in related research. To increase the readability of this paper for a reader who is unfamiliar with the subject of mortgage-backed securities portfolio management, we include a glossary of basic terms in Table 1 (at the end of the paper), and Table 2 in Section 8 provides at-a-glance information about the structure of our hypothesis tests, models and data.

## 2. DATA QUALITY AND MODEL PERFORMANCE

A number of researchers in finance have emphasized the importance of high quality data for accurate predictions of future states of the world, but specifying the nature of the relationship between data quality variation and the performance of forecasting systems has remained an elusive goal. Some interesting work is available, however, to guide our attention in this area. For example, Meyer and Pifer [MEYE70] did not obtain good results from a classification model that predicted failures for commercial banks. They later realized that they had considered a data set which was limited in coverage of the proper time frame. Martin [MART77] analyzed alternative types of early-warning models for predicting bank failures. One of the major criticisms of his work [ALTM81] is the choice of poor data quality; the sample size was too small. Altman [ALTM77] developed a performance- predictor system for the savings and loan industry. Data quality related to sample representativeness is one of the main criticisms of this work.

Beaver [BEAV67] emphasized the importance of data frequency in prediction of bankruptcy. Bankruptcy observations are not very frequent; therefore, unless the frequency of incoming information on the financial performance of a firm is high, vital data to predict bankruptcy may be missed. Deakin [DEAK72] proposed an alternate business failure model to assess the impact, frequency, and nature of bankruptcy misclassification. One of the major limitations of his work was a small data set collected over a very long period of time. Deakin also emphasized the importance of the cost of a wrong decision.

Data can vary on a number of dimensions, including frequency, accuracy, and response time. Ballou and Pazer [BALL85] showed how the effect of some data quality variations could be combined to yield a single and simple measure of data quality variation, i.e., the overall inaccuracy in the data. The argument they present is based on the fact that any kind of data quality variation ultimately makes data inaccurate, hence it is reasonable to combine such variations into a single measure of accuracy variation. In previous work [BANS92] it was shown how two such data quality variations -- frequency and accuracy -- affect the utility and the predictive accuracy of an RMFS. Given the objectives for this research, we focus on a single measure of data quality -- accuracy -- to uncover the relative strengths of different forecasting models with data of varying quality.

Recently, the effect of data quality on the performance of an information system has been recast through the models of information economics. An information economics model values information in terms of its expected utility to a user, in view of the costs of obtaining it, less the utility that a user can obtain by making the same decision in the absence of the information system (and thus without additional information costs) [DEMS72]. Clearly, the higher the *net utility* that an information system user can derive from its use, the better is the performance of the system.

A number of researchers have shown how information economics can be used to gauge the effect of data quality variations under different decision making situations [HILT81, DEMS85, BARU89, AHIT89]. Barua, Kriebel and Mukhopadhyay [BARU89] applied an information economics model to gauge the effect of data quality variations in manufacturing, Cushing [CUSH74] determined the effect of propagation of errors in internal control systems, Wilson [WILS75] investigated the effect of inaccuracy on delivery times in production, and Ijiri and Itami [IJIR73] studied the effect of inaccuracies in demand estimation on a firm's profitability.

## 3. REGRESSION AND NEURAL NETWORK APPROACHES TO FORECASTING

### 3.1. Regression and Neural Nets: Preliminaries

*Regression.* Regression is a well-known and widely used approach to forecasting. Generally, regression can be linear or non-linear. Linear regression models utilize a linear function of independent variables to estimate the value of a dependent variable. A non-linear regression model generally uses either a multiplicative function, or even a more complex function of independent variables, to compute the values of a dependent variable. Although linear regression models are most common in practice, other types of models, such as non-linear models, are also used depending on the nature of the relationship between the dependent and the independent variables.

*Neural Nets.* The neural net model is a close approximation to non-linear regression. A neural network consists of several layers of units called *neurons*: an input layer, one or more hidden layers, and an output layer. An *input layer* contains neurons that are associated with perceived inputs considered relevant for classification purposes. The *output layer* corresponds to the results of the classification scheme, and the *hidden layers* are sandwiched between the input and the output layers [LIPP87, RUME86].

Neurons are elementary processing elements. Typically, each neuron in the input layer is connected to each and every neuron in the hidden layer and each neuron in the hidden layer, in turn, is connected to each and every neuron in the output layer. All these connections between neurons carry a weight which denotes the strength of the connection. *Excitatory connections* occur when the weights are positive and *inhibitory connections* are those where the weights are negative.

As with regression, the data set for the neural net approach is classified into independent and dependent variables. But the data are further partitioned into two categories: one is the "training" data set and the other is the "actual" data set. The purpose of the training data set is to condition the network so that it can develop a functional form of the model. Once the functional form is developed, the actual data set can be tested on this net for classification. In the training phase a neural net develops its model by repeatedly adjusting the weights of different links between the neurons [HINT89]. This phenomenon of weight adjustment is referred to as "learning" in neural net terminology.

The hidden layer neurons help the neural net to perform those classifications where the output is non-linearly dependent on the inputs. A neural net with two hidden layers can separate arbitrarily complex data, whereas a neural net with one hidden layer can separate data existing in a convex open or closed region. A neural net with no hidden layers can only separate data which falls on opposite sides of some hyperplane

[LIPP87].

The output layer of a neural net provides the results. The number of neurons in the output layer roughly corresponds to the different types of results expected from the net. In market share prediction, for example, this number would be one, corresponding with a single number: the percent market share expected. However, in a foreign-exchange forecasting system that provides directional advice to traders, there would be three neurons in the output layer: one for signalling an exchange rate increase, a second for a decline, and a third for an unchanged position.

### 3.2. Comparative Analysis with Regression and Neural Nets

Comparison of various forms of regression (simple linear, non-linear, etc.) with neural nets has been the focus of many studies in the recent past [DUTT88, WHIT88]. Some of the main differences between the two techniques of classification are evident from these studies. *First*, traditional statistical techniques, like regression, are not adaptive. Neural nets consistently improve during their training phase, adjusting weights each time a mistake is made. Regression techniques, on the other hand, process all training data simultaneously before being used with new data. *Second*, there are suggestions [LIPP87] that in non-linear classification models, neural nets may prove to be more robust. Although the author's critique is not based on empirical analysis, he is able to show that the robustness of neural nets in this case follows from the algorithms that are employed. *Third*, while regression models may be useful in determining the right set of independent variables [DUTT88], the absence of direct links in many practical neural nets makes it impossible to determine which inputs affect the outputs directly. Therefore, unlike regression analysis, where the estimated coefficients enable the analyst to readily assess the effects of incremental changes in the *value* of independent variables, with neural nets the focus shifts to the effects of incremental changes in the *number* of variables. *Fourth*, regression equations require model specification in advance [DUTT88]. If model specification is unknown, the use of a neural net may be a better option.

Although neural nets have been shown to produce more accurate predictions with good quality data than the equivalent regression models, not much information is available on how this comparison would change if the input data are inaccurate. Theoretically, since a regression model uses an explicit relationship between the causal variables and the predicted state, it should be less vulnerable to the inaccuracies in data. However, a counter-argument for the strength of neural nets may be that they are adaptive, hence minor data inaccuracies should generally be ignored by the neural net during the training process. And, the better the learning by a neural net, the higher should be the insulation from data inaccuracies. Which approach works better with imperfect data apparently remains an open question for research.

