ABSTRACT

DALEY, CAITLIN MARIE. Application of Data Mining Tools for Exploring Data: Yarn Quality Case Study. (Under the direction of Dr. Timothy Clapp and Dr. Jeff Joines).

Businesses are constantly striving for a competitive edge in the economy, and data-driven decision making is crucial to achieve this goal. Four data mining tools, principal component analysis, cluster analysis, recursive partitioning, and discriminant analysis, were used to explore the major factors that contribute to ends down in a rotor spinning manufacturing process. Principal component analysis was used to explore the research question about whether the large number of cotton properties used to classify cotton could be reduced to a significant few. Cluster analysis was used to gain insight about whether there were groups of gins, counties, or classing offices that produced better raw material than others and led to less ends down. The important research question of what raw material properties were affecting ends down was explored with both recursive partitioning and discriminant analysis. Additional research investigated the effect of cotton variety and atmospheric conditions on spinning productivity. Each of the four data mining tools used was informative and offered a different perspective to the overall research question. Several significant factors emerged including humidity, temperature, %DP 555, and uniformity in addition to micronaire and the color properties (+b and Rd). With these results the researcher developed an improvement plan for better control and increased spinning productivity in future operations. A designed experiment is necessary to thoroughly investigate the impact of certain factors beyond the exploratory conclusions obtained from this study.

Application of Data Mining Tools for Exploring Data: Yarn Quality Case Study

by
Caitlin Marie Daley

A thesis submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Master of Science

Textile Engineering

Raleigh, North Carolina

2008

APPROVED BY:

Dr. Timothy Clapp Committee Chair	Dr. Jeff Joines Committee Co-Chair

Dr. Jeff Thompson

BIOGRAPHY

Caitlin Daley started attending North Carolina State University in the fall of 2003. She completed a Bachelor of Science in Textile Engineering with a Minor in Spanish in 2007. She continued her education in the fall of 2007 by pursuing a Master's of Textile Engineering with a Minor in Statistics. Her background includes project work and training in Six Sigma, Design for Six Sigma, Lean methodology, errorproofing, new product development, statistics, and various software packages. Her research interests include analytic and quality improvement methods applied to industry and healthcare applications.

ACKNOWLEDGEMENTS

I would like to thank Dr. Clapp, Dr. Joines, and Dr. Thompson for their guidance and help throughout my research and for serving on my committee. I would also like to thank my project sponsor Matt Thomas for his process knowledge and advice. Finally, I would like to thank Marissa Langford and Dr. Lori Rothenberg for their statistical support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 Introduction

In today's modern world, advances in technology have led to enormous computing and data storage capabilities. A data management system within a company stores large amounts of data with multiple input and response variables. In fact, in 2003 WalMart recorded 20 million transactions per day into a 10-terabyte database (Shmueli, Patel, & Bruce, 2007). The analyst is faced with huge challenges in dissecting large amounts of data and forming valuable and significant conclusions that may lead to a business advantage. Data mining is defined as the process of selecting, exploring, and modeling large amounts of data to discover models and patterns (Giudici, 2003). More concisely, data mining is the process of extracting large amounts of information from large datasets.

Considerable amounts of data are available in all industries, and there is a huge need for tools and analysis techniques that quickly get the information in the hands of the practitioner. These tools should be straightforward enough that all workers from the manufacturing floor up to executive leaders can implement them and understand the results. Four data mining tools will be described including principal component analysis, cluster analysis, recursive partitioning, and discriminant analysis. A yarn quality case study will then be presented with actual data and results from using the four tools. The tool methodology and case study serve as an example for the use of these data mining tools in the textile industry, but the tools can be applied to any data-rich problem in any industry.

2.0 Literature Review & Background of Data Mining Tools

2.0.1 Introduction to Data Mining

The recent surge in data mining research and literature highlights the importance and interest related to this subject in today's business world. Businesses are constantly striving for a competitive edge in the economy, and data-driven decision making is crucial to achieve this goal. Both multivariate statistics and knowledge discovery in databases are closely related to data mining. Multivariate statistics is defined as analysis of more than one variable at a time and is made up of a series of tools and models to look at the relationships among many dependent and independent variables. This type of analysis includes statistical methods that aid in data mining and simplification of large amounts of information. In relation to statistics, data mining has been described as "statistics at scale and speed" (Shmueli et al., 2007). Knowledge discovery in databases (KDD) is another term for data mining, and it was derived from artificial intelligence research (Hand, Mannila, & Smyth, 2001). It encompasses methods to discover models and useful information in the data.

Using the four data mining tools, principal component analysis, cluster analysis, recursive partitioning, and discriminant analysis, conclusions can be formed from a data-rich problem in any industry. Principal component analysis is used in the beginning data preparation stage to reduce the number of predictor variables to a smaller set of uncorrelated variables. Cluster analysis is used to form groups of similar observations based on several variables to gain insight from large amounts of data. Recursive partitioning is an exploratory and interactive tool that determines optimal splits among input variables to best predict the response. Finally, discriminant analysis is a classification and profiling tool used on continuous variables. Each of these tools is informative and offers a different perspective to the research question of interest. A detailed review of the literature and case study examples was conducted for each of the data mining tools.

2.0.2 Overview of Data Mining Tools

The following table summarizes the four tools to be discussed in regards to the input data required, the output from the analysis, major assumptions that need to be verified, and examples of applications in the data mining process.

Table 2.1 Overview of Four Data Mining Tools

Tool	**Input Data**	**Analysis Output**	**Assumptions**	**Applications**
Principal Component Analysis (PCA)	Highly-dimensional continuous data, multicollinear	Principal components	No distribution assumptions	Variable reduction (data prep), exploratory
Cluster Analysis	Predictor and response variables (continuous/categorical)	Clusters of similar observations	No distribution assumptions	Exploratory
Recursive Partitioning	Predictor and response variables (continuous/categorical)	Decision tree with groupings of predictor variables that best predict response	No distribution assumptions	Exploratory, Classification & Regression Trees
Discriminant Analysis	Multiple independent continuous variables, single classification variable	Linear combination of covariates that best predict group membership	Predictors - multivariate normal distribution, groups - common variance-covariance matrix	Classification

2.1 Principal Component Analysis

Principal component analysis (PCA) is a dimension reduction process where a large number of continuous variables are reduced and transformed to a smaller set of variables that account for most of the variance in the original variables. A principal component is a linear combination of the optimally weighted observed variables (Lehman, O'Rourke, Hatcher, & Stephans, 2005) . By reducing the dimensionality (the number of input variables in the model) of the dataset in the data preparation step, highly correlated variables can be represented by a single variable and the overall accuracy of the fitted model improves. The resulting principal components are uncorrelated which is advantageous when forming a regression model because problems with multicollinear data will be avoided and the effects of individual variables can be separated (Shmueli et al., 2007). Also, these linear combinations that are used in modeling require less degrees of freedom than if all the original variables were used leading to better fitted estimates. Principal component analysis can also be used for visualization of many variables in a lower dimension to obtain more information about a large dataset (Giudici, 2003).

The number of principal components that are extracted from the data will equal the number of original variables, but the goal is to find the few variables that explain the same information as the original, larger set of variables. The first principal component is the linear combination of the standardized variables that lies along the direction of maximum variance of the observations, and the second component is in the direction of the next highest variance and orthogonal or uncorrelated to the first component (Esbensen, Schoenkopf, & Midtgaard, 1994). In technical terms, the first principal component is the linear combination of the eigenvector that is related to the largest eigenvalue of the covariance matrix and the original variable (Hand et al., 2001). The eigenvector gives the optimal weights or loadings, and the eigenvalue is the amount of variance described by a given component for each variable (Sall, Creighton, & Lehman, 2005). The resulting principal components are orthogonal to and independent of other predictors (Truxillo, 2005). If the measurement scales of the variables are different, standardization of the data must be done before using the covariance matrix to perform PCA so that the variability among the different measurement units of the variables does not affect the calculation of the principal component weights. Another option to obtain standardized variables when carrying out principal component analysis is to use the scaled correlation matrix (Yang & Trewn, 2004). Sample size must also be considered when performing PCA because it must be relatively large or at least five times the number of variables being analyzed (Lehman et al., 2005).

There are several methods of identifying the most significant components that include looking at the magnitude of the eigenvalues, using the scree plot, or comparing the percentage of variance represented by each component. Researchers often retain components with eigenvalues greater than one, which is known as the Kaiser criterion (Lehman et al., 2005). This means that that component explains more variance than any single variable in the analysis since the sum of the eigenvalues is equal to the number of variables in the analysis. A scree plot depicts the proportion of total variance, or the eigenvalues on the y axis, explained by each component on the x axis in decreasing order. This plot is used to identify the “elbow” in the graph, and the number of components above this break is kept in the analysis as they are the most important. Finally, researchers often determine prior to the analysis that a subset of components will be kept if their cumulative eigenvalue variance is at least 80% for example (Truxillo, 2005).

Once the components that account for the majority of the variability are identified using one of these three methods, the results must be interpreted by looking at the related weight or loading coefficient between the variables and the principal components. Loadings are a measure of the correlation between the original variable and the component and are used to compute the principal component scores (Hardle & Simar, 2003). Variables are said to load on specific components if their weight coefficients are greater than the absolute value of .40 (Lehman et al., 2005). The principal components can also be rotated to better interpret the solution (*JMP User's Guide,* 2007) . It is useful to the plot the principal components against the loading of each variable to assess the qualitative characteristics of each group (Hardle & Simar, 2003).

After the variables have been classified into common groups based on the patterns of their loadings, the components can be identified and named with the aid of a subject matter expert. For example, a six question survey to assess the volunteer attitudes of people was analyzed with principal component analysis and two components were identified that accounted for about 70% of the total variance. The loadings were interpreted and three variables loaded on the first principal component later labeled “financial giving,” and three variables loaded on the second component labeled “helping others” (Lehman et al., 2005). These findings enabled researchers to focus their survey efforts on these two categories.

The outcome of this analysis is a reduced set of variables that are easier to interpret and can be used for subsequent analyses. In addition, principal component analysis can be used to detect outliers in large datasets. With the majority of the variance condensed into a few variables, a visual inspection of a histogram of the principal component scores aids in the identification of multivariate outliers that cannot be detected in 3-D scatterplots (Truxillo, 2005).

Principal component analysis has been used in the development of surveys or questionnaires to decrease the number of questions to a smaller set that still accounts for the same content and variability. The response rate and efficiency of the survey will increase with a smaller number of questions that still effectively addresses the aims of the researcher's study. For example a survey was developed to assess health care providers' thoughts about domestic violence in order to improve treatment methods. By using PCA the original survey of 104 items was reduced to six principal components or content domains made up of 39 total items (Maiuro, 2000). This new survey more accurately and efficiently assessed health care providers' opinions. It extracted the same information that the original 104 item survey did, and the results led to a better treatment plan for victims of domestic violence.

Another application of principal component analysis is as a preparatory step for regression analysis. Collinearity among input variables leads to unstable models, so principal component analysis is a method to obtain independent variables to use in the analysis. An international math study was completed to investigate skills, perceptions, and behaviors of middle and high school age students. Data was collected on many variables with the goal of predicting the students' math test scores. Twelve variables were analyzed and found to be highly correlated, so principal component analysis was conducted to obtain a set of independent, surrogate predictors to use in the regression analysis. Two principal components were identified and input into the regression analysis to more accurately predict math test scores (Truxillo, 2005).

2.2 Cluster Analysis

Cluster analysis is the process of identifying unknown groups in observations by minimizing the within-cluster variation and maximizing the between-cluster variation. It is an exploratory method to identify similar objects in large datasets, and the number and names of the groups are not known beforehand (Sall et al., 2005). It is an unsupervised learning method to determine the natural groupings in the data as there are no reference variables with known levels or classifications. On the other hand, discriminant analysis discussed below is a supervised method because prior knowledge about group membership is known (Giudici, 2003 ;

Truxillo, 2005). Cluster analysis is often used in market based research for customer segmentation and structure analysis and is a very common descriptive data mining method (Shmueli et al., 2007 ; Giudici, 2003).

To complete cluster analysis the distance between observations and between clusters needs to be determined. Several measures exist to calculate the distance between observations including the Euclidean distance, standardized Euclidean distance (Pearson method), correlation-based similarity, Mahalanobis distance, Manhattan distance, and maximum coordinate distance. The standardized Euclidean distance is most commonly used although it doesn't account for correlation between variables and is very sensitive to outlying values (Shmueli et al., 2007). Linkage methods are used to calculate the distance between clusters and include minimum distance (single), maximum distance (complete), average distance, centroid distance, and Ward's methods (Yang & Trewn, 2004 ; Shmueli et al., 2007). Minimum distance (single), maximum distance (complete), and average distance calculate the distances between each observation of the two groups being compared. The centroid method calculates the centroid of each group for comparison, and Ward's method analyzes variance within and between the clusters rather than using a distance measure (Giudici, 2003).

Using these techniques to calculate the between observation and between cluster distances, there are then several methods to perform the actual cluster analysis that include normal mixtures, self-organizing maps, hierarchical cluster, and k-means (Sall et al., 2005). The objective of these methods is to get internal cohesion and external separation among the groups (Giudici, 2003). The two general types of clustering algorithms are hierarchical and nonhierarchical (k-means) methods that are discussed in detail below. Once the cluster analysis has been completed, the results are then interpreted to determine the names and characteristics of the clusters (Yang & Trewn, 2004). The researcher must work closely with an expert in the discipline when classifying and labeling the clusters.

The hierarchical method is commonly used and easier to implement with smaller datasets. In this method the number of clusters obtained may range from one to the number of observations present where each observation is in a different group. Agglomerative hierarchical clustering is where each observation is in a unique cluster at the beginning of the analysis. The number of clusters is then reduced by one at each step by combining observations with the closest distance until all observations are combined into one large cluster at the end. The divisive hierarchical method is the opposite and begins with one large cluster of all the observations. The large cluster is divided into smaller clusters at each step until each observation is in a cluster by itself (Shmueli et al., 2007). An output of hierarchical cluster analysis, either agglomerative or divisive, is a dendrogram which is a tree-like depiction of each observation, its clustering sequence, and the distances between clusters. Using a similarity measure or cutoff line on the chart, the number of clusters can be identified (Sall et al., 2005). The researcher must be cautious when blindly using a cutoff line or "tree cut" on the dendrogram because it may not accurately depict the number of clusters (Kettenring, 2006). A scree plot is also used to show the distances between clusters and a natural break in the curve is an estimate for the number of clusters that represent the data. Subject matter experts should review the cluster results for understanding and interpretation.

The nonhierarchical or k-means method is used for large data sets and starts with the identification of the number of clusters (k). Nonhierarchical methods are quicker than the hierarchical methods because a distance matrix does not have to be calculated for each partition (Giudici, 2003). The computer software calculates the k cluster centroids or seeds at random points in the dataset. Each observation is assigned to the closest cluster, the k centroids are recalculated based upon the new assignment, and the process repeats until no observations are reassigned. Because of the challenge of predicting the number of clusters beforehand in this method, the number of data clusters found in the hierarchical method could be used and input into the k-mean method for further analysis (Yang & Trewn, 2004). Different values of k can be iteratively analyzed to determine the best clustering results. Problems can occur with outliers because the k-means analysis may lead to small clusters of these extreme values with large clusters containing the majority of the observations (Giudici, 2003).

Both clustering methods were used to look at risk and protective factors among middle school students who lived in urban public housing developments. Researchers were interested in looking at the relationship between risk factors and different education and behavior outcomes. The variables used for the cluster analysis were carefully considered and selected on criteria important to the aims of the study. The data was standardized and hierarchical analysis was first performed. The results from the dendrogram were used to further analyze three, four, and five clusters in the subsequent k-means analysis. Four clusters were found to best describe the differences in the data among the students and were broadly labeled as protected, high-risk, coping, and disconnected. The researchers than validated the differences between these clusters through chi-squared and analysis of variance methods (Anthony, 2008). This application highlights the complementary nature of these two clustering algorithms, the hierarchical and k-means methods.

Applications of cluster analysis exist in many industries such as marketing to describe different customer segments, in psychology to discover personality types based on surveys, in psychiatry to create categorizations of mental illnesses, in meteorology to identify spatial pressure patterns for climate studies of the atmosphere, and many other examples given below (Hardle & Simar, 2003 ; Hand et al., 2001). Market and political forecasters commonly use zip codes to group neighborhoods by their preferences (Shmueli et al., 2007). Cluster analysis could be employed in the new product development process. Different brands and products on the market could be classified to identify the competitive and successful clusters of products. The business can then assess the impact of its new product on this market cluster structure and its potential success. Otherwise, the business can begin the design process anew for the development of a unique and innovative product that fills a niche in the cluster structure (Punj, 1983).

Cluster analysis has also been used in healthcare. Several years ago clustering analyses were done with nursing survey data to discover more information about stress and burnout. The researchers found that stress is more related to the work environment and overall workload than with the degree of specialization on the clinical service unit. Also it was determined that more psychological damage results from conflict with physicians than problems with other nurses (Hillhouse & Adler, 1997). Another study looked at chronic senility in patients to improve the medical care they were receiving. The researcher investigated patients' revisit frequency and disease correlation among several other variables. Cluster analysis was performed and led to eight clusters of which two clusters of patients were targeted in this study. The first cluster was identified as patients with poor health so efforts were directed at more effective treatment for overall improved health and wellbeing. The second cluster of patients had a low revisit frequency, and administrators targeted their efforts to increase these patients' customer satisfaction and business (Cheng, 2006). In this case researchers were able to use the results from the cluster analysis to better serve their patients.

2.3 Recursive Partitioning

Recursive partitioning is a method that splits data to find groupings of input variables that best predict the response. It is also referred to as decision trees or the CART™ program (classification and regression trees). This method recursively fits all possible splits to find the optimum partition that maximizes the difference in the responses between the two branches. If the response is continuous, the splits are made to maximize the difference between the means by looking at the squared errors. If the response is categorical, the estimated probability for each response level is calculated and the splits are made to maximize the likelihood-ratio chi-square statistic (*JMP User's Guide*, 2007) . The tree is made up of stepwise, univariate binary decisions that lead to simple classification rules that can be applied in practical applications (Hand et al., 2001). For instance, a set of questions to ask patients during triage can be formed from a table of symptom and disease diagnosis data. Once the data partitions have been made, the resulting "leaves" at the terminal nodes of the decision tree label the most likely class and its associated probability (Giudici, 2003). It should be noted that association not causality can be determined from this exploratory modeling method.

In order to avoid problems with overfitting the data, different stopping criteria and pruning methods have been developed to eliminate weak and unimportant splits (Shmueli et al., 2007). The ideal decision tree has minimal leaves to aid in interpretation of the predictive model and at the same time lots of leaves to increase the accuracy of the model. Stopping rules can include criteria about the number of leaves or the number of steps, or the analysis may stop when the decrease in impurity or heterogeneity is too small. In a study for example, researchers defined a terminal node if the leaf had less than 25 observations or the increase in the adjusted chi-squared value was minimal (Gaspar, 1997). In addition to stopping criteria, pruning methods are used to scale back the size of the decision tree and the number of splits. Pruning involves finding the optimal subtree that minimizes a loss function or cost-complexity measure (Giudici, 2003). The tradeoff between the misclassification error and the number of nodes in the tree is crucial in order to model the patterns but not the noise in the data (Shmueli et al., 2007). Therefore, researchers often fit several different models on the training data and then assess their performance with the validation data by looking at the R-squared value and the lift or ROC curves (Gaudard, Ramsey, & Stephens, 2006).

Recursive partitioning is a useful tool for variable subset selection and does not require any transformation of the variables since there are no distribution assumptions for the analysis. Outlying values may significantly alter the results, so data exploration is an important step to complete prior to the analysis (Giudici, 2003). This method requires a large dataset to form significant conclusions from a classifier, but recent research has investigated the use of random forests to create many classification trees from a smaller dataset. These results could then be combined for an improved conclusion (Shmueli et al., 2007). In addition, exploratory recursive partitioning can be a helpful method to determine factor levels for a subsequent designed experiment by assessing the partition split values and probabilities (Gaudard et al., 2006).

A specific application of recursive partitioning was to look at a sample of patients that underwent treatment in a Radiation Therapy Oncology Group (RTOG) brain metastases trial. Cancer patients often develop brain metastases and doctors struggle with selecting the patients that will benefit the most from an aggressive treatment. A sample set of 1200 patients with twenty-one variables each were analyzed using recursive partitioning to evaluate the effectiveness of the treatment interventions. Two significantly different groups of patients were found that had survival rates of 7.1 and 2.3 months respectively. Based on the characteristics and variable levels of these two groups obtained from the partition analysis, future clinical trials will be redesigned and doctors will target patients that have a better survival rate for aggressive treatments (Gaspar, 1997).

Another application of recursive partitioning was in the field of meteorology. Scientists in Australia used recursive partitioning to look at many variables such as sea level pressure to identify weather states of dry or wet conditions. They developed a weather generator model for maximum accuracy and parsimony and minimum cost-complexity. This model was then applied in case studies throughout Australia to assess its effectiveness in different climate regions (Schnur, 1998).

Recursive partitioning was used in a manufacturing setting to investigate factors that may have contributed to banding defects in a printing process. Banding occurs when there are irregularities in the cylinder that transfers ink to the paper and therefore causing imperfections in the final product. The process team brainstormed thirty-nine potential factors that included both nominal and continuous variables. Due to the number of variables, their type, and the difficulty in using a specific statistical tool with several assumptions, recursive partitioning was used to explore the conditions that may lead to a printing defect (Ramsey, Stephens, & Gaudard, 2005).

2.4 Discriminant Analysis

Discriminant analysis, also known as Fisher's rule, classifies units into predefined groups based on similar characteristics (Giudici, 2003). The goal of this analysis is to determine which variables best predict group membership. This analysis is often considered a multivariate generalization of logistic regression (Truxillo, 2005). A linear discriminant function is used as the prediction criteria or "cutoff" score for classifying two groups, and Mahalanobis distance is used when categorizing multiple groups (Yang & Trewn, 2004). The discriminant function obtained is very useful to predict the classification of future observations.

The major assumptions in discriminant analysis are that the predictors are from a multivariate normal distribution, linearly related, and the categorical groups have a common variance-covariance matrix (Esbensen et al., 1994). Skewed data can considerably affect the resulting analysis so data transformations may be needed, however the method is more robust for large sample sizes. Quadratic discriminant analysis can be used instead of linear discriminant analysis if the covariance structure is significantly different (Shmueli et al., 2007). The quadratic discriminant analysis uses a separate estimate of covariance for each group rather than the pooled estimate that is used in the linear method (Truxillo, 2005). Research has shown that discriminant analysis is reasonably robust even when there are modest violations of these assumptions (Lachenbruch, 1975).

The general method of discriminant analysis is to classify observations into groups that are the closest distance away. Different distance measures can be used to calculate the distance from the observation to the mean (for a single variable class) or the centroid (vector of means for two or more variables) of the class. The Euclidean distance is commonly used although it is not a standardized distance, does not account for variance of the variables, and disregards any correlation between the variables. An alternative is to use the Mahalanobis distance which takes into account the variability and correlations of the variables. Once the distances have been calculated, a linear classification function with weighted discriminatory variables is found that maximizes group separation. A classification score is computed for each observation and the observation is grouped in the class for which it has the largest score, corresponding to the smallest distance from the class (Shmueli et al., 2007). The discriminant functions are determined from the test data and should be validated with a validation sample to assess the classification error rate and the fit of the model (Truxillo, 2005). A practical consideration the researcher should make when using discriminant analysis is the expected cost of misclassification and the complexity related to the classification decision. Information about unequal misclassification costs can be added to the functions in order to minimize the expected misclassification costs (Shmueli et al., 2007).

Stepwise discriminant analysis can be used to find the order of the variables that best predict group membership. Statistical criteria such as the alpha level are used to add or remove variables to the model (Truxillo, 2005). These results can be used when forming a regression model or simply to reduce the number of variables to the few, statistically significant predictors.

An application of discriminant analysis is in categorizing current and potential credit card holders into "good" and "bad" groups based on variables such as education level, salary, family size, and credit rating among others. By developing a discriminant function based on these variables, potential credit card holders can be screened and classified as a "good" or "bad" customer. Businesses can adjust their marketing and recruiting plans to increase revenue for the company by targeting their efforts towards desirable customers (Yang & Trewn, 2004). Another example is to use discriminant analysis to look at survey question results to classify the subjects. In a specific study a researcher was interested in predicting the type of gambler a person is based on responses from a 12-item questionnaire related to gambling. The participants were identified as one of the following types of gambler: binge, steady, or control. Two discriminant functions were identified from the analysis that maximized the distances between the group centroids. The first discriminant function seemed to divide the binge and steady gamblers while the second function separated the control and steady gamblers (Truxillo, 2005). These resulting functions based on the twelve variables were extremely useful for researchers in identifying people with potential gambling problems.

2.5 Evaluation of Data Mining Tools

When analyzing classification and prediction results, there are several methods to assess the performance of the tool employed. The classification or confusion matrix shows the correct and incorrect classifications for a specific dataset by assessing the actual and predicted class of observations. The number of correct classifications is displayed on the diagonal, and the overall error and accuracy can then be calculated from this matrix. Other common metrics calculated include sensitivity and specificity. Sensitivity is the probability that the predictor correctly classifies an observation, and specificity is the probability that the predictor correctly rules out an observation. The Receiver Operator Characteristic (ROC) curve is often used to depict the relationship between false-positive or false alarm on the x-axis (1-specificity) and true-positive or hit rate on the y-axis (sensitivity) rates, and the area under the curve is used to evaluate this relationship. An area of 1 corresponds to a perfectly fitting model with accurate discriminatory power while an area of .5 corresponds to a model fitting chance alone. The ROC curve highlights the predictive accuracy of the model and is based on the confusion matrix (Giudici, 2003). Both the classification matrix and ROC curve are used to assess discriminant analysis results, and the ROC curve is also used to find the best fitting model for recursive partitioning.

