AgEcon SEARCH
RESEARCH IN AGRICULTURAL & APPLIED ECONOMICS

The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search
http://ageconsearch.umn.edu
aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

Journal of Agricultural and Applied Economics, 31,1(April 1999):109–122
© 1999 Southern Agricultural Economics Association

Application of Recursive Partitioning to Agricultural Credit Scoring

Michael P. Novak and Eddy LaDue

ABSTRACT

Recursive Partitioning Algorithm (RPA) is introduced as a technique for credit scoring analysis, which allows direct incorporation of misclassification costs. This study corroborates nonagricultural credit studies, which indicate that RPA outperforms logistic regression based on within-sample observations. However, validation based on more appropriate out-of-sample observations indicates that logistic regression is superior under some conditions. Incorporation of misclassification costs can influence the creditworthiness decision.

Key Words: finance, credit scoring, misclassification, recursive partitioning algorithm

Many agricultural banks and lending institutions are beginning to recognize the advantages of credit scoring in conjunction with human analysis. Several institutions are currently using such models on at least a subset of their portfolio. Credit scoring models hold the promise of reducing the variability of credit decisions, adding efficiencies to credit risk assessment, establishing better loan pricing policies, and improving the safety and soundness of agricultural lending. Improved financial information systems have allowed agricultural lenders to more readily collect and retain data regarding the creditworthiness of borrowers. As such databases are populated the ability to monitor changes in creditworthiness overtime improves, and the need to explore new methods to estimate credit-scoring models increases.

Within the agricultural financial literature various nonparametric and parametric methods have been used to estimate credit-scoring models, such as experience-based algorithms (Alcott; Splett et al.), mathematical programming (Hardy and Adrian; Ziari, Leatham, and Turvey), logistic regression (Mortensen, Watt, and Leistritz), probit regression (Lufburrow, Barry, and Dixon; Miller et al.), discriminant analysis (Hardy and Weed; Dunn and Frey; Johnson and Hagan), and linear probability regression (Turvey). There is not unanimous agreement as to the best method for estimating credit-scoring models and new methods continue to be researched.

Most recently, the logistic regression has dominated the agricultural credit-scoring literature (Miller and LaDue, Turvey and Brown, Novak and LaDue, Splett et al.). 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-scoring models using four parametric methods with a single data set. He recommends logistic regression over probit regression, discriminant analysis, and linear probability regression based on predictive accuracy and ease of use, in addition to the favorable statistical proper-

Michael P. Novak is Manager of Agricultural Finance; Federal Agricultural Mortgage Corporation; Washington, DC. Eddy LaDue is W.I. Myers Professor of Agricultural Finance; Department of Agricultural, Resource, and Managerial Economics; Cornell University; Ithaca, NY.

ties previously mentioned. Logistic regression improves on some of the statistical properties of discriminant analysis and linear probability regression; however, it still possesses numerous statistical problems common to most parametric methods. These problems include (1) the need to pre-select the exact explanatory variables without well-developed theory, (2) inability to identify an individual variable's relative importance, (3) reduction of the information space's dimensionality, and (4) limited ability to incorporate relative misclassification costs.

Non-agricultural studies have used the Recursive Partitioning Algorithm (RPA) to classify financially stressed firms. 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 based on 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. Friedman originally developed RPA. A thorough theoretical exposition of RPA is presented in Breiman, et al. A more practical exposition of the computational aspects of RPA and a comprehensive bibliography of research using RPA are presented in the CART software documentation (Steinberg and Colla). RPA has been applied to many areas of research, such as behavior economics (Carson, Hanemann, and Steinberg), wildlife management (Grubb and King), and livestock management (Tronstad and Gum), but it has not been applied to agricultural credit-scoring.

Several non-agricultural financial stress classification studies indicate RPA outperforms the other parametric and judgmental models based on predictive accuracy. Marais, Patell, and Walfson compare RPA with a polytomous probit regression to classify commercial loans for publicly and privately held banking firms. Frydman, Altman, and Kao compare RPA with discriminant analysis to classify firms according to their degree of financial stress. Srinivasan and Kim compare RPA with discriminant analysis, logistic regression, goal programming, and a judgmental model (the Analytic Hierarchy Process) to evaluate the corporate credit granting process. Each of these studies uses cross-validation and the associated expected cost of misclassification to evaluate the RPA models. A shortcoming of these studies is that they do not use intertemporal (*ex ante*) predictions to compare and evaluate the models. Prediction is the basic objective of credit-scoring models (Joy and Tofeson). Credit-scoring models should not be limited to classifying borrowers in the same time period. The "true" test is their ability to classify borrowers in the future.