While theoretically it is difficult to prove the superiority of one approach over the other, we can demonstrate empirically the performance of both models under varying input data quality. To do so, we need to develop a methodology which allows comparisons and to define the relevant measures of performance. In the description of prepayment forecasting in mortgage-backed securities portfolio management that we will shortly present, the reader will note that the measure of utility that we employ for comparing the performance of alternate models in the presence of varying data quality is application-dependent. In fact, we expect that this will be the case in whatever business forecasting or data quality modeling context is selected when utility is used as a performance measure. For this reason, we give special care to the development presentation of the comparison metrics, to ensure that the reader will understand enough about the nature of the application to appreciate its quality as a testbed for the new ideas and findings that we present.

## 4. THE APPLICATION: DATA QUALITY VARIATIONS IN MORTGAGE-BACKED SECURITIES PORTFOLIO MANAGEMENT

### 4.1. Financial Risk Management

Financial services firms are perennially big spenders on information technologies deployed to improve operating performance and to increase money market prowess. But during the 1980s, the industry was plagued with difficulties that led numerous firms into bankruptcy, and others into unwanted mergers, while the weakest ongoing concerns were acquired outright, in arrangements that were mandated by the Federal Reserve Bank. (See STEV87 and SHAL89 for a discussion of specific examples of dramatic losses related to intentional business decisions; in addition, the interested reader may wish to refer to the "Inside Risks" column in the September 1992 issue of Communications of the ACM for a discussion of accidental financial mistakes, some of which entailed real losses [NEUM92].) Although by the end of the decade "the bloom" of information technology spending was "off the rose" in the industry, the expenditures that did remain were increasingly devoted to applications which were meant to increase corporate control and improve the management of financial risk.

*Financial risk management* is the management of the financial resources and commitments of a firm so as to maximize its value, taking into account the myriad of hard-to-predict events that can unexpectedly alter its ability to control cost and revenue flows. In this context, we define *risk* as "the lack of predictability of outcomes" affecting the set of financial assets and liabilities that constitute the firm's balance sheet and determine the manner in which its cash flows change [DOHE85, p. 15].

Financial services industry information technology specialists refer to the information systems that are meant to monitor and measure global risk as *financial risk management systems* (**FRMS**). The cost of building these systems is high. In large money center banks similar to those that we have worked with in New York City, a typical financial risk management system will cost on the order of $10 million or more.

Moreover, the required maintenance and periodic enhancements to keep up with changes in the industry can add a significant amount to this figure each year [SCHM90B]. As a result, there has arisen a cottage industry of firms which act as value-added systems vendors for integrated hardware and software, working in conjunction with large and frequently updated databases of information relevant to specific financial risk management problems (e.g., mortgage lending, foreign exchange rates, futures and options, and derivative instruments).

The largest firms normally track the market by acquiring real-time data feeds of two sorts. *Video data feeds* contain video images of fixed format pages of data about the market or a group of financial instruments. These cannot be decomposed into individual elements, in the same way that a viewer would be unable to pull stock prices off his television screen and into a spreadsheet program. *Digital data feeds*, on the other hand, contain digital data. This data can be unbundled, and then transmitted and manipulated to support computations for real-time financial decision modeling and early warning alarms of key changes in the market. Examples of digital data sources relevant to a money market bank's sales and trading or capital markets functions include the Chicago Board Options Exchange, and the New York Stock Exchange ticker (and related data), as well as market-specific wire services from firms such as Reuters, Telerate, Bloomberg, Knight-Ridder and Dow Jones [AREN89, SCHM90A]. These "quote vendors" consolidate data from the exchanges, from central banks worldwide, and from other governmental or private sector sources, and repackage it for digital transmission.

In other cases, large financial firms purchase data sets for infrequent or customized analyses. When this occurs, the project-specific data costs are usually very high (a quote vendor may specially package data for use by a specific client), and it may require careful consideration to determine the cost-benefit relationship in acquiring the data before such analyses should be conducted.

The popular press has reported on a number of firms that have made special financial commitments to obtaining the hardware, software and data feeds necessary to develop and deploy integrated trading and financial risk management systems. For example, Merrill Lynch and Shearson Lehman have set up special risk management units charged with developing systems solutions to monitor risk on a product-by-product basis, by currency and by geographic region. (The popular press magazines, *Wall Street Computer Review* and its successor, *Wall Street Technology*, make especially worthwhile reading to gain insight into the problems and systems solutions meant to solve them.)

Moves to redirect spending towards the management of risk are interesting, but we note that it is often the case that senior executives too quickly jump to the conclusion that as the quality of the hardware, software and data improves, so should the quality of the decisions that are made on the basis of them. Data quality considerations, of special interest to us in this paper, are often glossed over, though it should be clear that the selection of data vendor services (along with the models that use the data) involves a clear trade-off between expected or potential decision quality and the costs associated with obtaining (and perhaps maintaining in-house) high quality data.

For example, infrequent market indicator updates are almost always cheaper than frequent updates. But in the absence of any formal guidelines it is quite difficult to estimate exactly what frequency is appropriate. One common response on the part of management is to obtain substantially more data than is usually necessary on a routine basis, leading to excessive costs. (This ends up as a "knee-jerk" response for highly risk-averse managers.) This situation may be further complicated by the presence of other dimensions of data quality that we noted earlier, such as response time and accuracy. Thus, the design of an *effective* financial risk management system should treat the problem of detecting risk in the same manner as it crafts policies to secure firm profitability in risky markets: from the perspective of cost-benefit analysis.

### 4.2. Mortgage-Backed Securities

In order to explicitly evaluate the relationship between input data quality and model output, we selected the domain of prepayment forecasting in MBS portfolio management. This is an appropriate application area because of the importance that bankers place upon making useful forecasts to further their goal of successfully managing portfolios so that they produce spreads that contribute to bank profitability. The mortgage industry has two markets: a primary market and a secondary market [PINK87]. Lending institutions directly deal with the customers (or borrowers) in the primary market. Borrowers are obliged to pay the principal and the interest to the lender. The lender, in turn, is interested in raising capital and spreading the risk of making such loans with the help of federal agencies. Federal agencies purchase a bank mortgage portfolio or individual loans through a process called *securitization*. The securities created in this way are called mortgage-backed securities, and they trade freely in the secondary market just like any other financial instrument. The lenders or the servicing agencies collect payments from customers and, after deducting their servicing fees, pass them to the current holder of the MBS created to securitize the loan [HAYR89].

MBSs are considered to be fixed-income investments. However, the borrower retains an option to prepay the loan at any time and this makes investments in MBSs risky. Prepayments act as a *call option* on a fixed income security, adding uncertainty during the period of investment. In the case of rising interest rates, the holder of an MBS can expect good returns if prepayment occurs; the holder can invest prepayment cashflows in another security at a higher interest rate. On the other hand, prepayments occurring in a time of falling interest rates will lead to a loss. We define the *prepayment rate* as the percentage of total customers that prepay a loan at that specific point of time. Thus, whatever the case, a knowledge of the prepayment rate and an ability to predict future interest rates are essential for assessing risk [PINK87].

Currently, forecasts of future prepayment rates are made using regression models

[HAYR89]. These models estimate future prepayments based on current and past values of several relevant variables. These include: the age of the mortgage; the difference between the coupon rate of the mortgage and the current interest rate; and macroeconomic indicators, such as individual well-being, consumer confidence and GNP. The financial industry needs information on each of these indicators to forecast prepayments in the future to reduce the risk of managing portfolios of mortgage-backed securities. However, the question of determining *optimal data quality* (e.g., the interval between two successive reports, the delay in receiving reports and permissible inaccuracies) is important as recurring data feed requirements can be very expensive.

### 4.3. Data

The data for the study was gathered from two primary sources: a large money center bank in New York City and CITIBASE, a widely available electronic database that provides data on macroeconomic indicators. The macroeconomic data include: the annualized change in GNP, the ratio of personal income to personal expenditure, current mortgage rates, and consumer sentiment data. The remaining items in the model were taken from published reports that a participating bank in the study provided [SALO87]. The data set contained 1170 observations on 38 mortgage-backed securities. The time span of the data was from April 1987 to February 1990. The accuracy of this data set was assumed to be perfect and thus we employed it for future comparisons of data quality.

