The Korean Communications
in Statistics Vol. 7, No. 2, 2000
pp. 605-616

Application of Statistical Models for Default Probability of Loans in Mortgage Companies

Jin-Whan Jung[1)]

Abstract

Three primary interests frequently raised by mortgage companies are introduced and the corresponding statistical approaches for the default probability in mortgage companies are examined. Statistical models considered in this paper are time series, logistic regression, decision tree, neural network, and discrete time models. Usage of the models is illustrated using an artificially modified data set and the corresponding models are evaluated in appropriate manners.

Keywords : Statistical Consulting, Default Probability, Time Series Models, Logistic Regression, Decision Tree, Neural Network, Survival Analysis, and Discrete Time Model

1. Introduction.

In the area of the statistical consulting, we often encounter the clients from finance industries. Those industries include mortgage companies, mortgage insurance companies, credit cards companies, and insurance companies. One of the primary interests for them is to identify those who tend to be default (or bankruptcy). The default occurs quite rare. However, once it happens, the amount of loss for the company is not trivial at all, relative to their profit from a customer. Thus, predicting default probability has been major area of interest. Various statistical approaches depending on the primary interests are applicable in regarding with the default behaviors. Among many interests for the default behavior, two questions are focused in the first two sections. The first one is "What is an overall default rates for next twelve months?" . The second is "Who is more likely to be default in the first year of loan?".

In the middle of the loan period, the default risk of a loan can be instantly changed when extra information is added such as the delinquency patterns and the old information is updated such as employment status and income. In this situation, default risk probabilities in the middle of loan

1) Analytical Consultant, SAS Institute, Cary, NC 27513
E-mail : JinWhan.Jung@sas.com

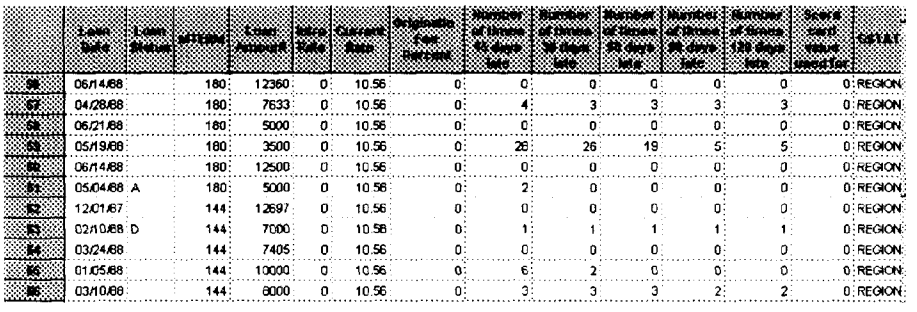
period are of interest. The serious delinquency for certain time periods is often considered as a surrogate measure of the default risk. The status of the serious delinquency for each loan is defined if a borrower does not pay the monthly payment for a long time such as 120 days or more. The third question considered in this paper is "What is the instant hazard (the first serious delinquency) rate for each loan at this time?".

The purpose of this paper is to help statistical consultants to understand the business problems in the finance industry and to provide a prototype to solve the corresponding problems in the statistical manners. In a way of answering the first question, several time series models are examined for the crude overall default rates using past three-year data and their predicted values for next twelve months are calculated using a selected model. The second question is assessed through the three well-known prediction models; they are the logistic regressions model (McCullagh and Nelder, 1989), the neural network with multi-layer perceptrons (Bishop, 1995), and the hybrid tree model from CART (Brieman et al, 1984) and CHAID (Kass, 1980). Finally, the last question for the instant hazard rates is addressed using the discrete time model (Allison, 1982). These applications are illustrated using a data set artificially modified from an actual mortgage company data set.

2. Initial Data Preparation

The information collected from the mortgage companies is usually very big unlike that from typical experimental designs or clinical trial with a few patients. In many situations, the mortgage companies have a lot of customers or accounts (e.g.100,000 or more). The information related with the loan of each customer (or account) is typically updated monthly, so the amount of information becomes even bigger and bigger as time goes on. For example, if the average age of the accounts is 5 years old, then the total cases (or records) are approximately $100,000 \times 12 \times 5 = 6,000,000$. Since the data are so big, they are often stored on the Main Frame or Unix platforms with structuring relational databases to minimize the disk spaces. The databases include ORACLE, SYBASE, and PROGRESS etc. The data structure of the relational database is not suitable for statistical analysis in many cases, thus appropriate data preprocessing steps are required to apply the corresponding statistical models. Each model has one or more appropriate data structures. Based on the previous consulting experience with mortgage companies, an initial data set after joining several relational tables (data sets) has the form like that in Figure 1:

Figure 1: Initial mortgage data structure after joining several relational data sets



Simple descriptions of a few selected variables are listed in below:

- Loan date: the date of the account is opened
- Loan Status: "A" for Active, "C" for Canceled and "D" for Defaulted
- Term: Term of the loan (month)
- Loan amount: Original loan amount
- Introductory rate: Introductory rate applied for some periods before actual interest rate begins
- Interest rate: Actual interest rate
- Origination fee percent: Origination fee percentage that was included in the original loan balance
- Late 15: Number of times the account has been 15 days past due
- Late 30: Number of times the account has been 30 days past due
- Late 60: Number of times the account has been 60 days past due
- Late 90: Number of times the account has been 90 days past due
- Late 120: Number of times the account has been 120 days past due (serious delinquency)
- Score: Score value used for credit decision
- Score cut-off: minimum score for each individuals
- Program: Type of loan services
- State : Location of borrower (grouped ten regions)

3. Time series models for overall default rates and its prediction

The first thing considered here is to help the question "What is the overall (or crude) default rates in next twelve month?". There are several statistical approaches available and time series models are one of them. The structure of the data set in Figure 1 is not appropriate to apply

time series analysis for predicting the crude default rates in next twelve months. The data set is aggregated to form the monthly default rates for time series models. For the simplification of illustrating the proposed method, the default rates after March 1996 are considered. Several time series smoothing models are examined using SAS/ETS (SAS institutes, 1993) and those models include Linear Trend, Simple Exponential Smoothing, Double Exponential Smoothing, Linear Exponential Smoothing, Seasonal Exponential Smoothing, Additive Winters model, and Multiplicative Winters model. After fitting these models, Additive Winters model is selected based on the root mean square criteria (see Table 1).

Table 1: Root mean square errors for six time series models

Time Series Models	Root Mean Square Error
Linear Trend	0.0003473
Simple Exponential Smoothing	0.0003608
Double Exponential Smoothing	0.0004030
Linear Exponential Smoothing	0.0003527
Additive Winters model	0.0003012
Multiplicative Winters model	0.0004067

The Winters method with additive terms has the form as the following:

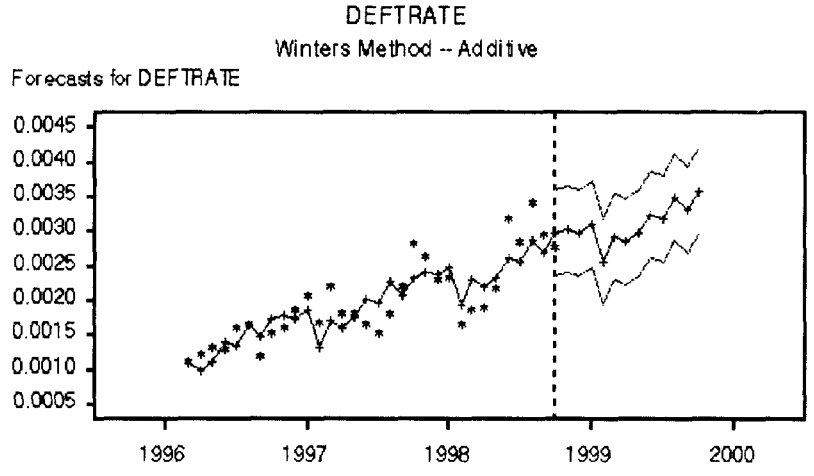
$$Y_t = \mu_t + \beta_t t + S_t + \varepsilon_t$$

where μ_t, β_t, S_t and ε_t denote, the intercept, the linear trend, the seasonal component for the season corresponding to time t and the random errors, respectively. The parameter estimates for the model and the corresponding twelve month predictions with 95 percent confidence interval are shown in Table 2 and Figure 2. The default rates in February, March and April are relatively lower than those in other months in Figure 2 and Table 2. This pattern explains that tax return in U.S. typically happens during these periods. As a result, it makes the default rates in these periods lower than usual. The default rates during the summer (June through September) have the highest default rates. This result may come from a large amount of expenditures during vacation periods. In addition, the default rates during November and January are relatively high due to the holiday season in which a lot of money are usually spent for gift, travel etc.

