USE OF RECURSIVE PARTITIONING IN THE DEVELOPMENT OF CREDIT SCORING MODELS

Ed LaDue and Mike Novak

Proceedings of a Seminar sponsored by
North Central Regional Project NC-207
"Regulatory, Efficiency and Management Issues Affecting Rural Financial Markets"
New York, NY
September 8-9, 1996

Department of Agricultural and Consumer Economics
College of Agricultural, Consumer and Environmental Sciences
332 Mumford Hall
Urbana, IL 61801

June 1997

Use of Recursive Partitioning in the Development of Credit Scoring Models

Eddy L. LaDue and Michael P. Novak[1]

The farm financial crisis in the mid 1980s brought increased interest in credit evaluation models. Many agricultural lenders and financial advisors have adopted formal credit evaluation models to monitor and forecast financial performance. Various non-parametric and parametric methods have been utilized to estimate the models, such as: experience-based algorithms (Alcott, Splett et al.); mathematical programming (Hardy and Adrian, and Ziari et al.); logistic regression (Mortensen et al.); probit regression (Lufburrow et al. and Miller et al.); discriminate analysis (Hardy and Weed, Dunn and Frey, and Johnson and Hagan); and linear probability regression (Turvey). There is not unanimous agreement on the best method for estimating credit evaluation models and new methods continue to be researched.

Most recently, the logistic regression has dominated the agricultural credit evaluation literature (Miller and LaDue, Turvey and Brown, Novak and LaDue, Splett et al.). The logistic regression succeeded discriminant analysis as the parametric method of choice, primarily based on its more favorable statistical properties (McFadden). Turvey reviews and empirically compares agriculture credit evaluation models using four parametric methods with a single data set. He recommends the logistic regression over the probit regression, discriminant analysis and the linear probability regression based on predictive accuracy and ease of use, in addition to the more favorable statistical properties. While the logistic regression improves on some of the statistical properties of the discriminant analysis and linear probability regression, it still possesses numerous statistical problems common to most parametric methods. These problems include: 1) the need for prior selection of the explanatory variables; 2) difficulty in identifying the individual variable's relative importance; 3) reduction of the information space's dimensionally by limiting variables considered; and 4) difficulty or inability to explicitly incorporate misclassification costs.

One way to circumvent parametric problems is to use a nonparametric classification method, such as Recursive Partitioning Algorithm (RPA). RPA is a computerized, nonparametric classification method that does not impose any a priori distribution assumptions. The essence of RPA is to develop a classification tree that partitions the observations according to binary splits of characteristic variables. The selection and partitioning process occurs repeatedly until no further selection or division of a characteristic variable is possible, or the process is stopped by some predetermined criteria. Ultimately the observations in the terminal nodes of the classification tree are assigned to classification groups. RPA was originally developed by Friedman. A thorough theoretical exposition of RPA is presented in Breiman, et al. A more practical exposition of the computational aspects of RPA is presented in the CART software documentation (Steinberg and Colla).

---

[1] Eddy L. LaDue is a professor of agricultural finance, Department of Agricultural, Resource, and Managerial Economics, Cornell University and Michael P. Novak is an economist with Farmer Mac.

This paper is divided into five sections. The first section presents the specifics of the RPA. This is followed by a section that discusses the advantages and limitations of RPA. The third section describes the data. The fourth and fifth sections present the creditworthiness models and empirical results, respectively. The final section summarizes the paper's results.

Recursive Partitioning Algorithm

In the discussion that follows, a hypothetical RPA tree growing process is presented and terminology is introduced. Following this brief introduction a detailed explanation of the selection of appropriate tree complexity is presented.

To understand the tree growing process, a hypothetical tree is illustrated in Figure 1. It is constructed using classification groups i (bad loans) and j (good loans), and characteristic variables A (equity/asset ratio), B (Return on Assets {ROA}) and C (Current ratio). The specific variables and split values used in this discussion are provided for expository purposes. Throughout the paper the classification groups will be limited to two. But, in general, classification groups can be greater than two. To start the tree growing process all the observations in the original sample, denoted by N, are contained in the parent node which constitutes the first subtree, denoted T_0 (not really a tree, but we will call it one anyway). T_0 possess no binary splits and can be referred to as the naive classification tree. All observations in the original sample are assigned to either group, j (good loans) or i (bad loans), based on an assignment rule. Assuming a special case, when misclassification costs are equal and prior probabilities are equal to the sample proportions of the groups, T_0 would be assigned to the group with the greatest proportion of observations, thus minimizing the number of observations misclassified. When misclassification costs are not equal and prior probabilities are not equal to the sample proportions of the groups, T_0 would be assigned to the group that minimizes the observed expected cost of misclassification[2].

To begin the tree growing process, RPA searches each individual characteristic variable and split value of the characteristic variable in a methodological manner. The computer algorithm then selects a characteristic variable, in this case A (equity/Asset ratio) and a split value of characteristic variable A, in this case a_1 (say .3), based on the optimal splitting rule. The optimal splitting rule implies that no other characteristic variable and split value can increase the amount of correctly classified observations in the two resulting descendent nodes. In this particular illustration, observations with a value of characteristic variable A less that a_1 (.3) will "fall" into the left node and the observations with a value of characteristic variable A greater than a_1 (.3) will "fall" into the right node. The resulting subtree, denoted by T_1, consists of a parent node and, a left and right terminal nodes, denoted by t_L and t_R, respectively. The terminal nodes in each subtree that are then assigned to groups, i (bad loans) or j (good loans), based on the assignment rule of minimizing observed expected cost of misclassification. T_0 and T_1 are the

[2] Observed expected cost of misclassification is calculated using the misclassification rate of group assignments, sample probabilities, cost of type I errors (classifying a less creditworthy borrower as creditworthy) and cost of type II errors (classifying a creditworthy borrower as less creditworthy).

beginning of a sequence of trees that ultimately concludes with T_{max}. However, in some cases, T_1 may also be T_{max} depending on the predetermined penalty parameters specified. If T_1 is not T_{max} then the recursive partitioning algorithm continues.

In this illustration, T_1 is not T_{max}, so the partitioning process continues. Now, B (Return on Assets) is the characteristic variable selected and b_1 (5%) is the "optimal" split value selected by the computer algorithm. The right node becomes an internal node and the observations within it are partitioned again. Observations with a value of characteristic variable B (ROA) less that b_1 (5%) "fall" into a new left node and observations with a value of characteristic variable B greater that b_1 (5%) "fall" into a new right node. The new left and right nodes become terminal nodes in T_2, while the left node in T_1 still remains a terminal model in T_2. All three terminal nodes in T_2 are then assigned to classification groups, i and j, based on the assignment rule of minimum observed expected cost of misclassification. Here again, T_2 does not minimize the observed expected cost of misclassification of the original sample, therefore the partitioning process continues. Variable C (Current ratio) is selected to develop T_3. As stated above, when the recursive partitioning process is finished, the resulting classification tree is known as T_{max}. In this illustration, $T_3 = T_{max}$. T_{max} is the tree that minimizes the expected observed cost of misclassification of the original sample. Obviously the development method will over fit the tree. Therefore, a method is needed to prune back the tree. Once the classification tree is developed and pruned back, it can be used to classify observations from outside the original sample.