The primary purpose of this study is to introduce RPA as a method for classifying creditworthy and less creditworthy agricultural borrowers, and compare RPA to the logistic regression. This study also challenges the RPA's superior prediction accuracy, as purported in the financial stress classification literature. In this study, RPA models are evaluated based on minimizing the expected cost of misclassification for creditworthy and less creditworthy borrowers in out-of-sample periods.

The remainder of the paper is divided into five sections. The first section presents the specifics of the RPA. The second section discusses the advantages and disadvantages of and the differences between the RPA and logistic regression. 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 this section, a hypothetical RPA tree growing process is presented and the terminology is introduced. To understand the tree growing process, a hypothetical tree is illustrated in Figure 1. It is constructed using classification groups i and j, and characteristic variables A

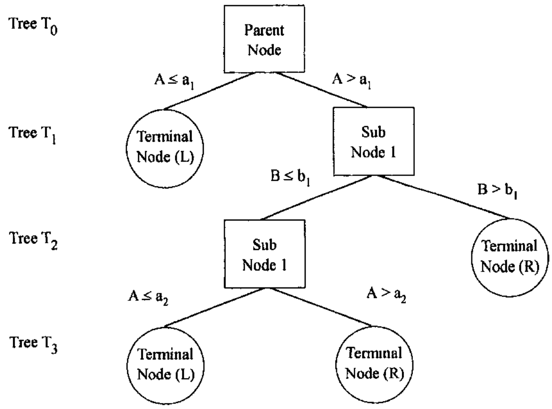


Figure 1. Hypothetical Recursive Partitioning Algorithm Tree

and B.[1] Throughout the paper the classification groups are 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 possesses no binary splits and can be referred to as the *naive classification tree.* All observations in the original sample are assigned to group j or i, based on an assignment rule. The assignment of T_0 to either group i or j depends on misclassification costs and prior probabilities. When misclassification costs are equal to each other and prior probabilities are equal to the sample proportions of the groups, T_0 is assigned to the group with the greatest proportion of observations, 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 is assigned to the group that minimizes the observed expected cost of misclassification.[2]

To begin the tree-growing process, RPA methodically searches each individual characteristic variable and split value of the characteristic variable. The computer algorithm then selects a characteristic variable, in this case A, and a split value of the characteristic variable A, in this case a_1, based on the optimal univariate splitting rule.[3] The optimal splitting rule implies that no other characteristic variable and split value can decrease the impurity or, in other words, the misclassified observations, taking into account misclassification costs and prior probabilities in the two resulting descendent nodes. In this particular illustration, A is the characteristic variable selected and a_1 is the "optimal" split value selected by the computer algorithm. Observations with a value of characteristic variable A less than or equal to a_1 will "fall" into the left node and the observations with a value of characteristic variable A greater than a_1 will "fall" into the right node. The resulting subtree, denoted by T_1, consists of a parent node and a left and right terminal node. The right terminal node is labeled Sub-Node 1 in Figure 1 because the tree continues from that node. The terminal nodes in the subtree are then assigned to groups, i or j, based on the assignment rule of minimizing observed expected cost of misclassification. T_0 and T_1 are the 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 is the characteristic variable selected and b_1 is the "optimal" split value selected by the computer algorithm. The right node becomes an internal node and the observations within it are partitioned. Observations with a value of char-

[1] Characteristic variables are analogous to independent variables in a parametric regression.

[2] The observed expected cost of misclassification $= c_{ij}\pi_i n_{ij}(T)/N_i + c_{ji}\pi_j n_{ji}(T)/N_j$, where c_{ij} (c_{ji}) is the cost of misclassifying a group i (j) observation as a group j (i) observation; π_i (π_j) is the prior probability of an observation belonging to group i (j); $n_{ij}(T)$ ($n_{ji}(T)$) is the total number of group i (j) observations misclassified as j (i) in the entire tree T; and N_i (N_j) is the number of original observations from group i(j).

[3] The univariate splitting rule implies splitting an axis of one variable at one point. This study is limited to univariate splitting rules; however, CART has the capability to split variables using linear combinations of variables. The resulting classification trees are usually very cumbersome and difficult to interpret when linear combination splitting rules are used.

acteristic variable B less than or equal to b_1 "fall" into a new left node and observations with a value of characteristic variable B greater than b_1 "fall" into a new right node. The new left (labeled Sub-Node 2 in Figure 1) and right nodes become terminal nodes in T_2, and 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 A is selected again to develop T_3. 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. Some suggested methods are v-fold cross-validation, jackknife, expert judgement, boot strapping, and holdout samples. Once the classification tree is developed and pruned back, it can be used to classify observations from outside the original sample.