# 5. MODELS

In this section, we report on the estimation of two prepayment models: a regression model and a neural net. The results were developed using a reference set of perfect data.

## 5.1. The Regression Model

To predict mortgage prepayments (thereby insulating investors from prepayment risk), analysts use macroeconomic and investment-specific information [HAYR89]. The *investment-specific factors* on which the prepayment rates depend are presented below [HAYR89, PETE85, ARAK85].

* *Coupon spread* is the difference between the coupon on the underlying mortgage and the prevailing market mortgage rate. The higher the coupon rate, the greater are the chances that a borrower will refinance the loan.

* *Strength of the spread* is what determines whether the coupon spread is strong enough to attract a borrower to refinance, even when some transaction costs are involved.

* *Security type* is usually defined in terms of the three different organizations that offer securitization services -- the General National Mortgage Association (GNMA, pronounced Ginny Mae); the Federal National Mortgage Association (FNMA, and referred to popularly as Fanny Mae); and the Federal National Home Loan Mortgage Corporation (FHLMC, and usually called Freddy Mac). Each of these consists of a pool of borrowers with common characteristics. That helps an MBS investor to determine the risk of the security. Note that the data set used in this research does not include FNMA securities.

* *Maturity of the security* is the total life of the security, which is generally 15 or 30 years for housing loans.

* *Age of the mortgage.* Studies show that prepayments rates are highest for new mortgages around 2 to 3 years old.

The *macroeconomic factors* that affect the prepayment rates are:

* *Change in the GNP value since last period.* Increases in GNP and good economic conditions enable a borrower to postpone refinancing a loan.

* *Consumer sentiment* is an indicator of how a consumer feels about the future economic conditions on a scale of 0 to 100. Higher consumer sentiment generally means that a home buyer is confident about his future earnings, and thus would be willing to buy a better house and to repay a current loan.

* *Ratio of personal income to personal expenditure.* This has the same effect as the change in GNP.

The prepayment rate is expressed in terms of the *constant annual prepayment annual percentage* (**CPP**), which is a single year prepayment rate for a certain mortgage pool. A *mortgage pool* is a group of mortgage loans which have same characteristics, e.g., same maturity, same coupon, same average prepayment rates).

A linear regression model, such as the following, is commonly used to predict prepayment rates by banking industry analysts [ARAK85, HAYR89]:

$$\begin{aligned} CPP_{pt} = \beta_0 + \beta_1\, MAT_p + \beta_2\, SPREAD_{pt} + \beta_3\, GXNP_t + \beta_4\, CONSENT_t \\ + \beta_5\, TYPE_p + \beta_6\, SMALL_{pt} + \beta_7\, TERM_{pt} + \beta_8\, RATIO_t + e_{pt}. \end{aligned} \tag{1}$$

The variables in the model parallel the investment-specific and macroeconomic factors discussed above, and are defined as follows:

$CPP_{pt}$ = *Constant prepayment expressed in annual percent for pool p at time t.*

| | | |
|---|---|---|
| $MAT_p$ | = | *Maturity class of secutiry, a qualitative variable with the value 1 if the security matures in 30 years, and 0 otherwise.* |
| $SPREAD_{pt}$ | = | *Difference between the coupon of pool p and the prevailing market rates for a similar mortgage at time t.* |
| $GXNP_t$ | = | *Annualized percentage change in GNP at time t.* |
| $CONSENT_t$ | = | *Consumer sentiment at time t on a scale of 0 to 100.* |
| $TYPE_p$ | = | *Type of security, a qualitative variable with the value 1 if the security type is GNMA, and 0 if it is FHLMC.* |
| $SMALL_{pt}$ | = | *Small spread size, a qualitative variable with the value 1 if SPREAD < 2% for pool p at time t, and 0 otherwise.* |
| $TERM_{pt}$ | = | *Age of mortgage pool p at time t in years.* |
| $RATIO_t$ | = | *Ratio of personal income to expenditure at time t.* |
| $\beta_i$ | = | *Coefficient of regression for independent variable i in the prepayment forecasting model.* |
| $e_{pt}$ | = | *Normally distributed regression residuals.* |

The following estimates for the coefficients were obtained using a statistical package (**SAS**) and the above-mentioned data.

$$\begin{aligned} CPP_{pt} = \underset{(13.030;.0001)}{190.794} - \underset{(-2.950;.0003)}{1.113}\, MAT_p + \underset{(20.390;.0001)}{2.325}\, SPREAD_{pt} - \underset{(-5.890;.0001)}{0.697}\, GXNP_t + \underset{(2.280;.0230)}{0.110}\, CONSENT_t \\ - \underset{(-5.580;.0001)}{2.084}\, TYPE_p + \underset{(9.990;.0001)}{6.375}\, SMALL_{pt} + \underset{(11.850;.0001)}{0.533}\, TERM_{pt} - \underset{(-14.080;.0001)}{174.029}\, RATIO_t. \end{aligned} \tag{2}$$

Note: The resultant $R^2$ for the model was 0.7032; corrected $R^2$ was 0.7012; the number of observations were 1170. The numbers in parentheses under the coefficients are the coefficients' t-statistics and their corresponding significance levels.

This model provides estimates of prepayment rates given perfect data. However, with imperfect data we expect some variations in both the predictive accuracy and utility of the model.

### 5.2. The Neural Net Model

A back propagation neural net model with three layers -- an input layer, an output layer and a single hidden layer -- was used for the study. Although many other types of models, such as the Hinton, Kohonen and Hopfield models, have been tried in practice, the back propagation model seems to work well for many financial forecasting situations [DUTT88]. The back propagation algorithm minimizes the mean-squared difference between the desired and the actual outputs from the net. The net is trained by initially selecting small random weights and internal thresholds, and then presenting all training data repeatedly. Weights are adjusted after every trial using side information specifying the correct class until the weights converge and the cost function is reduced to an acceptable value.

The input data employed for the neural net were the same as those used for regression. Thus, eight input nodes, corresponding to the eight independent variables, were created. For the neural net the determination of an appropriate number of hidden layers, and the number of nodes in each layer, is a function of judgment and trial and error. However, one hidden layer should suffice for partitioning the data existing in a convex open or closed region and two are needed for arbitrarily complex data [LIPP87].

Because our data are not expected to be very complex (there are proven relationships between the input data and the prepayment rate), we used only a single hidden layer for our net. The number of nodes in the hidden layer is generally selected on the basis of three criteria [LIPP87, JOHN92]:

(1) The number should be big enough to reduce the occurrence of local minima.

(2) The number should not be very high or else the net will memorize the training data without gaining the capability to forecast when new data are presented.

(3) The number should be chosen such that the convergence of the net occurs reasonably fast.

After varying the number of hidden nodes in the net and keeping in view the above three principles, we selected the number of nodes to be five.

The actual execution and testing of the neural net model was divided into two parts: a training part and a testing part. We trained the model on one half of the data set, approximately 585 observations. The training phase involved 6000 cycles. After training the net, we tested it using the complete data set of 1170 observations. We used an "off-the-shelf" software package called *Neuralware Professional.*

## 6. ACCURACY VARIATIONS

Many types of inaccuracies are possible in a data set. A few examples are: an operator's typing mistake, the imprecise measurement of subjective data (e.g., about consumer sentiment in our current example) or the lack of proper updates (e.g., data on most macroeconomic indicators are difficult to update on a day-to-day or month-to-month basis and thus mostly are extrapolated). For this study we assume that most types of inaccuracies only affect a part, and not the complete data set. For example, typing errors usually appear in just a fraction of a data set. Similarly, imprecision in subjective data and lack of proper updates may also affect only a part of the data set.