T_{max} usually over fits the data, therefore, the resubstitution risk underestimates the "true" risk. In other words, T_{max} classifies the original sample the best, but it is not necessarily representative of the population structure. Therefore, a method is needed to select the appropriate tree complexity from the nested sequence of trees. Some of the methods suggested are v-fold cross-validation, jackknifing, expert judgment, bootstrapping and holdout samples. In this study, cross-validation is used to select that appropriate correct tree complexity and minimize the resubstitution risk. That is, the tree with the smallest cross-validation resubstitution risk estimation is selected and used to predict out-of-sample observations.

Cross-validation randomly divides all the observations in the original sample into V groups of approximately equal size. The observations in V-1 groups are used to grow a tree corresponding to the range of penalty parameters values for which the tree, based on the original sample, was optimal. The observations withheld are then passed though the newly constructed tree and classified. The procedure is repeated V times. Each time the group withheld is passed through the newly constructed tree and classified. The misclassification risk is summed and averaged for all the V-fold cross-validation trials to obtain an overall cross-validation resubstitution risk estimate. The appropriate tree is typically less complex than T_{max}.

While statistical resampling schemes, such as cross-validation, are sufficient to alleviate the statistical over-fitting bias, however, they do not account for intertemporal (ex ante) predictions, the basic objective of credit evaluation models (Joy and Tofeson). That is, credit evaluation models should not only be used to classify borrowers, but to classify borrowers over time and predict future creditworthiness. Previous RPA financial distress studies have only

evaluated the models based on their cross-validated resubstitution risk. In this study, the tree with the minimum cross-validation resubstitution risk will be used to predict creditworthy and less creditworthy borrowers in the forthcoming period. The predicted borrower classifications will then be compared to actual borrower classifications in an out-of-sample period. The misclassification risk of the out-of-sample predictions are calculated and the models are ultimately evaluated with regards to the out-of-sample misclassification risk.

Advantages of RPA

One basic difference between RPA and most other widely used methods is the selection of variables. For credit evaluation models there is no well-developed theory to guide the selection of financial and economic variables. Most proceed heuristically, selecting variables suggested by expert opinions or based on previous credit evaluation models. Credit evaluation models developed using RPA do not require the variables to be selected in advance. The computer algorithm selects the variables from a predetermined group of variables. This feature is especially advantageous if there are a large number of variables. In this context, RPA is somewhat analogous to forward stepwise regression, except RPA is not limited by the mathematical tractability or assumptions of conventional statistics, like forward stepwise regression.

In addition to selecting a group of variables, RPA analyzes univariate attributes of individual variables, providing insight and understanding to their predictive structure. The algorithm selects the variable that best classifies the observations and the optimal spilt value of the variable. When selecting a variable RPA places no limit on the number of times a variable can be utilized. Often the same variable can appear in different parts of the tree. Furthermore, RPA is not significantly influenced by outliers, since splits occur on non-outlier values. Once the split value is selected, the outlier is assigned to a node and the RPA procedure continues.

In addition to selecting variables and their optimal split value, the CART software also provides competitive and surrogate variables and cut-off values listed in order of importance, for each node in the classification tree. Some variables may not appear in the final classification tree, but still can rank high as a competitive and surrogate variable. This list of competitive and surrogate variables give additional insight to the variable's usefulness. Competitive variables are alternative variables with slightly less ability to reduce impurity in the descendent nodes. Surrogate variables are variables that mimic the selected variables and split values, not only on size and composition of the descendent nodes, but also with respect to which observation lands in the left and right descendent node.

Another difference between the two methods is the way they divide the information space into classification regions. RPA repetitiously partitions the information space as the binary tree is formed. In contrast, the logistic regression if implemented as a binary qualitative choice model, partitions the information space into two regions. The logistic regression usually partitions the observations with respect to a prior probability, say c.

The two models also differ in the manner in which they incorporate misclassification cost and prior probabilities. RPA uses misclassification costs and prior probabilities to simultaneously determine variable selection, optimal split value and terminal node group assignments. Changes in the misclassification cost and prior probabilities can change the variables selected and the optimal split value, and, in turn, alter the structure of the classification tree. In contrast, the logistic regression is usually estimated without incorporating misclassification cost and prior probabilities. However, after the logistic regression is estimated a prior probability is used to classify borrowers as creditworthy/less creditworthy. Changes in the prior probability value can affect the predictive accuracy of the logistic regression (Mortensen et al.). Maddala (1983 and 1986) argues that prior probabilities should be taken to be the sample rate for the two groups, even with unequal sampling rates. In general, RPA appears to be more sensitive to misclassification costs and prior probabilities than the logistic regression.

Disadvantages of RPA

A limitation of RPA's variable selection method is that once a variable is selected all the succeeding variables are predicated on the original selected variable, again similar to forward stepwise regression. The tree growing process is intentionally myopic. RPA never looks ahead to see where it is going nor does it try to assess the overall performance of the tree during the splitting process.

In addition RPA provides limited summary statistics. Indications of goodness of fit and significance are modest. This is a disadvantage relative to logistic regression's ability to assign predicted probabilities of creditworthiness to each borrower. RPA can only classify observations into creditworthy or less creditworthy classes, and can not estimate an overall credit score. The predicted probabilities of creditworthiness provide additional quantitative information regarding a borrower. Furthermore, the predicted probabilities can also be converted to a binary creditworthy/less creditworthy classification scheme when a prior probability is specified. Often lenders want to assess the predicted probability of creditworthiness, not only classify borrowers as creditworthy/less creditworthy.

Data

The data for this study were collected from New York State dairy farms in a program jointly sponsored by Cornell Cooperative Extension and the Department of Agricultural, Resource and Managerial Economics at the New York State College of Agriculture and Life Sciences, Cornell University. Seventy farms have been Dairy Farm Business Management (DFBS) cooperators from 1985 through 1993. Data for these seventy farms are analyzed in this study. Such a data set is critical in studying the dynamic effects of farm creditworthiness[3]. The

[3] Two types of estimation biases that typically plague credit evaluation models are choice bias and selection bias. Choice bias occurs when the researcher first observes the dependent variable and then draws the sample based on that knowledge. This process of sample selection typically causes an "oversampling" of financial distress firms. To over come choice bias, this study selects the sample first and then calculates the dependent variable.

The other type of bias plaguing credit evaluation models is selection bias. Selection bias is a function of the

farms represent a segment of New York State dairy farms which value consistent contribution of financial and management information. The financial information collected includes the essential components for deriving a complete set of sixteen financial ratios and measures as recommended by the FFSC. Table 1 exhibits all sixteen mean values of the financial ratios and measures for the seventy farms over the sample period[4]. Additional farm productivity, cost management and profitability statistics for these farms are summarized in Smith, Knoblauch, and Putnam (1993).