RPA and Logistic Regression Comparison

In this section the advantages and disadvantages of and the differences between RPA and logistic regression are discussed. One basic difference between RPA and logistic regression is the way RPA selects variables. A credit-scoring model developed using RPA does not require the variables to be selected in advance. The computer algorithm can select variables from the predetermined group of variables, without subjective influences or violating parametric assumptions.

Other differences are that RPA places no limit on the number of times a variable can be selected; the same variable can be selected numerous times and appear in different parts of the tree. All selected variables are predicated on the preceding variables. 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. The tree growing process is intentionally myopic. Furthermore, outlier values do not significantly influence RPA: all splits occur on non-outlier values. Once the optimal split value for a variable is selected, the outlier observation is assigned to a node and the RPA procedure continues. In contrast, logistic regression allows each variable only to appear once in the model and can be severely affected by outlier values.

An advantage of RPA over the logistic regression methods is that RPA analyzes the univariate attributes of individual variables. RPA selects the optimal split value of the characteristic variables, and surrogate and competitive variables, along with their optimal split values listed in order of importance. The lists of surrogate and competitive variables provide additional insight and understanding to the predictive structure of the individual variables. Surrogate variables mimic the selected variable's ability to replicate the size and composition of the descendent nodes. Competitive variables are defined as alternative variables to the selected variables with slightly less ability to reduce impurity in the descendent nodes.

While lacking in variable selection and insight, logistic regression does have advantages. Logistic regression provides an overall summary statistic. The overall summary statistic can be used to evaluate and compare models. Logistic regression also assigns a predicted probability of creditworthiness to each individual borrower. Often lenders want a quantitative assessment of the borrower's creditworthiness, not just a method of classifying borrowers as creditworthy or less creditworthy. RPA can classify observations into creditworthy or less creditworthy groups, but cannot estimate a credit score for each individual borrower.

The two methods differ in the way they divide the information space into classification regions. RPA repetitiously partitions the information space as the tree is formed. A graphical illustration is presented in Figure 2; it is based on the hypothetical RPA tree in Figure 1. RPA partitions the information space into four rect-

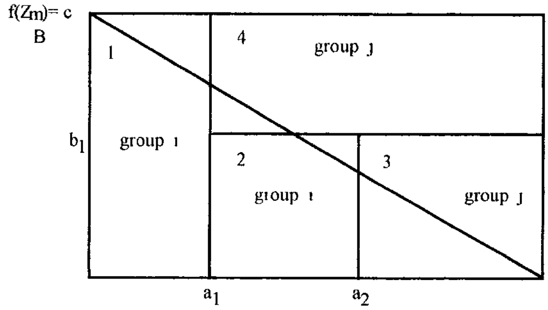


Figure 2. Observation Space

angular regions according to characteristic variables, A and B, and their respective optimal split values, a_1 and b_1. Observations falling in regions 1 and 2 are classified as group i and those falling in region 3 and 4 are classified as group j. Logistic regression, if implemented as a binary qualitative choice model, partitions the information space into two regions based on a prior probability, say c. The example line $f(Z_m) = c$ divides the information space. Z_m is a linear function of variables A and B corresponding to observation m, and $f(\times)$ is the cumulative logistic probability function. The observations are assigned to class i if $f(Z_m) => c$ or group j if $f(Z_m) < c$.

The two methods also differ in the manner in which they incorporate misclassification costs and prior probabilities. RPA uses misclassification costs and prior probabilities to simultaneously determine variable selection, optimal split values, and terminal node assignments. Changes in the misclassification costs and prior probabilities can change the selected variables and the optimal split values, and, in turn, change the structure of the classification tree. In contrast, logistic regression is usually estimated without incorporating misclassification costs and prior probabilities. However, after the logistic regression is estimated, a prior probability can be used to classify borrowers as creditworthy/less creditworthy.

Despite the differences in the two methods, the RPA and logistic regression methods can be integrated. RPA can select the relevant variables from a predetermined set of variables. The variables then can be employed in the logistic regression. In addition, the predicted probabilities from the logistic regression can be used as a variable in the predetermined group of variables from which the RPA model selects. Whether, and at what level, RPA selects the predicted probability variable to be part of the classification tree can provide evidence for or against logistic regression.

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.[4] The farms represent a segment of New York State dairy farms, which value consistent, annual financial and manage-

[4] 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 distressed firms. To overcome 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 nonrandomness of the data and can asymptotically bias the model's parameters and probabilities (Heckman). Selection bias typically can affect credit evaluation models in two ways. First, financially distressed borrowers are less likely to keep accurate records; therefore, these borrowers would tend not to be included in the sample (Zmijewski). Second, when panel data are employed there may be attrition of borrowers from the sample. In this study, some borrowers probably participated in the DFBS program during the earlier years of sample period, but exited the industry or stopped submitting records to the database before the end of the sample period. In analyzing financial distress models, Zmijweski 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.

ment information. The financial information collected includes the essential components for deriving a complete set of the sixteen financial ratios and measures recommended by the Farm Financial Standard Council (FFSC).[5] Additional farm productivity, cost management, and profitability statistics for these farms are summarized in Smith, Knoblauch, and Putnam.