To reflect and simulate this assumption in an empirical test, we varied only a certain fraction of the original data values. We capture the inaccuracy variations by varying a *fraction* (**FRAC**) of all actual data values by some randomly determined *amount* (**AMT**). The random inaccuracies vary between -10% to +10% of the original values of the data in steps of 1%. (The actual values that were altered were selected randomly by a program that we wrote for this specific purpose.) For example, if a certain inaccuracy was determined to be -8% and the original data value was 100, then the resultant inaccurate value will be 92 (100+(-.08*100)).

Our selection of this range is meant to cover two types of errors: errors in subjective measures and errors due to lack of updates. However, the range of typing errors is difficult to estimate. But, with a normal data entry screen with proper range checks, it is possible to place an upper limit on typing errors. Since FRAC and AMT together determine the extent of inaccuracies in a data set, we measured inaccuracy by multiplying the fraction of observations affected (FRAC) and the absolute amount of variation in each selected observation (AMT) to yield **AMTFRAC**.

# 7. MEASURING ACCURACY, MEASURING UTILITY

## 7.1. Measuring the Accuracy of a Risk Management Forecasting System (RMFS)

The *predictive accuracy* of a forecasting system can be measured by several different statistical indicators. Here we use the *coefficient of determination* (commonly known as $\mathbf{R^2}$). The $R^2$ measure reports the magnitude of the forecast error as a fraction of the magnitude of the dependent variable. This is a useful measure for this application context because it is readily employed by managers who are interested in judging the cost-quality trade-off associated with the design of their RMFS. The measure is given by

$$R^2 = \frac{SSTO - SSE}{SSTO} \tag{3}$$

where:

| | | |
|---|---|---|
| ***SSTO*** | = | *sum of squared totals; and* |
| ***SSE*** | = | *sum of squared errors.* |

SSTO measures the variation in the dependent variable when independent variables are not included. For the MBS portfolio prepayment problem, it is equal to:

$$SSTO = \sum_{j=1}^{N} ( CPP_j - \overline{CPP} )^2 \tag{4}$$

where:

| | | |
|---|---|---|
| $CPP_j$ | = | *Prepayment rate CPP for observation j.* |

$\overline{CPP}$ = *Mean prepayment rate for all the observations in a set.*

$N$ = *Total number of observations.*

SSE measures the variation in the dependent variable when a regression model utilizing a set of independent variables is employed. For our example, we have:

$$SSE = \sum_{j=1}^{N} ( CPP_j - CPP(e)_j )^2 \tag{5}$$

where:

$CPP(e)_j$ = *Estimated prepayment rate, CPP, for observation j.*

Thus, the $R^2$ measure can also be interpreted as the proportional reduction of total variation associated with the use of a set of independent variables.

While the value of SSTO will not change with changes in data quality, we should expect a change in the value of SSE. The value of $R^2$ is determined by comparing the predicted value of the prepayment rate using data of a certain quality with the actual prepayment rate. When the data quality degrades, we expect a corresponding decrease in the value of $R^2$.

For our reference data set (perfect data), the $R^2$ for the neural net model was 0.785; for the regression model with perfect data, $R^2$ was .7032.

### 7.2. Measuring the Utility of a Risk Management Forecasting System (RMFS)

Although $R^2$ is a good measure of fit between the predicted value and the actual value of the dependent variable in a forecasting model, this summary value alone does not fully describe the relationship between data quality and the overall *utility* of a forecasting system to a risk manager. For example, the $R^2$ measure does not tell us how a portfolio manager who invests on the basis of predictions from a forecasting system is going to be affected if the predictions are inaccurate. It is not hard to imagine that certain businesses are affected more than others when poor quality data are used for forecasting important business indicators. Changes in $R^2$ with the degradation of data quality may not correspond well with changes in the risk and reward calculations of a business investor. For instance, even a small error in the estimation of a financial indicator (such as *option delta*, the ratio of the change in price of an option to unit change in price of the underlying instrument) may lead to big losses for an investor.

*The DAMU Measure.* Given this limitation of $R^2$, we considered the use of a utility measure to investigate the relationship between data quality and a portfolio manager's performance. Of use would be a measure of the decrease in utility resulting when imperfect data were employed (as opposed to perfectly accurate data). We define a new metric for this purpose, *deviation around the maximum utility* (**DAMU**), given by where:

$\boldsymbol{U_{perfect,\,k}}$ = *Utility from an RMFS that employs "perfect" data for financial instrument k.*

$$DAMU_k = \frac{(\mid U_{perfect,k} - U_{imperfect,k} \mid)}{U_{perfect,k}} * 100 \tag{6}$$

$U_{imperfect,\,k}$ = *Utility from an RMFS that employs "imperfect" data for financial instrument k.*

The following discussion describes a methodology for quantifying DAMU in the MBS application, by first introducing the concept of effective duration and then applying it to create an efficient hedge, from which utility is measurable.

*The Effective Duration Concept.* To monitor the performance of mortgage-backed securities, portfolio managers use the measure of *effective duration*. Effective duration calculates the risk associated with investments in MBSs [FABO88], and is given by

$$Effective\ Duration = \frac{P_- - P_+}{(P_0)\,(y_+ - y_-)} \tag{7}$$

where:

$P_0$ = *Initial price of the financial instrument.*

$P_-$ = *Price if yield is decreased by x basis points (one basis point equals 0.01% or .0001).*

$P_+$ = *Price if yield is increased by x basis points.*

$y_-$ = *Initial yield minus x basis points.*

$y_+$ = *Initial yield plus x basis points.*

The prices in this equation are the present value of cash flows generated by the mortgage-backed security during its lifetime. Because the prepayments affect the cash flows, they will affect the prices and, hence, the effective duration also. As described earlier, changes in the market interest rates affect the prepayment rates for mortgage-backed securities. In general, the lower the market interest rate, the higher should be the prepayment rate: more and more borrowers will find refinancing economically advantageous. Thus, MBSs are believed to carry *interest rate risk* (i.e., the risk arising from interest rate fluctuations).

Portfolio managers (or the institutions that maintain their portfolios) generally evidence an upper bound on the trading and position risk that they can tolerate. Each has a different preference as to just how much risk is tolerable. With knowledge of the risk that is associated with a specific MBS in the form of effective duration, the portfolio manager can devise combinations of financial instruments in such a way that the overall risk of the position can be reduced to an acceptable level.

The technique of reducing risk in this manner is called "hedging". In hedging, the portfolio manager buys (is long in) some positions and simultaneously sells (is short in) others. The key risk that the portfolio manager faces that makes hedging especially worthwhile is the extent to which margin interest income will erode if future mortgage prepayments are unexpectedly large, freeing up cash that can be invested, albeit in a world with a greatly reduced set of opportunities to achieve such large returns.

Two factors are important in designing efficient hedges (i.e., the combination of positions that yield highest return for a specified level of risk):

(1) the types of instruments selected for the hedge; and,

(2) the number of securities which should be bought for each security that is sold.