Table 2: Parameter estimates and the twelve months predictions

Paraneter Estimates		Twelve month Predictions and 95% Confidence Intervals			
Descriptions	Estimates	Descriptions	Predicted Values	95% Lower Limits	95% Upper Limits
Intercept	0.00278	NOV1998	0.003027	0.003647	0.002407
Trend	0.00005	DEC1998	0.002989	0.003609	0.002369
Seasonal Factor 1	0.00017	JAN1998	0.003104	0.003724	0.002484
Seasonal Factor 2	-0.00041	FEB1998	0.002572	0.003193	0.001952
Seasonal Factor 3	-0.00009	MAR1998	0.002940	0.003561	0.002320
Seasonal Factor 4	-0.00024	APR1998	0.002845	0.003466	0.002225
Seasonal Factor 5	-0.00016	MAY1998	0.002976	0.003597	0.002356
Seasonal Factor 6	0.00006	JUN1998	0.003251	0.003871	0.002630
Seasonal Factor 7	-0.00004	JUL1998	0.003197	0.003818	0.002576
Seasonal Factor 8	0.00022	AUG1998	0.003510	0.004131	0.002889
Seasonal Factor 9	-0.00002	SEP1998	0.003322	0.003943	0.002701
Seasonal Factor 10	0.00019				
Seasonal Factor 11	0.00020				
Seasonal Factor 12	0.00011				

Figure 2: Scatter plots with the predicted values and 95% confidence interval



4. Default probability model for individual loans using the information at the time of application

When the customers apply for the loan or just after they open their accounts, the default risk of each account for the first several months or several years is often of interest. Under this consideration, the available information is only baseline information. With the limited information, the default probability of each account in the first year can be considered in several ways. In data mining area, this is a typical set up for the predictive modeling of the risk scores in an individual loan. There are many analytical approaches for the predictive modeling but we consider three well-known approaches (Logistic Regressions, Neural Network, Hybrid Tree Modeling) with this data in this paper.

The original data set is modified for the purpose of fitting these models. Some loans, which do not satisfy the criteria, are excluded. Those loans are

- Non-default loans which are active but do not reach the minimum of the specified periods.
- Loans whose terms are less than one year

Each account has one record with the status of default experience during the first year and other baseline information. The modified data set is randomly split into two data sets (a training data set 70% and a validation data set 30 %). The models are evaluated on the validation data in terms of the percentages of the response in deciles.

The logistic regression is a popular analytical approach when the response (or target) variable is categorical. A standard logistic regression for binary response (i.e., default or non-default) can be expressed as the following:

$$Logit(P_i) = \log\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

where P_i is the default probability of the i-th borrower in the first year and βs are the corresponding parameters for the available inputs (independent variables). For the logistic regression, stepwise selection methods with $\alpha = 0.20$ is applied.

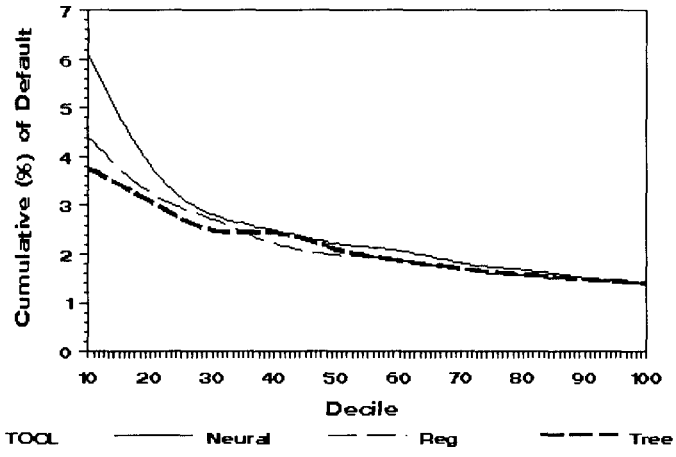
Artificial neural networks are originally developed by researchers who were trying to mimic the neurophysiology of the human brain (Bishop, 1995 and Ripley 1996). Neural networks are especially useful for prediction problems when no mathematical formula is known that relates inputs to outputs (targets). There are several types of the neural networks available such as the feed forward neural network, the radial basis function network, the normalized radial basis neural network, Kohonan network etc. Here, one of the most well known neural networks, the feed forward neural network (a.k.a., multilayer perceptrons), is considered in this paper and the mathematical expression can have the form:

$$g_0^{-1}(P_i) = w_0 + w_1 H_1 + w_2 H_2 + \cdots + w_M H_M$$
$$H_k = \tanh(w_{0k} + w_{1k} x_{1i} + w_{2k} x_{2i} + \cdots + w_{pk} x_{pi})$$

where $k(=1,2,\cdots,M)$ denotes the index of the hidden layers and H_k denotes the k-th hidden layer, w_{jk} denotes the coefficients of the input x_j for the k-th hidden layer, and g_0^{-1} denotes the inverse activation function ($\tanh^{-1}$). Multilayer perceptron neural networks with several neurons (3, 5, 7, 9) are examined as an initial step. The preliminary run with five different seeds and weight decays are used to stabilize the global optimization (Ripley, 1996). When the neural networks with various neurons (3, 5, 7, 9) are compared, the performances of them are pretty much the same in terms of the percentages of response in deciles. But the neural network with seven hidden neurons provides slightly better predictions than other models in terms of the default rates at 10% decile. The detailed results of the neural network comparison steps are omitted because the similar steps are conducted to compare the result from the seven-neuron neural network with those from the logistic regression and tree.

Decision tree is one of the popular and descriptive tools to solve the classification problems. There are a variety of algorithms for building decision trees. Three of the most popular approaches are CART (Breiman et al, 1984), CHAID (Kass, 1980) and C4.5 (Quinlan, 1993). Each approach has its own characteristics such as the ways of handling missing values, splitting criteria, pruning methods, etc. In this paper, a hybrid tree model with Chi-Square splitting criteria (p=0.20) and binary splits in SAS Enterprise Miner is applied (SAS Institute, 1999). Like the comparison of neural networks in the above, the tree, the logistic regression, and the selected neural network models are evaluated on the validation data set in terms of the percentage of default loans on the overlaid decile plot. The cumulative percentages of the default loans in each decile for each model are displayed on the Figure 3.

Figure 3: Cumulative Percentage of Default Loans in each Decile Groups



The cumulative percentages of the default loans in each decile are calculated after sorting the predicted values from each model (tool) with the descending order, and binning the sorted predicted values into 10 groups. In this validation data set, the tree and the logistic regression models do not behave as good as the neural network with seven neurons. Cross-validation can be applied as an alternative but it is not considered in this paper because it is computationally too expensive (Efron, 1983).

5. Default probability model using delinquency history

The model in the previous section is useful when one wants to know about the default probability at the baseline. However, the risk of being default for each loan is often of interest in the middle of the loan period because it can be changed in many reasons such as the unemployment of the borrower, the death of the borrower, minor delinquency patterns and current incomes etc. With this updated information, the risk of the first serious delinquency, (not paying monthly payment more than 120 days), is addressed as a surrogate measure (or symptom) of the default. The time of the serious delinquency for each loan is recorded with monthly scale and this event can be regarded as a discrete time point. In this situation, one approach for the delinquency structure is a discrete time model (Allison, 1982, Allison, 1995). The discrete time model has a functional form:

$$Logit(P_{ti}) = \log\left(\frac{P_{ti}}{1-P_{ti}}\right) = \beta_0 + \cdots + \beta_p x_{ip} + \eta_1 t + \cdots + \eta_j t^j + \delta_1 Z_{t1} + \cdots + \delta_q Z_{tq} + \gamma_1 t x_{i1} + \cdots + \gamma_k t x_{ik}$$

$$= \text{Baseline info} + \text{Loan age terms} + \text{Delinquecy history} + \text{Interactons between loan age and baseline info}$$

where βs denote the corresponding parameters for the fixed information (x), ηs denote the parameters for the polynomial terms for loan age (t), δs denote the parameters for the time varying information such as minor delinquency history and γs denote the interactions between loan age and baseline information. Other interaction terms are not considered in the model.

The original data set does not have the right form for this type of analysis as well. After converting the original data set into using an appropriate data set, it can have the form in Figure 4. Most loans have several records because the serious delinquency status as well as other time varying information is updated monthly.