Creditworthiness Measures

A key value available in this data set was the planned/scheduled principal and interest payment on total debt. This variable reflects the borrower's expectations of debt obligations for the up-coming year. Having this component facilitates the calculation of the coverage ratio,[6] an essential element of this study. The coverage ratio approximates whether the borrower generates enough income to meet all expected payments and is an indicator of creditworthiness. The coverage ratio is based on actual financial statements and has been introduced to credit-scoring models as a measure of creditworthiness, an alternative to loan classification and loan default models[7] (Novak and LaDue (1994); Khoju and Barry). This indicator of creditworthiness is aligned with cash-flow or performance-based lending, as opposed to the more traditional collateral-based lending, and its use has been facilitated by improvements in farm records and computerized loan analysis systems.

The coverage ratio, a quantitative indicator of creditworthiness, needs to be converted to a binary variable in order to assist the lender in making a decision to grant or deny a credit request. Therefore in this study an a-priori cut-off level of 1 is used. A coverage ratio greater (less) than 1 indicates that the borrower did (not) generate enough income to meet all expected debt obligations. Thus, a coverage ratio greater (less) than 1 indicates a creditworthy (less creditworthy) borrower.[8]

In addition to the standard annual coverage ratio, two-year and three-year average coverage ratios are employed in this study. The two-year and three-year average coverage ratios were found to provide a more stable, extended indicator of creditworthiness (Novak and LaDue, 1997). Using the annual, two-year average, and three-year average measures of creditworthiness and an a-priori cut-off value

[5] 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 bounded to the -4 to 15 interval. The bounded interval of the coverage ratio indicates both extremes of debt repayment capacity.

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

[7] 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, Barry, and Dixon; Hardy and Adrian; Hardy et al., Turvey and Brown, 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 to bank examiners' or credit reviewers' loan classification. These problems are not limited to agricultural credit scoring models (Maris et al.; Dietrich and Kaplan). Some agricultural credit scoring studies have used default (Miller and LaDue, and Mortensen, Watt, and Leistritz). Default is inherently a more objective measure. However, lenders and borrowers can influence default classifications by decisions to forebear, restructure, or grant additional credit to repay a delinquent loan. Borrowers can influence or delay default by selling assets, depleting credit reserves, seeking off-farm employment, and other similar activities. Default is based on a single lender's criteria. Borrowers with split credit can be current with one lender and delinquent or in arrears with another lender. Additionally, the severity of some types of default such as loan losses makes it less than adequate. A lender would be better served to identify these borrowers before such action occurs. Because of these ambiguities surrounding default, an alternative cash-flow measure of creditworthiness is used.

[8] 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. The sample represents borrowers from Farm Service Agency, Farm Credit and various private banks. The various lending institutions can be translated into varying degrees of creditworthiness among the borrowers in the sample. Creditworthiness to one lender may be less creditworthy to another. The data can be viewed as a compilation of lenders' portfolios.

of one, the seventy farms are classified as creditworthy or less creditworthy. The number found to be creditworthy in any one year varied from 50 to 66 based on annual data. Using two-year averages, the number of creditworthy farms increased from 57 to 66 depending on the two-year period chosen. For three-year periods the number of creditworthy farms was 68, 65, and 57 for 1985–87, 1988–90, and 1991–93, respectively. The number of borrowers considered creditworthy decreases over time. Identifying a borrower with diminishing debt repayment ability prior to any serious financial problems exemplifies the usefulness of the creditworthiness indicator and should be of value to lenders when evaluating a borrower's credit risk or monitoring his/her overall loan portfolio.[9]

Development of the Creditworthiness Model

In this section the annual, two-year average, and three-year average credit-scoring models are discussed. The annual model uses lagged characteristic values to classify creditworthy and less creditworthy borrowers. That is, the annual model is developed with pooled data using characteristic values for each year from 1985–89 to classify creditworthy and less creditworthy borrowers for the following year of 1986–90, respectively. The models are evaluated using 1990, 1991 and 1992 characteristic values to predict 1991, 1992, and 1993 creditworthy and less creditworthy borrowers' classifications, respectively. Finally, the predicted creditworthy classifications for 1991, 1992, and 1993 are compared to the actual classifications for the same time period to determine the intertemporal efficacy of the model.