The selection of financial instruments for a hedge is an important factor for designing good hedges, and this is the subject of much discussion in finance textbooks. To simplify matters, however, we assume that an MBS is being hedged against a 30-year treasury bond with an effective duration of 0.08. Thus, *the problem is to determine the ratio of the number of MBSs to be bought for each treasury sold.* This ratio is commonly known as the *hedge ratio*, a can be calculated as follows:

$$Hedge\ Ratio = \frac{RISK_{Permissible} + DUR_{Treasury}}{DUR_{MBS}} \tag{8}$$

The variables are defined as:

| | | |
|---|---|---|
| $\boldsymbol{Risk_{Permissible}}$ | = | *Total allowable risk for the hedge.* |
| $\boldsymbol{DUR_{MBS}}$ | = | *The effective duration of the MBS.* |
| $\boldsymbol{DUR_{Treasury}}$ | = | *The effective duration of the treasury.* |

Based on discussions we had with managers at our research site, we assume the values of $Risk_{Permissible}$ to be 0.06. With the values of the other variables given, equation (8) shows that the hedge ratio is dependent on the effective duration estimate for the MBS. Hence, predictions from the forecasting system will affect the overall hedge ratio and the risk associated with investments in the MBSs.

Given that the hedge ratio is a suitable measure for evaluating the utility of an RMFS used for monitoring investment risk in MBSs, the DAMU for MBSs is

$$DAMU_k = \frac{|(Hedge\ Ratio_{predicted,k} - Hedge\ Ratio_{perfect,k})|}{Hedge\ Ratio_{perfect,k}} * 100 \qquad (9)$$

where:

| | | |
|---|---|---|
| $Hedge\ Ratio_{predicted,k}$ | = | *Predicted hedge ratio for security k based on imperfect data.* |
| $Hedge\ Ratio_{perfect,k}$ | = | *Actual hedge ratio for security k based on perfect data.* |

As an example, consider that if the effective duration of an MBS is forecasted to be 0.06 with perfect data, then the hedge ratio will be 0.023. This implies that for each 1000 treasury bonds 23 MBSs of this particular type will be needed to form an efficient hedge. The hedge reduces the resultant exposure to risk so that it can be contained within the .06 limit.

Now, suppose the data used for predicting prepayments rates are inaccurate. Further assume that the predicted prepayment rates by the inaccurate data generate a duration of .05 instead of .06. The duration of .05 results in a hedge ratio of 0.028. This hedge ratio is around 21.74% higher (equating with a DAMU of 21.74) than the one that is based on perfect data. The miscalculated hedge ratio affects both the risk and the reward for an investment, thus making the hedge ratio a good measure for the utility of a forecasting system.

## 8. THE HYPOTHESES

To test if one model performs better than the other under varying accuracy of input data, we develop two hypotheses: one fore predictive accuracy and a second for forecast utility. Our methodology is summarized in Table 2 to assist the reader in understanding the structure of this research. Information is presented on: the hypotheses, the relevant models and variables, the hypothesis tests and test statistic, the data we used, the forecast performance measures, and the predicted results.

### 8.1. Comparing Predictive Accuracy of the Models Under Varying Data Accuracy

The first hypothesis tests the effect of the forecasting model on the predictive accuracy of the RMFS, given that the data are inaccurate. A simple linear model, implying risk neutrality on the part of the decision maker, is employed in this analysis:

$$R_i = \beta_0 + \beta_1 \varepsilon_i + \beta_2 \, model_i \, \varepsilon_i \tag{10}$$

where:

| | | |
|---|---|---|
| $R_i$ | = | *Predictive accuracy of the forecasting system using data set i.* |
| $\varepsilon_i$ | = | *Amount of inaccuracies in the data set i.* |
| $model_i$ | = | *Qualitative variable which equals 1 when a neural network is used for forecasting using data set i, and 0 otherwise (i.e., when regression is used, our base case).* |
| $model_i\ \varepsilon_i$ | = | *Interaction effect between the model used and the inaccuracy corresponding to the data set i.* |

$\beta_i$ = *Regression coefficient for variable i.*

When data inaccuracies are not present, the predictive accuracy of the model should be at its maximum, and this is given by the intercept term. When the data are inaccurate, the predictive accuracy of the RMFS is expected to decline. However, the rate of the decline of predictive accuracy with an increase in data inaccuracies will depend on the choice of the forecasting model -- *if the estimate for $\beta_2$ is found to be significant.*

If we find $\beta_2$ to be positive and significant, then this would show that neural nets, are able to predict more accurately than regression when data inaccuracies are present. On the other hand, if $\beta_2$ is found to be significant and negative, we should expect regression to provide us with higher predictive accuracy. Finally, if $\beta_2$ is found to be insignificant, suggesting that the interaction effect between the forecasting model and the data inaccuracies is not present, then the choice of the model should not make any difference.

***Hypothesis 1: $H_0$, the null hypothesis is $\beta_2 = 0$, and the alternate hypothesis, $H_a$, is $\beta_2 \neq 0$.***

The test that we employ for this hypothesis is based on the test statistic $t^* = b_2/s(b_2)$, where $b_2$ is an estimate for the coefficient $\beta_2$, and $s(b_2)$ is the standard deviation for the coefficient $\beta_2$. The decision rule for this hypothesis is the standard t-test: $|t^*| \leq$

t(1-$\alpha$/2; df), where $\alpha$ is the degree of confidence and df is the total number of degrees of freedom. If this relation is true, then we conclude that the null hypothesis, $H_0$, also is true. Otherwise we conclude that the null hypothesis, $H_a$, is probably true.

It should be noted that in testing the significance of $\beta_1$ we are examining the sole effect of data inaccuracies on model predictive accuracy -- irrespective of the model-specific differences in the results. In previous work conducted by Bansal and Kauffman [BANS92] it was shown that an increase (decrease) in data inaccuracies will result in a decrease (an increase) in $R^2$, when the regression approach to prepayment forecasting for MBSs is used. In this paper, however, we are more interested in the interaction of data quality and the model.

### 8.2. Comparing Predictive Utility of the Models Under Varying Data Accuracy

An accurate data set has the potential to deliver maximum utility to the user of a risk management forecasting system. But, with inaccuracies incorporated in the data set, the utility of the system should vary. With more inaccuracies, the expected difference between the utility obtained when accurate versus inaccurate data are used should be greater. This difference in utility was operationally defined in Section 7 as the DAMU measure. Thus, DAMU should generally increase with an increase in inaccuracy. And, because utility, in effect, is the negative of DAMU, utility should decline with an increase in inaccuracies.

Here we test whether such a decline in utility depends on the forecasting model used. To test this hypothesis, we examined the model (shown below) that predicts the utility of a risk management forecasting system.

$$U_i = \beta_3 + \beta_4 \varepsilon_i + \beta_5 \, model_i \, \varepsilon_i \tag{11}$$

where:

| | | |
|---|---|---|
| $U_i$ | = | *The utility of the forecasting system using data set i.* |
| $\varepsilon_i$ | = | *The amount of inaccuracies in data set i.* |
| $model_i$ | = | *A qualitative variable which equals 1 when a neural network is used for forecasting using data set i, and 0 otherwise (i.e., when regression is used, our base case).* |
| $model_i * \varepsilon_i$ | = | *The interaction effect between the model used and the inaccuracy corresponding to the data set i.* |
| $\beta_i$ | = | *Regression coefficient for variable i.* |

Clearly, if $\beta_4$ is significantly negative, it can be said that the utility of the RMFS will decrease with an increase in data inaccuracies. However (again), a more important concern here is the effect of the choice of the forecasting model on the utility of the RMFS. Thus, if $\beta_5$ is found to be significant and positive, then we can conclude that the neural net provides higher utility *even with an inaccurate input data set*. On the other hand, if $\beta_5$ is found to be negative, the reverse should be true, i.e., the regression outperforms neural nets with inaccurate input data sets.

***Hypothesis 2:** $H_0$, **the null hypothesis, is** $\beta_5 = 0$**; the alternate,** $H_a$, **is** $\beta_5 \neq 0$.*

We next present the complete regression results for the predictive accuracy and utility models so that the reader can develop additional perspective and better interpret the results.