Development of the Creditworthiness Models

The RPA and logistic regression methods are used to estimate models of creditworthiness. Each model classifies creditworthy and less creditworthy borrowers using one period lagged characteristic values. The annual models are developed using 1985, 1986, 1987, 1988, and 1989 characteristic values to classify 1986, 1987, 1988, 1989 and 1990 creditworthy and less creditworthy borrowers, respectively. To evaluate the RPA and logistic regression models 1990, 1991 and 1992 characteristic values are used to predict 1991, 1992 and 1993 creditworthy and less creditworthy borrowers, respectfully. Finally, the predicted creditworthy classifications are compared to the actual classifications for 1991, 1992 and 1993 to determine the intertemporal efficacy of the models. Similarly, the two year average RPA and logistic regression models, used 1985-1986 and 1987-88 averages of the characteristic values to classify creditworthy borrowers in the average periods 1987-88 and 1989-90, respectfully. The models were evaluated using 1989-90 average characteristics values to predict 1991-92 average creditworthy and less creditworthy borrowers. The three-year average model used 1985-86-87 average characteristic variables to classify 1988-89-90 average creditworthy and less creditworthy borrowers. Similarly, 1988-89-90 average characteristic values were used to predict 1991-92-93 average creditworthy and less creditworthy borrowers. In both the two-year and three-year average models the predicted creditworthy and less creditworthy percentages were compared to actual classifications, again to determine the intertemporal efficiency of the models.

The dependent variable in these models is creditworthiness as measured by the debt coverage ratio (or more formally, the term debt and capital lease coverage ratio). A key

---

nonrandomness of the data and can asymptotically bias the model's parameters and probabilities (Heckman). There are typically two ways selection bias can affect credit evaluation models. First, financially distressed borrowers are less likely to keep accurate records, therefore, these borrower are not included in the sample (Zmijewski et al.). And secondly, through the attrition rate of borrowers, because panel data are usually employed. In this study, there were borrowers who participated in the DFBS program during the earlier years of sample period, but exited the industry or stopped submitting records to the data base before the end of the sample period. In analyzing financial distress models, Zmijweski et al. found selection bias causes no significant changes in the overall classification and prediction rate. Given Zmijweski's results the study does not correct for selection bias and proceeds to estimate the credit evaluation models with the data presented.

[4] Some of the borrowers reported zero liabilities, therefore, their current ratio and coverage ratio could not be calculated. To retain these borrowers in the sample and avoid values of infinity, the current ratios were given a value of 7, indicating strong liquidity, and the coverage ratio value was bound to -4 to 15 interval. The bounded internal of the coverage ratio indicates both extremes of debt repayment capacity.

component of this data set was the planned/scheduled principal and interest payment on debt collection in the prior year. This component reflects the borrowers expectations of principal and interest repayments in the up-coming year. Having these components allows the calculation of the coverage ratio.[5] The coverage ratio estimates whether the borrower generated enough income to meet all expected payments and is used as an indicator of creditworthiness in this study. The coverage ratio as an indicator of creditworthiness, based on actual financial statements, has been introduced to credit scoring models as an alternative that overcomes some of the difficulties associated with loan classification and loan default models[6] (Novak and LaDue (1994), and Khoju and Barry). Creditworthiness models are aligned with cash-flow or performance-based lending.

Basically, this definition of creditworthiness assumes that a "good" borrower is one who generates sufficient funds to make the loan payments. If the collateral has to be called upon to pay off the loan, it is unprofitable because of the legal fees, large loan officer time requirement and associated costs.

The coverage ratio, a quantitative indicator of creditworthiness, is not always a sufficient indicator of creditworthiness, because lenders have to ultimately make a credit decision to grant or deny a credit request, therefore, the coverage ratio needs to be converted to a binary variable. Typically, when converting the coverage ratio to a binary variable an a priori, cut-off level of one is specified. That is, a coverage ratio greater (less) then one identifies a borrower as creditworthy

---

[5] If not specified otherwise, the coverage ratio refers to the term debt and capital lease coverage ratio as defined by the FFSC.

[6] Historically, agricultural credit evaluation models have been predicated on predicting bank examiners' or credit reviewers' loan classification schemes (Johnson and Hagan; Dunn and Frey; Hardy and Weed; Lufburrow et al.; Hardy and Adrian; Hardy et al., Turvey and Brown, and Oltman). These studies have assessed the ability of statistical, mathematical or judgmental methods to replicate expert judgment. However, these models present some problems when credit evaluation is concerned. It is difficult to determine whether the error is due to the model or bank examiners' or credit reviewers' loan classification. These problems are not limited to agricultural credit evaluation models (Maris et al., and Dietrich and Kaplan).

To improve on loan classification models, some agricultural credit evaluation studies have used default (Miller and LaDue, and Mortensen et al.). However, while default is inherently a more objective measure, then loan classification, it has problems of its own. First, what is default? Is default measured on a monthly, quarterly, or annual basis. Second, lenders and borrowers can influence default classification. Lenders can influence default classifications by decisions to forebear, restructure, or grant additional credit to repay a delinquent loan. A borrower can influence or delay default by selling assets, depleting credit reserves, seeking off-farm employment, and other similar activities. Third, default is also usually based on a single lender's criteria, therefore, borrowers with split credit can be current with one lender and delinquent or arrears with another lender. Because of these ambiguities in definition, default can lead to the same problems incurred when modeling a bank examiners' or credit reviewers' loan classification scheme. The source of error in the model is indistinguishable, possibly resulting from the model and/or possibly resulting from the bank's definition of default. Fourth, default occurs after a credit request has been granted. A creditworthiness model can be used to determine if a credit request should be granted or denied.

(less creditworthy)[7]. Furthermore, the two-year and three-year average coverage ratio has been found to provide a more stable, extended indicator of creditworthiness (Novak and LaDue, 1996).

RPA does not require the characteristic (dependent) variables to be predetermined. All sixteen FFSC recommended ratios and measures, and lagged classification variables were included in the population set. Many of the variables represent similar financial concepts, but were still included in the population set, allowing RPA to select the most appropriate variables. In addition, the predicted probability of creditworthiness from the logistic regression model was included as a possible characteristic variable.

The logistic regression model requires the characteristic or explanatory variables to be selected in advance. This study follows previous studies and specifies a parsimonious credit scoring model, where a borrower's creditworthiness is a function of solvency, liquidity and lagged debt repayment capacity (Miller and LaDue, Miller et al. and Novak and LaDue, 1996). The specific variables used in the model are debt/asset ratio, current ratio and a binary, lagged dependent variable[8].

The utilization of both estimation methods requires the specification of an appropriate prior probability. In the RPA the prior probability enters into the development of the model. The specified prior probability is essential in the development of the tree and the variables selected. The logit model development is not explicitly dependent on the prior probability. However, a prior probability needs to be specified in order to classify the observations. In this study, the total sample's prior probability of being creditworthy is used. The values are 0.852, 0.896 and 0.905 for the annual, two-year average and three-year average periods, respectfully. The prior probabilities demonstrate that it is easier to be classified as a creditworthy borrower as the average period lengths, using a prior cut-off value of one.