The two-year average model is developed using 1985–1986 and 1987–88 averages of the characteristic values to classify creditworthy borrowers in the average periods 1987–88 and 1989–90, respectively. The evaluation process then uses 1989–90 average characteristic values to predict 1991–92 average creditworthy and less creditworthy borrowers' classifications. The three-year average model is developed using 1985–86–87 average characteristic variables to classify 1988–90 average creditworthy and less creditworthy borrowers. The three-year average model is evaluated using 1988–90 average characteristic values to predict 1991–92–93 average creditworthy and less creditworthy borrowers. In both the two-year and three-year average models, the predicted classifications are compared to actual classifications for the same time period to determine the intertemporal efficiency of the models.

RPA does not require individual characteristic variables to be selected in advance. It does, however, require selecting a predetermined group of variables. In this study, the 16 FFSC recommended ratios and measures were selected as the predetermined group of variables.[10] Many of the variables in this predetermined group of variables represent similar financial concepts, but are still included in the population set, allowing RPA to select the appropriate variables. In addition, the predicted probability of creditworthiness from the logistic regression model and the lagged classification variables were included in the predetermined group of variables.

The logistic regression model requires that the characteristic or explanatory variables be selected in advance. As a result, 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.; Novak and LaDue, 1997). The specific variables

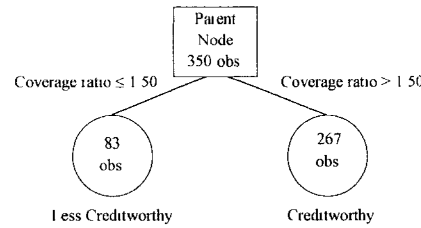
[9] Granted other factors—such as collateral offered and a borrower's credit history, personal attributes, and management ability—also influence credit risk. Many of the other factors listed have to be evaluated, in conjunction with the model, by the loan officer. Creditworthiness models are designed to assist, not replace, the loan officer in lending decisions.

[10] All 16 FFSC recommended ratios and measures were included in the analysis even though two of the variables, debt/asset ratio and equity/asset ratio, are identical. The choice to include all 16 ratios and measures was based on consistency and completeness.

used in the model are debt-to-asset ratio, current ratio, and lagged dependent variable.[11]

Both estimation methods require the specification of a prior probability. In this study, the proportion of creditworthy borrowers in the total sample determines the prior probability. The values are 0.852, 0.896, and 0.905 for the annual, two-year average and three-year average periods, respectively. The prior probabilities for average periods demonstrate that the percentage of creditworthy borrowers in the sample data set increases as the average period lengthens.

In addition to prior probabilities, misclassification costs also need to be specified. Previous agricultural credit-scoring models, except for Ziari, Leatham, and Turvey,[12] either ignore misclassification costs or assume they are equal. It is not reasonable to assume that the misclassification costs are equal for all types of decisions. The cost of granting, or renewing, a loan to a less creditworthy borrower is typically greater than the cost of denying, or not renewing, a loan to a creditworthy borrower. Estimating these misclassification costs is beyond the scope of this study and the data, but the study does illustrate the classification sensitivity of these costs. The relative costs of Type I and Type II misclassification errors are varied accordingly from 1:1, 2:1, 3:1, 4:1, and 5:1, with the relatively higher misclassification cost put on the Type I error.[13] While the less creditworthy measure used in this model may not be as serious as actual loan losses or bankruptcy of a borrower, there is still a higher cost associated with loan servicing and payment collection for less creditworthy borrowers.



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 3. RPA Tree Using Annual Data

Comparison of RPA and Logit Model Results

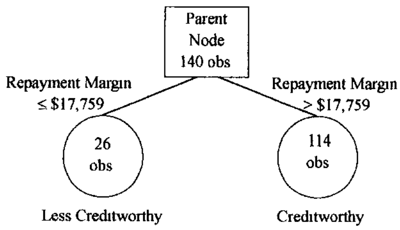
Figure 3 presents the classification tree generated from the RPA for the annual time period when the misclassification cost of a type I error is three times greater than a type II error (i.e. 3:1). The model is simple. It is comprised of the coverage ratio lagged one period. Borrowers with a coverage ratio greater than 1.50 a year prior are classified as creditworthy and borrowers with a coverage ratio less than 1.50 a year prior are classified as less creditworthy. Put differently, to ensure 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

[11] Two other logistic regression models, a stepwise and an "eight variable" model (the latter, 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 signs and the within-sample and out-of-sample prediction rates were lower than RPA's and the paramoninous (three variable) logit model's prediction rates for all the comparable time periods.

[12] Ziari, Leatham, and Turvey assume the misclassification cost for a noncurrent loan is twice as much as that for a current loan.