## 9. RESULTS AND DISCUSSION

The two hypotheses were separately tested with the prepayment data discussed earlier. For the test of Hypothesis 1, we used both a regression and a neural net to predict mortgage prepayment rates for a number of data sets, each of which varied in terms of the data accuracy dimension. Then, both the predictive accuracy and the utility were calculated using each of these data sets. The results were applied to the linear models in equations (10) and (11). The outcomes obtained for the hypothesis tests are discussed below.

### 9.1. Hypothesis Test Results

*Hypothesis 1.* The results of the model we employed to examine predictive accuracy when data quality varies is shown below.

We note that the maximum predictive accuracy of the model that we used was estimated

$$R_i = \underset{(41.624;.001)}{\underset{+0.7195}{b_0}} + \underset{(-5.417;.001)}{\underset{-0.0007}{b_1\ \varepsilon_i}} + \underset{(13.709;.001)}{\underset{+0.000386}{b_2\ model_i\ \varepsilon_i}} \tag{12}$$

to be in the neighborhood of 72%, as shown by the first parameter estimate, $b_0$. In addition, our results show that imperfect data reduce predictive accuracy. The sign of $b_2$ is negative (as one might expect), and the estimate of -.0007 is highly significant.

Of primary interest to us, however, was the estimate, $b_2$, for $\beta_2$. This was found to be 0.000386, and the standard error for the coefficient was 0.0000669. $t^*$ equals $b_2/s(b_2)$ or 5.763. Because $|t^*| > t(0.975;93) = 1.99$ for the 95% confidence level and 93 degrees of freedom, we conclude the alternate hypothesis $H_a$ to be true, contrary to our expectations. This implies that the two models have a significant difference in their impact on predictive accuracy when inaccurate input data set are used. Neural nets seem to perform better than regression in this case.

*Hypothesis 2.* The results of the model that enabled us to test how utility varies as data quality degrades is shown below.

$$U_i = \underset{(41.624;.001)}{\underset{+2.0067}{b_3}} + \underset{(-5.417;.001)}{\underset{-0.0072}{b_4\ \varepsilon_i}} + \underset{(13.709;.001)}{\underset{+0.018527}{b_5\ model_i\ \varepsilon_i}} \tag{13}$$

The estimate, $b_5$ for the coefficient $\beta_5$ was 0.018527, and the standard error for the coefficient was 0.001351. In this case, $t^*$ equals $b_5/s(b_5)$ or 13.709. Here, $|t^*| > t(0.975;3645) = 1.98$ for the 95% confidence level and 3645 degrees of freedom. So, we conclude the alternate hypothesis $H_a$. This implies that the two models exhibit a significant difference in their impact on utility when an inaccurate input data set is used. Neural nets seem to perform better than regression in this case also. Finally, we note that the signs of $b_1$ and $b_2$, the estimates of $\beta_3$ and $\beta_4$, are in the expected direction.

### 9.2. Discussion

The above results show that neural nets perform better than regression -- even with inaccurate input data -- when forecasting mortgage prepayments. So, on average, if inaccuracies up to 10% of the correct values are present in up to 10% of the observations (AMTFRAC = 100) in the forecasting system, a neural nets-based system would be more useful and would have a .04 higher predictive accuracy as compared to a regression-based system. Generally, 5% to 10% errors in 5% to 10% of the observations are common, and if a large sum of money is invested in particular portfolios, then even a small error in the hedge ratio computations can heavily expose a firm to risk. Thus, in such situations, it may be beneficial for a risk manager to consider alternate models for forecasting to improve the usefulness of these predictions.

Why did these results emerge? A possible explanation of our results derives from the method employed by neural nets to create a model. With neural nets forecasting is done in two phases: a learning phase and an actual testing phase. It is possible that the net in the first phase learns how to differentiate errors from the actual data. However, such learning cannot take place in regression-based models, which depend on averaging the readings from all observations.

Another possible explanation is that the neural nets started with a better predictive accuracy than regression even without the errors in the input data set. Their superior performance over regression may simply be a result of a better modeling strategy. In fact, it has been shown that if a forecasting model is non-linear, then a neural net model usually performs better than a linear regression model [LIPP87, DUTT88] . In our example of prepayment predictions, it is possible that a non-linear model may describe prepayments better than a linear model (though this has not been reported by papers we have seen that review relevant industry experience). Hence, the neural net may be the better choice. Moreover, a neural net has the capability to generate underlying relationships between the dependent variables and the independent variables from the input and the output data sets. On the other hand, a regression model requires these relationships in advance; only then can a regression model be used for forecasting. So developing these relationships from the data themselves seems to be a better forecasting strategy.

A reduction in the difference between the actual and the predicted prepayment rate was an important aim for using a neural net. The actual prepayment rate has been shown to follow a typical time-series pattern with a number of peaks, especially during a period of falling interest rates, and an equally large number of troughs, representing the high interest rate points [HAYR89]. Because neural nets are good at capturing both the local and the global maxima and minima, we believe that they were able to predict the peaks and the troughs accurately, thus producing highly reliable hedge ratios. On the other hand, in linear regression, the peaks and troughs of a time-series may never be accurately predicted, because the goal of linear regression is to fit a line that represents their average.

Although it appears that neural nets may be a better choice than regression for MBS forecasting, this may not be the case, especially when the errors in the input datasets are not very high. The key issue is one of cost. Neural nets are generally more expensive to use; they require two sessions compared to one in regression. Also, training a network may take a very long time, depending on the nature of the application. So, if we expect the data to have very low inaccuracy -- say around 1% to 2% inaccuracies in 1% to 2% of the observations (corresponding to AMTFRAC = 4 at the maximum), then a regression-based forecasting may be more cost-efficient. Clearly, a risk manager needs to examine both the cost and the benefit sides before investing a large amount of money in a system that employs one particular model.

## 10. CONCLUSION

In real-world applications, the data that are used cannot be guaranteed to be error-free. This work indicates how the performance of two alternate forecasting approaches compare when data quality varies. Our results suggest that, for the case of

MBS prepayments at least, neural nets have potential to perform more accurate predictions than regression. The primary result we obtained, however, is that neural nets performed better than regression models -- even when the input data were inaccurate.

This research also introduces the idea of using a utility approach for evaluating forecasting systems. In the past, most researchers have compared forecasting systems in terms of their predictive accuracy. However, predictive accuracy optimization alone may not always lead to system designs that deliver business value. It is beneficial to attempt to measure and evaluate the value that accrues, and to recognize that there will likely be many situations where high predictive accuracy may not result automatically in high utility. Predictive accuracy is measured in terms of an average for all observations (or variables), and if one observation (or variable) is more critical than others, then that observation (or variable) may need a higher degree of accuracy.

Thus, a manager who needs to forecast investment risk in the future needs to do three things when deciding on a model/data quality combination to build into an RMFS:

(1) Define the application and a meaningful, implementable utility function, which will be used to compare different forecasting systems.

(2) Simulate the predictive performance and the utility associated with the model/data quality combinations under consideration.

(3) Then, employ the intuition (if not the full details) of information economics to

arrive at an estimate of when the net expected utility of a forecasting system is maximized for the choice variables, model selection and data quality, assuming that data costs are known.

We stress that both the data acquisition costs and the utility need to be taken into account to optimize MBS prepayment RMFS design. And, it is highly likely that for small data errors, a regression-based forecasting system may be more cost-efficient than one based on a neural-net.

This work opens up several avenues for future research. An obvious first step is to extend the range of the errors (maximum AMTFRAC > 100), and assess the performance of alternate models. More interesting to examine empirically would be how to arrive at an optimal model and data quality combination when the decision maker is not risk neutral. In this case, a non-linear utility function would be assumed (unlike the present research, where we have assumed risk neutrality).