A lift chart is similar to the ROC curve and is used to look at the specificity versus the cumulative number of observations to compare the performance of the model with different portions of the population (Shmueli et al., 2007). The lift chart is compared to the baseline model or the response in absence of a model to show the relative improvement of the fitted model. Lift curves are commonly used to assess the many different splits and fitted models used in recursive portioning.

A consideration when selecting a model that accurately describes the data is the bias-variance trade-off. Bias is the systematic error between the expected value and true observation and is related to the model approximation. Variance is the random and data-driven component of error (Hand et al., 2001). Cross-validation can be used to find solutions to the bias-variance dilemma through different model selection criteria. For this method the dataset is usually divided into three parts for analysis: training, validation, and testing. The training data is used to fit the model, and the model is then selected using the validation data. The prediction error will be calculated using the test data. Another approach is the k-fold cross-validation in which k subsets of data are formed from the dataset. The model is fit k times and one subset is excluded each time which is used to calculate the prediction error (Giudici, 2003). Cross-validation is commonly used with both recursive partitioning and discriminant analysis to assess the fit of the training model on the validation or hold-out data, and classification error statistics can be calculated.

In addition to these methods for assessing the results from the classification and prediction tools, some basic methods or statistics can be used to evaluate the exploratory tools. For PCA the results can be quantified by looking at the cumulative percent variation represented by the eigenvalues. For cluster analysis, the clusters can be verified by looking at the spread of the cluster observations across several variables and an analysis of variance conducted to see if the clusters are statistically different. An R-squared value, the proportion of variance explained by the model, is often shown for recursive partitioning, and a graph showing this value versus the number of splits taken is helpful when determining the optimal classification tree without overfitting.

3.0 Yarn Quality Case Study Problem Definition

3.1 Definition of Case Study and Research Questions

After the researcher established an understanding of the four data mining tools and how they have been used in different applications, a case study was chosen to use these methods to solve a real problem in industry. A specific textile manufacturing plant was studied with the research aim to improve yarn productivity by focusing on ends down in the production of cotton yarn in the open end spinning process. An ends down is defined as a yarn break in the final production process where cotton sliver is spun into cotton yarn. Ends down lead to decreased efficiency and profit because they are an interruption in the manufacturing process that must be repaired before production can be continued. By exploring and analyzing the many different variables concerning the raw materials and spinning processes, the researcher hoped to obtain an understanding of the major factors that contribute to ends down to develop an improvement plan for better control and productivity in future operations.

The overall objective of the case study was to determine what factors were driving ends down by using the four data mining tools to improve the overall efficiency in the spinning process. Specific research questions were identified that could be addressed by each of the statistical tools and result in data-driven decisions to guide continuous improvement. Principal component analysis was used to explore the research question about whether the large number of cotton properties used to classify cotton could be reduced to a significant few. Cluster analysis was used to gain insight about whether there were groups of gins, counties, or classing offices that produced better raw material than others. The important research question of what raw material properties were affecting ends down was explored with both recursive partitioning and discriminant analysis. Recursive partitioning helped define a decision tree and factor levels that would lead to decreased ends down, and discriminant analysis yielded a classification function with material properties as inputs to predict the number of ends down as the outcome. Additional research investigated the effect of cotton variety and atmospheric conditions on spinning productivity using the exploratory recursive partitioning tool and discriminant analysis.

3.2 Textile Yarn Spinning Background

Open-end rotor spinning is the process from which a bale of raw cotton is blended, cleaned, processed, and spun into yarn. The figure below depicts the yarn spinning process. First, laydown machines uniformly blend ninety bales of cotton with varying properties to achieve an “average” laydown with a wide distribution of properties. Multimixers and cleaning machines open and blend the cotton to decrease trash and any foreign particles that may be leftover from the ginning process. The laydown is then processed in the carding machines where the fibers are further cleaned, rid of entanglements, and made parallel into cotton sliver, a ropelike extension of the cotton fibers. The drawing machines orient and cross-blend the fibers and make the sliver more uniform through doubling. The doubling process is where six slivers are combined to decrease the coefficient of variation of the output sliver strand; a low coefficient of variation value corresponds to a more uniform thickness profile of the sliver. Finally, the sliver is processed in the open-end rotor spinning machines where yarn is produced and wrapped up in packages. The actual yarn formation occurs in the spinbox, and the spinning and winding steps are combined into one step. The specific manufacturing plant in this study consumes about 90,000 bales of cotton per year and produces mainly 100% cotton with a cotton count of 18 hanks per pound (Ne=18). A hank is equivalent to 840 yards, and a larger cotton count corresponds to a finer yarn. The open-end spun cotton is ultimately used for knitting and jersey applications.

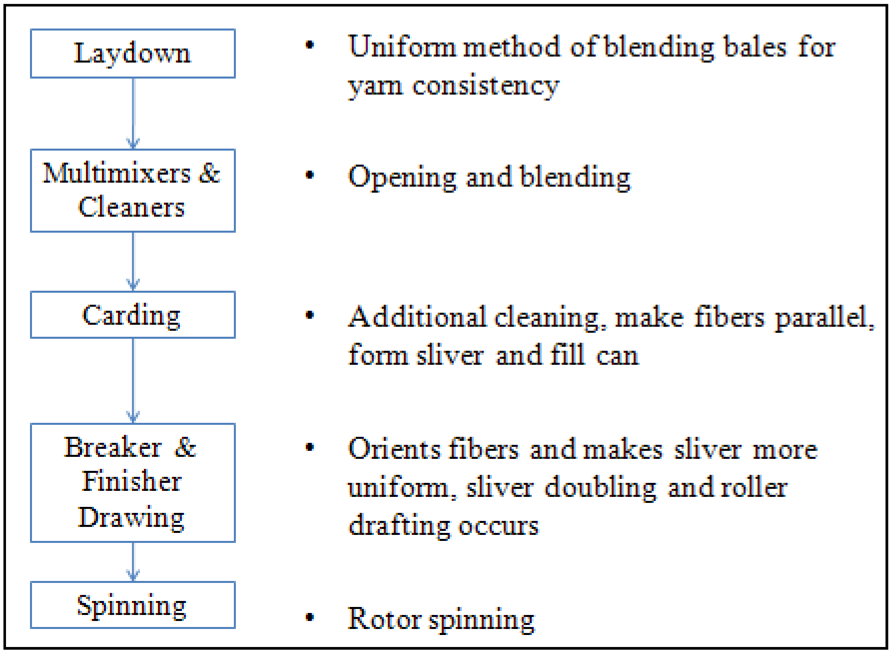


Figure 3.1 Yarn Spinning Process

3.3 Textile Spinning Research

The researcher studied the textile literature for the influence of cotton characteristics on yarn quality and factors that may affect spinning productivity. First, the researcher focused on studies that relate cotton fiber characteristics to yarn quality. For 100% cotton yarns strength, fineness, and length are the most important fiber properties for rotor spun yarn (Bradow, 2000). A study was done to specifically investigate the effects of fineness and strength on open-end yarn properties, and it was found that low micronaire cotton was more important than strength to the resulting yarn strength (Simpson & Murray, 1978). For blended yarns friction is an additional important property because it aids in fiber cohesion. In regards to strength, fineness, and length researchers have shown that fiber length is primarily a genetic trait that is directly related to yarn fineness, strength, and spinning efficiency (Bradow, 2000).

A study was completed to look at the effect seven fiber properties (micronaire, length, length uniformity, strength, greyness, yellowness, and trash grade) had on the strength of open-end rotor spun yarns. The researchers found strength, fineness, and length uniformity to be the most important properties affecting yarn strength. In addition they found there is a strong positive relationship between micronaire and length uniformity, strength and length, and a strong negative relationship between micronaire and length. The researchers highlighted the lack of significance of the effect of fiber length on yarn strength (Ethridge, Towery, & Hembree, 1982).

Next, the researcher focused on known factors that affect spinning productivity defined as the number of ends down in the production. An experiment was conducted to investigate ends down in a ring spinning process of cotton yarn. It was found that large amounts of short fiber content was the most significant factor affecting ends down followed by span length, micronaire variability, and the number of bales per laydown. Short fiber content was defined as the percentage of fibers shorter than 12.5 mm in length (Backe, 1986). Another study was done to determine the effect of blend, twist factor, and fineness of cotton on spinning productivity. The largest ends down rate was seen with 100% coarse and 100% low-strength cotton at low twist factors, and lower ends down were related with fine fiber cotton (Simpson & Murray, 1978). Micronaire is an important factor that affects both the yarn quality and spinning productivity where fine fiber cotton leads to lower ends down. Finally, a research study was conducted to investigate the effect of feed sliver moisture content on rotor spinning performance. It was found that high moisture content in the feed sliver leads to decreased ends down but negatively impacts yarn evenness and elongation as the number of imperfections increases (Basal & Rust, 2001).

The impact of the atmospheric conditions on spinning productivity was then researched. A study was done to examine the effects of a direct conditioning system on spinning productivity under two different scenarios of cold/dry and warm/humid. The warm/humid conditions were found to lead to more ends down because of fluctuations in the abrasion between the fiber and metal of the open end machine. In addition, the number of yarn imperfections and cotton stickiness increased. It was concluded that temperature and humidity have a larger effect on yarn performance than yarn quality (Artzt P., Herter T., & Preininger H., 1990). The temperature and humidity dependent properties of the sugars that are present on cotton lead to stickiness, which is a worldwide problem as the average degree of contamination of stickiness of cotton has reached about 22%.

The stickiness of cotton and sugar accumulation on the machinery leads to decreased spinning productivity because the frictional forces are increased resulting in more machine wear (Hequet, 2005).

Fiber color can be traced back to how the cotton was grown and ultimately related to fiber quality and even spinning productivity. For example, white fibers with high Rd values are more mature and of higher quality than dull and yellow fibers that resulted from bad environmental conditions. The ultimate cotton fiber is white as snow, strong as steel, as fine as silk, and as long as wool (Bradow, 2000). Factors that negatively affect the ideal color properties include weather, insect damage, foreign matter contamination, and excessive moisture and temperature levels during storage. These damaging effects of the environment on the color of cotton later lead to decreased spinning productivity. The color properties can also affect other processes such as dye uptake and the application of finishing treatments (*The Classification of Cotton* 2001) .

3.4 Data Structure

After definition of the research questions and review of the textile yarn spinning process, eleven total datasets were obtained with around 500,000 observations and nearly one hundred different variables detailing information about the cotton classification properties, production process, and quality of the end yarn product (see Figure 3.2 below). There were approximately 350,000 cotton bales, 4300 laydowns, and 300 gins studied in the cotton data from October 2003 to May 2008. The researcher was faced with lots of information but used data mining tools to explore the data to find patterns and useful information for quality improvement at the spinning plant.

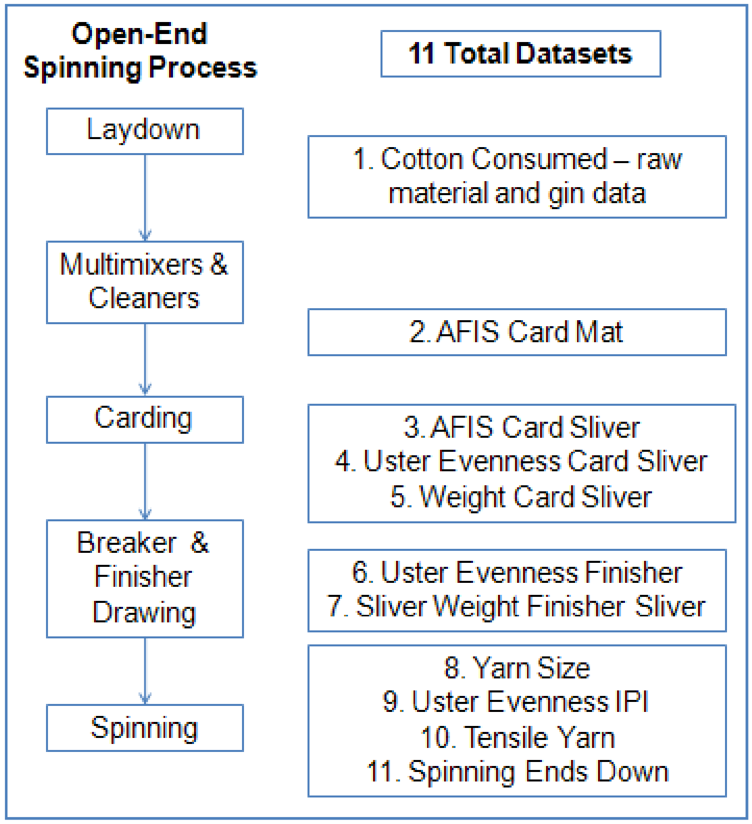


Figure 3.2 Structure of Datasets Related to Yarn Spinning Process

3.4.1 Data Variable Definitions

The majority of the research focused on the cotton classification properties and their effect on spinning ends down in the final process step. Before analyzing the data, an understanding of the variables and their definitions was necessary in order to comprehend the results and form meaningful conclusions. The United States Department of Agriculture has several Classing offices throughout the US that classify the bales of cotton based on several parameters including fiber length, length uniformity, strength, micronaire, color, trash area, and classer leaf grade. Two samples of fibers are taken from each bale for the testing which primarily occurs on the high volume precision instrument (HVI) machines, although the leaf grade is determined manually (*The Classification of Cotton*.2001) . See Table 3.1 below for an overview of the major variables studied.

Table 3.1 Summary of Variable Definitions

Variable	Type	Definition
Length	Independent	inches, upper half mean length
Uniformity Index	Independent	%, ratio of upper half mean length to mean
Strength	Independent	g/Tex
Micronaire	Independent	measure of fiber fineness and maturity
Rd	Independent	% reflectance or brightness
Plus b	Independent	degree of color pigmentation
Trash Area	Independent	percentage of the surface area occupied by trash particles is calculated
Classer Leaf	Independent	visual estimate of the amount of cotton plant leaf particles in the cotton, 1-7 scale
Ends Down	Dependent	ED/MRh, number of yarn breaks per 1000 rotor hours

Fiber length is the upper half mean length (the average length of the longer half of fibers) and is reported in inches with a short fiber defined as less than 0.99 inch. Length uniformity or the uniformity index is the ratio of the mean length to the upper half mean length of the fibers, and 77% or less is a very low degree of uniformity. The strength of a fiber is the amount of force in grams necessary to break one tex unit of fibers, and a strength value less than 23 grams per tex characterizes a weak fiber. A tex is the weight in grams of 1,000 meters of fiber. Micronaire is a gauge of the fineness and maturity of the fibers with a smaller micronaire value corresponding to finer fibers. The color of fibers is measured by the degree of reflectance and yellowness. Reflectance (Rd) is the percentage reflectance or brightness of the fibers, and the yellowness measure (+b) is the amount of color pigmentation. These two values are plotted on the Nickerson-Hunter cotton colorimeter diagram, and the fiber sample's three-digit color code is determined by the intersection of these two values on the diagram (see Appendix A). The trash area is the percentage of the sample surface area that is made up of trash particles. Finally, the classer leaf grade is a visual approximation of the amount of cotton leaf content present in the sample on a scale from one to seven with seven corresponding to higher trash content. Although the trash area and classer leaf grade are measured separately, there are observable correlations between the two measures (see Appendix B) (*The Classification of Cotton* 2001) .

Short Fiber Index was an additional variable present in the material data that the USDA Classing offices do not measure. It is a function of the uniformity and length of the fiber and therefore an indirect measurement that is highly correlated with both of these variables. Therefore, it was not considered in the analyses because the uniformity and length variables sufficiently described the cotton material properties.

Ends down in the manufacturing plant are measured as the number of yarn breaks per 1000 rotor hours, and this variable was used as the main outcome measure of the process (dependent variable). Other available data included information from intermediate processes such as the card and drawing sliver and the physical yarn testing data.

3.4.2 Dataset Estimates

To address the main research question of the impact of raw material on spinning productivity, two datasets had to be joined, Cotton Consumed (raw material) and Spinning Ends Down. There was no common identifier between the two datasets and only limited data available about the process flow rate; therefore, the researcher attempted to define an approximate time between laydown and spinning in order to merge the datasets based on production time. The average production rate of the spinning plant was 855,000 pounds per week with one laydown weighing approximately 45,000 pounds, and so the weekly production was about nineteen laydowns. A continuous spinning operation was assumed with 168 working hours per week which led to the approximation that each laydown takes about 8.8 hours from laydown to spinning.

Considering the production and efficiency rates and work in production (WIP) levels, a more detailed analysis was done that led to a calculated 8.4 production hours per laydown. From these results and due to the fact that the data is only averaged and available daily, the researcher assumed a two day moving average estimate from the laydown to ends down. This estimate was confirmed with the project sponsor and then used to merge the data in order to determine the impact the raw material had on spinning productivity.

Due to the vague estimate that was made for the time between the raw material and ends down data and the many other potential factors that could impact the process, only speculative evidence can be determined for the analyses and all conclusions are exploratory. In the future a designed experiment should be conducted to track a specific laydown of cotton through the spinning process to collect more reliable data. Precise labeling and tracking of the cotton material would be necessary throughout the process. This tracking requires considerable effort due to the nature of the spinning process that involves lots of blending and different steps to yield uniform properties.

In addition to the time estimate, the researcher made an assumption about "good" and "bad" levels of ends down in order to recode the continuous values into a categorical variable. For each crop year, the number of ends down were coded as "bad" if they were greater than the 75th percentile (see table). This classification of the spinning outcome as "good" or "bad" enabled a categorical response to be used in fitting models for recursive partitioning and discriminant analysis.

Table 3.2 "Bad" Ends Down Quartiles by Crop Year

Crop Year	**75% Quartiles ("Bad" Ends Down)**
2003	284.9
2004	284.2
2005	285.0
2006	308.0
2007	302.0

3.4.3 Cotton Variety Data

At the suggestion of the project sponsor, the researcher investigated the effects of a specific cotton variety on US cotton properties in general and then in relation to the specific case study. Based on the sponsor's process knowledge and industry experience, he hypothesized that increasing percentages of Deltapine brand 555 BG/RR cotton variety led to decreasing quality of the raw material and could therefore impact the spinning performance of the cotton yarn. BG stands for the trademark Bollgard technology that offers built-in pest control to cotton, and RR (Roundup Ready) is the genetic engineering process to enable cotton to be safely treated with herbicides (glyphosate) that kill harmful weeds. The amount of Deltapine variety grown has steadily increased over the past several years (see figure below). It represents a significant portion of the total US Upland cotton acreage and is the most planted cotton product (*Deltapine*, 2008) .

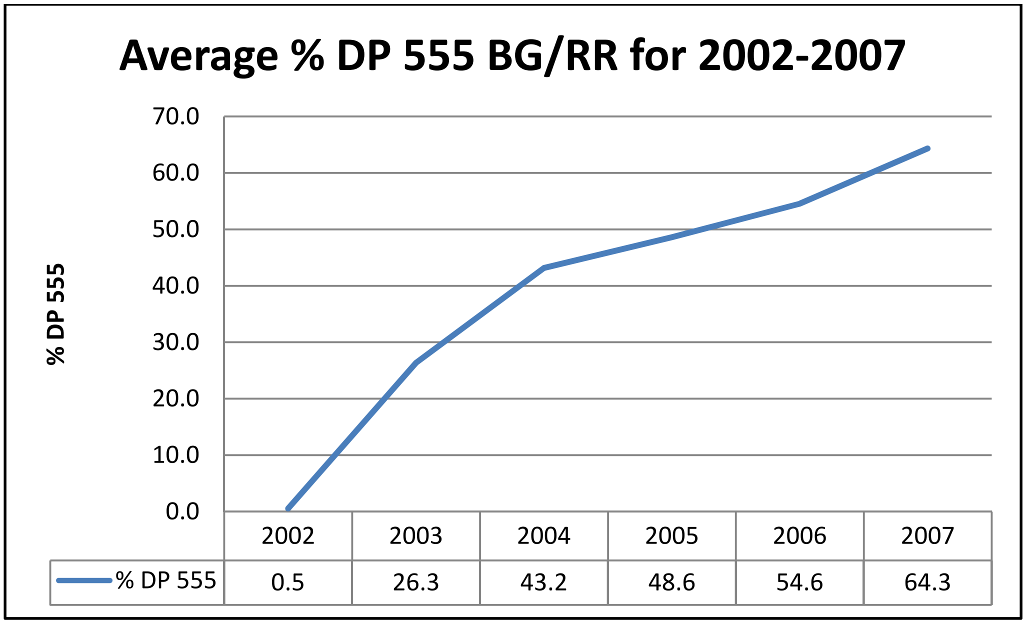


Figure 3.3 Average %DP555 by Crop Year(Agricultural Marketing Service, 2002-2007)

Six years of USDA cotton variety data by state for AL, FL, GA, LA, MS, and SC was analyzed and compared to the resulting crop quality summary data from Cotton Incorporated. The researcher was ultimately looking for any correlations and relationships between this specific Deltapine cotton variety factor and the resulting cotton classification properties. After the collection of the USDA and Cotton Incorporated data, the percentage of DP 555 cotton variety data by state was related to the material and spinning ends data for the specific case study. The USDA data concerning the percentage of DP 555 grown per crop year per state was added to the cotton data based on the state where the cotton bale was ginned. The cotton variety data was added for all ten ginning states represented in the case study dataset: AL, AR, FL, GA, LA, MO, MS, NC, SC, and

TN. The percentage of DP 555 and its effect on the cotton classification properties and the number of ends down was considered in recursive partitioning, discriminant analysis, and a separate regression analysis.

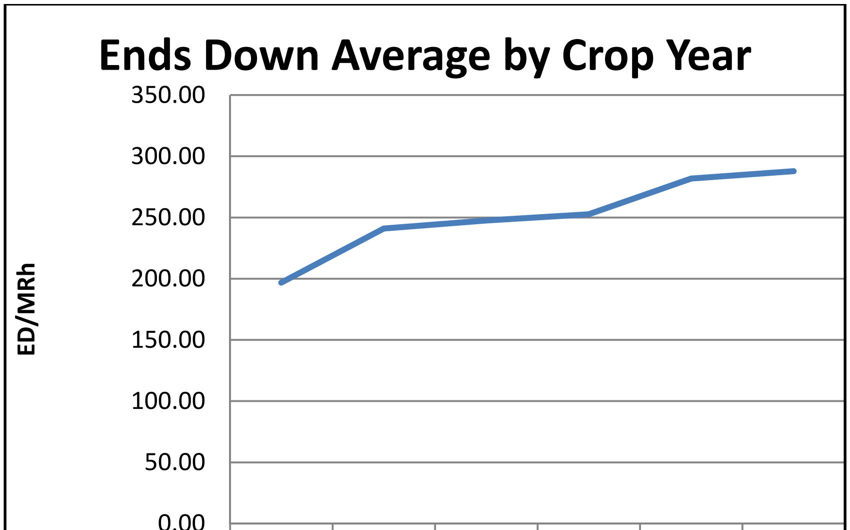
3.4.4 Temperature and Humidity Data

Temperature and humidity data later became available as part of the case study. The researcher wanted to determine if the temperature and humidity inside and outside the manufacturing plant had any influence on spinning productivity in terms of ends down. The data was measured every minute in the plant but it was compiled to a daily average to match with the spinning data. This environmental data was available from the cleaning, carding, drawing, and spinning zones within the plant and the researcher chose to focus on the zones associated with the spinning production (see Appendix C). The average temperature and humidity values in spinning were calculated for these three zones. The effects due to these spinning conditions in addition to the outside temperature and humidity variables were considered in both recursive partitioning and discriminant analysis.

3.5 Data Preparation

After determination of the research questions, data structure, and variable definitions the researcher began preparing and exploring the data for analysis. This preliminary data preparation step is critical for any conclusions that are ultimately made because of the well-known reality that bad data input into the analysis tools will only lead to bad results, the “garbage-in, garbage-out” phenomenon. The data should be cleaned by identifying outliers and missing data. It is helpful to graphically explore the data via histograms, scatterplots, or other methods to observe the spread of the data and to detect any collinearity or correlation among variables. Specific statistical assumptions may need to be checked such as multivariate normality and equality of the variance-covariance matrices. Depending on the results of the normality tests, the data may need to be transformed using square root, log, and reciprocal transformations for example (Hand et al., 2001). If the analysis techniques used require supervised training then the dataset should be separated into training and validation datasets, typically a 70/30 split. The models can then be tested on the validation data to avoid overfitting problems.

The researcher initially explored the distribution of the data using histograms, scatterplot matrices, and pairwise correlations between factors. As mentioned above, the Short Fiber Index variable was collinear with uniformity and length so it was not included in further analyses. By definition, variables are collinear if they are highly correlated (Truxillo, 2005). The researcher also looked at the raw material and ends down trends over time and noticed significant differences by crop year, the year the cotton was grown. Therefore, the five years of data was divided by crop year into manageable groups to more accurately investigate important factors (see figure below). Since there were so few observations and only four days of available data for the 2002 crop year, it was eliminated from further crop year analyses.