In addition to prior probabilities, misclassification cost also need to be specified. Previous agricultural credit evaluation models either ignore misclassification costs or assume they are equal. However, it is evident that the misclassification costs may not be the same for all types of decisions. The cost of granting or renewing a loan to a less creditworthy borrower is incurred through loan losses, legal fees and loan servicing, which is not the same as not granting

---

[7] The terminology less creditworthy is used instead of not creditworthy, because it is recognized that the farms in the data sample have been in operation over a nine year period and most of them have utilized some form of debt over this period. However, the sample includes FmHA, Farm Credit and various private banks borrowers. The various lending institutions bring varying degrees of creditworthiness amongst the borrowers to the sample. Creditworthiness to one lender may be less creditworthy to another. The data set can be viewed as a compilation of lender's portfolios.

[8] Two other logistic regression models, a stepwise and "eight variable" model (the later, was presented in Novak and LaDue, 1994) were also estimated for annual, two-year and three-year average periods. The results are not reported, because the parameters did not always have the expected sign and the within sample and out-of-sample prediction rates were lower than RPA's and paramoninous (three variable) logit model's prediction rates for all the comparable time periods.

or renewing a loan to a creditworthy borrower where the costs are the foregone profits. This study does not estimate the cost of these misclassifications, but demonstrates the classification sensitivity of these costs. The relative cost of Type I and Type II misclassification errors are varied accordingly from 1:1, 1:2, 1:3, 1:4 and 1:5, with the relatively higher misclassification cost put on Type I error[9]. While the less creditworthy measure used in this model may not be as serious as a actual loan losses or a bankrupt borrower. There is still a higher cost associated to loan servicing, payment collection, and loan analysis for less creditworthy borrowers.

Comparison of RPA and Logit Model Results

Figure 2 presents the tree model generated from the RPA for the annual time period when the misclassification cost of a type I error is three times greater than that of a type II error. The model is simple, comprised of the coverage ratio lagged one period. Borrowers with a coverage ratio greater than 1.50 are classified as creditworthy and borrowers with a coverage ratio less than 1.50 are classified as less creditworthy. In other words, to insure all payments will be made by the borrower in the next year the current coverage ratio needs to be greater than 1.50.

In the same figure, below the classification tree, five surrogate variables are listed. These variables were selected on their ability to mimic the selected variable, the coverage ratio, and split value, 1.50. Repayment ability measured by the repayment margin and binary, lagged dependent variable are included in the list and appear to be good surrogate variables. Another variable selected by the algorithm, as a surrogate variable, is the borrower's predicted probability of creditworthiness resulting the logistic regression. The selection of predicted probability adds some additional validity to the use of the variable as a credit score. Also noteworthy is that the split value of the predicted probability is very similar to the prior probability of the annual sample period. A list of competitor variables are also listed near the bottom of the figure. The repayment margin was listed as the first competitor variable. This implies that if the coverage ratio was restricted or eliminated from the sample the repayment margin would have been selected as the primary variable in the classification tree. In addition to the selection of the surrogate and competitor variables the split values are listed to provide additional univariate insight to the classification process.

In figure 3 the two-year average classification tree is presented, again for a 3:1 relative cost of misclassification, with the higher misclassification cost attributed to a type I error. In this classification tree the repayment margin was selected as the primary characteristic variable and the coverage ratio was selected as the first competitor and surrogate variable. Selection of the repayment margin is somewhat surprising since it is generally perceived to be strongly influenced by farm size. Similar to the annual model the binary lagged dependent variable was also selected as surrogate and competitor variables and the predicted probability was also selected as a surrogate variable. The other variables selected were net farm income, return on equity and

[9] Type I error is a less creditworthy borrower classified as a creditworthy borrower and a Type II error is a creditworthy borrower classified as a less creditworthy borrower.

operating expense ratio. These variables represent a borrower's financial efficiency and profitability.

In figure 4 the classification tree for the three-year average period is presented. Again as a comparison to the previous two trees, a 3:1 relative misclassification cost ratio is used. In the tree the repayment margin was selected as the primary variable characteristic and the coverage ratio was selected as the surrogate and competitor variables. In this time period, the binary lagged dependent variable or predicted probability were not selected as either competitor or surrogate variables. The selected competitor and surrogate variables were operating expense ratio, net farm income, rate of return on assets, operating profit margin ratio and interest expense ratio.

All the ratios and measures selected as surrogate or competitive variables in the two-year and three-year time periods represent a borrower's repayment capacity, financial efficiency or profitability. A borrower's solvency and liquidity position does not appear as useful in classifying two-year and three-year average indicators of borrowers' creditworthiness. However, solvency was useful in the annual classification tree as a competitor variable.

The estimated logistic regression models are presented in Table 2. All the parameters, for each of the models have the expected sign. In the annual model the debt/asset ratio and the binary lagged dependent parameters are significant at the 95% level. In the two-year average model the binary lagged dependent variable is significant at the 99% level.

The within- and out-of-sample misclassification rates of the RPA and logistic regression models are presented in table 3 and 4, respectively. Historically, agricultural credit evaluation models have been evaluated on their misclassification rates. A comparison of the within sample misclassification rates indicates that the RPA model classifies the observation better than the logistic regression for relative misclassification cost of 1:1 and 2:1. In this case, the RPA model is also the naive model. That is, the computer algorithm concluded that if relative misclassification cost were equal or occurring at a 2:1 ratio that a lenders should classify all borrowers as creditworthy. When the relative misclassification cost ratios are assumed to be higher (i.e. 3:1, 4:1 and 5:1) the logistic regression does better at classifying the borrowers. Note that the logistic regression model does not change with the misclassification costs scheme as does the RPA model, because the logistic model does not consider misclassification costs. Furthermore, comparison based on misclassification rates does not account for misclassification costs. A comparison of the out-of-sample misclassification rates indicates that the RPA does better at classifying borrowers in 1991 for all relative misclassification cost scenarios. The logistic regression models does better at classifying borrowers in 1992 and 1993 for all the relative misclassification cost scenarios.

When the two-year average data are used to develop the models, the misclassification results indicate that the RPA model does better at classifying the within-sample borrowers for all relative misclassification cost scenarios. Again, when relative costs are equal, the computer algorithm concludes that all borrowers should be classified as creditworthy in order to minimize the cost of misclassification. In contrast, the logistic model does better when classifying out-of-sample borrowers for the two-year average period. The three-year average data indicate that the RPA does best at classifying both the within- and out-of-sample borrowers for all relative classification cost scenarios.

Next, the models are evaluated on minimizing expected cost of misclassification. It can be seen from the misclassification results that as the relative cost of misclassification rate increases, RPA takes these costs into account when developing a classification tree and the classification tree can be altered. In some cases, the misclassification rate increases as the misclassification costs increase, but the overall expected cost of misclassification is lower.

In Table 5 the expected cost of misclassification for each model and relative misclassification cost is presented. The RPA model does best at minimizing the expected misclassification cost for the annual, two-year average and three-year average time periods for all relative misclassification costs scenarios. This is not surprising since this is the objective of RPA is to minimize the expected cost of misclassification, while the objective of the logistic regression is to maximize the likelihood function for the specific data set. Previous nonagricultural financial stress studies with similar results have concluded, at this point in their research, that RPA is a better model for minimizing expected misclassification cost.