Although this research provides a new line of thinking for system design, we should remind the reader of one important caveat: the results of this research were derived in a highly specialized domain of business, prepayment predictions (even though the data set used was large and rich). It is possible that in other applications, and for other utility functions, regression may perform better than neural nets with imperfect input data. Interested readers are encouraged to apply our methodology to ascertain the generalizability of the results. Risk management forecasting for foreign exchange trading is one possible domain. Fluctuations in foreign exchange rates previously have been predicted successfully using both the neural net and the regression approaches, among others [COLI92].

## REFERENCES

[AHIT89] Ahituv, N. "Assessing the Value of Information: Problems and Approaches." In *Proceedings of the Tenth International Conference on Information Systems*, Boston, Massachusetts, 1989, pp. 315-325.

[ALTM77] Altman, E. I., Haldeman, R. and Narayanan, P. "Zeta Analysis: A New Model for Identifying Bankruptcy Risk." *Journal of Banking and Finance*, Vol. 60, pp. 51-58.

[ALTM81] Altman, E. I., Avery, R. B., Eisenbeis, R. A. and Sinkey, J. F., Jr. *Applications of Classification Techniques in Business, Banking and Finance.* JAI Press Inc., Greenwich, Connecticut, 1981.

[ARAK85] Arak, M. and Goodman, L. S. "Prepayment Risk in Ginnie Mae Pools." *Secondary Mortgage Markets*, Spring 1985, pp. 20-25.

[AREN89] Arend, M. "Financial Databases Flag Overseas Opportunities." *Wall Street Computer Review*, October 1989.

[BALL85] Ballou, D. P. and Pazer, H. L. "Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems." *Management Science*, Vol. 31, No. 2, February 1985, pp. 152-162.

[BANS92] Bansal, A. and Kauffman, R. J. "Risk Management and Data Quality Selection: An Empirical Approach." Working Paper, Center for Research on Information Systems, Stern School of Business, New York University, September 1992.

[BARU89] Barua, A., Kriebel, C. H. and Mukhopadhyay, T. "MIS and Information Economics: Augmenting Rich Description with Analytical Rigor in Information Systems Design." In *Proceedings of the Tenth International Conference on Information Systems*, Boston, Massachusetts, 1989, pp. 327-339.

[BEAV67] Beaver, W. H. "Financial Ratios As Predictors of Failure." In *Empirical Research in Accounting: Selected Studies 1966*, a supplement to the *Journal of Accounting Research*, Vol. 4, 1967, pp. 71-111.

[COLI92] Colin, A. "Machine Learning Techniques for Foreign Exchange Trading." In *Proceedings of the Third Annual International Conference on Advanced Trading Technologies*, New York, July, 1992.

[DEAK72] Deakin, E. B. "A Discriminant Analysis of Predictors of Business Failure." *Journal of Accounting Research*, Vol. 10, No. 1, pp. 167-179.

[DEMS85] Demski, J. S. *Information Analysis*. Second Edition, Addison-Wesley Publishing, Reading, MA, 1985.

[DOYL79] Doyle, J. "A Truth Maintenance System." *Artificial Intelligence*, Vol. 12, 1979, pp. 231-272.

[DOHE85] Doherty, N. A. *Corporate Risk Management*. McGraw-Hill, Inc., New York, NY, 1985.

[DUTT88] Dutta, S. and Shekhar, S. "Bond Rating: A Non-Conservative Application of Neural Networks." In *Proceedings of the Second International Conference on Neural Networks*, 1988, pp. 443-450.

[FIGL90] Figlewski, S., Silber, W. L. and Subrahmanyam, M. G. (editors). *Financial Options: From Theory to Practice*. Business One Irwin, Homewood, Illinois, 1990.

[HAUG87] Haugen, R. A. *Introductory Investment Theory*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1987.

[HAYR89] Hayre, L. S., Lauterbach, K. and Mohebbi C. "Prepayment Models and Methodologies." In *Advances and Innovations in the Bond and Mortgage Markets*, edited by Frank J. Fabozzi, 1989.

[HILT81] Hilton, R. W. "The Determinants of Information Value: Synthesizing Some General Results." *Management Science*, Vol. 27, No. 1, 1981, pp. 57-64.

[HINT89] Hinton, G. E. "Connectionist Learning Procedures," *Artificial Intelligence*, September, 1989, pp. 185-234.

[LIPP87] Lippmann, R. P. "An Introduction to Computing with Neural Nets." *IEEE Acoustic Speech and Signal Processing Magazine*, April 1987, pp. 4-22.

[MART77] Martin, D. "Early Warning of Bank Failure: A Logit Regression Approach." *Journal of Banking and Finance*, Vol. 1, pp. 249-276.

[MEYE70] Meyer, P. A. and Pifer, H. W. "Prediction of Bank Failures," *Journal of Finance*, Vol. 25, 1970, pp. 853-868.

[NEUM92] Neumann, P. G. "Inside Risks: Accidental Financial Losses." *Communications of the ACM*, Vol. 35, No. 9, September, 1992, p. 194.

[PINK87] Pinkus, S. M., Hunter, S. M. and Roll, R. "An Introduction to the Market and Mortgage Analysis." In *Advances in Bond Analysis & Portfolio Strategies*, Frank J. Fabozzi and T. Dessa Garlicki (editors), Probus Publishing Company, Chicago, IL, 1987.

[RUME86] Rumelhart, D. E. and McClelland, J. L. *Parallel Distributed Processing*. D. E. Rumelhart, J. L. McClelland, and the PDP Research Group Eds. MIT Press, Cambridge, Mass., 1986.

[SALO87] Salomon Brothers, Inc. "Mortgage Security Prepayment Rate Profile." Daniel Walsh and Michael Beveridge (editors), various issues beginning in April 1987 and ending in February 1990.

[SCHM89] Schmerken, I. "CIOs Orchestrate Trading, Teamwork as New Portfolio Analysis Theme." *Wall Street Computer Review*, November 1989.

[SCHM90A] Schmerken, I. "Software to Ride Out Interest-Rate Turbulence." *Wall Street Computer Review*, March 1990.

[SCHM90B] Schmerken, I. "Wall Street Struggles to Balance Global Risk." *Wall Street Computer Review*," August 1990, pp. 12-56.

[SHAL89] Shale, T. "The Great Risk-Management Systems Failure," *Euromoney*, February 1989.

[STEV87] Stevenson, M. "International Banking a Game of Skill As Well." *The Economist*, March 21, 1987.

[WALD85] Waldman, M. and Gordon, M. "Determining the Yield of a Mortgage Security." In *The Handbook of Mortgage-Backed Securities*," edited by Frank J. Fabozzi, Probus Publishing, Chicago, 1985.

[WHIT88] White, H. "Economic Prediction Using Neural Networks: The Case of IBM Daily Stock Returns." In *Proceedings of Second IEEE International Conference on Neural Nets*, 1988, San Diego, pp. II-451:II-458.