	2002	2003	2004	2005	2006	2007
# Bales	1080	76396	102499	85407	107969	18375
Ends Down	196.80	241.00	247.60	252.70	281.79	287.85

Figure 3.4 Ends Down Average by Crop Year

3.6 Verification of Assumptions for Data Mining Tools

Principal component analysis, cluster analysis, and recursive partitioning do not require any specific statistical assumptions because they are exploratory tools but significant outliers were eliminated from the data prior to use of these analyses. Discriminant analysis does require the assumptions of multivariate normality and a common variance-covariance structure. SAS® software was used to check these assumptions for each dataset by crop year, and specifically the %MULTINORM macro to test multivariate normality, the CALIS procedure to check the multivariate kurtosis and skewness, and the DISCRIM procedure to test the homogeneity of the covariance matrices (see Appendix D for SAS code used).

The %MULTINORM macro is a procedure to check the multivariate normality of several variables. The assumption of multivariate normality means that the joint distribution of the variables is normal in addition to each variable being univariate normal (Truxillo, 2005). First, the univariate procedure was used to check univariate normality by examining the quantile-quantile plots of each of the eight cotton classification variables (mic, length, strength, +b, Rd, trash area, uniformity, classer leaf). These plots showed varying trends and distributions due to the nature of some of these variables, and univariate normality could not be assumed. However, because of the exploratory nature of this data mining analysis, the overall normality assumption was not rejected and further checks on the multivariate distribution were completed. The %MULTINORM macro was then used to look at the multivariate normality. Output from the macro includes a multivariate quantile-quantile plot of the squared Mahalanobis distances of the observations from the centroid against the chi-squared quantiles. The data is multivariate normal and distributed as a chi-squared distribution when these plotted values are close to a diagonal straight line. For example the figure below for data from the 2003 crop year shows a multivariate normal distribution because the points are approximately centered on a diagonal line. Based on the multivariate quantile-quantile plots for each crop year (see Appendix E), no obvious deviations were observed and approximate multivariate normality could be assumed.

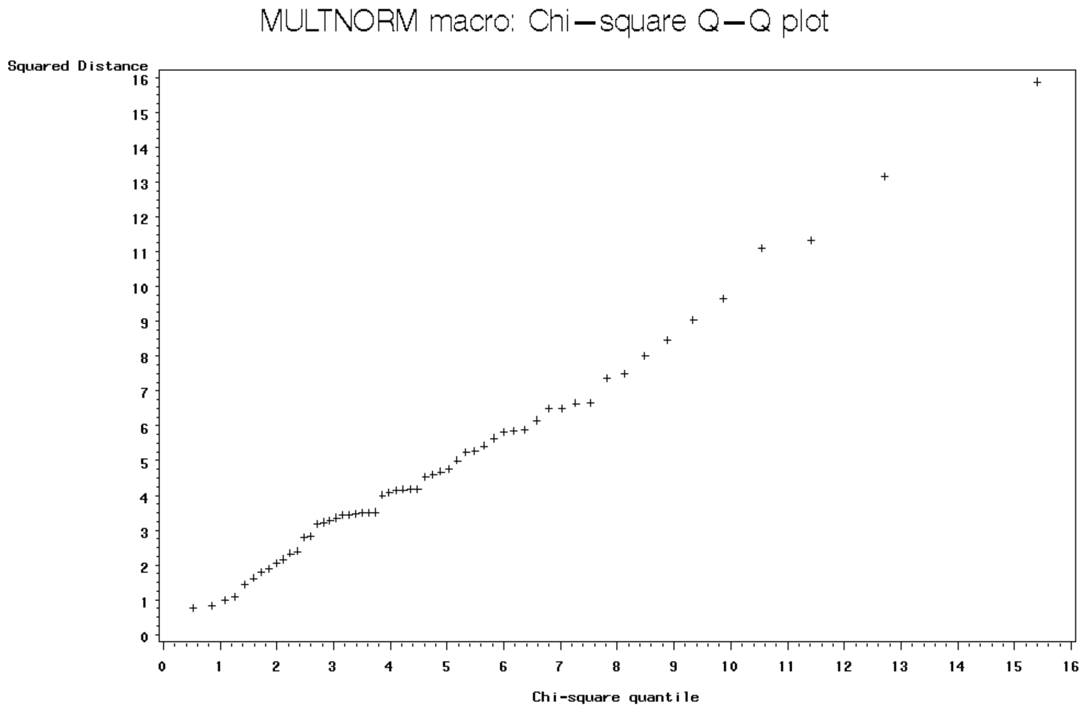


Figure 3.5 Multivariate Quantile-Quantile Plot for 2003 Crop Year Data

In addition to checking the univariate and multivariate quantile-quantile plots, the skewness and kurtosis of the data were investigated. Skewness is a measure of the lack of symmetry of the data, and kurtosis is a measure of the peaked or flat nature of the data relative to the normal distribution. Kurtosis problems have a larger impact on the multivariate normality of a dataset than skewness (Truxillo, 2005). The CALIS (Covariance Analysis and Linear Structural equations) procedure was used in SAS® to assess these multivariate measures. The univariate kurtosis and skewness values showed no significant nonzero values or obvious departures from normality (see Appendix F). The table below shows Mardia's multivariate kurtosis statistics for each crop year, and in general there were no significant nonzero values except for the outlier in 2005. A detailed analysis of this dataset was later conducted and it was found that all values for the classer leaf variable were the same so this variable had no discriminatory power and should not be included as a factor. Analyses will be presented later with and without this variable. In conclusion, the data was assumed to be approximately multivariate normal and the exploratory data analysis could be carried out without major data transformations based on the multivariate quantile-quantile plots and the kurtosis and skewness statistics.

Table 3.3 Mardia’s Multivariate Kurtosis Values by Crop Year from CALIS Procedure

Crop Year	**Mardia's Multivariate Kurtosis Value**
2003	2.0515
2004	-0.2739
2005	85388.04
2005 (without classer leaf variable)	0.043
2006	-1.6823
2007	0.1356

Finally, the DISRIM procedure was evaluated for each crop year dataset to check the homogeneity of the spinning outcome (“good”/”bad”) covariance matrices. The results showed that the Chi-squared tests were significant at the .05 alpha level for each crop year. This meant that significant differences existed between the variance-covariance matrices and the within rather than the pooled covariance matrix should be used in the discriminant function (Truxillo, 2005). Therefore, quadratic discriminant analysis was used which takes into account separate covariance estimates for each group rather than the pooled matrix. After checking the multivariate normality and common variance-covariance structure assumptions, discriminant analysis could then be performed.

4.0 Results

Once the data was prepared the four data mining tools, principal component analysis, cluster analysis, recursive partitioning, and discriminant analysis, were used to find information and models in the data to address the research questions. Most of the analysis was done in JMP® software with additional testing for discriminant analysis completed in SAS®. See Appendix G for a detailed guide of the methodology and steps used in JMP®. Each of the tools was iteratively used and the results analyzed for importance to the overall investigation of factors that lead to decreased yarn quality and increased number of ends down. For recursive partitioning and discriminant analysis the data was divided into training and validation datasets to prevent overfitting and to assess the fitted model. The researcher also studied the effect of cotton variety and atmospheric conditions on spinning productivity.

4.1 Principal Component Analysis

First, principal component analysis was used as a preliminary step to create a small number of surrogate variables from the many original variables. The practical question being explored was whether the many cotton classification properties used could be reduced to a significant few variables. These uncorrelated principal components could then be used in future regression analyses or simply aid in the interpretation of a large number of variables.

The analysis was done on the correlation matrix which is already normalized so scaling of the data was not necessary. The cumulative percentage of variation represented by the eigenvalues, the scree plot, and the eigenvalue-1 criterion were all examined to estimate the number of principal components. The number of principal components was specified, and varimax rotation was done to visualize the component membership and solution. The rotated factor pattern was made into a chart with the principal component loadings on the y axis and the variables as the x axis categories. This chart depicted which variables loaded on which component, and the principal components could be qualitatively labeled if possible.

The following figure shows the four rotated principal components identified from the eight cotton classification properties. By rotating the components and then making a chart of the results, the component membership was easier to determine. For instance, the first principal component (red bars) is made up of the classer leaf and trash variables, and the remaining principal components can be determined in a similar manner.

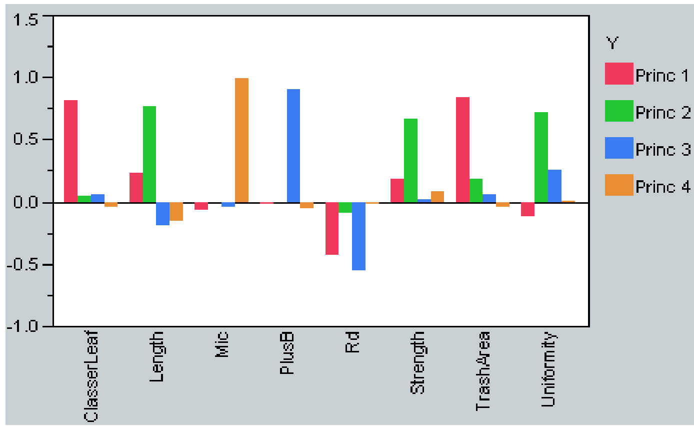


Figure 4.1 Rotated Factor Pattern of Cotton Classification Variables

This dimension reduction process was done on several different datasets. PCA was first used on data from the card sliver that was measured on the AFIS (Advanced Fiber Information System) machine. By using information from the eigenvalue-1 criterion and the scree plot below, the five tested variables were reduced to two principal components that accounted for nearly 70% of the variation. By considering a chart of the component loadings and variables (see Appendix H), the first component was identified as “particle defects” made up of dust, trash, and VFM (Visible Foreign Matter) variables. The second component was made up of neps and SFC (Short Fiber Content) variables and labeled “fiber defects”. Future analyses could study the effects of these two transformed factors in comparison to the original five variables.

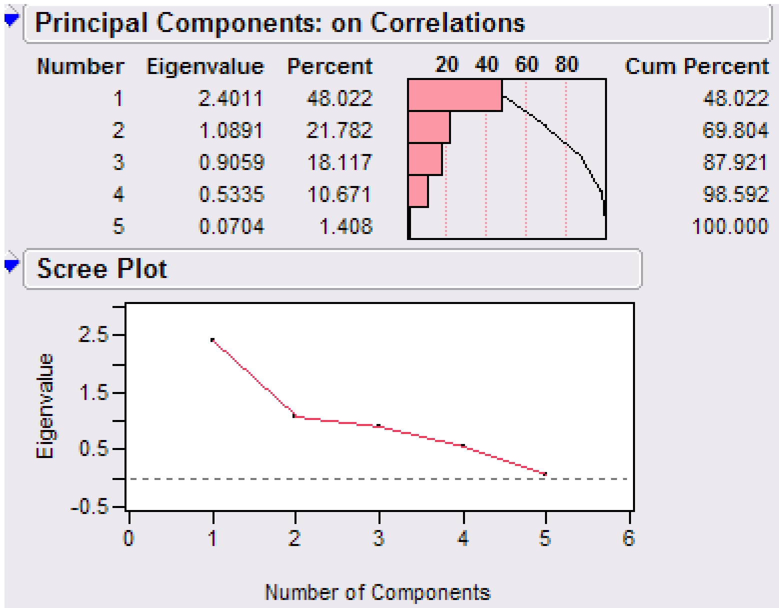


Figure 4.2 Eigenvalue Results and Scree Plot for AFIS Card Sliver Data

PCA was also done on final yarn testing data. The yarn is tested for evenness and imperfections on the Uster testing machine, and four parameters are tested at various levels for a total of twelve different testing variables. The parameters of interest are CV (coefficient of variation), thin and thick places, and neps which are clusters of fibers in the yarn. The twelve tested variables were highly correlated, and the principal component analysis showed that two components accounted for 77% of the variation of the original twelve variables. From the scree plot below, there is a clear break in the graph at three components indicating that the two components above this break were the most significant. These two components could be used in future analyses rather than the original twelve variables. See Appendix I for the rotated factor pattern detailing the component membership.

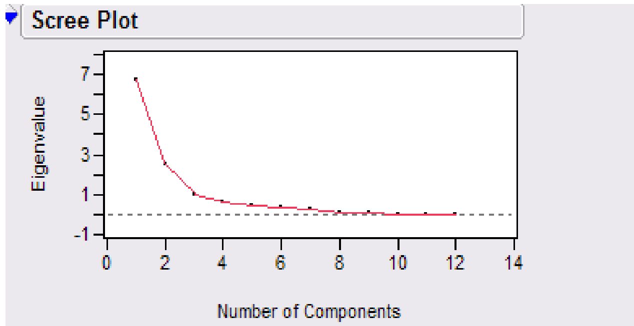


Figure 4.3 Scree Plot of Final Yarn Testing Data

Finally, several iterations of principal component analysis were done on the eight cotton variables (micronaire, length, strength, +b, Rd, trash area, uniformity index, and classer leaf) to address the major research question of this data mining tool. Looking at all observations from all crop years, four principal components were identified that accounted for 68% of the variation. Based on the rotated factor pattern above in Figure 4.1, the following table shows the four principal components, their qualitative name, and the original variables that make up each component.

Table 4.1 Principal Component Membership of Cotton Classification Variables

Principal Component #	**Name**	**Variables**
1	"Trash"	classer leaf, trash area
2	"Critical Fiber Properties"	length, strength, uniformity index
3	"Color Grade"	+b, Rd
4	"Fineness"	micronaire

PCA was also done for each crop year individually and similar results were obtained with the number of components ranging from three to five and the cumulative variation represented by the eigenvalues between 64% and 76% (see Table 4.2 below). The resulting principal components from these analyses could be used in future regression and recursive partitioning analyses as uncorrelated input variables. In addition the reduced set of variables is easier to interpret and visualize at one time than the many original variables.

Table 4.2 PCA Results of Cotton Classification Data by Crop Year

Crop Year	# Components	% Variation	PC #1	PC #2	PC #3	PC #4	PC #5
2003	4	70.3	+b & Rd	Classer Leaf & Trash	Length & Strength & Uniformity	Mic	
2004	4	66.5	Classer Leaf & Trash	Length & Strength	+b & Uniformity	Mic & Rd	
2005	5	76.0	Length & Mic & Uniformity	+b & Rd	Trash	Strength	Classer Leaf
2006	4	72.8	Classer Leaf & Trash & Rd	Length & Strength & Uniformity	+b	Mic	
2007	3	64.4	Classer Leaf & Length & Trash	Mic & Strength & Uniformity	+b & Rd		

4.2 Cluster Analysis

Cluster analysis is a method to find collections of observations with similar values. The researcher investigated different clusters in the data by analyzing average cotton classification properties by gin, county, state, and classing office. The cluster analysis was then expanded to investigate natural groupings of counties and states and their effect on the number of ends down. If the process owners were able to determine clusters of counties or states with “good” cotton that lead to increased spinning efficiency, they could alter their buying practices to purchase only higher quality cotton. Both hierarchical and k-means clustering algorithms were used.

Hierarchical clustering was first used since the researcher had no prior knowledge about the number of groups that could be present in the data, and because this clustering method is more efficient with smaller datasets (less than several thousand rows). The data was averaged by gin, county, state, and classing office for this analysis. The default Ward minimum variance method was used to calculate the distances between clusters, and a label variable such as gin or county was selected to help identify the clusters. An estimate of the number of clusters was obtained by analyzing the dendrogram of the distances between the clusters and the scree plot. Cluster membership was studied to determine major similarities and differences within and among the clusters, and the clusters were labeled if possible. Three-dimensional scatterplots were used to look at the spread of the clusters across three properties.

First, cluster analysis was done to look at natural groupings of the cotton classification properties by gin name. Approximately 330 different gins were present in the data from ten different states. From the dendrogram and scree plot, the researcher identified four clusters of gins in the data (see Appendix J). Due to the large number of different gins, no observable trend or identifier was seen among the clusters that enabled them to be labeled. An ANOVA was done of the eight cotton properties by cluster to verify the clusters were statistically different. As an example, the ANOVA results below show that cluster three has a significantly lower amount of trash and highlights a group of gins to purchase cotton from with a desirable property.

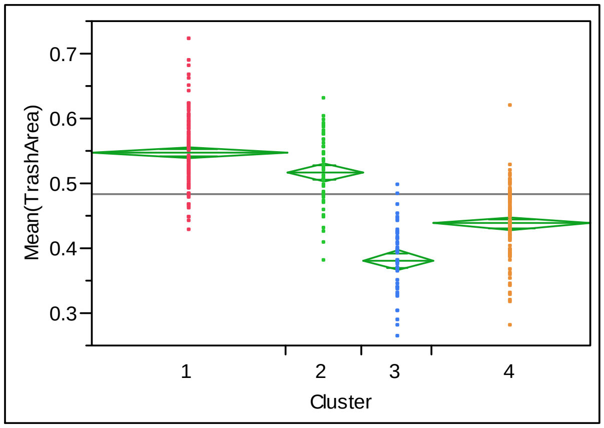


Figure 4.4 ANOVA Results of Trash Area by Gin Cluster

Additionally, a 3D scatterplot was use to further explore the gin clusters. This interactive plot enables the user to select three factors and visualize the contour ellipsoids of the clusters. In the figure below, length, strength, and uniformity were selected as an example and the four cluster ellipsoids are displayed (Legend: 1 = red, 2 = blue, 3 = green, 4 = orange). The user can rotate the plot and visually determine the cluster with the required properties.

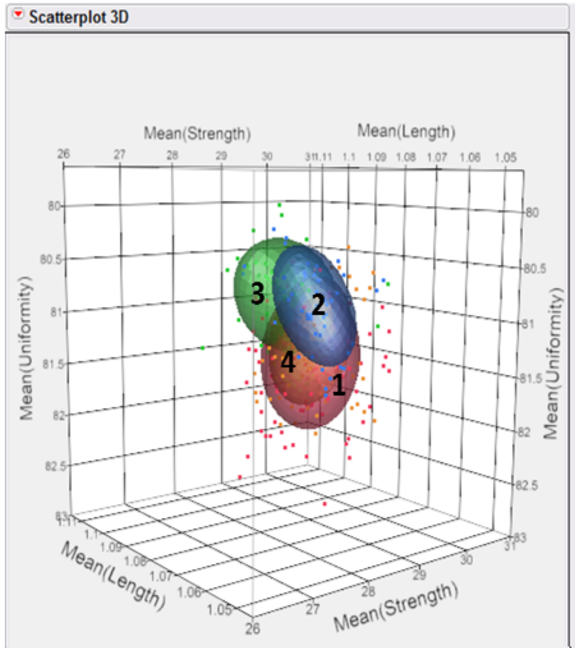


Figure 4.5 3D Scatterplot of Strength, Length, and Uniformity by Gin Clusters

Next, cluster analysis was done to explore the 140 different counties where the cotton was ginned. It was discovered that several counties had the same name, so the county names were recoded as "county_state" to eliminate confusion. Like the results from the gin cluster analysis, four clusters of counties were found but no descriptive cluster names could be identified (see Appendix K for detailed results). To further explore the natural groupings of counties in the data and their relationship to the number of ends down, hierarchical analysis was done by "county_state" for all cotton classification properties (mic, length, strength, plus b, Rd, trash area, classer leaf, uniformity), the environmental conditions (inside and outside spinning temperature and humidity variables), and the number of ends down. Five clusters emerged that were statistically different, and cluster one had the highest number of ends down. The following table shows the mean value by cluster for each of the variables, and characteristics of each cluster can be determined. This table could be very informative to a yarn manufacturer considering which county to buy cotton from.

Table 4.3 Cluster Results by County_State with Mean Values of Variables

Cluster	1	2	3	4	5
# Observations	20	33	36	44	7
Mean Ends Down	293.06	279.10	265.24	245.46	242.43
Mean Mic	4.48	4.49	4.49	4.44	4.52
Mean Length	1.08	1.08	1.07	1.07	1.08
Mean Strength	28.92	28.89	28.19	28.72	28.59
Mean Plus B	7.80	7.99	8.12	8.19	8.12
Mean Rd	75.96	74.64	75.77	75.23	76.13
Mean Trash Area	0.49	0.51	0.43	0.53	0.45
Mean Uniformity	80.95	81.07	80.88	81.36	81.29
Mean Classer Leaf	3.54	3.49	3.41	3.71	3.33
Mean Spinning Temp	77.14	78.86	78.56	78.54	78.73
Mean Spinning Humidity	53.11	50.99	54.06	52.64	51.41
Mean Outside Temp	52.77	67.70	64.79	64.44	56.77
Mean Outside Humidity	54.10	50.46	54.37	51.83	69.11

The ten ginning states were then investigated, and three clusters emerged from the dendrogram (see Figure 4.6 below). The first cluster was made up of AL, AR, MS, TN, and MO, the second cluster contained FL, GA, and SC, and the third cluster made up of LA and NC. These three clusters roughly correspond to geographical regions in the United States and could tentatively be labeled as such. The first cluster was identified as the Mid-South, the second cluster was the South, and the third cluster was a combination of both regions. Finally, the seven classing offices were analyzed and two clusters were identified. The first cluster contained Birmingham AL, Dumas AR, Memphis TN, Memphis/Dumas, and Rayville LA, and the second cluster was Florence SC and Macon GA. Since the classification properties were averaged by gin, county, state and classing office for these aggregate cluster analyses, the sample size was therefore dependant on the number of gins, counties, states, and classing offices for each. As a result the sample size was very small for the state (n=10) and classing office (n=7) analyses. The clustering algorithm was not as efficient at finding statistically different clusters, and the results are only exploratory.

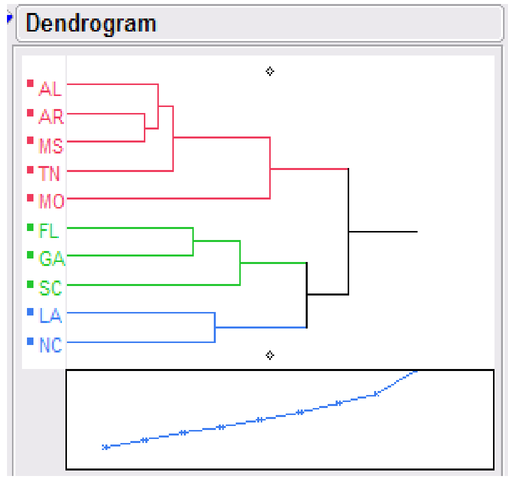


Figure 4.6 Hierarchical Cluster Dendrogram and Scree Plot by State

The project sponsor hypothesized that the state where the cotton was ginned ultimately affected the number of ends down in the spinning process. To further evaluate this effect, an overall analysis of the ends down by state was conducted for the entire study. From the figure below, SC, FL, and GA were the three ginning states that resulted in the highest number of ends down. Interestingly enough, these three states belong to the second cluster of states identified above as the South cluster. The third cluster (LA and NC) follows with medium ends down levels, and the first cluster (AL, AR, MS, TN, MO) has the lowest ends down levels according to the chart below. There are many potential factors and growing conditions of the states that affect the quality of the grown cotton and eventually the spinning productivity in the manufacturing plant.

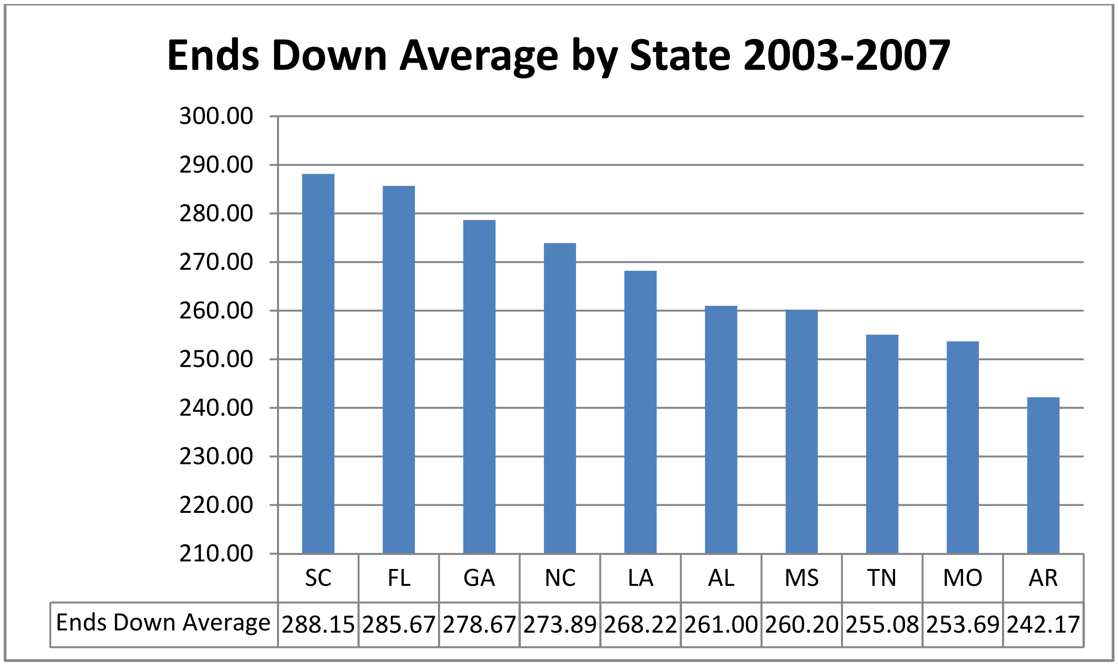


Figure 4.7 Ends Down Average by State for 2003-2007

After the initial hierarchical clustering results the researcher tried the k-means clustering algorithm to further investigate the natural groupings in the entire cotton classification dataset as the k-means method can be used on much larger tables than the hierarchical clustering algorithm. Using a cluster number estimate from the hierarchical clustering results (based on the gin and county analyses), the researcher input four clusters into the k-means algorithm to calculate the cluster centroids to begin the iterative analysis. In the software each iteration can be stepped through or the entire analysis run. The analysis output gives a summary of the number of observations in each cluster, the maximum distance to another cluster, the number of iterations the analysis took, and the mean and standard deviation of each cluster.