Using the annual time period data, the RPA model minimized expected misclassification cost in 1991 for all relative misclassification costs scenarios, and in 1992 and 1993, when the misclassification costs are equal. Recall that the annual RPA model was also the naive model when the misclassification costs are equal. Again this implies that all the borrowers should be classified as creditworthy when misclassification costs are equal. In other words, this is the best model when misclassification costs are assumed to be equal. However, the assumption that misclassification costs are equal is not very realistic in credit screening models. Using the same data, the logistic regression model does best at classifying observations from 1992 and 1993, when misclassification costs are unequal. Likewise, the logistic regression does better at minimizing the expected cost of misclassification for the two-year average out-of-sample observations when relative misclassification are unequal. When misclassification costs are equal, the RPA model, represented by the naive model, does better. Lastly, the three-year average RPA model does better at minimizing the expected costs of misclassification when predicting out-of-sample observation for each of the relative misclassification scenarios.

Conclusion

In sum, there is obviously no dominate model for classifying borrowers in this data set. The RPA model dominates some of the time, while the logistic regression dominates other times. However, each model has its own attributes. The RPA model selects individual characteristics

variables and split values that are most useful in classifying the borrowers, and provides a list of competitor and surrogate variables with corresponding split values. This type of analysis leads to greater univariate insight of the data. The logistic regression model provides parameters and predicted probabilities of creditworthiness in which to quantitatively score the borrowers. However, these scores need to be converted to a binary variable in order to decide whether to grant or deny a credit request. Furthermore, the RPA model indicated some additional validity to the logistic regression's predicted probability of creditworthiness, by selecting it as a competitor or surrogate variable.

In addition misclassification cost can have an affect on which model is selected. This data set indicates that if the assumption of equal misclassification cost is assumed, the best model is the naive model and all the observations are classified as creditworthy. However, it is unrealistic for a lender to assume all borrowers are creditworthy, therefore misclassification costs need to be considered when evaluating these models.

This study also concurs with previous nonagricultural financial stress studies that RPA does minimize the expected cost of misclassification in comparison to other classification methods when the within sample data are utilized. However, the results do not concur when intertemporal out-of-sample observations are utilized, which is the obvious application of credit evaluation models.

References

Betubiza, E. and D.J. Leatham. "A Review of Agricultural Credit Assessment Research and Annotated Bibliography." Texas Experiment Station, Texas A&M University System, College Station, Texas, June 1990.

Breiman, L., J.H. Friedman, R.A. Olshen, and C.J. Stone. Classification and Regression Trees. Belmont CA: Wadsworth International Group 1984.

Dietrich, J.R. and R.S. Kaplan. "Empirical Analysis of the Commercial Loan Classification Decision." The Accounting Review. 57(1982):18-38.

Dunn, D.J. and T.L. Frey. "Discriminant Analysis of Loans for Cash Grain Farms." Agr. Fin. Rev. 36(1976):60-66.

Farm Financial Standard Council. Financial Guidelines for Agricultural Producers: Recommendations of the Farm Financial Standards Council (Revised) 1995.

Friedman, J.H. "A Recursive Partitioning Decision Rule for Nonparametric Classification." IEEE Transactions on Computers, April (1977): 404-09.

Hardy, W.E. Jr. and J.L. Adrian, Jr. "A Linear Programming Alternative to Discriminant Analysis in Credit Scoring." Agribusiness 1 (1985):285-292.

Hardy, W.E., Jr., S.R. Spurlock, D.R. Parrish and L.A. Benoist. "An Analysis of Factors that Affect the Quality of Federal Land Bank Loan." Southern Journal of Agricultural Economics. 19 (1987):175-182.

Hardy, W.E. and J.B. Weed. "Objective Evaluation for Agricultural Lending." Southern Journal of Agricultural Economics. 12(1980):159-64.

Johnson, R.B. and A.R. Hagan. "Agricultural Loan Evaluation with Discriminant Analysis." Southern Journal of Agricultural Economics. 5(1973):57-62.

Khoju, M.R. and P.J. Barry. "Business Performance Based Credit Scoring Models: A New Approach to Credit Evaluation." Proceedings North Central Region Project NC-207 "Regulatory Efficiency and Management Issues Affecting Rural Financial Markets" Federal Reserve Bank of Chicago, Chicago Illinois, October 4-5, 1993.

Lufburrow, J., P.J. Barry and B.L. Dixon. "Credit Scoring for Farm Loan Pricing." Agr. Fin. Rev. 44 (1984):8-14.

Maddala, G.S. Limited-Dependent and Qualitative Variables in Econometrics. Cambridge University Press. 1983

Madalla, G.S. "Econometric Issues in the Empirical Analysis of Thrift Institutions' Insolvency and Failure." Federal Home Loan Bank Board, Invited Research Working Paper 56, October 1986.

McFadden, D. "A Comment on Discriminate Analysis versus LOGIT Analysis." Annuals of Economics and Social Measurement 5(1976):511-23.

Miller, L.H., P. Barry, C. DeVuyst, D.A. Lins and B.J. Sherrick. "Farmer Mac Credit Risk and Capital Adequacy." Agr. Fin. Rev. 54 (1994):66-79.

Miller, L.H. and E.L. LaDue. "Credit Assessment Models for Farm Borrowers:A Logit Analysis." Agr. Fin. Rev. 49(1989):22-36.

Mortensen, T.D., L. Watt, and F.L. Leistritz. "Predicting Probability of Loan Default." Agr. Fin. Rev. 48(1988):60-76.

Novak, M.P. and E.L. LaDue. "An Analysis of Multiperiod Agricultural Credit Evaluation Models for New York Dairy Farms." Agr. Fin. Rev. 54(1994):47-57.

Novak, M.P. and E.L. LaDue. "Stabilizing and Extending, Qualitative and Quantitative Measure in Multiperiod Agricultural Credit Evaluation Model." Agr. Fin. Rev. (1996) (forthcoming)

Oltman, A.W. "Aggregate Loan Quality Assessment in the Search for Related Credit-Scoring Model." Agr. Fin. Rev. 54 (1994):94-107.

Smith, S.F., W.A. Knoblauch, and L.D. Putnam. "Dairy Farm Management Business Summary, New York State, 1993" Department of Agricultural, Resource, and Managerial Economics, Cornell University, Ithaca, NY. September 1994. R.B.94-07.

Splett, N.S., P.J. Barry, B.L. Dixon, and P.N. Ellinger. "A Joint Experience and Statistical Approach to Credit Scoring." Agr. Fin. Rev. 54 (1994):39-54.

Steinberg, D. and P. Colla. CART Tree-structured Non-Parametric Data Analysis. San Diego, CA: Salford Systems, 1995.

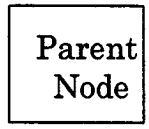
Turvey, C.G. "Credit Scoring for Agricultural Loans: A Review with Application". Agr. Fin. Rev. 51 (1991):43-54.