**Table 1. Glossary: Terms Related to Mortgage-Backed Securities Portfolio Management**

| **TERM** | **BRIEF DEFINITION** |
|---|---|
| *Mortgage-backed security (MBS)* | A mortgage-backed security is a bond which is backed by a pool (group) of loans. The principal and interest payments received from the underlying loans are passed through to the bondholders. |
| *Mortgage pool* | A mortgage pool is a group of mortgage loans which have the same underlying characteristics, such as the maturity date, coupon and the origin of the loans, etc. |
| *Prepayment rate* | Most residential mortgages can be prepaid at any time prior to maturity. The percentage of borrowers that choose to exercise this option of prepayment in any given year is measured by *constant percent prepayment* or *CPP*. |
| *Deviation around the maximum utility (DAMU)* | The absolute value of the difference between utility obtained from perfect data versus imperfect data, measured as a percentage of the utility gained from the perfect data. |
| *Fixed-income security* | Fixed-income securities, like bonds, are a pre-defined, limited dollar claim. The dollar receipts from these investments will never exceed this promised claim, though they can fall short when default occurs. |
| *Effective duration* | The effective duration of a fixed-income security is the ratio of the proportional drop in bond price to a small increment in the yield. |
| *Efficient hedge selection* | The efficient hedge formation involves investment in two or more portfolios in such a way so that the overall exposure to investment risk is minimized. To form an efficient hedge an investor generally maintains a long (buy) positions in one portfolio and a short (sell) positions in the others. Thus, whenever some financial states change, one portfolio experiences a gain and the other a loss, resulting in a controlled risk in the total investment. *Effective duration* is the basis of fixed-income hedging applications. |
| *Hedge ratio* | For fixed-income securities, a hedge ratio is defined as the proportional change in prices with respect to interest rate changes. Individual hedge ratios of different portfolios determine their optimal combinations for an efficient hedge formation. |

*Note: For additional background, see ARAK85, FIGL90, HAUG87 and WALD85.*

**Table 2. Methodology Overview for Two Hypothesis Tests**

<table>
<tr><th>TOPIC</th><th>HYPOTHESIS 1</th><th>HYPOTHESIS 2</th></tr>
<tr><td>Effect Explored</td><td>The interaction effect between the type of the model and the amount of inaccuracies on the predictive accuracy.</td><td>The interaction effect between the type of the model and the amount of inaccuracies on the utility of the system.</td></tr>
<tr><td>Models</td><td>$R_i = \beta_0 + \beta_1 \varepsilon_i + \beta_2 \, model_i \, \varepsilon_i$</td><td>$U_i = \beta_3 + \beta_4 \varepsilon_i + \beta_5 \, model_i \, \varepsilon_i$</td></tr>
<tr><td>Variables</td><td colspan="2">Dependent Variables<br>$R_i$ = Predictive accuracy of forecasting system using data set i.<br>$U_i$ = Utility of the forecasting system using data set i.<br>Independent Variables<br>$\varepsilon_i$ = Amount of inaccuracies in data set i.<br>$model_i$ = Qualitative variable indicating model employed in forecast relative to data set i (1 = neural net; 0 = regression).<br>$model_i * \varepsilon_i$ = Interaction effect between the model and the inaccuracy of data set i.<br>Parameters to be Estimated<br>$\beta_0$ ($\beta_3$) = Maximum predictive accuracy (utility) obtained by using the perfect data.<br>$\beta_1$ ($\beta_4$) = Rate of change of predictive accuracy (utility) with a corresponding change in the accuracy of input data.<br>$\beta_2$ ($\beta_5$) = Rate at which the interaction between model type and inaccuracy in the input data affects the predictive accuracy of the forecasting system.</td></tr>
<tr><td>Hypothesis Tests</td><td>To explore the effect, the null hypothesis is $H_0$: $\beta_2 = 0$; the alternate is $H_a$: $\beta_2 \neq 0$.</td><td>The null hypothesis is $H_0$: $\beta_5 = 0$; the alternate is $H_a$: $\beta_5 \neq 0$.</td></tr>
<tr><td>Hypothesis Test Statistic</td><td colspan="2">In both cases, the test is a standard t-test, with $t^* = b_j/s(b_j)$, where<br><br>$b_j$ = An estimate for the coefficient, $\beta_j$.<br>$s(b_j)$ = Standard deviation of the coefficient, $\beta_j$.<br><br>The decision rule for our test is $|t^*| \leq t(1-\alpha/2; df)$, conclude $H_0$, otherwise conclude $H_a$. $\alpha$ is the degree of confidence and df is the total number of degrees of freedom.</td></tr>
<tr><td>Data Used</td><td colspan="2">Prepayment data on 38 MBSs, amounting to 1170 observations total, covering the period from April 1987 to February 1990, were used. Each data set included 51 simulated data accuracy variations, from -10% to +10% of the base value of the variable. In the regression analyses, all observations were included. For neural nets, half of the observations in each of these 51 sets were used to train the data set. After training, the model was tested on the complete data set.</td></tr>
<tr><td>Forecast Performance Measures</td><td>The key measure for evaluation is $R_i$ (actually $R^2$ in regression), the predictive accuracy of the forecast, in the presence of data quality variation.</td><td>Here, the key measure of performance is based on DAMU -- the deviation around the mean utility. We obtained DAMUs for each MBS and for each regression, yielding 38 * 51 = 1958 values in all.</td></tr>
<tr><td>Predicted Results</td><td>Both regression and neural nets yield the same predictive accuracy with imperfect data. Thus, we expected $\beta_2 = 0$.</td><td>Both regression and neural nets yield the same utility, even with imperfect data. Thus, we believed the null hypothesis $\beta_5 = 0$ would hold.</td></tr>
</table>

**Table 3.** **Regression $R^2$s with Data Quality Variations in Terms of FRAC and AMT**

| AMT (in % of the observation value) | FRAC (in % of total number of observations) | | | | | |
|---|---|---|---|---|---|---|
| | 0 | 2 | 4 | 6 | 8 | 10 |
| 0 | 0.706 | --- | --- | --- | --- | --- |
| 1 | --- | 0.705 | 0.705 | 0.705 | 0.705 | 0.705 |
| 2 | --- | 0.705 | 0.704 | 0.703 | 0.702 | 0.702 |
| 3 | --- | 0.704 | 0.702 | 0.700 | 0.699 | 0.698 |
| 4 | --- | 0.703 | 0.699 | 0.697 | 0.695 | 0.693 |
| 5 | --- | 0.701 | 0.697 | 0.693 | 0.691 | 0.689 |
| 6 | --- | 0.702 | 0.695 | 0.690 | 0.687 | 0.684 |
| 7 | --- | 0.699 | 0.692 | 0.686 | 0.683 | 0.680 |
| 8 | --- | 0.697 | 0.689 | 0.683 | 0.679 | 0.676 |
| 9 | --- | 0.696 | 0.687 | 0.679 | 0.676 | 0.673 |
| 10 | --- | 0.694 | 0.684 | 0.676 | 0.673 | 0.670 |

**Table 4.** **Neural Net $R^2$s with Data Quality Variations in Terms of FRAC and AMT.**

| AMT (in % of the observation value) | FRAC (in % of total number of observations) | | | | | |
|---|---|---|---|---|---|---|
| | 0 | 2 | 4 | 6 | 8 | 10 |
| 0 | 0.757 | --- | --- | --- | --- | --- |
| 1 | --- | N/A | 0.747 | 0.748 | 0.748 | 0.748 |
| 2 | --- | N/A | 0.730 | 0.731 | 0.731 | 0.731 |
| 3 | --- | N/A | 0.721 | 0.721 | 0.721 | 0.722 |
| 4 | --- | N/A | 0.714 | 0.719 | 0.748 | 0.748 |
| 5 | --- | N/A | 0.710 | 0.711 | 0.710 | 0.710 |
| 6 | --- | N/A | 0.706 | 0.707 | 0.706 | 0.706 |
| 7 | --- | 0.736 | 0.702 | 0.703 | 0.703 | 0.703 |
| 8 | --- | 0.732 | 0.698 | 0.699 | 0.699 | 0.699 |
| 9 | --- | 0.729 | 0.694 | 0.695 | 0.695 | 0.692 |
| 10 | --- | 0.726 | 0.691 | 0.656 | 0.691 | 0.692 |