Figure 4: A sample data structure to fit the discrete time model

	Unique Loan Number	LOANDATE	Loan Amount	Annual Income of Primary Applicant	DDEFAULT	Introductory Rate	LAST30	LAST60	LAST90	AGE_MO
1301	703010	13247	3681 04	35000	0	0	0	0	0	1
1302	703010	13247	3681 04	35000	0	0	0	0	0	2
1303	703010	13247	3681 04	35000	0	0	0	0	0	3
1304	703010	13247	3681 04	35000	0	0	0	0	0	4
1305	703010	13247	3681 04	35000	0	0	0	0	0	5
1306	703010	13247	3681 04	35000	0	0	1	0	0	6
1307	703010	13247	3681 04	35000	0	0	2	1	0	7
1308	703010	13247	3681 04	35000	0	0	3	2	1	8
1309	703010	13247	3681 04	35000	1	0	4	3	2	9
1310	704083	13160	3158	32365	0	0	0	0	0	1
1311	704083	13160	3158	32365	0	0	0	0	0	2
1312	704083	13160	3158	32365	0	0	0	0	0	3
1313	704083	13160	3158	32365	0	0	0	0	0	4
1314	704083	13160	3158	32365	0	0	0	0	0	5
1315	704083	13160	3158	32365	0	0	0	0	0	6

With the data set in the above, the discrete time model can be applied using PROC GENMOD in SAS (SAS Institute, 1996). Minor delinquency (not paying monthly due in 30 days, 60 days and 90 days) information is used as a time varying information in the model. Also, other baseline information such as income, state, program etc. is added to the model. Loan age and its polynomial terms up to 5 degrees are included in the model. After fitting this initial model with the interaction terms between baseline information and loan ages, the backward elimination

method with a 0.05 significance level is applied sequentially. After applying the backward elimination, the final model includes the following variables (grouped program, difference of the risk score from the cut-off, the number of 60 day delinquency, the number of 90 day delinquency, regions, loan age and interactions of loan age with grouped programs and regions). The interaction terms in the final model indicate that the proportional hazards (in terms of odds) assumption is not satisfied for regions and grouped programs.

Two graphs for default rate on the basis of the chosen model are displayed in Figure 5 and Figure 6. The graph in Figure 5 shows the crude default rates across the loan age. The distribution of risk is right skewed and the loans around eight months old have the highest risk in overall. The distributions of default risk across the loan age are examined separately for each loan program in Figure 6 because the interactions between the loan age and the program are significant. The variable GPRG with A through F indicates the program in a mortgage company. Each program has different services such as loan amount, purpose of loans, terms and interest rates. Default risks of some loans (B and C) decrease slowly as the loan age gets older. This might be the loans with long term and relatively lower interest rates. The default risk for other loans (D) rapidly rise and drops, and the loan E and F has very similar shape. These include the loans for tuition of Kinder Garden to 12^{th} grade and for the computer which tends to have high interests. The interpretation about the loan programs is limited in this paper due to the lack of the knowledge in each program.

Figure 5: Crude default rates across the loan age

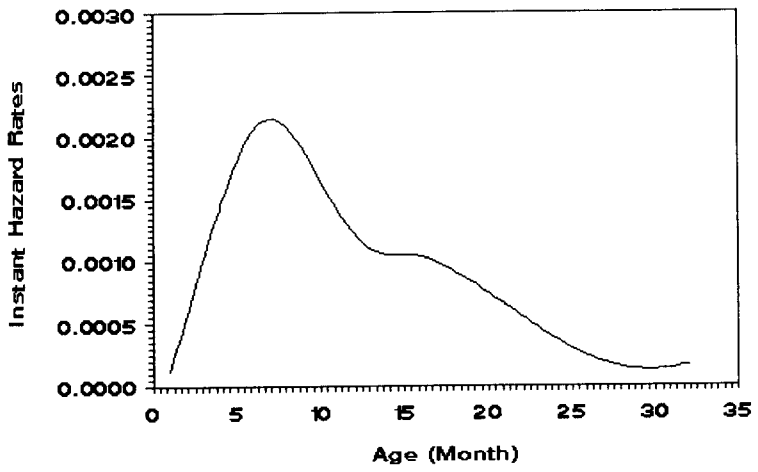
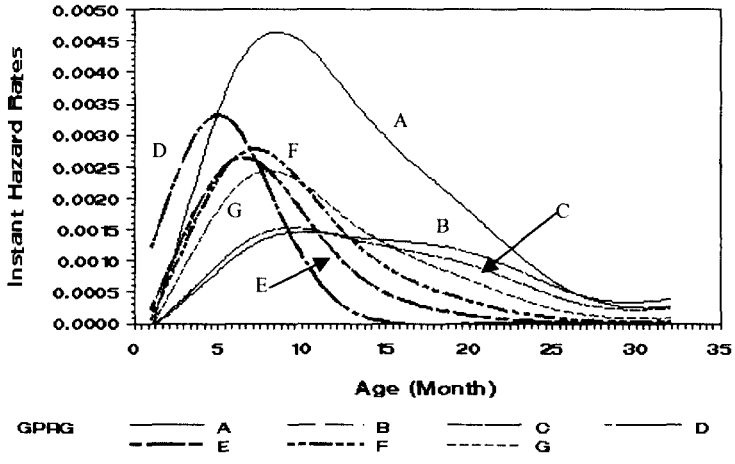