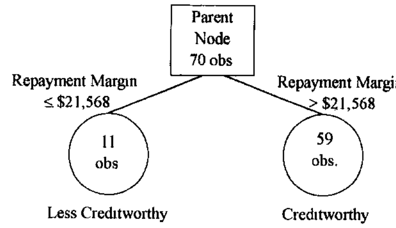
[13] Type I error is a less creditworthy borrower classified as a creditworthy borrower and a Type II error is a creditworthy borrower classified as a less creditworthy borrower.



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 4. RPA Tree Using Two-Year Average Data



Surrogate Variables	Split Values
1 Term Debt and Capital Lease Coverage 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 5. RPA Tree Using Three-Year Average Data

mimic the selected variable, the coverage ratio, and its optimal split value of 1.50. The repayment margin, net farm income from operations, binary lagged dependent variable, predicted probability of creditworthiness, and operating expense ratio were identified as surrogate variables. The selection of the predicted probability of creditworthiness from the logistic regression adds some additional validity to the use of this variable as a credit score. Also noteworthy is that the split value of the predicted probability of creditworthiness is very similar to the prior probability for the annual sample period.

A list of competitor variables is also presented in the same figure. The repayment margin was listed as the first competitor variable. The competitor variable implies that if the selected variable (i.e. coverage ratio) was restricted or eliminated from the sample, the repayment margin—the first competitor variable—would have been chosen as the selected variable in the classification tree. The other competitor variables selected were debt-to-equity ratio, debt-to-asset ratio, operating expense ratio, and operating profit margin ratio.

Figure 4 presents the two-year average classification tree, again using a 3:1 relative misclassification costs ratio, with the higher misclassification cost attributed to a type I error. In this classification tree the repayment margin was selected as the characteristic variable and the coverage ratio was selected as a competitor and surrogate variable. Similar to the annual model, the binary lagged dependent variable and predicted probability of creditworthiness were selected as surrogate variables. The other surrogate and competitive variables selected were net farm income, interest expense ratio, operating expense ratio, and return on equity.

Figure 5 presents the classification tree for the three-year average period. Similar to the previous two trees, a 3:1 relative misclassification cost ratio is used. The repayment margin was selected as the primary characteristic

Table 1. Logistic Parameter Estimates of Creditworthiness Models

Variables	Annual	Two-Year	Three-Year
Intercept	2.02	0.70	0.39
	(0.01)[a]	(0.59)	(0.09)
Debt/Asset Ratio	−1.90	−1.72	−0.92
	(0.03)	(0.26)	(0.73)
Current Ratio	0.03	0.15	0.13
	(0.78)	(0.51)	(0.72)
Lagged Dep. Var.	0.96	2.26	2.36
	(0.05)	(0.01)	(0.21)
Model X^2	14.26	18.71	6.16
Prior Probabilities	0.852	0.896	0.905

[a] p-values are reported in parentheses.

variable and the coverage ratio was selected as the first surrogate and competitor variable. In this average time period, the binary lagged dependent variable and predicted probability were not selected as either competitor or surrogate variables. The selected surrogate and competitor variables were operating expense ratio, net farm income, rate of return on assets, current ratio, interest expense ratio, and operating profit margin ratio.

The results are consistent with expectations. In general, most of the surrogate or competitive variables, especially in the two-year and three-year time periods, represent a borrower's repayment capacity, financial efficiency or profitability. The best indicator of creditworthiness is repayment capacity and the repayment capacity is predicated on operating profits and losses, hence profitability and financial efficiency.

The actual classification trees may at first appear to be a concern. The classification trees have a low number of characteristic variables and in some cases the naive model is selected when relative misclassification costs are low.[14] However, this is consistent with other studies. Frydman, Altman, and Kao found the naive model also did best in classifying their data when misclassification costs were assumed equal, and found that the cross-validation classification trees had considerably fewer splits than the non-cross-validation classification trees. The largest cross-validation classification tree they estimated had a maximum of three splits. In their study, for exposition purposes the non-cross-validation trees were presented. These trees are aesthetically more appealing. They are not pruned, have considerably more characteristic values and classify more observations, but of course have less generalization outside the sample data.

The parameters of the logistic regression models are presented in Table 1. All the parameters for each of the models have the expected sign. In the annual model the debt-to-asset ratio and the lagged dependent parameters are significant at the 95% level. In the two-year average model the lagged dependent variable is significant at the 99% level. None of the variables is statistically significant in the three-year average model.

Table 2 presents the expected costs of misclassification for each model and level of relative misclassification cost. The RPA model, not surprisingly, does best at minimizing the expected misclassification cost for the within-sample time periods for all relative misclassification costs scenarios. 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, regardless of misclassification costs. Based on the RPA objective, the nonagricultural financial stress studies

[14] RPA selects the naïve model when the annual data are used and misclassification costs are 1:1 and 1:2, and when the two-year average data are used and misclassification costs are 1:1.