The k-means clustering algorithm was run with the eight classification properties, environmental variables (temperature and humidity), and the number of ends down to find the four clusters with the most similar observations. The analysis took 136 steps to iteratively assign observations to clusters, and the following summary output was obtained about each cluster (see Figure 4.8). Cluster 1 has the highest number of ends down and cluster 4 has the lowest number of ends down so differences in the variable levels between these two clusters can be explored. The parallel coordinate plot was also available to visualize the spread of the observations across the variables in each cluster (see Figure 4.9 below). From these plots it appears that certain variables consistently have a large range of values, including micronaire and plus B, and cluster 4 has large variations across all the variables which could have an averaging effect that leads to the lowest number of ends down.

Cluster Summary

Step	Criterion
136	0

Cluster	Count	Max Dist
1	81071	7.78956396
2	106432	6.20027465
3	101965	7.87925052
4	86779	8.07439241

Cluster Means

Cluster	Mic	Length	Strength	PlusB	Rd	TrashArea	Uniformity	ClasserLeaf	Ends Down (2 day lag)	Spinning Temp	Spinning Humidity	A2OAT	A2OAH
1	4.44087282	1.06551702	28.2111174	8.18024078	74.9317759	0.53948021	80.6682143	3.88365753	293.336595	79.5006137	54.2459405	76.2689177	51.7159012
2	4.4658317	1.07585162	29.0753307	8.31536944	75.3447337	0.55906118	81.199253	3.9522606	244.479811	77.224042	54.7636172	52.460256	58.2716687
3	4.53487275	1.06155053	27.9473927	8.04586672	76.5931251	0.35256902	80.856389	3.04083754	262.635865	78.1914184	53.1391728	59.9689225	56.8119883
4	4.47078786	1.09282154	29.1316989	8.02164809	74.2956049	0.56295763	81.7180608	3.64120352	241.319912	79.3155947	49.0359048	71.6901726	43.3101527

Cluster Standard Deviations

Cluster	Mic	Length	Strength	PlusB	Rd	TrashArea	Uniformity	ClasserLeaf	Ends Down (2 day lag)	Spinning Temp	Spinning Humidity	A2OAT	A2OAH
1	0.29102194	0.02016085	1.32021156	0.4444124	1.80453499	0.13889579	0.78153399	0.34092325	52.2539273	2.01759418	2.92487273	7.28372735	11.3667433
2	0.28597821	0.01982423	1.50290517	0.46961692	1.93623586	0.13869102	0.89146767	0.21374348	49.6451642	1.03686393	3.11175201	11.4362734	15.9737773
3	0.27120623	0.01878202	1.24401568	0.49846092	1.84501195	0.10070948	0.92074059	0.26032393	54.8271244	1.16253313	2.9006788	13.9884107	13.7388918
4	0.260594	0.01934214	1.46300552	0.42292087	1.70869259	0.16064261	0.91058196	0.51195702	51.6175338	1.25757094	2.72076015	10.9894822	11.2633685

Figure 4.8 K-Means Results for Four Clusters

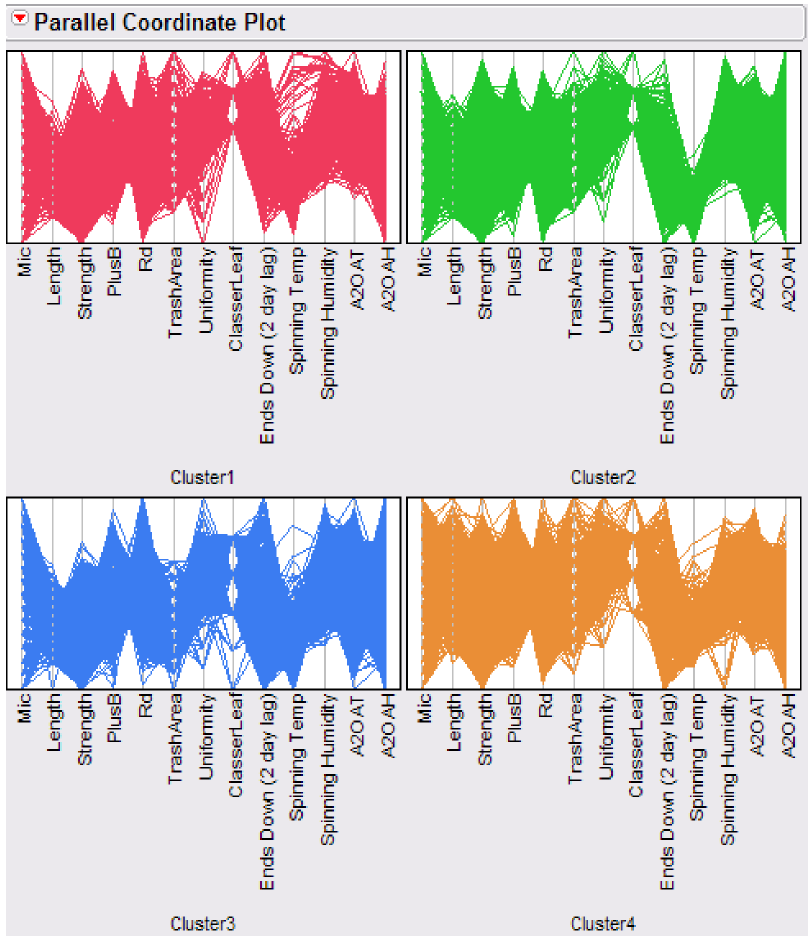


Figure 4.9 Parallel Coordinate Plot of K-Means Clustering Results

A clustering option available is to use an ordering variable that sorts the data by a particular column. For example the first principal component could be used to order the data from small to large values across all variables. This ordering can lead to more efficient clustering when dealing with large datasets that are randomly organized.(*JMP User's Guide*, 2007) . This option was tried with the k-means method and four clusters with the first principal component selected as the ordering variable. However, the same results were obtained whether the ordering variable was used or not, and the number of iterations and analysis time was not improved with the addition of the ordering variable. In this case, the ordering option did not lead to more efficient clustering.

4.3 Recursive Partitioning

Recursive partitioning is an exploratory tool used to find levels of the input factors that best predict the response. The researcher was interested in determining what raw material properties were affecting ends down. The response is the number of ends down per 1000 rotor hours, and this variable was analyzed as both a categorical and continuous factor. The independent factors investigated were micronaire, length, strength, +b, Rd, trash area, uniformity index, and classer leaf grade. The effects of cotton variety and environmental conditions including temperature and humidity were also considered. To assess the fit of the model the ROC curve, lift curve, and R-squared value were consulted.

Before analyzing the partition results, it is helpful to recall the definitions of the variables and the general goal for each variable to understand any relationships to spinning efficiency that may emerge (see table below). For instance if the partition analysis yields a leaf result of "good" spinning outcomes for Trash Area< .20, this is concurrent with the general goal of decreasing the Trash Area variable.

Table 4.4 Variable Definitions and Improvement Goal

Variable	**Definition**	**Improvement Goal**
Micronaire	measure of fiber fineness and maturity	Target
Length	upper half mean length	Increase
Uniformity Index	ratio of upper half mean length to mean	Increase
Strength	g/Tex	Increase
Plus b	degree of color pigmentation	Decrease
Rd	% reflectance or brightness	Increase
Trash Area	percentage of the surface area occupied by trash particles is calculated	Decrease
Classer Leaf	visual estimate of the amount of cotton plant leaf particles in the cotton, 1-7 scale	Decrease

To begin the researcher looked at aggregate data by laydown (90 cotton bales per laydown) to determine overall trends and important factors with the categorical ends down response of “good” or “bad”. The average of each of the eight fiber properties and the spinning outcome for each laydown was investigated, and uniformity found to have the biggest effect on “bad” spinning outcomes. The categorical output below shows the hierarchical partitions among the variables to divide the “good” and “bad” outcomes, with uniformity being the first split. In fact if the uniformity level was less than 80.72, this would result in extreme levels of ends down 52% of the time. After twelve total splits (R-squared .222) it was found that 76% of the time a “bad” spinning outcome was obtained with the following factor levels: *Mean(Uniformity)<80.76 & Mean(Mic)<4.54 & Mean(ClasserLeaf)>=3.1 & Mean(Length)<1.05 & Mean(PlusB)<8.35 & Mean(TrashArea)>=0.41*. These levels of the cotton properties represent a large proportion of “bad” ends down, and this information could be extremely useful when considering different cotton suppliers from which to purchase raw material from.

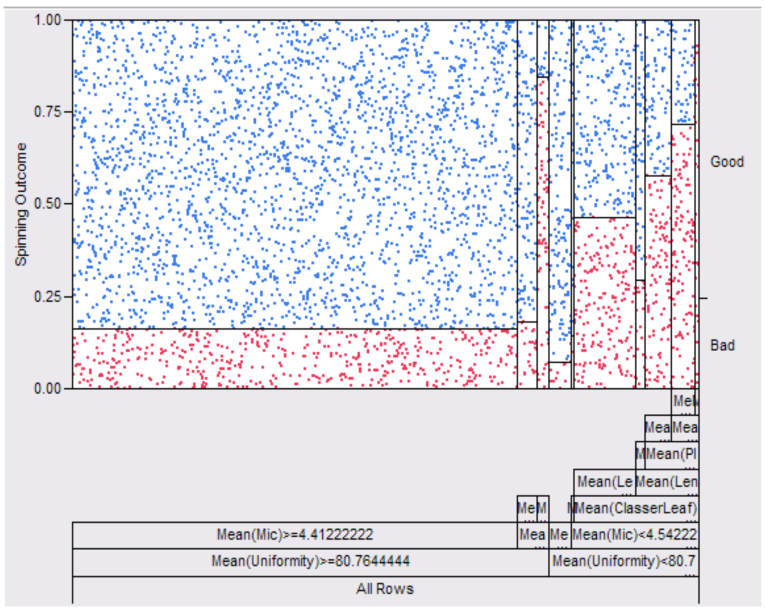


Figure 4.10 Recursive Partitioning Output of Variable Splits

With the addition of the average spinning temperature and humidity variables in the manufacturing plant to the laydown partition analysis, humidity emerged as an influential factor and lower humidities appeared to lead to “bad” spinning outcomes. The following leaf result is associated with 95% “bad” ends down: *Mean(Uniformity)<80.76 & Mean(Mic)<4.54 & Mean Spinning Humidity<53.64 & Mean(PlusB)>=8.14.* These results are for nine partition splits with an R-squared value of .259. Although this is a low R-squared value compared to an R-squared value from a regression analysis for example, recursive partitioning is an exploratory tool to look at optimal splits among the variables and not used for an exact prediction model. R-squared is a good quantitative measure to compare different fitted models, but it will not be the only determination of a classification tree and its usefulness to the research question.

The ROC curve below depicts the number of true positives (y-axis) plotted against the number of false positives (x-axis) to show the sorting efficiency of the model. If the model is a perfect fit, the ROC curve would move all the way to the top and then all the way to the right with an area of 1.0 under the curve. This partition model with nine splits is a good fitting model with the curve in the upper left quadrant and an area of .81 underneath the curve.

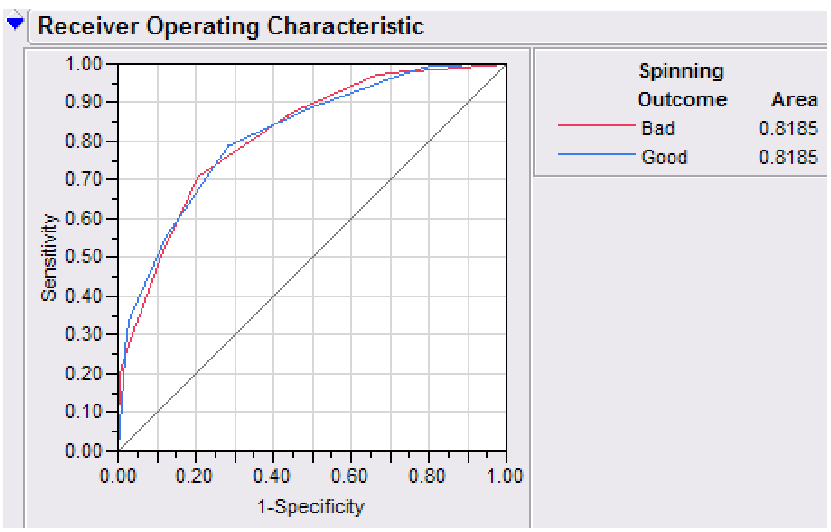


Figure 4.11 ROC Curve of Recursive Partitioning Results by Laydown

The lift curve for the same classification tree is below and shows similar results to the ROC curve. This chart is an assessment of the "lift" the model gives over no predictive model with different portions of the population and ordered fitted probabilities on the x axis. The greater the area between the baseline (y=1) and the lift curve, the better the model is. From the curve, this classification tree better models "bad" versus "good" outcomes for the top-rated 20% of fitted probabilities of the population and has a lift of nearly four times that of no fitted model.

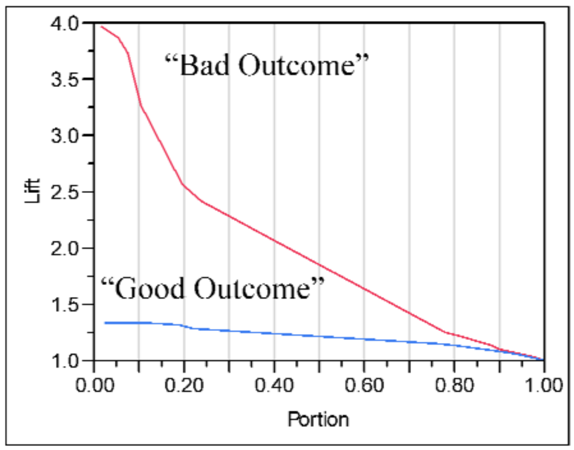


Figure 4.12 Lift Curve of Recursive Partitioning Results by Laydown

Specifically investigating the effect of temperature for the laydown analysis, a "bad" spinning outcome results 100% of the time when the spinning temperature in the plant is greater than 85 degrees Fahrenheit. On the other hand, less than 2% "bad" outcomes are achieved (98% "good") when the spinning temperature is less than 80 degrees Fahrenheit and the spinning humidity is greater than 58%. A relative humidity of 60% if typically preferred in spinning plants (Artzt P. et al., 1990), and this industry standard can be directly applied to these partition results where nearly all "good" outcomes occur when the humidity is greater than 58%. When the outside air temperature and humidity were considered, 90% "bad" spinning outcomes result when the outside temperature is greater than 85 degrees Fahrenheit and the outside humidity is greater than 46%. The results are similar for both the outside environmental conditions and the conditions inside the manufacturing plant with higher temperatures and lower humidity resulting in more ends down.

Next, recursive partitioning was used by crop year to investigate the effects of the cotton classification properties, cotton variety, and environmental conditions on the spinning outcome, ends down. Many different iterations were performed and it's difficult to obtain one overall and definitive result because of the exploratory nature of this analysis tool. Recursive partitioning is somewhat subjective in the number of splits that can be taken and the variables to split on, so different stopping criteria and a standard splitting methodology should be used. Thirteen input variables were investigated per cotton bale: micronaire, length, strength, +b, Rd, trash area, uniformity index, classer leaf grade, spinning temp, spinning humidity, outside temp, outside humidity, and %DP 555 (average percentage by state). The ends down variable was analyzed as both a categorical and continuous response in the partition.

The categorical response was first considered with "good" and "bad" spinning outcomes (ends down by crop year coded as "bad" if greater than the 75^{th} percentile). The researcher chose a standard measure of twenty splits to enable comparisons across crop years (see Table 4.5 below for results). Note that the 2007 crop year data was only split eleven times because twelve splits completely fit the data with an R-squared value of 1.0. This general splitting method showed some similarities among the crop years. Considering the variable with the largest splitting statistic (higher G^2 likelihood-ratio chi-square statistics correspond to a more optimum split), four out of the five crop years had temperature as the most important factor affecting the spinning outcome. The outside temperature is highly correlated with the spinning temperature inside the manufacturing plant so a general statement could be made that nearly all bad spinning outcomes occur when the temperature exceeds 79 degrees on average. It was surprising that the temperature and humidity variables were so influential and accounted for the majority of the splits in the classification tree. When this analysis was previously run without the environmental variables, the factors uniformity, Rd, +b, and micronaire emerged as being important but they did not surface in this particular analysis. The areas under the ROC curves are considerably high with relatively significant R-squared values that indicate good fitting models. One specific leaf result was chosen per crop year to highlight an example of an interesting result.

Table 4.5 Recursive Partitioning Results by Crop Year with Categorical Response for 20 Splits

Crop Year	**Variable with Largest Splitting Statistic**	**R-squared**	**ROC Area**	**Significant Leaf Result**	**Probability "Bad" Spinning Outcome**
2003	Outside Temp	0.609	0.997	OutsideTemp>=68.07 & SpinningTemp>=79.23 & OutsideHumidity<34.79	1
2004	Outside Temp	0.498	0.926	OutsideTemp<79.09 & Rd<76 & SpinningHumid<50.42 & SpinningTemp<77.50	0.92
2005	Spinning Temp	0.793	0.985	OutsideHumidity>=45.11 & SpinningTemp>=80.27	1
2006	Spinning Humidity	0.453	0.899	OutsideTemp>=79.41 & SpinningHumid<51.78	0.90
2007*	Spinning Temp	0.960	0.999	SpinningTemp>=77.15 & SpinningHumid>=52.19 & OutsideHumidity<40.08	1

*Results for only 11 splits

The continuous ends down response was then considered in the partition analysis (see Table 4.6 below). All datasets were analyzed with twenty splits and similar results to the categorical response analysis were obtained. Temperature again emerged as an important factor, but several other variables were also significant including %DP 555, Rd, and humidity. The sum of squares is calculated for the spitting statistic in the continuous response method, rather than the G^2 calculated for the categorical response method. It should be noted that the probability of the response and the ROC calculations can only be done for categorical responses, and thus they are not included in the table below.

Table 4.6 Recursive Partitioning Results by Crop Year with Continuous Response for 20 Splits

Crop Year	Variable with Largest Splitting Statistic	R-squared
2003	% DP 555	0.604
2004	Rd	0.463
2005	Outside Humidity	0.809
2006	Outside Temp	0.338
2007	Outside Temp	0.625

To complement this analysis, pairwise correlations were calculated for all variables by crop year to compare these associations to those from the partition analysis. The following table highlights the most significant positive and negative correlations among all of the factors with the ends down continuous variable. These results highlight the cotton variety and humidity variables as being more important than temperature. The color variables (+b and Rd) and uniformity also emerge as being moderately important which is what was observed in the initial analyses without the environmental variables. Note that all the relationships reported below are significant for the crop year dataset but the correlation values are still relatively low (<.45).

Table 4.7 Pairwise Correlation Results for the Number of Ends Down with the Independent Factors

Crop Year	Largest Positive Factor	Correlation Value	Largest Negative Factor	Correlation value	2nd Largest Positive Factor	2nd Largest Negative Factor
2003	%DP555	0.45	Rd	-0.25	Spinning Humidity	Uniformity
2004	%DP555	0.24	Rd	-0.41	Length	Plus B
2005	Outside Humidity	0.47	Spinning Humidity	-0.31	Outside Temp	Plus B
2006	%DP555	0.14	Spinning Humidity	-0.27	Outside Temp	Uniformity
2007	%DP555	0.33	Outside Humidity	-0.19	Rd	Strength

After the analysis of the aggregate laydown and crop year data, the researcher used the principal component tool to identify a smaller set of variables to use as input factors in the partition analysis. The researcher hoped that these uncorrelated variables would lead to a better fitting model and ultimately could be used in a regression analysis. The overall data averaged by laydown was used for analysis. Three principal components were found that accounted for about 71% of the variation of the eight cotton variables. The first component was made up of classer leaf, strength, and trash area, the second component was made up of length, Rd, uniformity, and the third component was made up of mic and +b. These three components were used as factors in a partition model with the spinning outcome as the categorical response. A model was fit with three splits that had an R-squared value of .026 and an area underneath the ROC curve of .54. The following leaf result is associated with 78% “bad” ends down: *Prin1< -3.16 & Prin3>= -0.46 & Prin2>= -0.32*. The results with the principal components as input factors were considerably worse than the partition results with all eight variables in terms of the R-squared value and ROC area under the curve. The partition results were also harder to interpret since the three transformed components accounted for the eight original fiber variables and could not be directly related to one specific factor. The method of using principal components as input factors to an analysis may be more applicable in predictive regression studies where only general information is needed about the transformed components and their effects. However in this study, the surrogate variables led to a worse-fitting model that gave less information than using all the original variables.

For the actual recursive partitioning analysis, the researcher used several different options in the software to explore the data. The “Lock Columns” option was frequently used to interactively lock columns so they weren’t considered for splitting. Therefore, when building a classification tree more variables and their splitting levels could be explored rather than continuously splitting on the same factor with different levels. In addition, the “Show Split Probabilities” option was used to quantify the probabilities of a “good” and “bad” outcome at each split when fitting a categorical response. The Split History table was a helpful visualization to see the effect of the number of splits on the R-squared value. This table served as a guide to determine when overfitting could occur when additional splits don’t increase the R-squared value. Finally, the Columns Contributions table below showed how each input factor contributed to the overall fit of the model in terms of the number of splits and the split statistic (G^2 for categorical responses and sum of squares for continuous responses).

Column Contributions

Term	Number of Splits	G^2
Total	8	704.318647
Mean(Uniformity)	1	315.164894
Mean(Mic)	2	211.81578
Mean(Rd)	1	99.4972648
Mean(PlusB)	1	30.3099912
Mean(Length)	1	20.0203707
Mean(ClasserLeaf)	1	14.1908215
Mean(TrashArea)	1	13.3195257
Mean(Strength)	0	0

Figure 4.13 Column Contributions of Input Factors

In addition to gaining insight about the effect of different factors on spinning ends down, the partition results can be used to plan for future experimental designs. Since recursive partitioning is only an exploratory tool that shows possible associations between selected variables and the spinning outcome, a designed experiment is necessary to determine causation between the factors. The important variables from the classification tree can be selected for study, and the splitting levels of these factors can be used to set up the levels for a design of experiments.

4.4 Discriminant Analysis

Discriminant analysis is a method to predict a classification function based on continuous variable measurements. The question of interest for this study is what raw material properties are affecting ends down. In this case, the spinning outcome is classified as “good” or “bad” based on the number of ends down per 1000 rotor hours (“bad” if greater than the 75th percentile), and the discriminating variables include cotton classification properties, environmental conditions, and the percentage of a specific cotton variety present.

Prior to running the discriminant analysis, the assumptions of normality, independence, and constant variance within each group were checked. Each observation was from a different cotton bale and therefore an independent measure. From the SAS® procedures, quantile-quantile plots, and multivariate statistics of the data by crop year, an approximate normal distribution was assumed. The check on homogeneity of the variance-covariance matrices showed that there were significant differences among the groups and the within covariance matrix should be used in a quadratic discriminant function. Because the primary purpose of using this analysis was to explore which variables predict “good” or “bad” levels of ends down and it has been shown that this method is robust for large sample sizes, the assumptions were verified and quadratic discriminant analysis performed (Lachenbruch, 1975).

First, aggregate data by laydown (90 bales of cotton) for all crop years was investigated to look for overall trends. The eight cotton classification properties (+b, Rd, uniformity, strength, length, micronaire, trash area, classer leaf) were input as the continuous covariate factors and the spinning outcome (“good” or “bad”) was the classification variable. Within JMP® a Mahalanobis distance formula is used to calculate the distance from each observation to the centroid (multivariate mean) of both the “good” and “bad” outcome groups. The discriminant method then classifies each observation into the closest group. Using quadratic discriminant analysis, a misclassification error of 34.9% was obtained. The confusion matrix below shows the number of correctly classified observations on the diagonal (2808 observations), and the observations on the off-diagonal were misclassified (1504 observations).

Table 4.8 Confusion Matrix

	Bad	**Good**
Bad	842	228
Good	1276	1966

Further results assessed the fit of the model and explored a stepwise analysis. The figure below shows an area of .7640 underneath the ROC curve which indicates that this discriminant model was halfway between a model that only detects random variation (area=.5) and a perfectly accurate model (area=1). A stepwise analysis was done to rank the discriminatory power of the factors in predicting the spinning outcome. The following stepwise results were obtained in order from most to least significant: uniformity, micronaire, +b, strength, Rd, trash area, classer leaf, length. The uniformity index was found to be the most important factor which was equivalent to the findings from the initial recursive partitioning analysis.