Figure 6: Default rate across the loan age in each grouped program



6. Summary and further consideration

In this paper, three primary interests related with the default risk of the loan accounts are examined using various statistical approaches. Each objective is examined through simple specifications of an appropriate statistical model for illustration purpose. The crude default probabilities for the next twelve months are predicted using several time series smoothing methods. An extended approach such as ARIMA models with extensive specifications and/or other approach using available information such as economic conditions and events can be considered in the future. For the prediction of early default rate (first year) for each loan is examined using three well-known approaches in data mining. The models compared in the paper are somewhat limited. As an extension, one can think about ensemble models which combine the results from multiple models.

For most mortgage companies, the prediction of the default probability and the default amount for the default loan are two major important things. Thus, statistical approaches incorporating with the default amounts are attractive. This can be addressed either with the model with mixture distribution of default amount or with the two stage model (the first model for default probability and then the second model for default loan amount). As shown in section 2, the details of loan status have more than two categories at each time point (i.e., Default, delinquency 30, 60, 90, 120, and none). Since all delinquency status is curable, they are recurrent stage. An interesting approach with delinquency structure is a Markov transition probability model. The transition probability model can be constructed using nominal logistic regression models (Stokes, Davis and Koch, 1995) or neural network models with multiple target variables (Bishop, 1995) after

modifying the original data sets to an appropriate form.

Acknowledgements

I would like to thank Professor Lee Kwan-Jeh and Professor Huh Myung-Hoe for the initial suggestion and comments. I also appreciate the editor and two referees for the helpful suggestions and comments.

References

[1] Allison, P.D. (1982), Discrete-Time Methods for the Analysis of Event Histories *in Sociological Methodology*, 1982, ed. S Leinhardt, San Francisco, CA: Jossey-Bass, 61-98

[2] Allison, P.D. (1995), *Survival Analysis Using the SAS System*, Cary NC, SAS Institute Inc

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

[4] Breiman, L, Friedman, J.H., Olshen, R.A., Stone, C.J.(1984), *Classification and Regression Tree*, Wadsworth: Belmont, C.A.

[5] Efron, B. (1983), Estimating the Error Rate of a Prediction Rule: Improvement on Crossvalidation, *Journal of the American Statistical Association, 78*, 316-331

[6] Kass, G.V.(1980), An exploratory Technique for Investigating Large Quantities of Categorical Data, *Applied Statistics, 29*, 119-127

[7] McCullagh P. and Nelder, J.A. (1989), *Generalized Linear Models, Second Edition*, New York: Chapman Hall.

[8] Quinlan, J.R.(1993), *C4.5: Programs for Machine Learning*, Morgan Kaufman: San Mateo, CA

[9] Ripley B.D. (1996), *Pattern Regognition and Neural Networks*, Cambridge University Press

[10] Sarle, W.S. :(1994), Neural Networks and Statistical Models, *Proceedings of the nineteenth Annual SAS User Group International Conference*

[11] SAS Institute Inc (1993), *SAS/ETS User's Guide, Version 6, Second Edition*, Cary, NC: SAS Institute Inc.

[12] SAS Institute Inc (1996), *SAS/STAT Change and Enhancement, User's Guide, Version 6.12*, Cary, NC: SAS Institute Inc.

[13] SAS Institute Inc (1999), *Applying Data Mining Techniques using Enterprise Miner for version 4, SAS course note*, Cary, NC: SAS Institute Inc.

[14] Stokes, M., Davis, C., Koch, G. (1995), *Categorical Data Analysis Using the SAS System*, Cary, NC., SAS Institute, Inc.