Table 2. Expected Cost of Misclassification[a] for the RPA and Logistic Regression Models

Cost Based on Within-Sample Observations (1985–1990)							
	RPA				Logistic Regression[c]		
Relative Costs[d]	1-Year Model	2-Year Model	3-Year Model	Relative Costs[d]	1-Year Model	2-Year Model	3-Year Model
1:1	0.150[b]	0.100[b]	0.014	1:1	0.198	0.134	0.110
2:1	0.300[b]	0.122	0.014	2:1	0.303	0.184	0.164
3:1	0.314	0.131	0.014	3:1	0.408	0.234	0.218
4:1	0.364	0.139	0.014	4:1	0.512	0.284	0.272
5:1	0.414	0.147	0.014	5:1	0.617	0.334	0.326

Cost Based on Out-of-Sample Obervations (1991–1993)							
	RPA				Logistic Regression[c]		
Relative Costs	1-Year Model	2-Year Model	3-Year Model	Relative Costs[c]	1-Year Model	2-Year Model	3-Year Model
1:1	0.150[b]	0.100[b]	0.080	1:1	0.207	0.117	0.087
2:1	0.300[b]	0.234	0.129	2:1	0.332	0.171	0.143
3:1	0.314	0.295	0.177	3:1	0.457	0.225	0.198
4:1	0.364	0.357	0.226	4:1	0.582	0.279	0.254
5:1	0.414	0.418	0.274	5:1	0.707	0.332	0.309
	1992				1992		
1:1	0.150[b]			1:1	0.189		
2:1	0.300[b]			2:1	0.235		
3:1	0.338			3:1	0.282		
4:1	0.366			4:1	0.329		
5:1	0.395			5:1	0.376		
	1993				1993		
1:1	0.150[b]			1:1	0.151		
2:1	0.300[b]			2:1	0.233		
3:1	0.356			3:1	0.316		
4:1	0.401			4:1	0.398		
5:1	0.446			5:1	0.481		

[a] See endnote #2 for cost of misclassification calculation.
[b] Represents the naïve model.
[c] The logistic regression does not explicitly account for cost of misclassification during the development of the model. For comparison purposes, the expected costs of misclassification is calculated by keeping the number of misclassified borrowers constant and varying the relative misclassification cost scenarios for each model.
[d] Relative Cost of type I and type II misclassification errors (cost of granting credit to a less creditworthy borrower: Cost of not granting credit to a creditworthy borrower).

have concluded that RPA is a better model than other models. If this study were to conclude here, it would also conclude RPA is a better method of classification. However, this study continues by comparing intertemporal, out-of-sample observations.

Using the annual time period data, the RPA model performs best in 1991 for all relative misclassification costs scenarios, and in 1992 and 1993 when the misclassification costs are equal. The annual RPA model with equal misclassification costs is also the naive model. It is interesting to note that previous agricultural credit-scoring studies typically have assumed equal misclassification costs, but did not always compare the estimated model's results with the naïve model. In this case, the naïve model outperforms the logistic regression model. Nevertheless, the assumption that misclassification costs are equal is not very realistic in credit screening models.

Using the same annual data, the logistic regression model does best at minimizing expected cost of misclassification when misclassification costs are not assumed to be equal. Logistic regression also does best at minimizing the expected cost of misclassification using the two-year average out-of-sample data for each relative misclassification cost scenario, except when misclassification costs are equal. When misclassification costs are equal, then RPA, represented by the naive model, does better. Finally, RPA does best at minimizing the expected cost of misclassification using the three-year average out-of-sample data for each of the relative misclassification costs scenarios. From these results we cannot conclude that either model is superior using this data set. A different data set may have different results and would warrant exploration.

Conclusion

This study introduces RPA to agricultural credit-scoring. The study also demonstrates RPA's advantages and disadvantages in relation to logistic regression. The advantages of RPA include not requiring pre-selected variables, provision of the univariate attributes of individual variables, not being affected by outliers, provision of surrogate and competitive variable summary lists, and explicit incorporation of misclassification costs. On the other hand, logistic regression possesses some desirable advantages over RPA, such as the availability of overall summary statistics and an individual quantitative credit score for each observation.

More significantly, the study only partially corroborates the results of the non-agricultural credit classification studies. RPA outperforms logistic regression when the RPA models are selected and compared using cross-validation methods and expected cost of misclassification and the evaluation is based on within-sample observations. However, when the validation process is taken one step further and uses intertemporal (out-of-sample) minimization of expected cost of misclassification as the evaluation method, the same results are not achieved. In some cases RPA outperforms logistic regression and, in other cases, logistic regression outperforms the RPA model. Given the normal use of credit-scoring models, out-of-sample evaluation is most appropriate. These findings suggests that cross-validation may not be sufficiently effective to surmount potential overfitting the sample data which limits RPA's intertemporal predictive ability.