Turvey, C.G. and R. Brown. "Credit Scoring for Federal Lending Institutions: The Case of Canada's Farm Credit Corporations." Agr. Fin. Rev. 50(1990):47-57.

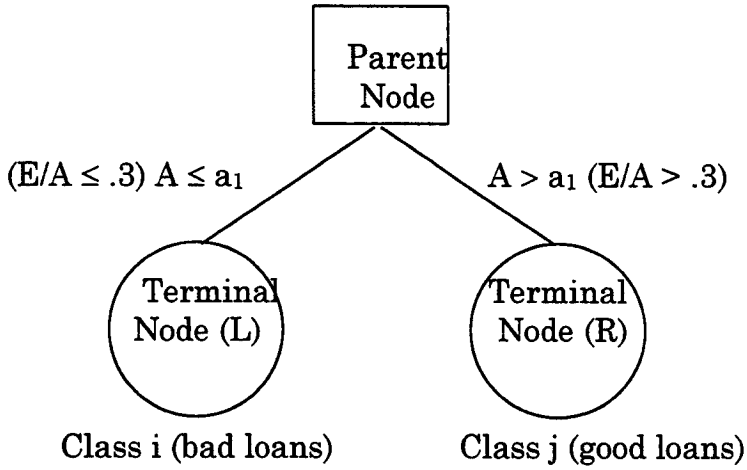
Ziari, H.A., D.J. Leatham, and Calum G. Turvey. "Application of Mathematical Programming Techniques in Credit Scoring of Agricultural Loans." Agr. Fin. Rev. 55(1995):74-88.

Zmijewski, M.E. "Methodological Issues Related to the Estimation of Financial Distress Prediction Models." Journal of Accounting Research Supplement 22 (1994):59-86.