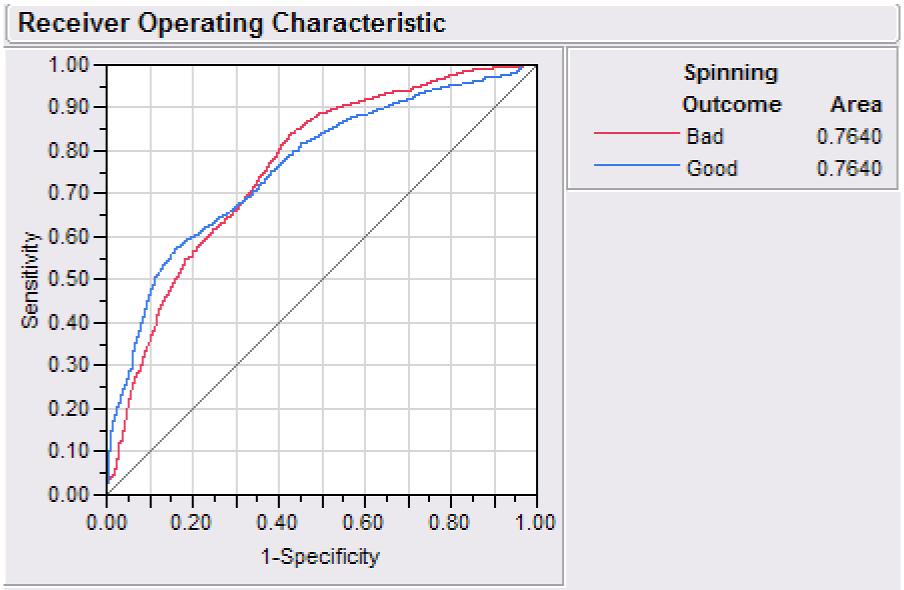


Figure 4.14 ROC Curve

Using the same aggregate analysis with the addition of the spinning temperature and humidity as predictor variables, the misclassification error decreased to 28.5%. The temperature and humidity variables appeared to be significant with the humidity variable having slightly more discriminatory power than the temperature variable. The analysis was taken one step further with the addition of the outside temperature and humidity variables, and the misclassification error decreased to 25.8%. The stepwise results for all the cotton classification and temperature factors were: uniformity, micronaire, spinning humidity, spinning temperature, outside humidity, strength, trash, outside temperature, +b, length, Rd, classer leaf.

After the initial analysis of the aggregate data by laydown, specific analyses were performed of cotton bales by crop year to further investigate any trends. Two sets of discriminant analyses were performed for each crop year to investigate the impact of different sets of predictor variables (see Table 4.9 below). First, quadratic stepwise discriminant analysis was performed with only the eight cotton properties and the spinning outcome. Both uniformity and the color variables (+b and Rd) appeared to be major drivers of “good” and “bad” ends down throughout all crop years. Then the analysis was performed with the average spinning temperature, spinning humidity, and the %DP555 by ginning state discriminant variables in addition to the cotton properties. The addition of these predictor variables decreased the misclassification error, as seen in the table below, and led to a better fitting model (see Appendix L for comparison of ROC areas across different crop years). In fact, for the 2007 dataset the area underneath the ROC curve increased from .71 to .86 indicating a more accurate model. The %DP555, humidity, and temperature were important factors in the discriminant functions and further analysis of these variables is presented later in the study.

Table 4.9 Comparison of Stepwise Quadratic Discriminant Analysis Results by Crop Year for Two Sets of Predictor Variables

Crop Year	**% Misclassified**	**Stepwise Order (8 fiber properties)**	**% Misclassified**	**Stepwise Order (8 fiber properties, spinning temperature and humidity, % DP555 by state)**
2003	38.0	Rd, Uniformity, Mic, Strength, Length, Trash, +b, Classer Leaf	25.6	Uniformity, Humidity, %DP555, Mic, Rd, Temp, Strength, Trash, Classer Leaf, +b, Length
2004	41.8	Rd, Uniformity, Strength, +b, Length, Classer Leaf, Trash, Mic	28.1	Rd, %DP555, Temp, Uniformity, Length, Humidity, +b, Strength, Classer Leaf, Trash, Mic
2005	75.2	+b, Rd, Uniformity, Strength, Length, Mic, Trash, Classer Leaf	75.2	+b, Temp, Rd, %DP555, Humidity, Uniformity, Strength, Trash, Mic, Length, Classer Leaf
2006	43.3	Uniformity, Classer Leaf, Rd, +b, Trash, Strength, Length, Mic	30.9	Humidity, %DP555, Temp, Uniformity, Classer Leaf, +b, Rd, Trash, Strength, Length, Mic
2007	35.5	Rd, +b, Trash, Strength, Uniformity, Mic, Classer Leaf, Length	27.8	%DP555, Rd, Humidity, Temp, +b, Length, Strength, Mic, Uniformity, Trash, Classer Leaf

The discriminant analysis method was mainly evaluated in JMP® but the assumptions verified and the analysis run in SAS® as well so the researcher thought it would be interesting to briefly compare the results of both software packages. Both are using slightly different calculations and approximations for the quadratic discriminant function but the results should be somewhat analogous. The table below shows the misclassification error when classifying a spinning outcome using the eight cotton fiber properties (micronaire, length, strength, +b, Rd, trash area, uniformity index, and classer leaf) and the spinning temperature and humidity variables. JMP® has a higher error rate for all crop years except for 2006 when it outperforms SAS® by 1.6%. There is a significant difference in the misclassification rates for the 2005 data which could be due to large deviations and trends in that crop year that the SAS® procedure modeled better.

Table 4.10 Comparison of JMP® and SAS® Quadratic Discriminant Analysis Misclassification Rates with Cotton Properties and Spinning Temperature and Humidity Predictor Variables

Crop Year	JMP % Misclassified	SAS % Misclassified
2003	29.1	28.3
2004	29.0	26.8
2005	75.2	50.0
2006	29.5	31.1
2007	28.7	25.1

After the researcher completed many different iterations of this analysis, the 2005 crop year data continually stood out with outlying values. Upon further investigation, it was found that all cotton from this crop year was classified as leaf grade 4. Therefore, the classer leaf factor had no discriminatory power and was in fact negatively influencing the power of the entire discriminant function. This variable was eliminated from the analysis, and the misclassification rate dropped from 75% to 21% and the ROC area rose dramatically from .14 to .81 with all the discriminant variables. See Appendix M for detailed results from the analysis with and without the classer leaf variable. This finding demonstrates how critical a complete and thorough investigation and preparation of the data is in order to avoid faulty conclusions.

4.5 Further Research

4.5.1 Cotton Variety

First the USDA cotton variety data and its effect on the overall crop quality summary data from Cotton Inc was explored. The percentage DP 555 by state was compared to the strength, uniformity, micronaire, length, trash, and classer leaf averages of the cotton ginned from that state. It was found that the percentage of Deltapine brand 555 BG/RR cotton variety was most closely correlated to strength and length uniformity index out of all factors considered including micronaire, length, trash, and classer leaf grade (see Figure 4.15 below with the pairwise correlations). The strength of the raw material is associated with an increase in the DP 555 cotton variety whereas the length uniformity of the raw material decreases with larger amounts of DP 555 cotton variety. Micronaire, length, trash percentage, and the classer leaf grade were not affected by the amount of DP 555 cotton variety present.

Pairwise Correlations

Variable	by Variable	Correlation	Count	Signif Prob
Len (32/inch)	Strength Average	0.8319	6	0.0400*
Strength Average	% DP 555 Average	0.7912	6	0.0609
Len (32/inch)	% DP 555 Average	0.3948	6	0.4385
Leaf Grade Index	Trash %	0.3492	6	0.4975
Leaf Grade Index	Mic	0.3339	6	0.5177
Leaf Grade Index	% DP 555 Average	0.3285	6	0.5249
Len (32/inch)	LUI Average	0.2138	6	0.6841
Trash %	Len (32/inch)	0.2123	6	0.6863
Trash %	LUI Average	0.2080	6	0.6926
Mic	LUI Average	0.0311	6	0.9534
Leaf Grade Index	Strength Average	-0.1025	6	0.8468
Trash %	Mic	-0.1195	6	0.8216
Trash %	Strength Average	-0.1334	6	0.8011
Strength Average	LUI Average	-0.2566	6	0.6236
Trash %	% DP 555 Average	-0.3205	6	0.5357
Leaf Grade Index	Len (32/inch)	-0.3707	6	0.4694
Mic	% DP 555 Average	-0.4096	6	0.4199
Mic	Strength Average	-0.6059	6	0.2024
Leaf Grade Index	LUI Average	-0.6899	6	0.1293
LUI Average	% DP 555 Average	-0.7247	6	0.1033
Len (32/inch)	Mic	-0.7771	6	0.0690

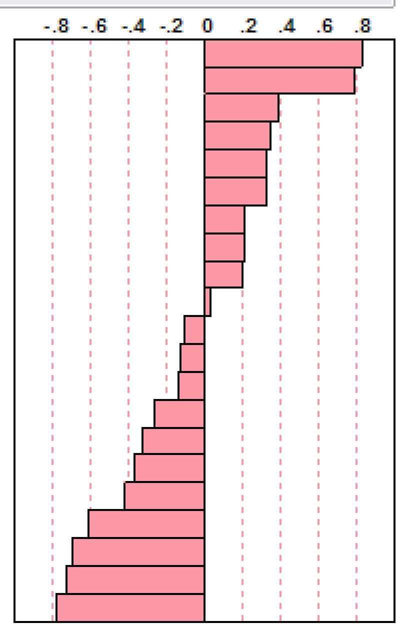


Figure 4.15 Pairwise Correlations of % DP 555 and Cotton Inc. Variables

After the initial analysis of the specific cotton variety and its effect on the overall US cotton quality, the percentage of DP 555 cotton variety data by state was related to the cotton material and productivity (ends down) data for the specific case study. Since the cotton variety data was only available on a yearly basis, the researcher began exploring its effect with the cotton classification properties and ends down averages by year by state. Since not all states grew this specific variety each year between 2003 and 2007, there was only data for around eight states per year resulting in a total sample size of 41. Although a smaller sample than prior analyses, noteworthy information was discovered as discussed below.

The relationship between the percentage of DP 555 and the cotton classification properties was first examined for any significant relationships and for comparison to the results with the Cotton Inc. crop quality data (see Appendix N for pairwise correlations). The only significant relationship with cotton variety was seen with uniformity by the percentage of DP 555 which was inversely related (as the percentage of DP 555 increases the cotton length uniformity index decreases), whereas the Cotton Inc. analysis showed significant relationships between DP 555 and both strength and uniformity. The effect of DP 555 on strength was not detected for the case study. In fact, the smallest relationship out of all the variables was seen with the strength and DP 555 factors.

Next, associations among the percentage of DP 555 and its effect on the cotton classification variables and the number of ends down were assessed (see Appendix O for pairwise correlations). Significant relationships were observed between the following in ranked order: ends down by %DP 555 (positive), micronaire by %DP 555 (inverse), uniformity by %DP 555 (inverse), and +b by %DP 555 (positive). Ends down and % DP 555 had the largest observed correlation and the largest amount of variation (R-squared) was explained by a linear model between these two variables with ends down as the dependent variable and % DP 555 as the independent variable. As the percentage of DP 555 increases, the number of ends down increases and this fact could be crucial for buyers that are trying to increase spinning efficiency. One of the smallest observed correlations was between ends down and strength. This is an interesting result because strength was not closely related to yarn quality nor yarn productivity in this case study even though research has shown that low-strength cotton leads to high levels of ends down (Simpson & Murray, 1978).

In addition to looking at the relationships among the variables, a regression model was fit to investigate the effect of the cotton properties and % DP 555 on the ends down response. An initial model with all the factors explained approximately 78% of the variation, and the % DP 555 and micronaire variables were found to be significant. The Pareto plot below depicts the importance of these two variables in comparison to the other predictors.

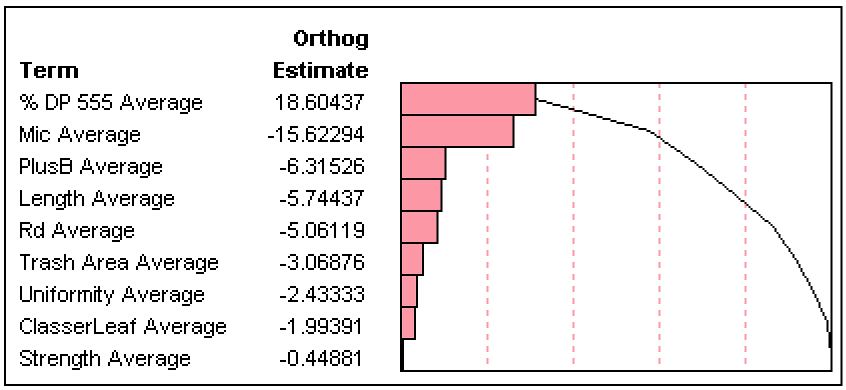


Figure 4.16 Pareto Plot of Predictor Estimates in Ends Down Regression Model

The researcher then examined the predicted levels of the factors that would lead to the minimum number of ends down, or the highest spinning efficiency. The JMP® prediction profiler below shows the nine variables (%DP 555 and eight cotton classification properties) and their settings to yield the lowest level of ends down.

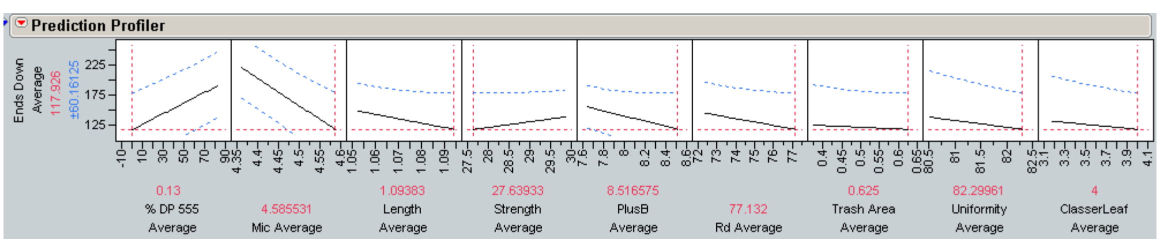


Figure 4.17 Prediction Profiler of Input Variables' Settings to Achieve Lowest Ends Down

After the initial analysis, a final model was fit with only the significant variables, % DP 555 and micronaire. The prediction equation obtained from this model to predict the number of ends down is as follows: *Y= 1723.1884 + 0.8953957(%DP555) - 331.4198(Micronaire).* The regression assumptions were verified by examining the residuals which were independent and normally distributed (see Figure 4.18 below).

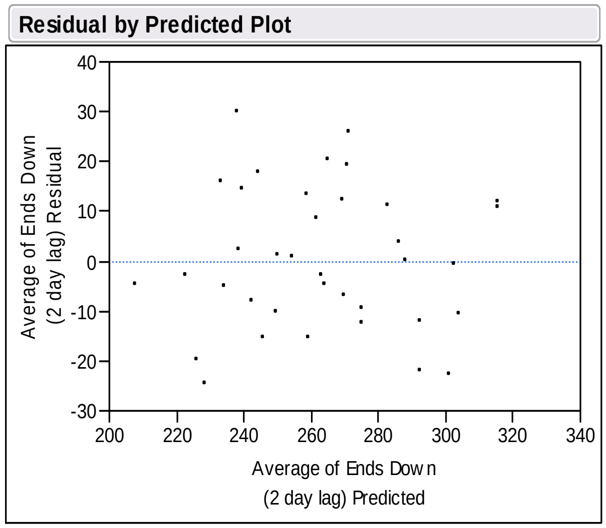


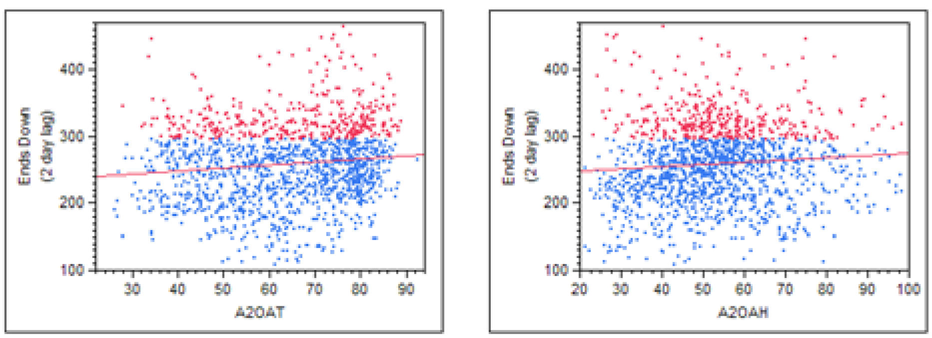
Figure 4.18 Residual Plot

4.5.2 Temperature and Humidity

From the recursive partitioning results it was discovered that higher temperatures and lower humidity generally resulted in a higher number of ends down. The temperature and humidity variables appeared to be the major drivers of ends down in recursive partitioning and were significant factors in discriminant analysis as well. The researcher was interested in a detailed exploration of these factors and their effect on ends down to eventually develop improvement plans to better control these environmental conditions.

First the relationship between the manufacturing temperature and humidity and the outside temperature and humidity was assessed. The spinning temperature and outside temperature were strongly associated with a correlation coefficient value of .65, but the humidity variables were less related with a correlation value of only .19. Temperature and humidity were inversely related (R=-.27). In a fully functioning HVAC system, the inside conditions should remain constant regardless of the outside temperature so modifications or maintenance may need to be done to the system. This analysis was done with the entire material dataset from all crop years so any trends and outliers may be skewing these overall results.

When aggregate data was analyzed for recursive partitioning and discriminant analysis, humidity emerged as slightly more important than the effect of temperature, but the opposite result was seen with the individual crop year analyses. For the overall dataset, a correlation analysis was done and both temperature and humidity were slightly associated with the number of ends down. A linear model was fit between ends down and the environmental conditions, and the resulting low R-squared values showed no significant relationship (see Figure 4.19 below). For the figure below, the red points are “bad” spinning outcomes and the blue points are “good” spinning outcomes. One can notice that the “bad” outcome observations are randomly spread across all temperature and humidity values and no observable trend is seen. However, there is a slight concentration of higher levels of ends down (“bad” outcomes) around 80 degrees and 50% humidity which corresponds to earlier results.



R-squared=.015 R-squared=.006

Figure 4.19 Linear Fit of Ends Down by Outside Temperature (A2OAT) and Outside Humidity (A2OAH)

In regards to the impact of humidity, research has shown that low moisture content leads to increased short fiber content and decreased length and strength properties in cotton. Low moisture content results in more ends down due to the higher amounts of static electricity that decrease the overall spinning efficiency (Cecil, 1998). The researcher hypothesized that a lower spinning humidity led to increased static electricity present in the spinning process and therefore more yarn breaks or ends down. From the recursive partitioning results, almost all "good" spinning outcomes were associated with higher humidities (greater than 58%) and lower temperatures (less than 80 degrees). Although it has been shown that temperature and humidity have a larger effect on yarn efficiency than yarn quality, the factors and their effects still need to be considered when looking at the overall yarn quality and performance of a spinning operation (Artzt P. et al., 1990). The environmental conditions impact the quality of the cotton material in terms of its moisture content, conditioning time, and stickiness levels for instance. As stickiness increases more particles accumulate on the spinning machines and can lead to decreased spinning productivity (Hequet, 2005).

The effect of the time of year and its resulting changes on the environmental conditions must be considered. Based on known cyclic weather patterns and weekly forecasts, the spinning manufacturer could proactively schedule the plant's spinning production to achieve minimal levels of ends down. For example, production could run only during the night in summer months to avoid high temperatures or run 24/7 during the cooler months with the aid of a humidifier machine to moderate the humidity levels. Based on this initial research of the impact of temperature and humidity on spinning efficiency, more recommendations could be investigated for tighter controls of the manufacturing plant and spinning equipment to mitigate the impact of the environmental conditions.

5.0 Case Study Discussion of Results

After completion of the case study analysis and many iterations of the data mining tools, the researcher gained a greater understanding of the complexities of large amounts of data and the necessary methods to prepare and then extract useful information. In fact, it has been shown that almost 80% of the work required for a data-rich problem is in the data preparation stage (Truxillo, 2005). By having actual data from the case study in which to apply the data mining tools, the researcher gained real world experience and understanding that cannot be learned from a book. In addition, communication between the data analyst and process owner is crucial. Throughout the analysis process there was constant communication between the researcher and the project sponsor, who was a subject matter expert. Without this ongoing consultation the analyst could potentially lead the analysis in a direction that would not be useful or logical to the overall study goals. In addition, the subject matter expert can aid in interpretation of the results and the application to his business objectives.

When discussing the results and forming conclusions, it is important to note some of the limitations of this study. There were several approximations made when first setting up the datasets that could affect the results. The researcher made a process time estimate to merge the raw material and ends down datasets in order to study the fundamental research question of the study, what factors were driving ends down. Since the majority of the data points were only available as a daily average, all data was condensed to this daily metric in order to join all the information and data tables. Therefore, considerable information was lost and not available for analysis. The researcher made an assumption about “good” and “bad” levels of ends down (“bad” if greater than 75th percentile) for each crop year to transform the continuous response to a categorical response, and this could potentially skew the analyses. Finally, since the data came from a normal operating process and not a designed experiment, all conclusions will be associative and exploratory in nature. In the future a designed experiment could specifically measure the impact of several factors and different levels on the output response, ends down. In addition, the many steps involved in the open end spinning process serve to “average” out the fiber properties so it is often difficult to see specific trends in the process data. Further analyses should examine the relationships between cleaning, blending, carding, and drawing and their specific effects on yarn productivity.

The four main data mining tools employed each provided a different perspective to the research problem at hand. The general research question was to determine what factors were driving ends down to improve productivity in a yarn spinning operation. Principal component analysis was used as a dimension reduction tool to yield a smaller set of transformed variables that could be used in future analyses. Cluster analysis was used to find natural groupings of the cotton classification observations that would lead to improved spinning productivity by examining trends by gin, county, or state. Recursive partitioning and discriminant analysis were used to investigate the relationship between cotton properties, cotton variety, atmospheric conditions, and the resulting number of ends down.

Due to the differences in and among the four data mining tools, there is no overall measure to compare the performance of the tools and to classify the "best" tool. However, certain tools provide more useful information than others and different analyses are better suited for specific data scenarios. PCA was probably the least useful of the four tools. The principal components are normally used in predictive models because they are uncorrelated and of reduced dimensionality, but a predictive model was not the major aim of this study. The case study PCA results were used in a recursive partition analysis that led to a worse-fitting model that was harder to interpret than using all the original variables. Cluster analysis provided interesting information about groups of observations with similar characteristics, specifically groups of extreme levels of ends down ("good" and "bad"). The analysis by state showed cotton from SC, FL, and GA was associated with the highest number of ends down which is congruent with an earlier study done by the process owner. The most challenging part of this data mining tool was identifying the differences among the clusters and assigning meaningful labels.

Recursive partitioning was the most useful tool applied to the case study data because of its exploratory and graphical nature. It is a straightforward method that leads to instantaneous results but care must be taken in their interpretation and issues with fitting the data too well (i.e. overfitting). Discriminant analysis provided some insights with the stepwise analysis that ranked the factors in terms of their effect on a "good" or "bad" spinning outcome. However, the statistical nature of this tool and the required assumptions often make it difficult to apply in exploratory analyses with non-normal data.

Recursive partitioning and discriminant analysis led to the most information about properties that affect ends down. Several factors surfaced as being significant to high levels of ends down including humidity, temperature, %DP 555, and uniformity in addition to micronaire and the color properties (+b and Rd). See Table 5.1 below for more detailed results. In order to increase spinning productivity, these factors need to be maintained and controlled. The manufacturer may need to alter their buying practices to buy quality cotton with set parameters from the very beginning to lead to increased spinning efficiency in the end. Likewise, adjustments may need to be made to obtain higher or lower values of certain influential variables in the laydown method. The importance of the environmental conditions highlights the need for an efficient and stable HVAC system for the manufacturing plant and as much control of the outside conditions as possible (i.e. weather resistant windows, closed doors, adequate insulation, etc.). There potentially could be other hidden factors that are strongly correlated with the important factors found in these analyses, and only further experimentation could rule out these extraneous factors and noise in the process to truly understand the significant factors.

Finally, the significance of the results must be considered with respect to the business. Do these results empower the project sponsor to make better business decisions that are based on actual data? Can the analysis tools aid the business in becoming more efficient and profitable? With any data-rich problem that is undertaken, an improvement plan must be developed to translate the data results to solutions that are implemented and maintained for continuous improvement.

Table 5.1 Summary of Results and Discussion

Data Mining Tool	**Research Question**	**Results**	**Discussion**
PCA	Can the large number of cotton properties used to classify cotton be reduced to a significant few?	3 or 4 significant components were identified from the original eight cotton classification variables (overall and by crop year analyses)	Used principal components as input factors for later partition analysis, difficult to interpret results
		Additional use of tool to reduce the dimensionality of two datasets: AFIS card sliver data and Uster yarn testing data	Dimension reduction process yields uncorrelated factors that can be used in further regression analyses, least useful tool in this study
Cluster Analysis	Are there groups of gins, counties, or classing offices that produce better cotton raw material than others?	Iterations of hierarchical and k-means analyses yielded average of four clusters in data	Select raw material from specific county or state based on cluster averages of desired properties, difficult to label clusters
Recursive Partitioning	What cotton raw material properties are affecting ends down (spinning efficiency)?	Most important factors: temperature and humidity, high temperatures (>79 degrees) and low humidities (<58%) lead to almost all “bad” spinning outcomes	Tighter methods and controls on environmental conditions
		Second most important factor: %DP 555	Large effect of cotton variety, other varieties need to be examined in future
		Other important factors: uniformity, color variables (+b and Rd), micronaire	Color variables may represent other lurking variables, future analysis to study effect of growing conditions on cotton quality and therefore spinning productivity
Discriminant Analysis	What cotton raw material properties are affecting ends down (spinning efficiency)?	Important factors from stepwise analysis: uniformity, micronaire, spinning humidity, spinning temperature	Similar results to recursive partitioning, discriminant function could be used to classify or predict ends down based on variable properties

6.0 Recommendations and Future Work

With the results obtained from the data mining tools in this case study, an improvement team at the spinning manufacturing plant should be formed to critically apply the results to their processes and implement solutions. Data-driven decision making is a very powerful improvement tool that can lead to successful results within an organization. The team should brainstorm potential solutions, fully develop these ideas and deploy them throughout the manufacturing plant, and finally assess the effectiveness of their improvement plan. Data analysis is ineffective unless it is translated to actionable results.

