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 (Lubburrow, 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-
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
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 \( T_{\text{max}} \). However, in some cases \( T_1 \) may also be \( T_{\text{max}} \) depending on the predetermined penalty parameters specified. If \( T_1 \) is not \( T_{\text{max}} \), then the recursive partitioning algorithm continues.

In this illustration, \( T_1 \) is not \( T_{\text{max}} \), 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 is
\[
E(C) = c_{ij} p_{ij}(T) N + c_{ji} p_{ji}(T) N,
\]
where \( c_{ij} \) is the number of original observations from group \( i \).

\(^3\) 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.
characteristic 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 ob-
erved expected cost of misclassification.

Here again, $T_2$ does not minimize the ob-
erved 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 ex-
pected observed cost of misclassification of
the original sample. Obviously the develop-
ment method will over fit the tree; therefore,
a method is needed to prune back the tree.
Some suggested methods are v-fold cross-val-
idation, jackknife, expert judgement, boot
strapping, and holdout samples. Once the clas-
sification 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 disadvan-
tages of and the differences between RPA and
logistic regression are discussed. One basic
difference between RPA and logistic regres-
sion is the way RPA selects variables. A cred-
it-scoring model developed using RPA does
not require the variables to be selected in ad-
-ance. The computer algorithm can select vari-
ables from the predetermined group of vari-
able's, 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 nu-
-merous 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 vari-
able only to appear once in the model and can
be severely affected by outlier values.

An advantage of RPA over the logistic re-
gression 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 se-
lected variable’s ability to replicate the size
and composition of the descendent nodes.
Competitive variables are defined as alterna-
tive variables to the selected variables with
slightly less ability to reduce impurity in the
descendent nodes.

While lacking in variable selection and in-
sight, logistic regression does have advantag-
es. Logistic regression provides an overall
summary statistic. The overall summary sta-
tistic can be used to evaluate and compare
models. Logistic regression also assigns a pre-
dicted probability of creditworthiness to each
individual borrower. Often lenders want a
quantitative assessment of the borrower’s
creditworthiness, not just a method of classi-
fying borrowers as creditworthy or less cred-
itworth. RPA can classify observations into
creditworthy or less creditworthy groups, but
cannot estimate a credit score for each indi-
vidual borrower.

The two methods differ in the way they
divide the information space into classification
regions. RPA repetitiously partitions the infor-
mation 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-
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 \( f(x) \) is the cumulative logistic probability function. The observations are assigned to class \( i \) if \( f(Z_m) \geq c \) 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. 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, Zmijewski found selection bias causes no significant changes in the overall classification and prediction rate. Given Zmijewski'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). 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, 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 (Novak and LaDue (1994); Khoju and Barry). This indicator of creditworthiness is aligned with cashflow 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.

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 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.

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. 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, 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.

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.

### 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
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

<table>
<thead>
<tr>
<th>Variables</th>
<th>Annual</th>
<th>Two-Year</th>
<th>Three-Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>2.02</td>
<td>0.70</td>
<td>0.39</td>
</tr>
<tr>
<td>(0.01)</td>
<td>(0.59)</td>
<td>(0.90)</td>
<td></td>
</tr>
<tr>
<td>Debt/Asset Ratio</td>
<td>–1.90</td>
<td>–1.72</td>
<td>–0.92</td>
</tr>
<tr>
<td>(0.03)</td>
<td>(0.26)</td>
<td>(0.73)</td>
<td></td>
</tr>
<tr>
<td>Current Ratio</td>
<td>0.03</td>
<td>0.15</td>
<td>0.13</td>
</tr>
<tr>
<td>(0.78)</td>
<td>(0.51)</td>
<td>(0.72)</td>
<td></td>
</tr>
<tr>
<td>Lagged Dep. Var.</td>
<td>0.96</td>
<td>2.26</td>
<td>2.36</td>
</tr>
<tr>
<td>(0.05)</td>
<td>(0.01)</td>
<td>(0.21)</td>
<td></td>
</tr>
<tr>
<td>Model X²</td>
<td>14.26</td>
<td>18.71</td>
<td>6.16</td>
</tr>
<tr>
<td>Prior Probabilities</td>
<td>0.852</td>
<td>0.896</td>
<td>0.905</td>
</tr>
</tbody>
</table>

*p-values are reported in parentheses.

The results are consistent with expectations. In general, most of the surrogate or competitive variables, especially in the two- 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. 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 parameters are significant at the 99% level. In the three-year average model none of the variables is statistically significant.

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...
Table 2. Expected Cost of Misclassification\(^a\) for the RPA and Logistic Regression Models

<table>
<thead>
<tr>
<th>Relative Costs(^d)</th>
<th>1-Year Model</th>
<th>2-Year Model</th>
<th>3-Year Model</th>
<th>Relative Costs(^d)</th>
<th>1-Year Model</th>
<th>2-Year Model</th>
<th>3-Year Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:1</td>
<td>0.150(^b)</td>
<td>0.100(^b)</td>
<td>0.014</td>
<td>1:1</td>
<td>0.198</td>
<td>0.134</td>
<td>0.110</td>
</tr>
<tr>
<td>2:1</td>
<td>0.300(^b)</td>
<td>0.122</td>
<td>0.014</td>
<td>2:1</td>
<td>0.303</td>
<td>0.184</td>
<td>0.164</td>
</tr>
<tr>
<td>3:1</td>
<td>0.314</td>
<td>0.131</td>
<td>0.014</td>
<td>3:1</td>
<td>0.408</td>
<td>0.234</td>
<td>0.218</td>
</tr>
<tr>
<td>4:1</td>
<td>0.364</td>
<td>0.139</td>
<td>0.014</td>
<td>4:1</td>
<td>0.512</td>
<td>0.284</td>
<td>0.272</td>
</tr>
<tr>
<td>5:1</td>
<td>0.414</td>
<td>0.147</td>
<td>0.014</td>
<td>5:1</td>
<td>0.617</td>
<td>0.334</td>
<td>0.326</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RPA</th>
<th>Logistic Regression(^c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1992</td>
<td>1992</td>
</tr>
<tr>
<td>1:1</td>
<td>0.150(^b)</td>
</tr>
<tr>
<td>2:1</td>
<td>0.300(^b)</td>
</tr>
<tr>
<td>3:1</td>
<td>0.338</td>
</tr>
<tr>
<td>4:1</td>
<td>0.366</td>
</tr>
<tr>
<td>5:1</td>
<td>0.395</td>
</tr>
<tr>
<td>1993</td>
<td>1993</td>
</tr>
<tr>
<td>1:1</td>
<td>0.150(^b)</td>
</tr>
<tr>
<td>2:1</td>
<td>0.300(^b)</td>
</tr>
<tr>
<td>3:1</td>
<td>0.356</td>
</tr>
<tr>
<td>4:1</td>
<td>0.401</td>
</tr>
<tr>
<td>5:1</td>
<td>0.446</td>
</tr>
</tbody>
</table>

\(^a\) 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 naïve 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 suggest 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

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


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.


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