Tree T_0



Tree T_1



Tree T_2

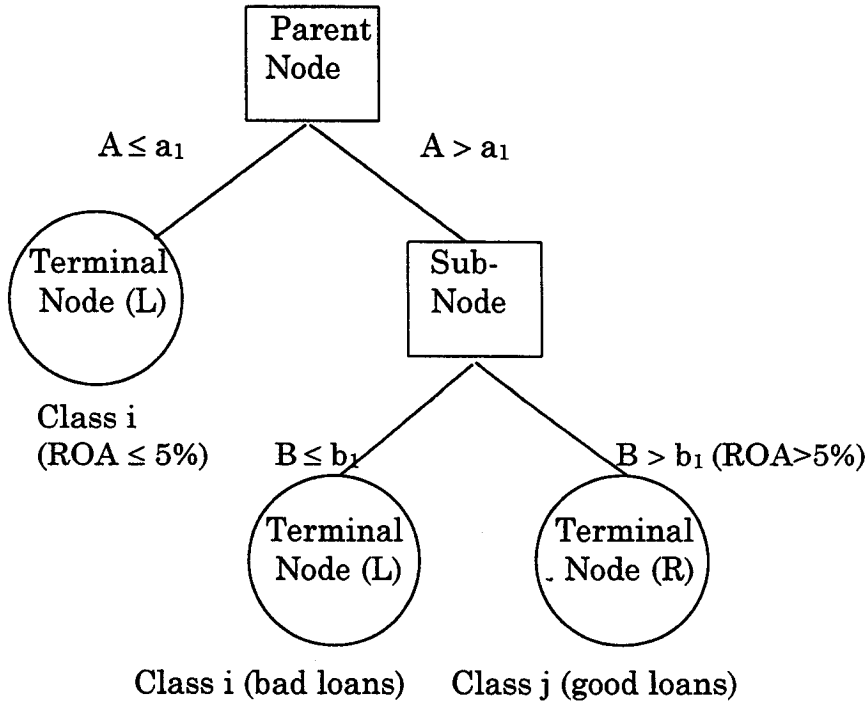


Figure 1. Hypothetical Recursive Partitioning Algorithm Tree

Tree T_3

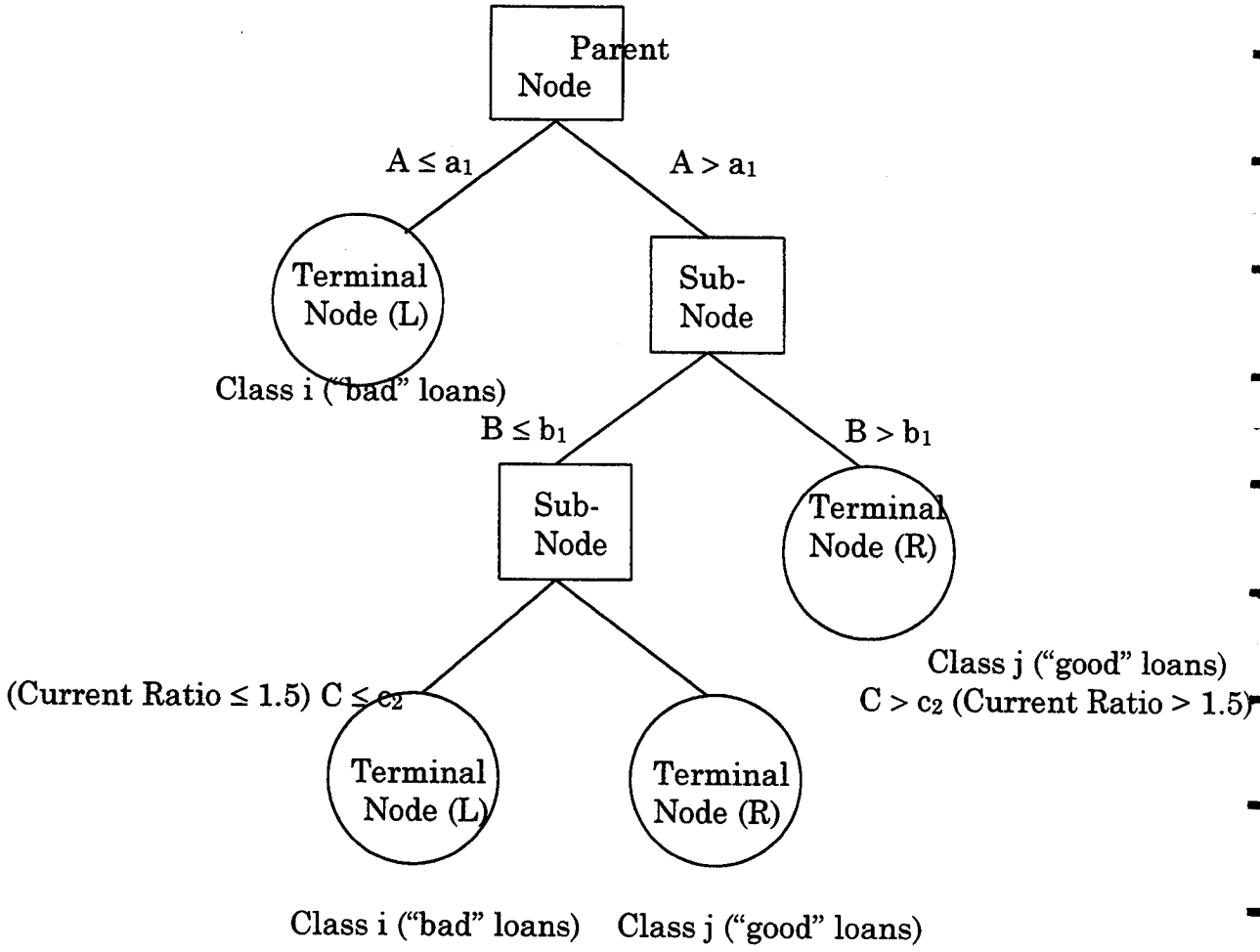
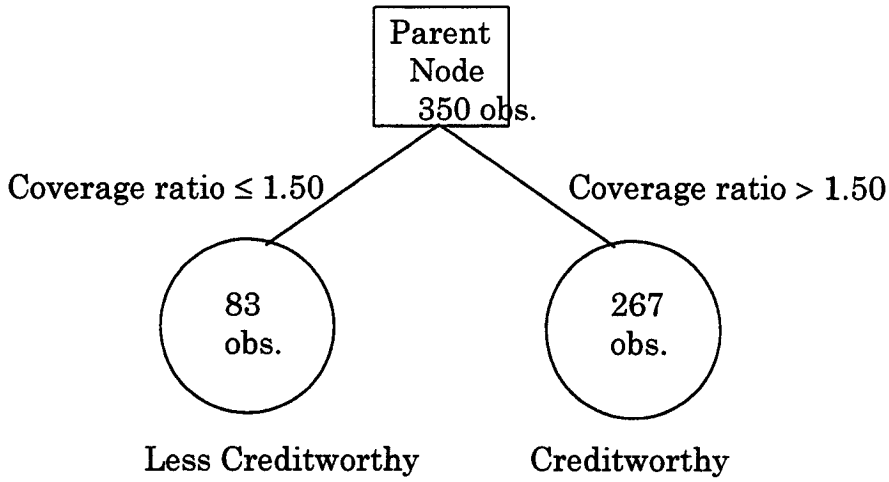
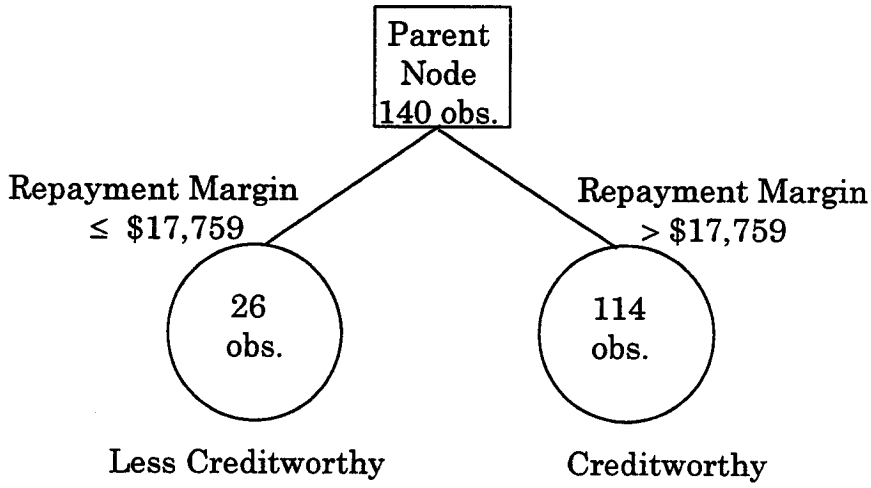


Figure 1 (Continued). Hypothetical Recursive Partitioning Algorithm Tree



Surrogate Variables	Split Values
1. Capital replacement and term debt repayment margin	$18,552
2. Net farm income from operations ratio	0.181
3. Binary lagged dependent variable	0.500
4. Predicted probability of creditworthiness	0.837
5. Operating expense ratio	0.747
Competitor Variables	
1. Capital replacement and term debt repayment margin	$18,419
2. Debt/equity ratio	0.408
3. Debt/asset ratio	0.290
4. Operating expense ratio	0.640
5. Operating profit margin ratio	0.152

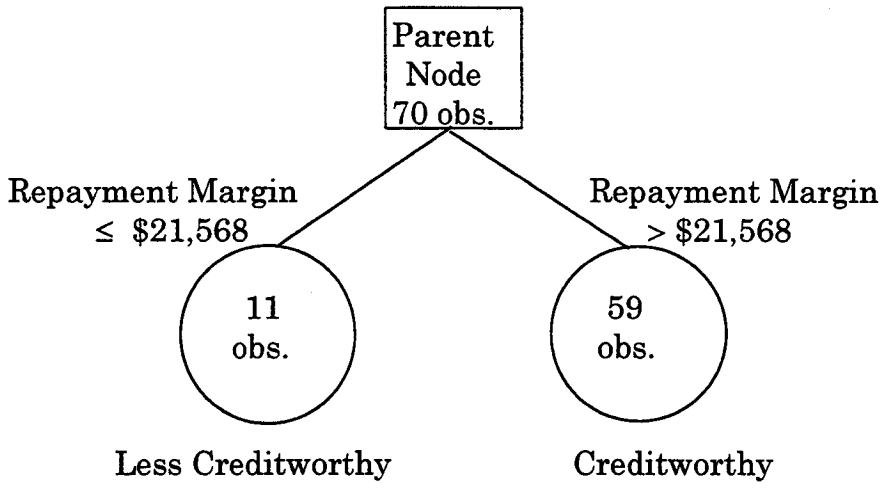
Figure 2. RPA Tree Using Annual Data



Surrogate Variables	Split Values
1. Term debt and capital lease coverage ratio	1.405
2. Predicted probability of creditworthiness	0.818
3. Binary lagged dependent variable	0.500
4. Net farm income	$ 22,922
5. Interest expense ratio	0.158

Competitor Variables	
1. Term debt and capital lease coverage ratio	1.698
2. Operating expense ratio	0.749
3. Predicted probability of creditworthiness	0.853
4. Rate of return on equity	0.013
5. Net farm income	$69,172

Figure 3. RPA Tree Using Two-Year Average Data



Surrogate Variables	Split Values
1. Term debt and capital lease overage ratio	1.429
2. Operating expense ratio	0.748
3. Net farm income	$22,265
4. Rate of return of assets	0.046
5. Current ratio	0.856

Competitor Variables	
1. Term debt and capital lease coverage ratio	1.663
2. Operating expense ratio	0.748
3. Rate of return on assets	0.046
4. Interest expense ratio	0.277
5. Operating profit margin ratio	0.158

Figure 4. RPA Tree Using Three-Year Average Data

Table 1. Mean Value of the Sixteen FFCS Recommended Financial Ratios and Measures, 70 New York Dairy Farms, 1985-93.