In regards to the case study data, a design of experiments is necessary to fully investigate the impact of the cotton factors on spinning ends down in a controlled environment. The experiment could be designed from the partitioning results in regards to the important factors and levels discovered. Also, a data collection plan and material tracking tool is essential to eliminate extra noise in the data. For this analysis a two-day process time estimate was made in order to merge the cotton data with the resulting ends down data. This was a very rough estimate that was necessary in order to determine any associations in the data, but the cotton must be tracked throughout the manufacturing process with labels and timestamps for a more thorough experiment and analysis.

After investigation of the effects of cotton variety, it is clear that this variable is influential to the yarn spinning process. More research is needed about the effects of other types of cotton variety, not just the DP 555 studied in this case. After studying the effects of all cotton varieties, it would be extremely helpful to cotton purchasers if the USDA required the farmers to specify the percentages of each cotton variety present in the bale. The manufacturers could have more control over the quality of the cotton they were purchasing and its impact on spinning productivity.

7.0 Conclusion and Key Learnings

The spinning manufacturing case study was an example of a situation in industry where large amounts of data are collected but where very little data analysis and data-driven decision making is currently in use. By knowing and understanding these four data mining tools, a practitioner in any industry can apply them to their own data to explore the underlying trends and patterns to enable continuous improvement in their processes. The researcher learned the following after completion of this case study:

- The actual implementation of the data mining tools is very different from simply reading books and research articles because the majority of the learning comes from actual hands-on application by doing the techniques and interpreting the results. It's relatively easy to follow the analysis steps in the software but the challenges occur when interpreting the results and actually applying solutions.
- The difficulty of issues related to data storage, preparation, and computational hardware needs to be addressed and considered. Businesses are collecting more and more amounts of data, but this data may not be accessible or of interest for research purposes. In addition, more resources need to be dedicated to hardware that is capable of handling large datasets and analysts who are able to find useful information within these huge databases.
- The research question must be clearly defined from the onset of the project to enable the analyst to be effective in their exploration of the data. Communication between the subject matter expert and research analyst is critical for success in the project. The project team must also identify what they intend to do with the results to add value to their processes by creating an implementation plan for example.
- The software tool must be used as a business advantage. The many graphical and exploratory tools available in JMP® for example enable any person from the machine operator to the CEO to quickly understand the data and obtain useful information. It is valuable for the analyst to understand and use several different software programs when exploring the data because not one package offers all the tools; they each provide a different "toolbox" that can be helpful for the analysis.

- Assumptions are not critical for data mining tools, but outliers and skewed data can significantly impact the results. Data preparation and initial exploration is very important to complete prior to the actual analysis, and data transformation methods vary by industry.
- These data mining tools should be used as part of a continuous improvement process. Initial results can be discovered, an improvement solution implemented, and then the data analyzed again to determine the effectiveness of the solution. The researcher had a difficult time knowing when to “stop” the analysis iterations, but in actuality the data is continually reassessed and investigated for critical factors.

REFERENCES

Agricultural Marketing Service. (2002-2007). *Cotton varieties planted*. Memphis, TN: USDA.

Anthony, E. K. (2008). Cluster profiles of youths living in urban poverty: Factors affecting risk and resilience. *Social Work Research, 32*(1), 6-17. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=aph&AN=31243643&site=ehost-live&scope=site

Artzt P., Herter T., & Preininger H. (1990). Possibilities of direct conditioning in rotor spinning. *International Textile Bulletin: Yarn Forming, 36*(3), 4. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=tdh&AN=TDH0542690199007518&site=ehost-live&scope=site

Backe, E. E. (1986). Effect of short fiber content in cotton on plant performance and quality'. *Textile Research Journal, 56*(2), 112-115. doi:10.1177/004051758605600207

Basal, G., & Rust, J. P. (2001). Effects of feed sliver moisture content on rotor spinning performance and rotor spun yarn properties. *Textile Research Journal*, *71*(3), 271.

Bradow, . (2000). Quantitation of fiber quality and the cotton production-processing interface: A physiologist's perspective. *The Journal of Cotton Science, 4*, 34.

Cecil, K. (1998). Cotton fiber properties as determined by ginning practices. (Textile Engineering, NCSU).

Cheng, . (2006). Using data mining to evaluate patient-oriented medical services for chronic senility outpatients. *Quality Quantity, 40*(6), 1079.

The Classification of Cotton. (2001). http://www.cottoninc.com/ClassificationofCotton/

Deltapine. (2008). Retrieved Sep 6, 2008, from http://www.deltaandpine.com/default.asp

Esbensen, K., Schoenkopf, S., & Midtgaard, T. (1994). *Multivariate analysis in practice*. Trondheim: CAMO. Retrieved from WorldCat

Ethridge, M. D., Towery, J. D., & Hembree, J. F. (1982). Estimating functional relationships between fiber properties and the strength of open-end spun yarns'. *Textile Research Journal, 52*(1), 35-45. doi:10.1177/004051758205200106

Gaspar, L. (1997). Recursive partitioning analysis (RPA) of prognostic factors in three radiation therapy oncology group (RTOG) brain metastases trials. *International Journal of Radiation Oncology, Biology, Physics, 37*(4), 745.

Gaudard, M., Ramsey, P., & Stephens, M. (2006). *Interactive data mining and design of experiments: The JMP partition and custome design platforms*North Haven Group.

Giudici, P. (2003). *Applied data mining : Statistical methods for business and industry*. Chichester; Hoboken, NJ: J. Wiley.

Hand, D., Mannila, H., & Smyth, P. (2001). *Principles of data mining.* Cambridge, MA: MIT Press.

Hardle, W., & Simar, L. (2003). *Applied multivariate statistical analysis*. Germany: Springer-Verlag.

Hequet, E. (2005). Processing sticky cotton: Effect of stickiness on yarn quality. *Textile Research Journal, 75*(5), 402.

Hillhouse, J. J., & Adler, C. M. (1997). Investigating stress effect patterns in hospital staff nurses: Results of a cluster analysis. *Social Science & Medicine, 45*(12), 1781-1788.

JMP User's Guide (2007). Cary, NC: SAS Institute Inc.

Kettenring, J. R. (2006). The practice of cluster analysis. *Journal of Classification, 23*(1), 3-30. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=aph&AN=21634033&site=ehost-live&scope=site

Lachenbruch, P. A. (1975). Discriminant analysis. (). New York: Hafner Press.

Lehman, A., O'Rourke, N., Hatcher, L., & Stephans, E. (2005). *JMP for basic univariate and multivariate statistics: A step-by step guide* (SAS Institute ed.). Cary, NC:

Maiuro, R. D. (2000). Development of a health care provider survey for domestic violence: Psychometric properties. *American Journal of Preventive Medicine, 19*(4), 245.

Punj, . (1983). Cluster analysis in marketing research: Review and suggestions for application. *Journal of Marketing Research, 20*(2), 134.

Ramsey, P., Stephens, M. & Gaudard, M. (2005). *Six sigma, data mining, and the JMP partition platform.*http://www.northhavengroup.com/pdfs/DataMiningAndDOE.pdf

Sall, J., Creighton, L., & Lehman, A. (2005). *JMP start statistics: A guide to statistical and data analysis using JMP and JMP IN software* (3rd ed.). Belmont, CA: Thomson.

Schnur, . (1998). A case study of statistical downscaling in Australia using weather classification by recursive partitioning. *Journal of Hydrology, 212*, 362.

Shmueli, G., Patel, N. R., & Bruce, P. C. (2007). *Data mining for business intelligence: Concepts, techniques, and applications in microsoft office excel with XLMiner.* Hoboken, New Jersey: Wiley.

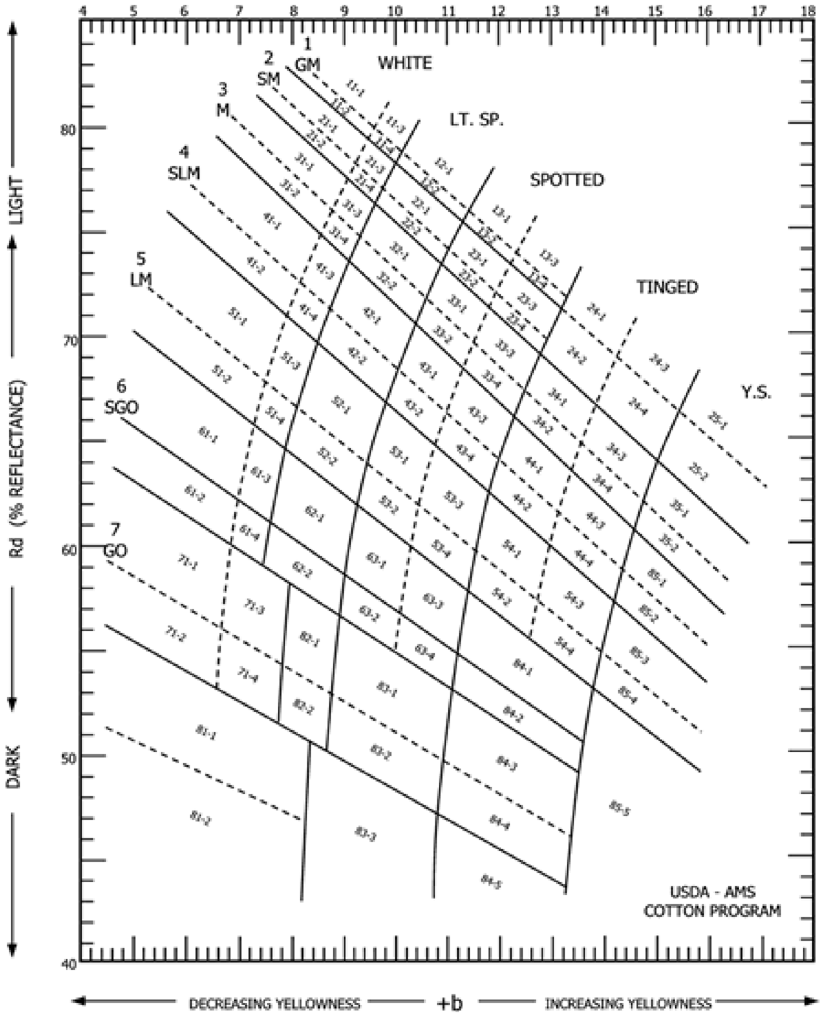
Simpson, J., & Murray, M. (1978). Effect of cotton fiber fineness and strength on mechanical processing and open-end spinning and yarn properties. *Textile Research Journal, 48*, 270.

Truxillo, C. (2005). *Multivariate statistical methods: Practical research applications.* Cary, NC: SAS Institute.

Yang, K., & Trewn, J. (2004). *Multivariate statistical methods in quality management.* New York: McGraw-Hill.

APPENDICES

Appendix A: Nickerson-Hunter Cotton Colorimeter Diagram for Upland Cotton (*The Classification of Cotton.*2001)



Appendix B: Relationship of Trash Area to Classer's Leaf Grade (*The Classification of Cotton.*2001)

Trash Measurement (4-yr. Avg.) (% area)	**Classer's Leaf Grade**
0.12	1
0.20	2
0.33	3
0.50	4
0.68	5
0.92	6
1.21	7

Appendix C: HVAC Data Definitions

Variable Names	Zone in Plant	Variable Definition
A1OAT		Airwash 1 outside air temperature
A1OAH		Airwash 1 outside humidity
A1OAE		Airwash 1 outside air enthalpy (temperature of air and moisture content)
A1Z1T	AW #1 - 1: Uniflocs 1 - 3 (cleaners, openers)	Airwash 1 zone 1 temperature (actual)
A1Z1TSP		Airwash 1 zone 1 temperature setpoint
A1Z1HUM		Airwash 1 zone 1 humidity (actual)
A1Z2T	AW #1 - 2: Card lines 1 - 3 and Drawframes 1 - 5	Airwash 1 zone 2 temperature (actual)
A1Z2TSP		Airwash 1 zone 2 temperature setpoint
A1Z2HUM		Airwash 1 zone 2 humidity (actual)
A1Z3T	AW #1 - 3: Card lines 4 - 6 and Drawframes 6 - 14	Airwash 1 zone 3 temperature (actual)
A1Z3TSP		Airwash 1 zone 3 temperature setpoint
A1Z3HUM		Airwash 1 zone 3 humidity (actual)
A1Z4T	AW #1 - 4: Cleaning lines 1 - 6	Airwash 1 zone 4 temperature (actual)
A1Z4TSP		Airwash 1 zone 4 temperature setpoint
A1Z4HUM		Airwash 1 zone 4 humidity (actual)
A2RAT		Airwash 2 return air temperature
A2RAH		Airwash 2 return air humidity
A2OAT		Airwash 2 outside air temperature
A2OAH		Airwash 2 outside humidity
A2OAE		Airwash 2 outside air enthalpy (temperature of air and moisture content)
A2Z1T	AW #2 - 1: Spinning frames 16 - 24	Airwash 2 zone 1 temperature (actual)
A2Z1TSP		Airwash 2 zone 1 temperature setpoint
A2Z1HUM		Airwash 2 zone 1 humidity (actual)
A2Z2T	AW #2 - 2: Spinning frames 9 - 15	Airwash 2 zone 2 temperature (actual)
A2Z2TSP		Airwash 2 zone 2 temperature setpoint
A2Z2HUM		Airwash 2 zone 2 humidity (actual)
A2Z3T	AW #2 - 3: Spinning frames 1 - 8	Airwash 2 zone 3 temperature (actual)
A2Z3TSP		Airwash 2 zone 3 temperature setpoint
A2Z3HUM		Airwash 2 zone 3 humidity (actual)

Appendix D: SAS Code

Check Univariate and Multivariate Normality

//First, library name declared where the datasets are located

Libname SpinningData 'E:\Crop Year Data';

//Reference to the MULTINORM macro code to evaluate multivariate normality

%include 'E:\Crop Year Data\multnorm.sas';

//Check univariate normality and graph quantile-quantile (Q-Q) plots

proc univariate data = SpinningData.*datasetname* noprint;

qqplot Mic Length Strength PlusB Rd TrashArea Uniformity ClasserLeaf /normal (mu = est sigma = est);

run;

//Check multivariate normality (plot of squared Mahalanobis distances of observations from centroid versus chi-squared quantiles)

//General form of %MULTINORM macro

%MULTINORM(data= *datasetname*, VAR=*variables)*

//Code to assess multivariate normality of crop year data

%multinorm(data = SpinningData.*datasetname*,

var = Mic Length Strength PlusB Rd TrashArea Uniformity ClasserLeaf)

//Proc calis to check multivariate kurtosis and skewness, options included to perform analysis on convariance matrix using weighted least squares (WLS) method to display multivariate kappa statistics and output kurtosis statistics

proc calis data = SpinningData.*datasetname* covariance method = WLS kurtosis;

var Mic Length Strength PlusB Rd TrashArea Uniformity ClasserLeaf;

lineqs

    Mic = e1,

```
        Length = e2,
        Strength = e3,
        PlusB = e4,
        Rd = e5,
        TrashArea = e6,
        Uniformity = e7,
        ClasserLeaf = e8;
 std
 e1=eps1, e2=eps2, e3=eps3, e4=eps4, e5=eps5, e6=eps6, e7=eps7, e8=eps8;
 cov
 e1=eps1, e2=eps2, e3=eps3, e4=eps4, e5=eps5, e6=eps6, e7=eps7, e8=eps8;
run ;
```

Discriminant Procedure to Check Variance-Covariance Matrices

//General form of discriminant procedure

proc discrim data=*datasetname* pool=test;

class *ClassificationVariable*;

var *ContinuousVariables(Y covariates)*;

run;

//Code to check Variance-Covariance Matrices of crop year data

proc discrim data= SpinningData.*datasetname* pool=test;

class YarnOutcome;

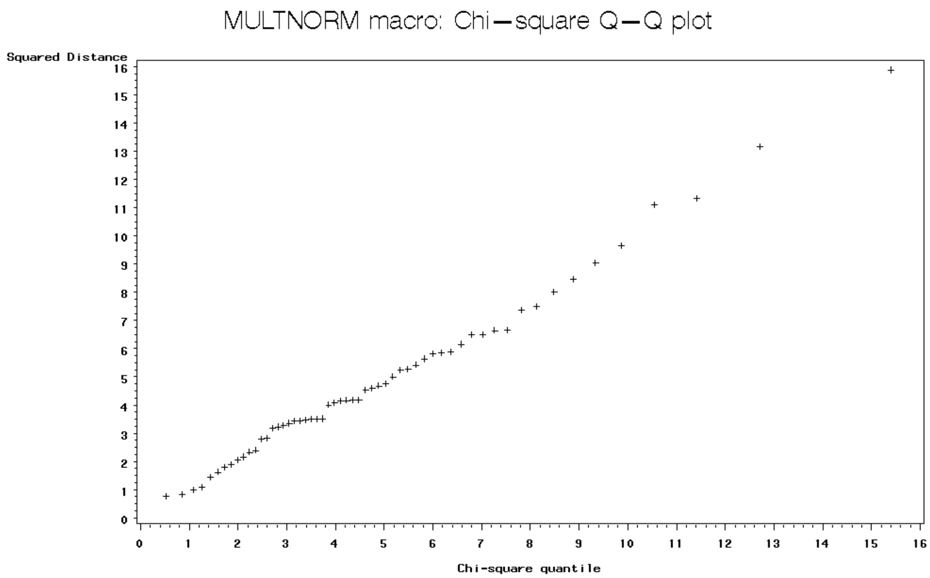
var Mic Length Strength PlusB Rd TrashArea Uniformity ClasserLeaf;

run;

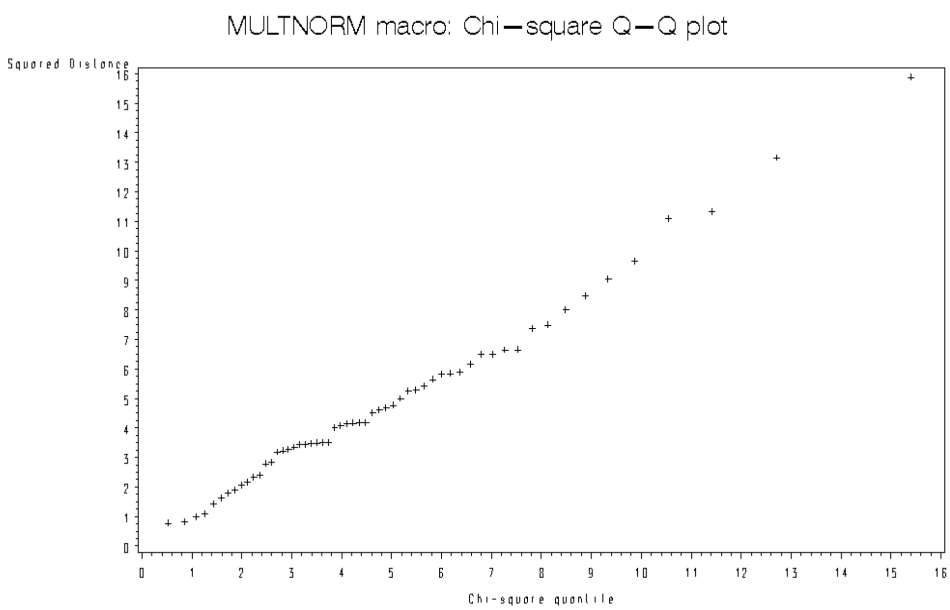
//pool=test option to determine whether groups have homogeneous covariance matrices at default alpha level of .05

Appendix E: Multivariate Plots of Cotton Properties by Crop Year from SAS MULTINORM Macro

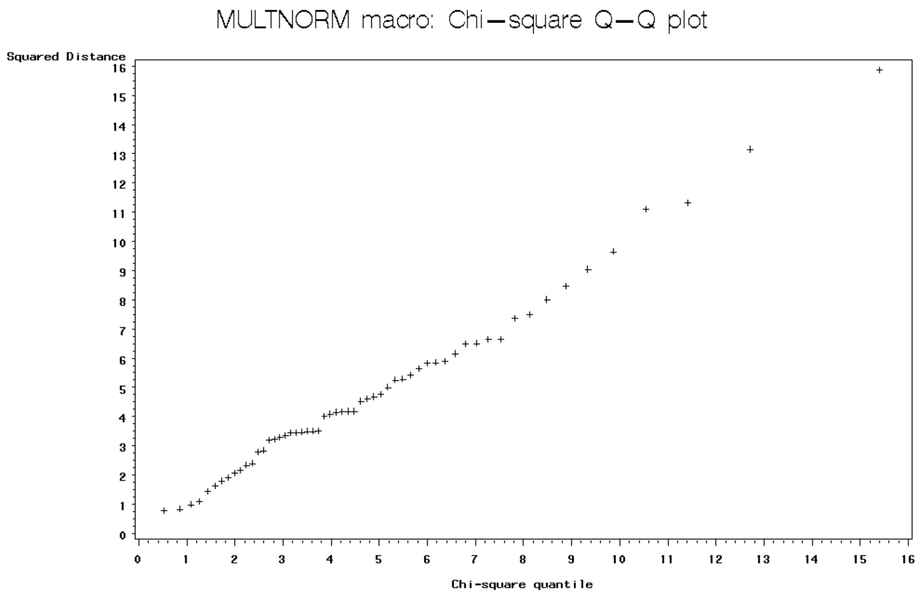
2003



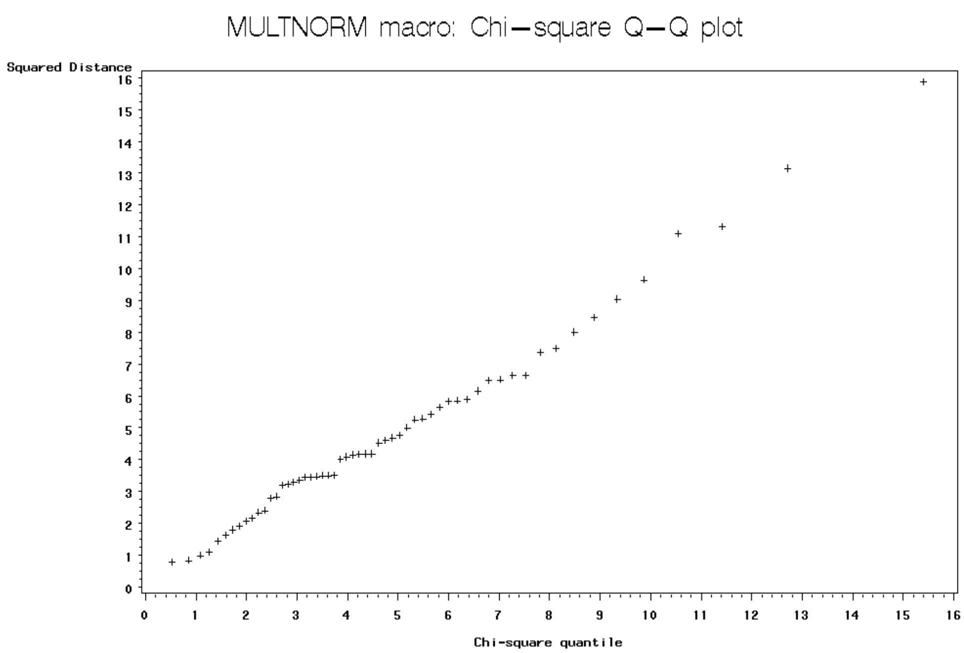
2004



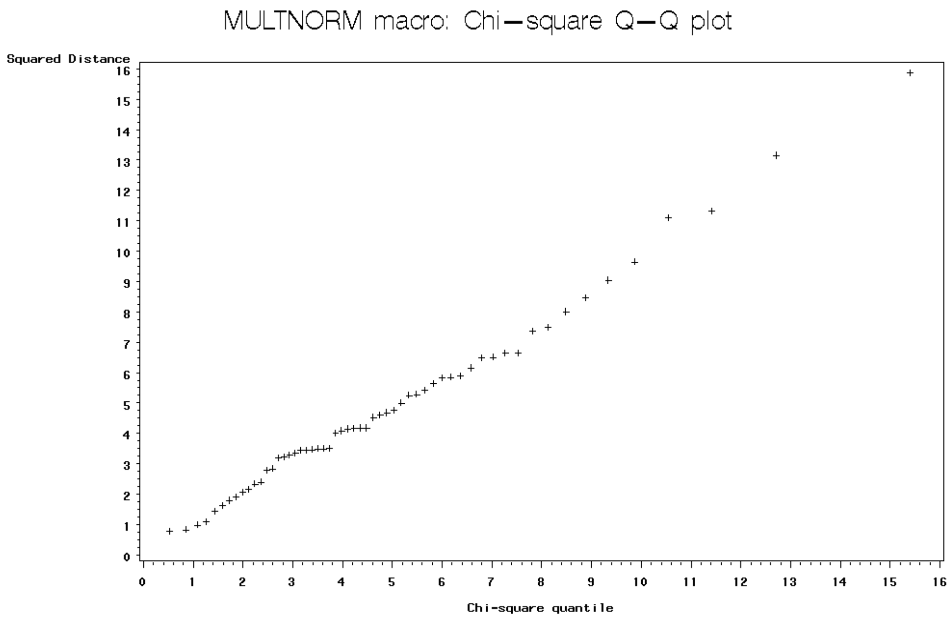
2005


MULTNORM macro: Chi—square Q—Q plot
Squared Distance
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Chi-square quantile