This study also considers relative misclassification costs. Previously, agricultural credit-scoring research has generally—except for Zairi, Leatham, and Turvey—evaluated models based on the number of misclassified observations, and has not considered minimizing expected costs of misclassification. The results of this study indicate that misclassification costs can affect the development of the RPA model. Future agricultural credit-scoring research should consider minimizing expected costs of misclassification, instead of minimizing misclassified observations, to evaluate models. Similarly, effort should be made towards calculating actual misclassification costs, instead of using relative misclassification costs.

Finally, while the study has taken strides in introducing RPA to agricultural credit-scoring, the conclusion of RPA's superior performance is not as convincing as the non-agricultural financial stress literature's results. However, RPA does appear to be superior in some situations. Further testing and model refinements are suggested. From a practical standpoint, RPA presents several attractive features and can be employed in conjunction with other existing methods.

References

Alcott, K.W. "An Agricultural Loan Rating System." *The Journal of Commercial Bank Lending,* February 1985.

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, TX, June 1990.

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

Carson, R., M. Hanemann, and D. Steinberg, "A Discrete Choice Contingent Valuation Estimate of the Value of Kenai King Salmon." *The Journal of Behavior Economics*, 19(1990):53–68.

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." *Agricultural Finance Review* 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):404–409.

Frydman, H., E.I. Altman, and D. Kao. "Introducing Recursive Partitioning for Financial Classification: The Case of Financial Distress." *The Journal of Finance* 40 (1985):269–291.

Grubb, T.G. and R.M. King. "Assessing Human Disturbance of Breeding Bald Eagles with Classification Tree Models." *The Journal of Wildlife Management* 55(1991):500–511.

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):159–164.

Heckman, J.J. "Sample Selection Bias as a Specification Error." *Econometrica* 47(1979):153–162.

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

Joy, O.M. and J.O. Tollefson. "On the Financial Applications of Discriminant Analysis." *Journal of Financial and Quantitative Analysis* 10(1975):723–740.

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, IL, October 4–5, 1993.

LaDue, Eddy L. Warren F. Lee, Steven D. Hanson, and David Kohl. "Credit Evaluation Procedures at Agricultural Banks in the Northeast and Eastern Cornbelt." *Agricultural Economics Resources* 92-3, Cornell University, Department of Agricultural Economics, February 1992.

Lufburrow, J., P.J. Barry, and B.L. Dixon. "Credit Scoring for Farm Loan Pricing." *Agricultural Finance Review* 44(1984):8–14.

Marais, M.L., J.M. Patell, and M.A. Walfson. "The Experimental Design of Classification Models: An Application of Recursive Partitioning and Bootstrapping to Commercial Bank Loan Classifications." *Journal of Accounting Research Supplement* 22(1984):87–114.

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):511–523.

Miller, L.H., P. Barry, C. DeVuyst, D.A. Lins, and B.J. Sherrick. "Farmer Mac Credit Risk and Capital Adequacy." *Agricultural Finance Review* 54(1994):66–79.

Miller, L.H. and E.L. LaDue. "Credit Assessment Models for Farm Borrowers: A Logit Analysis." *Agricultural Finance Review* 49(1989): 22–36.

Mortensen, T.D., L. Watt, and F.L. Leistritz. "Predicting Probability of Loan Default." *Agricultural Finance Review* 48(1988):60–76.

Novak, M.P. and E.L. LaDue. "An Analysis of Multiperiod Agricultural Credit Evaluation Models for New York Dairy Farms." *Agricultural Finance Review* 54(1994):47–57.

Novak, M.P. and E.L. LaDue. "Stabilizing and Extending, Qualitative and Quantitative Measure in Multiperiod Agricultural Credit Evaluation

Model." *Agricultural Finance Review* 57(1997):39–52.

Oltman, A.W. "Aggregate Loan Quality Assessment in the Search for Related Credit-Scoring Model." *Agricultural Finance Review* 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." *Agricultural Finance Review* 54(1994):39–54.

Srinivansan, V., and Y.H. Kim. "Credit Granting: A Comparative Analysis of Classification Procedures." *Journal of Finance* 42(1987):665–681.

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

Tronstad, R. and R. Gum. "Cow Culling Decisions Adapted for Management with CART." *American Journal of Agricultural Economics* 76(1994):237–249.

Turvey, C.G. "Credit Scoring for Agricultural Loans: A Review with Application." *Agricultural Finance Review* 51(1991):43–54.

Turvey, C.G. and R. Brown. "Credit Scoring for Federal Lending Institutions: The Case of Canada's Farm Credit Corporations." *Agricultural Finance Review* 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." *Agricultural Finance Review* 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.