Ratio/Measure	1985	1986	1987	1988	1989	1990	1991	1992	1993
Liquidity									
Current Ratio	2.89	2.94	3.06	3.25	3.48	2.90	2.77	2.59	2.50
Working Capital ($)	52,711	49,111	63,799	70,272	84,755	65,891	53,295	57,443	40,148
Solvency									
Debt/Asset Ratio	0.34	0.34	0.31	0.30	0.27	0.28	0.29	0.29	0.29
Equity/Asset Ratio	0.66	0.66	0.69	0.70	0.73	0.72	0.71	0.71	0.71
Debt/Equity Ratio	-0.58	2.17	0.73	0.63	0.51	0.51	0.56	0.56	0.53
Profitability									
Rate of Return on Assets	0.09	0.09	0.10	0.09	0.11	0.09	0.07	0.07	0.06
Rate of Return on Equity	0.10	0.10	0.11	0.10	0.13	0.10	0.06	0.08	0.07
Operating Profit Margin Ratio	0.21	0.20	0.22	0.21	0.23	0.20	0.16	0.17	0.15
Net Farm Income ($)	105,352	100,588	122,144	129,899	148,560	133,232	105,790	133,809	130,104
Debt Repayment Capacity									
TDACLCR[a]	2.70	3.28	3.73	3.40	3.76	3.59	3.29	2.73	2.32
CRATDRM[b] ($)	79,199	70,967	95,968	86,035	106,381	76,021	55,275	83,920	69,612
Financial Efficiency									
Asset Turnover Ratio	0.41	0.42	0.43	0.43	0.46	0.46	0.40	0.43	0.43
Operating Expense Ratio	0.60	0.62	0.60	0.62	0.60	0.64	0.68	0.67	0.69
Depreciation Expense Ratio	0.13	0.12	0.11	0.10	0.10	0.09	0.09	0.09	0.09
Interest Expense Ratio	0.08	0.07	0.06	0.06	0.05	0.05	0.06	0.05	0.05
Net Farm Income from Operation Ratio	0.27	0.26	0.29	0.28	0.30	0.27	0.23	0.24	0.23

[a] Term debt and capital lease coverage ratio
[b] Capital replacement and term debt repayment margin

Table 2.

Logistic Parameter Estimates of Creditworthiness Models

Variables	Annual[a]	Two-Year Average[b]	Three-Year Average[c]
Intercept	2.02 (0.01)[d]	0.70 (0.59)	0.39 (0.09)
Debt/Asset Ratio	-1.90 (0.03)	-1.72 (0.26)	-0.92 (0.73)
Current Ratio	0.03 (0.78)	0.15 (0.51)	0.13 (0.72)
Lagged Dependent Variable	0.96 (0.05)	2.26 (0.01)	2.36 (0.21)
Model X^2	14.26	18.71	6.16
Prior Probabilities	0.85	0.90	0.90

[a] 1985, 1986, 1987, 1988, 1989, 1990
[b] 1985-86, 1987-88, 1989-90
[c] 1985-86-87, 1988-89-90
[d] P-Values are reported in parenthesis.

Table 3. **Misclassification Results of the RPA Models**

Within-Sample

	Annual[a]			Two-Year Average[a]			Three-Year Average[a]		
Cost[b]	I	II	%	I	II	%	I	II	%
1	33	0	9	12	0	9	0	1	1
2	33	0	9	1	15	11	0	1	1
3	11	61	21	1	15	11	0	1	1
4	11	61	21	1	15	11	0	1	1
5	11	61	21	1	15	11	0	1	1

Out-of-Sample

	1991			1991-92			1991-93		
Cost[b]	I	II	%	I	II	%	I	II	%
1	18	0	26	13	0	9	7	2	13
2	18	0	26	8	7	11	7	2	13
3	6	10	23	8	7	11	7	2	13
4	6	10	23	8	7	11	7	2	13
5	6	10	23	8	7	11	7	2	13

	1992		
	I	II	%
1	16	0	23
2	16	0	23
3	4	16	29
4	4	16	29
5	4	16	29

	1993		
	I	II	%
1	20	0	29
2	20	0	29
3	6	13	27
4	6	13	27
5	6	13	27

[a] I = the number of type I misclassifications, II = the number of type II misclassification, % = percent of observation misclassified.

[b] Cost of type I misclassification (a less creditworthy borrower classified as creditworthy). Cost of a type II misclassification remains at 1 for all alternatives.

Table 4. **Misclassification Results of the Logistic Regression Models**

Within-Sample

	Annual[a]			Two-Year Average[a]			Three-Year Average[a]		
Cost[b]	I	II	%	I	II	%	I	II	%
1	23	35	17	6	12	13	3	4	10
2	23	35	17	6	12	13	3	4	10
3	23	35	17	6	12	13	3	4	10
4	23	35	17	6	12	13	3	4	10
5	23	35	17	6	12	13	3	4	10

Out-of-Sample

	1991			1991-92			1991-93		
Cost[b]	I	II	%	I	II	%	I	II	%
1	15	5	29	7	4	8	8	2	14
2	15	5	29	7	4	8	8	2	14
3	15	5	29	7	4	8	8	2	14
4	15	5	29	7	4	8	8	2	14
5	15	5	29	7	4	8	8	2	14

	1992		
	I	II	%
1	5	9	20
2	5	9	20
3	5	9	20
4	5	9	20
5	5	9	20

	1993		
	I	II	%
1	11	4	21
2	11	4	21
3	11	4	21
4	11	4	21
5	11	4	21

[a] I = the number of type I misclassifications, II = the number of type II misclassification, % = percent of observation misclassified.

[b] Cost of type I misclassification (a less creditworthy borrower classified as creditworthy). Cost of a type II misclassification remains at 1 for all alternatives.

Table 5. **Expected Cost of Misclassification of RPA and Logistic Regression**

Within-Sample

RPA				Logistic Regression			
Cost	One-Year	Two-Year	Three-Year	Cost[a]	One-Year	Two-Year	Three-Year
1	0.150	0.100	0.014	1	0.198	0.134	0.110
2	0.300	0.122	0.014	2	0.303	0.184	0.164
3	0.314	0.131	0.014	3	0.408	0.234	0.218
4	0.364	0.139	0.014	4	0.512	0.284	0.272
5	0.414	0.147	0.014	5	0.617	0.334	0.326

Out-of-Sample

RPA				Logistic Regression			
Cost	1991	1991-92	1991-93	Cost[a]	1991	1991-92	1991-93
1	0.150	0.100	0.080	1	0.207	0.117	0.087
2	0.300	0.234	0.129	2	0.332	0.171	0.143
3	0.314	0.295	0.177	3	0.457	0.225	0.198
4	0.364	0.357	0.226	4	0.582	0.279	0.254
5	0.414	0.418	0.274	5	0.707	0.332	0.309
	1992				1992		
1	0.150			1	0.189		
2	0.300			2	0.235		
3	0.338			3	0.282		
4	0.366			4	0.329		
5	0.395			5	0.376		
	1993				1993		
1	0.150			1	0.151		
2	0.300			2	0.233		
3	0.356			3	0.316		
4	0.401			4	0.398		
5	0.446			5	0.481		

[a] To make a comparison with the RPA classification the misclassification rate is held constant, but the relative costs of misclassification is varied. The logistic regression does not explicitly account for cost of misclassification during the development of the model.