2006


MULTNORM macro: Chi—square Q—Q plot
Squared Distance
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Chi-square quantile

2007



MULTNORM macro: Chi−square Q−Q plot
Squared Distance
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Chi-square quantile

Appendix F: Univariate Skewness and Kurtosis Values by Crop Year for Eight Cotton Classification Properties

Crop Year	Statistic	Mic	Length	Strength	PlusB	Rd	Trash Area	Uniformity	Classer Leaf
2003	Skewness	-0.67	0.55	0.64	0.23	-0.45	0.51	-0.14	-0.29
	Kurtosis	0.50	0.58	0.66	0.38	-0.10	0.06	-0.10	-1.77
2004	Skewness	-0.61	0.24	0.45	-0.15	0.17	0.45	0.11	0.11
	Kurtosis	0.16	-0.15	0.29	0.29	-0.75	0.00	-0.49	-0.94
2005 (without classer leaf)	Skewness	-0.51	-0.73	0.32	-0.35	-0.09	0.31	0.34	N/A
	Kurtosis	-0.34	-0.36	0.62	0.17	-0.08	-0.32	-0.38	N/A
2006	Skewness	-0.65	0.49	0.45	0.09	-0.16	0.53	0.29	0.22
	Kurtosis	-0.07	-0.14	0.11	-0.67	-0.23	-0.11	-0.67	-0.16
2007	Skewness	-0.40	-0.61	-0.07	-0.10	0.28	-0.04	-0.02	-1.58
	Kurtosis	-0.54	-0.72	-0.53	0.46	0.11	0.58	-0.44	0.81

Appendix G: How-To Guide for PCA, Cluster Analysis, Recursive Partitioning, and Discriminant Analysis Data Mining Tools in JMP® Software
(*JMP User's Guide* 2007)

1. **PCA**

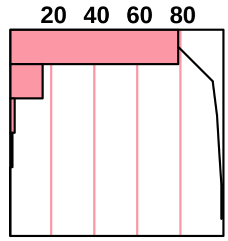
- <u>What is it?</u>
  - Dimension reduction process to reduce a large number of continuous variables to a smaller set of variables that account for most of the variance
  - PC → linear combination of optimally weighted observed variables
- Where is it used?
  - Data preparation stage to reduce number of redundant variables in surveys/questionnaires, social science research
- <u>Getting Started:</u>
- Perform analysis on correlation matrix, the covariance matrix, or on unscaled data
- Example: Different chemical compounds were measured for solubility in different solvents
  - Solubility.jmp file in JMP Sample Data directory
- Analyze>Multivariate Methods>Principal Components
- In Y, columns, select variables to analyze
- Initial report shows components with eigenvalues (portion of the total variation in the multivariate space), percent (eigenvalue as a % of the total eigenvalues), and the cumulative % of variation represented by eigenvalues

Multivariate

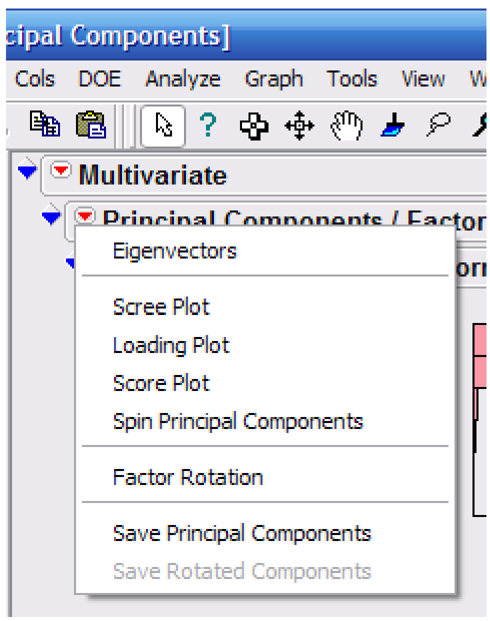
Principal Components / Factor Analysis

Principal Components: on Correlations

Number	Eigenvalue	Percent	Cum Percent
1	4.7850	79.750	79.750
2	0.9452	15.754	95.504
3	0.1399	2.331	97.835
4	0.0611	1.018	98.853
5	0.0471	0.785	99.638
6	0.0217	0.362	100.000



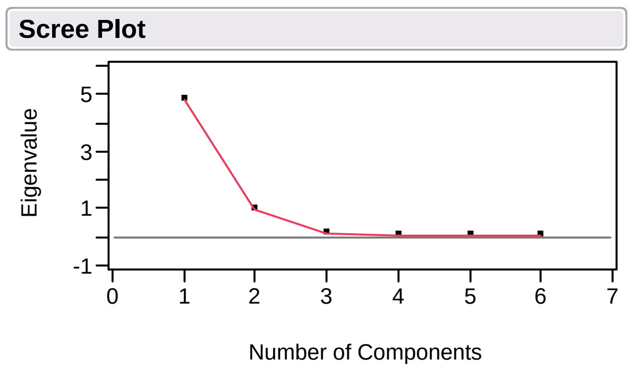
- Note: Eigenvalues sum to the # of variables being analyzed because each variable contributes one unit of variance to the analysis
- <u>Options & Eigenvectors:</u>
- PC Hot Spot Options > Eigenvectors, Scree Plot, Loading Plot, Score Plot, Spin Principal Components, Factor Rotation, Save Principal/Rotated Components



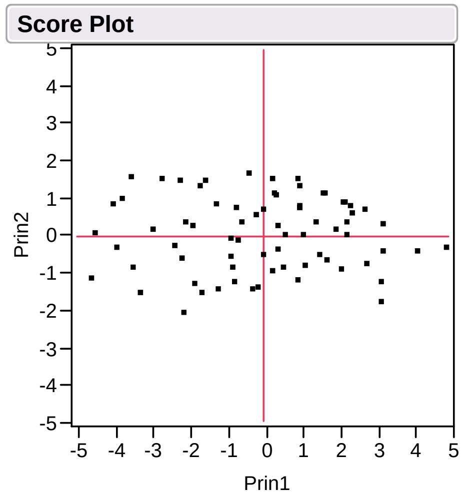
- Eigenvectors - coefficients to form a linear combination of the original variables to calculate the principal component

Eigenvectors						
1-Octanol	0.37441	0.55987	-0.11070	-0.65842	0.31660	0.01874
Ether	0.34834	0.64314	0.11973	0.62764	-0.20890	0.11456
Chloroform	0.41940	-0.29864	-0.64850	0.30599	0.43061	0.18793
Benzene	0.44561	-0.14756	-0.21904	-0.09455	-0.49849	-0.68865
Carbon Tetrachloride	0.43102	-0.29736	0.18487	-0.24135	-0.45965	0.64968
Hexane	0.42217	-0.27117	0.68608	0.10831	0.45926	-0.23426

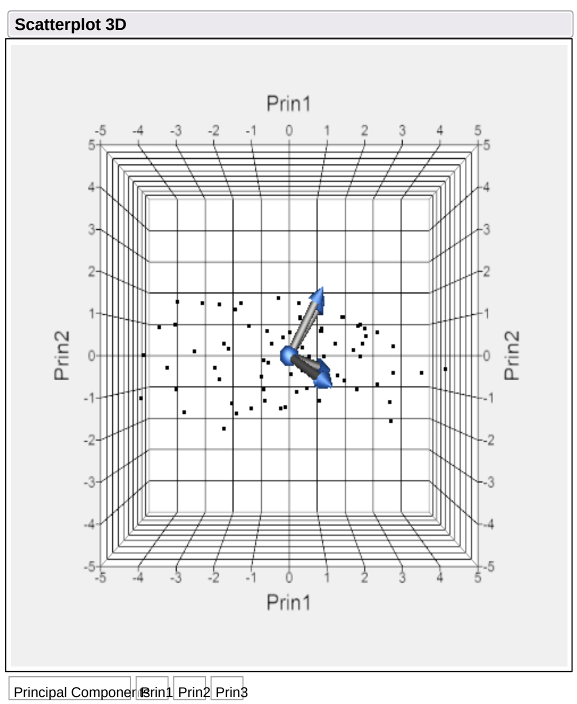
- <u>Scree and Loading Plot:</u>
- Scree Plot – eigenvalues vs. the number of components, used to determine # of meaningful components (before break in plot)



- From example Scree plot, 2 components are seen before the break in the plot and are kept in the analysis
- <u>Score Plot and 3D Plot:</u>
- Score Plot – shows scores for each pair of principal components



- Spin Principal Components – 3D plot of 3 PC's, biplot rays are function of PC's on axes
    - Show PC's, Rotated Components or data columns on axes
    - Select specific axis contents



- Rotation to Solution:
- Rotation – linear transformation on the PC's to interpret the solution

- o Factoring method → Principal Components
- o # of factors to rotate
- o Rotation method

- Save Principal/Rotated Components – save # of components that you specified as new columns to data table

Factor Rotation: Varimax

Prior Communality Estimates:1

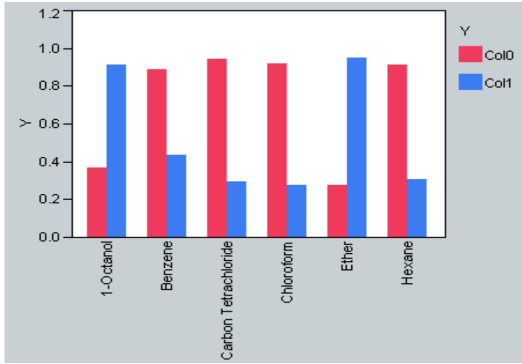
Unrotated Factor Pattern

1-Octanol	0.819019	0.544318
Ether	0.761978	0.625283
Chloroform	0.917422	-0.290351
Benzene	0.974761	-0.143460
Carbon Tetrachloride	0.942849	-0.289099
Hexane	0.923483	-0.263641

Rotated Factor Pattern

1-Octanol	0.3686763	0.9116749
Ether	0.2758844	0.9462967
Chloroform	0.9211420	0.2783219
Benzene	0.8855353	0.4319341
Carbon Tetrachloride	0.9414227	0.2937115
Hexane	0.9110652	0.3037894

- Example – 2 factors, varimax (orthogonal) rotation (maximize variance of a column of factor pattern matrix)
  - o 1st rotated component loads (>.40) on chloroform, benzene, carbon tetracholoride, and hexane and 2nd component loads on octanol and ether (circled values)
  - o Right click on rotated factor pattern (above) to make into data table and create bar chart to show factor membership (Y values: principal components, X value: variable names)
  - o Conclusion: 6 variables transformed to two principal components, surrogate variables to use in future regression analyses



Principal Component #	Variables
1	chloroform, benzene, carbon tetracholoride, and hexane
2	octanol and ether

2. Cluster Analysis

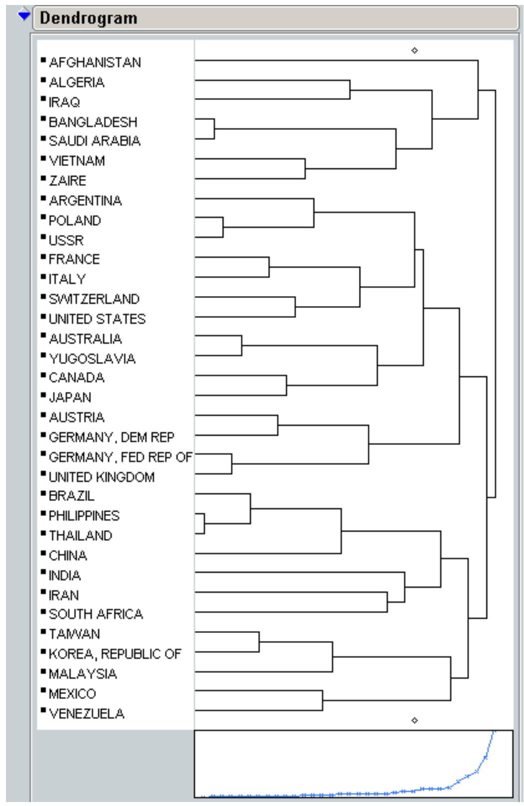
- What is it?
- Exploratory method to identify similar objects in large datasets, number and names of groups not known beforehand
- Where is it used? → Market based research for customer segmentation and structure analysis
- Getting Started:
- 2 methods:
- Hierarchical →small tables (up to several thousand rows), combines rows in an hierarchical sequence by combining the two clusters that are closest together, dendrogram output
- K-means clustering →larger tables (hundreds of thousands of rows), specify # of clusters beforehand and cluster seed points are randomly calculated, points iteratively assigned to clusters
- Analyze > Multivariate Methods > Cluster
- Select Y, Columns then Hierarchical or K-means method
- Hierarchical:
- Character data can be analyzed
    - Ordinal data: data value used is index of category
    - Nominal data: categories must match to contribute a distance of 0 or standardized distance of 2 otherwise
- Select Label variable to aid in identifying clusters of observations
- 5 distance measures between clusters available: Average, Centroid, Ward, Single, and Complete
- Option to add ordering variable to sort by mean value - ex. Use principal component
- Example - 1976 Crude Birth and Death Rates per 100,000 people
    - Birth Death Subset.jmp file in JMP Sample Data directory
    - Select Birth and Death variables as cluster variables and Ward's default method (distance measure)

- <u>Hierarchical Output:</u>
- Clustering History table

Clustering History

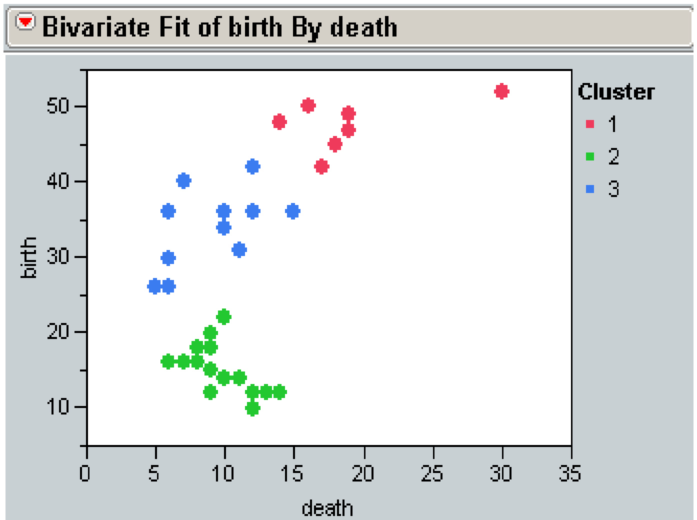
Number of Clusters	Distance	Leader	Joiner
33	0.000000000	PHILIPPINES	THAILAND
32	0.104271097	BANGLADESH	SAUDI ARABIA
31	0.104271097	POLAND	USSR
30	0.104271097	GERMANY, FED REP OF	UNITED KINGDOM
29	0.104271097	AUSTRALIA	YUGOSLAVIA
28	0.120401892	BRAZIL	PHILIPPINES

- Dendrogram tree – depiction of clustering history, diamond shaped handle to identify # of clusters

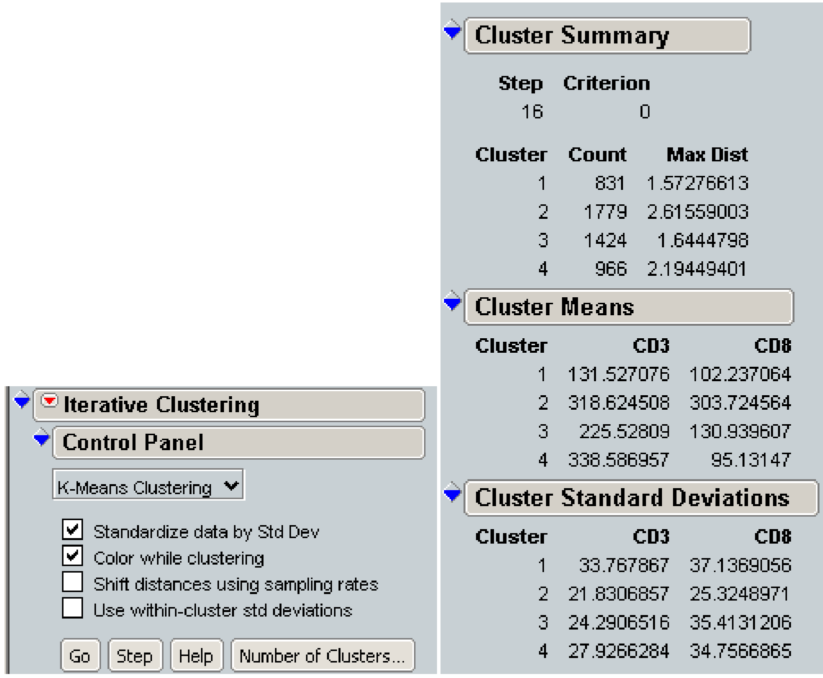


- Scree plot (distance graph at bottom of dendrogram) – shows distances between clusters and natural break in curve is cutting point for # of clusters to use
- <u>Hierarchical Options:</u>
- Color & Mark clusters
- Save cluster – cluster # of each row in data column

- Color map (heat map) – shows colored values of data across range
- Dendrogram Scale – even spacing, geometric spacing, distance scale
- Hierarchical Visualization:
- Save clusters from dendrogram analysis
- Analyze>Fit Y by X to look at clusters of observations using color and marker options, ANOVA to look at differences among clusters



- K-Means:
- Choose number of clusters to form
- Select Centers variable as cluster identifier where the # of levels = # of clusters
- Control Panel – run complete clustering process or step through, change # of clusters
- Output - Cluster Summary, Means, and Standard Deviation
- Example - Table with eight numeric variables and 5000 observations about the measurement or counting of cells
    - Cytometry.jmp file in JMP Sample Data directory
    - Select two variables CD3 and CD8 as Y variables and the number of clusters as 4, output below



Iterative Clustering

Control Panel

K-Means Clustering

- [x] Standardize data by Std Dev
- [x] Color while clustering
- [ ] Shift distances using sampling rates
- [ ] Use within-cluster std deviations

Go | Step | Help | Number of Clusters...

Cluster Summary

Step	Criterion
16	0

Cluster	Count	Max Dist
1	831	1.57276613
2	1779	2.61559003
3	1424	1.6444798
4	966	2.19449401

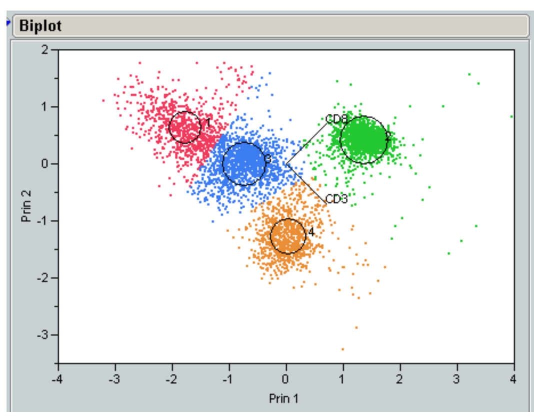
Cluster Means

Cluster	CD3	CD8
1	131.527076	102.237064
2	318.624508	303.724564
3	225.52809	130.939607
4	338.586957	95.13147

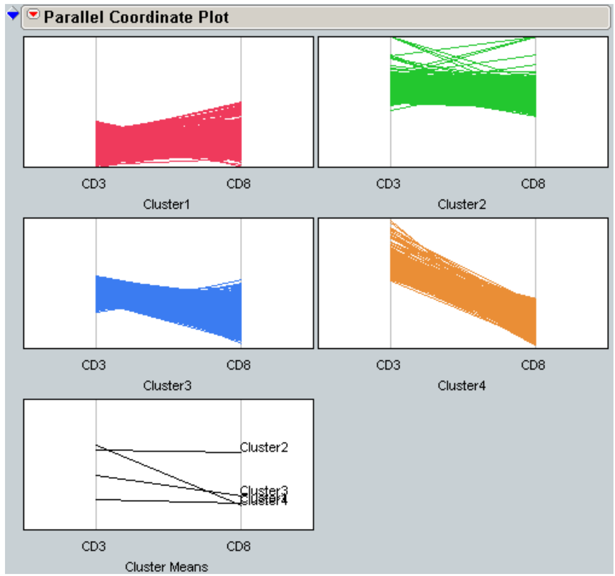
Cluster Standard Deviations

Cluster	CD3	CD8
1	33.767867	37.1369056
2	21.8306857	25.3248971
3	24.2906516	35.4131206
4	27.9266284	34.7566865

- <u>K-Means Visualization:</u>
- Iterative Clustering hot spot option>Biplot and Parallel Coord Plots
- Biplot - observe clustering process in scatterplot of first two principal components



- Parallel Coordinate Plot – each separate cluster and cluster mean shown for the two variables across the 4 clusters

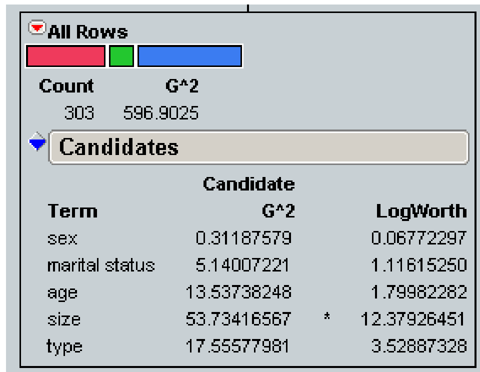


- K-Means Options:
- Mark Clusters – different mark for each cluster
- Seed with Selected Rows – select the rows in the data table to start the cluster centers at and then choose this option
- Save Mixture Probabilities – a column for each cluster is created and the probability an observation belongs to that cluster is saved

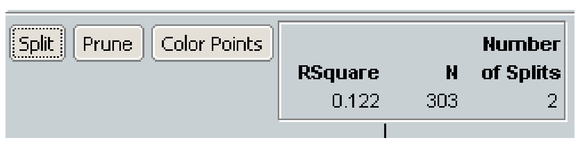
3. Recursive Partitioning

- What is it?
  - Method to split data to find groupings of input variables that best predict the response
  - No distribution assumptions
- Where is it used?
  - To find hidden patterns in large datasets, form decision tree, medical and social science applications
- Getting Started:
- Recursively fit all possible splits to find the optimum partition that maximizes the difference in the responses between the two branches
- Factor variables
  - Continuous – splits are created by cutting value
  - Categorical – splits are created by two groups of levels
- Response
  - Continuous – splits maximize the difference between the means (look at squared errors)
  - Categorical - estimated probability for each response level is calculated and the splits are made to maximize the likelihood-ratio chi-square statistic
- Analyze>Modeling>Partition
- Select Y, Response and X, Factors
- Categorical Response:
- Example – Data about car preferences of people to determine which factors affect the country of purchased cars
  - Car Poll.jmp file in JMP Sample Data directory
  - Select country as the Y variable, and the other columns (sex, marital status, age, size, and type) as X factors

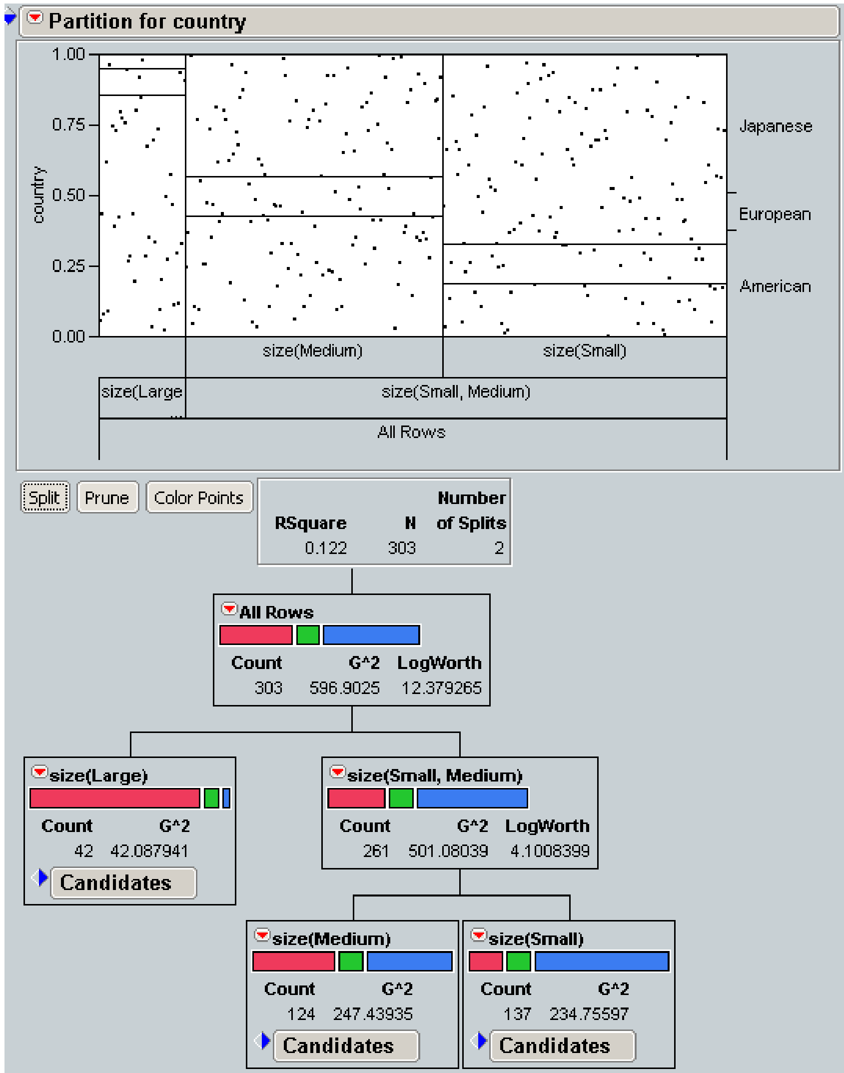
- Initially one group, then the Split command calculates a splitting value (or groups) for the X column that best splits the response
- Candidates report shows each column, * marks optimum split with largest G^2 value, LogWorth is negative log of adjusted p-value (larger value = more significant)



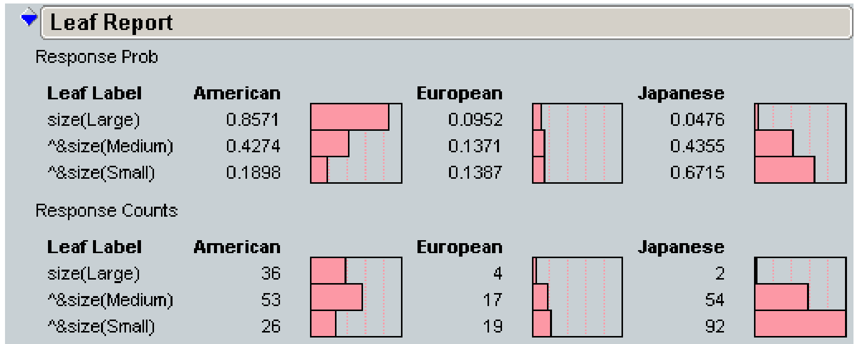
- Imputes appears when there are missing values on the splitting variable and JMP then randomly generates numbers
- Split command to divide into two groups
- Prune command to combine the most recent split back into one group



- <u>Categorical Output:</u>



- Known also as a double-decker plot
- Hierarchical partitions
- Probabilities sum to 1
- <u>Options:</u>
- Leaf Report – shows the mean and count or rates for the bottom-level leaves (output for Car Poll example below)



Leaf Report

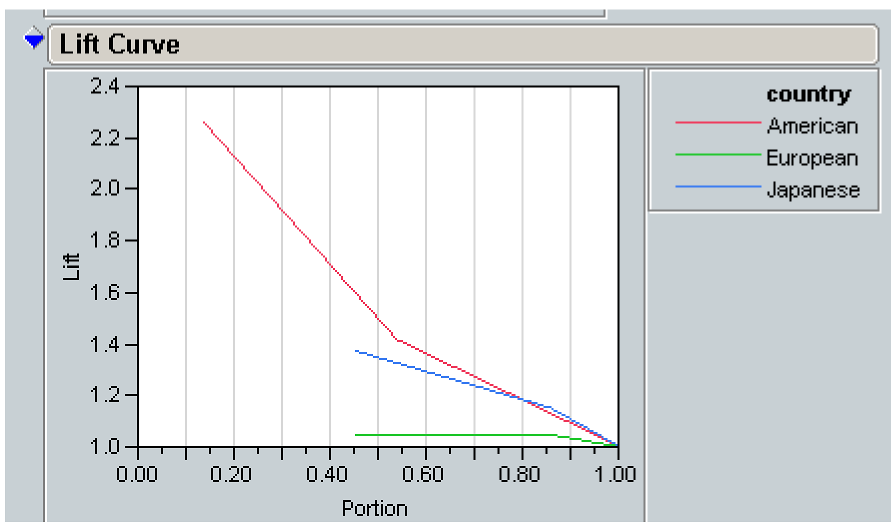
Response Prob

Leaf Label	American	European	Japanese
size(Large)	0.8571	0.0952	0.0476
^&size(Medium)	0.4274	0.1371	0.4355
^&size(Small)	0.1898	0.1387	0.6715

Response Counts

Leaf Label	American	European	Japanese
size(Large)	36	4	2
^&size(Medium)	53	17	54
^&size(Small)	26	19	92

- Lock Column or Lock Node from analysis if not important
- Assess model with Lift and ROC curves (explained in Discriminant Analysis section below)
- <u>Lift Curve:</u>
- Assess goodness of fit
- Specificity versus the cumulative number of observations to assess the “lift” the model gives over a baseline model



- More the curve is decreasing → better the model
- <u>Stopping Criteria:</u>
- Test data is commonly overfit with recursive partitioning

- Different stopping criteria include: number of leaves, number of splits, number of cases/split, very small decreases in impurity
- Alternative is to prune full-grown tree and assess classification rates of sub-trees
- Continuous Response:
- Example – information collected by the U.S Census concerning housing in the area of Boston
    - Boston Housing.jmp file in JMP Sample Data directory
    - Select mvalue as the Y variable, and the other columns as X factors
- Candidates report shows each column, * marks optimum split with largest Sum of Squares (SS) value

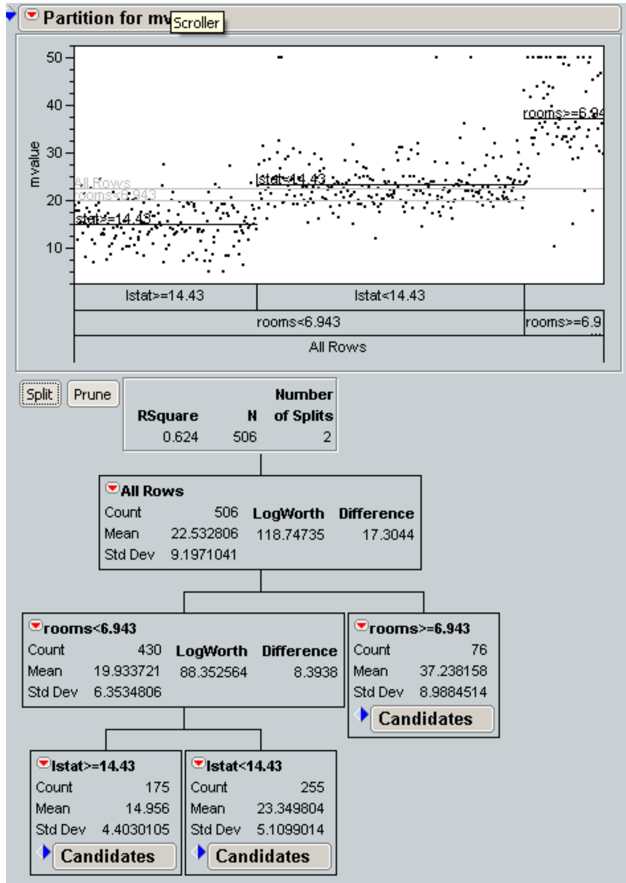
All Rows

Count	506
Mean	22.532806
Std Dev	9.1971041

Candidates

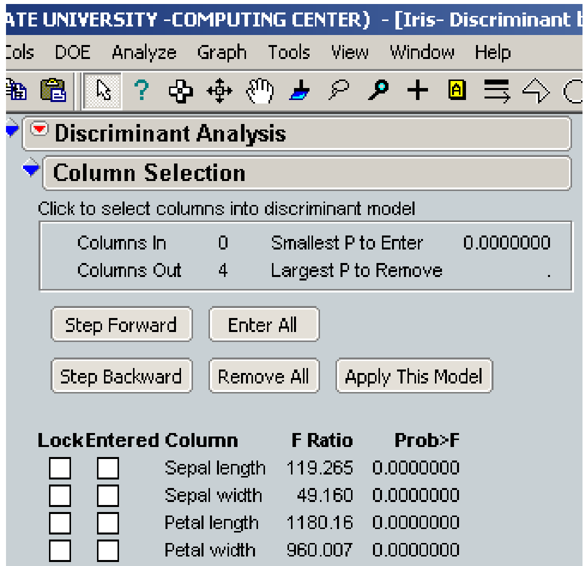
Term	Candidate SS		LogWorth
crim	8266.17273		32.6638216
zn	6669.06251		24.9773486
indus	11083.22547		48.7519537
chas	1312.07927		4.1110954
nox	9536.22405		39.5670978
rooms	19339.55503	*	118.7473483
age	5573.64765		19.6751451
distance	4994.54054		17.1453361
radial	6708.64333		24.6205659
tax	8618.08428		34.5266980
pt	10438.69478		44.8775094
b	5259.31980		18.2910466
lstat	18896.19401		113.7427626

- Output - lines for each split with name of variable and level of split

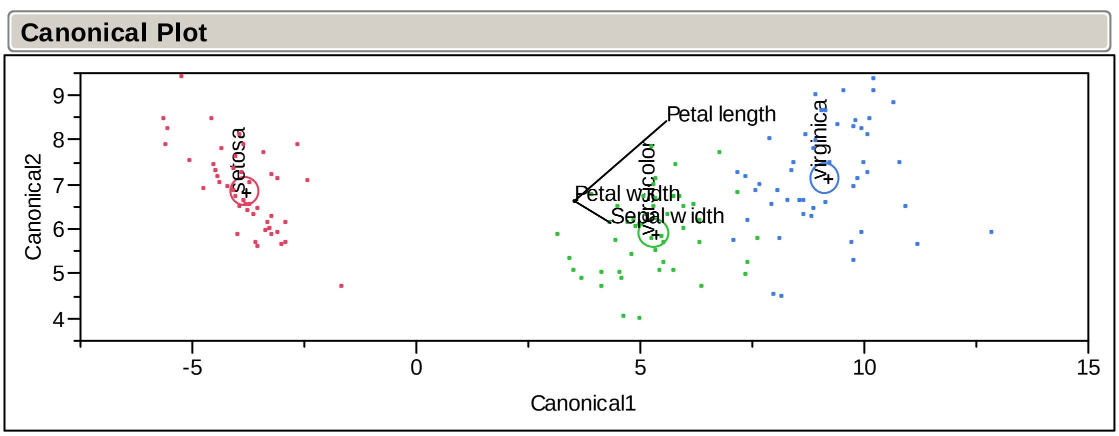

Partition for mv
Scroller
mvalue
50
40
30
20
10
All Rows
rooms<6.943
lstat>=14.43
lstat<14.43
rooms>=6.94
lstat>=14.43
lstat<14.43
rooms<6.943
rooms>=6.9
All Rows
Split
Prune
RSquare N Number of Splits
0.624 506 2
All Rows
Count 506 LogWorth 118.74735 Difference 17.3044
Mean 22.532806
Std Dev 9.1971041
rooms<6.943
Count 430 LogWorth 88.352564 Difference 8.3938
Mean 19.933721
Std Dev 6.3534806
rooms>=6.943
Count 76
Mean 37.238158
Std Dev 8.9884514
Candidates
lstat>=14.43
Count 175
Mean 14.956
Std Dev 4.4030105
Candidates
lstat<14.43
Count 255
Mean 23.349804
Std Dev 5.1099014
Candidates

4. Discriminant Analysis

- What is it?
  - Method to classify units into known categorical groups based on similar characteristics of the predictor variables
  - Multivariate version of logistic regression
  - Assumptions: normal distribution, independence, and constant variance within each group
- Where is it used?
  - Market research, classification applications
- Getting Started:
- Check assumptions beforehand: predictors are normally distributed, linearly related, and groups have equal within-group variances
- Example – Fisher Iris data with sepal length, sepal width, petal length, and petal width, and goal is to identify species type based on these four measured values
  - Iris.jmp file in JMP Sample Data directory
  - Species is X category and 4 other values are Y covariates
- Analyze > Multivariate Methods > Discriminant
- Select continuous variables as Y, Covariates and classification variable as X, Categories
- Types of DA – linear (Y's are normally distributed with the same variances and covariances), quadratic (covariances different), combination of two types
- Method measures Mahalanobis distance from observation to each group's centroid (multivariate mean) and classifies to closest group
- Stepwise Selection:
- Choose Stepwise Variable Selection to find which variables discriminate well
- Analyze F ratios and p-values to control which columns are entered into the model
- Select variables to add, and then apply model



- <u>Canonical Plot:</u>
- Displays observations and multivariate means in the two dimensions that best separate the groups
- Biplot of points and the variable direction rays
- 3D Plot is available when there are four or more groups



- <u>Scoring Report:</u>
- Shows classification performance
- DA Hot Spot>Score Options>Show Interesting Rows Only (misclassified or fitted probabilities between 0.05 and 0.95)

Discriminant Scores

Columns

Sepal width

Petal length

Petal width

Number Misclassified	3
Percent Misclassified	2
-2LogLikelihood	7.246

Row	Actual	SqDist(Actual)	Prob(Actual)	-Log(Prob)	Predicted	Prob(Pred)	Others
69	versicolor	6.67417	0.7974	0.226	versicolor	0.7974	virginica 0.20
71	versicolor	5.41762	0.5528	0.593	versicolor	0.5528	virginica 0.45
73	versicolor	4.87093	0.7165	0.333	versicolor	0.7165	virginica 0.28
78	versicolor	4.40923	0.4520	0.794	* virginica	0.5480	
84	versicolor	5.60225	0.3517	1.045	* virginica	0.6483	
107	virginica	6.32934	0.5655	0.570	virginica	0.5655	versicolor 0.43
120	virginica	8.11284	0.7425	0.298	virginica	0.7425	versicolor 0.26
124	virginica	2.48676	0.9413	0.061	virginica	0.9413	
127	virginica	3.22058	0.8567	0.155	virginica	0.8567	versicolor 0.14
128	virginica	3.30397	0.7814	0.247	virginica	0.7814	versicolor 0.22
130	virginica	8.35644	0.9250	0.078	virginica	0.9250	
134	virginica	7.23535	0.1892	1.665	* versicolor	0.8108	
135	virginica	12.67539	0.6003	0.510	virginica	0.6003	versicolor 0.40
139	virginica	4.06831	0.6874	0.375	virginica	0.6874	versicolor 0.31
150	virginica	2.21526	0.9043	0.101	virginica	0.9043	

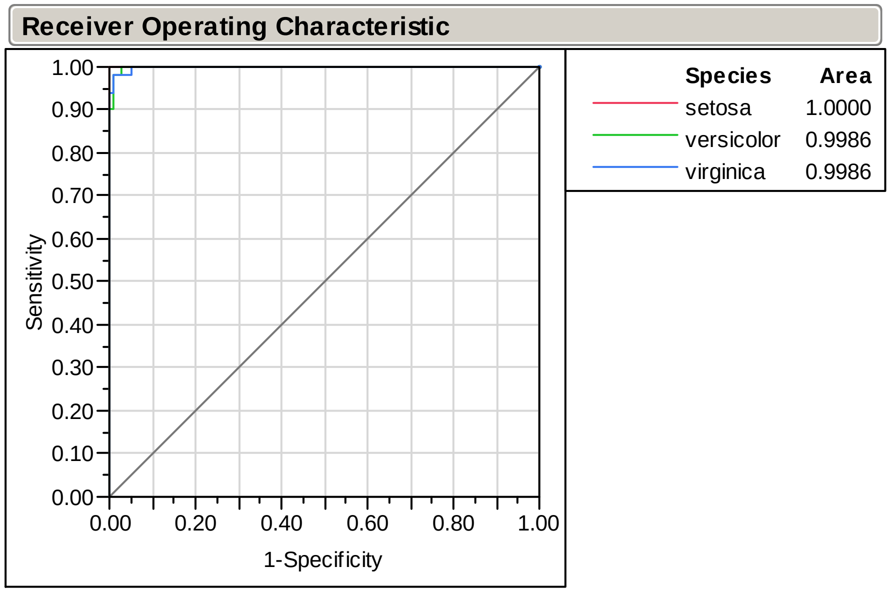
'*' indicates misclassified

Counts: Actual Rows by Predicted Columns

	setosa	versicolor	virginica
setosa	50	0	0
versicolor	0	48	2
virginica	0	1	49

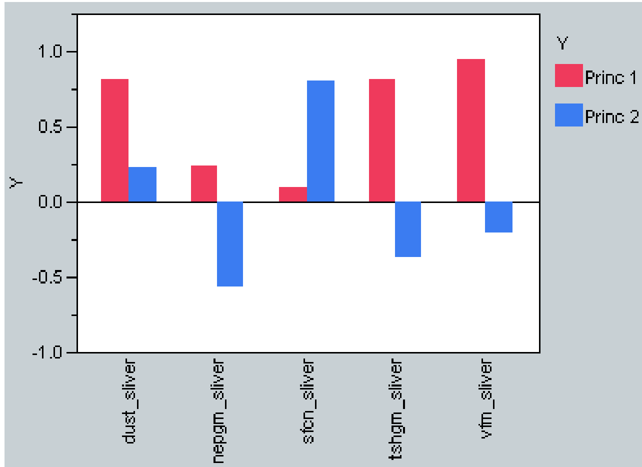
- Asterisk (* - left of Predicted column in above table) – misclassified observation
- Confusion matrix (red box at bottom of above table) – misclassifications show on off diagonal (if everything classified correctly → values of 0 everywhere except the diagonal)
- <u>ROC Curve (Receiver Operator Characteristic):</u>
- DA Hot Spot>Score Options>ROC Curve
- Show the predictive accuracy and sorting efficiency of the model, ideal curve coincides with vertical axis
- Area under the curve is measure of goodness of fit – between 0.5 (no predictive value) and 1 (perfect fit)
- Sensitivity – probability of an event among events (true positives)
- Specificity - probability of a non-event among non-events

- 1-Specificity – false positives



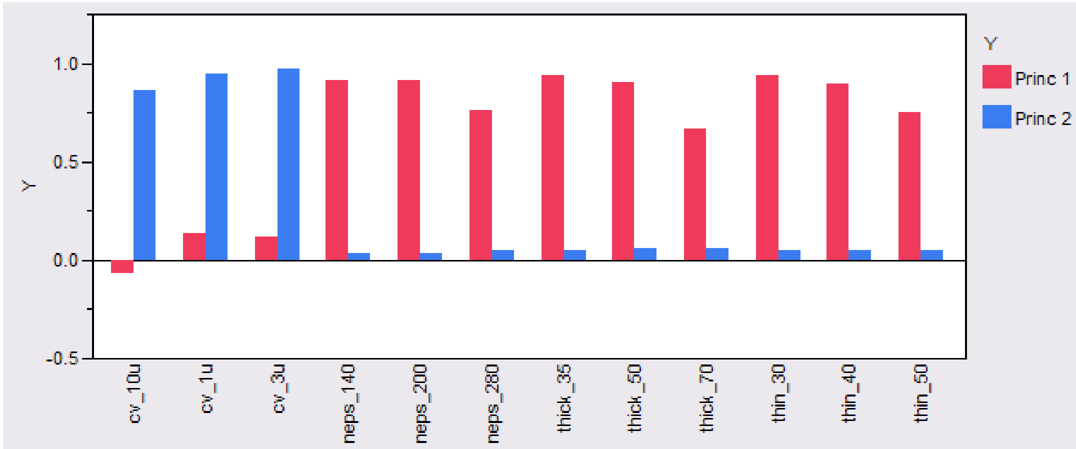
- Test Model on Validation Data:
- Score Options>Save Formulas to save the discriminant equations, column of probabilities and predicted membership added to data table
- Add new observations to data table to test prediction probabilities, look at contingency table to assess classification error rate (Analyze>Fit Y by X with Actual Group Response as Y and Predicted Group Response as X)

Appendix H: PCA Results from AFIS Card Sliver Data



Principal Component #	Name	Variables
1	“Particle Defects”	dust, trash, VFM (visible foreign matter)
2	“Fiber Defects”	neps, SFC (short fiber content)

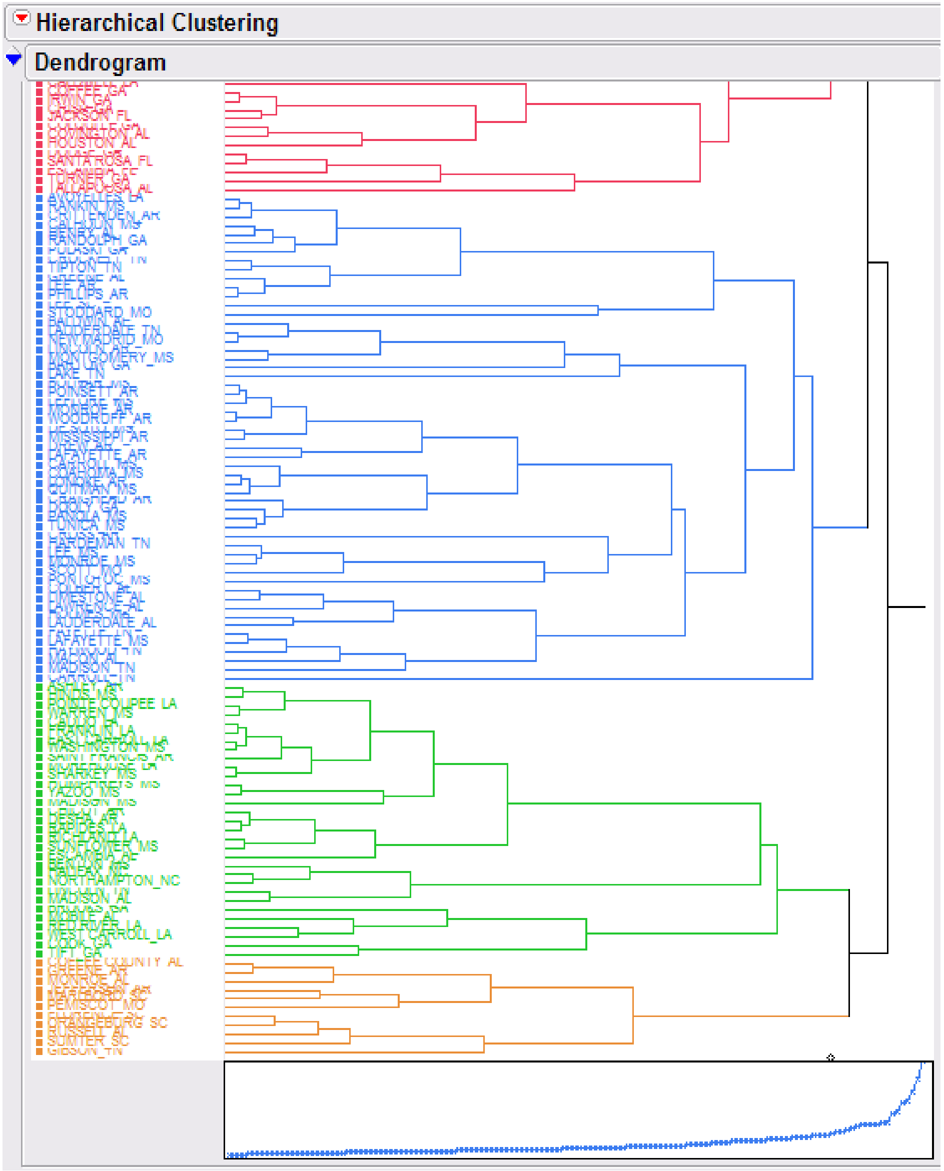
Appendix I: PCA Results from Final Yarn Testing Data



Principal Component #	**Variables**
1	neps +140%, neps +200%, neps +280%, thick +35%, thick +50%, thick +70%, thin -30%, thin -40%, thin -50%
2	CV at 10 meters, CV at 1 meter, CV at 3 meters

Appendix J: Hierarchical Cluster Dendrogram Results by Gin

Appendix K: Hierarchical Cluster Dendrogram Results by County_State



Appendix L: Comparison of ROC Areas for Different Crop Years Using Discriminant Analysis

Crop Year	**ROC (8 fiber properties)**	**ROC (8 fiber properties, spinning temp and humid, % DP555 by state)**
2003	0.66	0.37
2004	0.71	0.81
2005	0.14	0.14
2006	0.61	0.75
2007	0.71	0.86

Appendix M: Comparison of 2005 Crop Year Data With and Without Classer Leaf Discriminant Variable

Table - Comparison of Stepwise Quadratic Discriminant Analysis Results for Two Sets of Predictor Variables

Crop Year	JMP % Misclassified	Stepwise (fiber properties)	JMP % Misclassified	Stepwise (fiber properties, spinning temp and humidity, % DP555 by state)
2005	75.2	Plus b, Rd, Uniformity, Strength, Length, Mic, Trash, Classer Leaf	75.2	Plus b, Temp, Rd, %DP555, Humidity, Uniformity, Strength, Trash, Mic, Length, Classer Leaf
2005 (without classer leaf)	41.8	Plus b, Rd, Uniformity, Strength, Length, Mic, Trash	21.8	Plus b, Temp, Rd, %DP555, Humidity, Uniformity, Strength, Trash, Mic, Length

Table - Comparison of ROC Areas

Crop Year	ROC (fiber properties)	ROC (fiber properties, spinning temp and humid, % DP555 by state)
2005	0.14	0.14
2005 (without classer leaf)	0.64	0.81

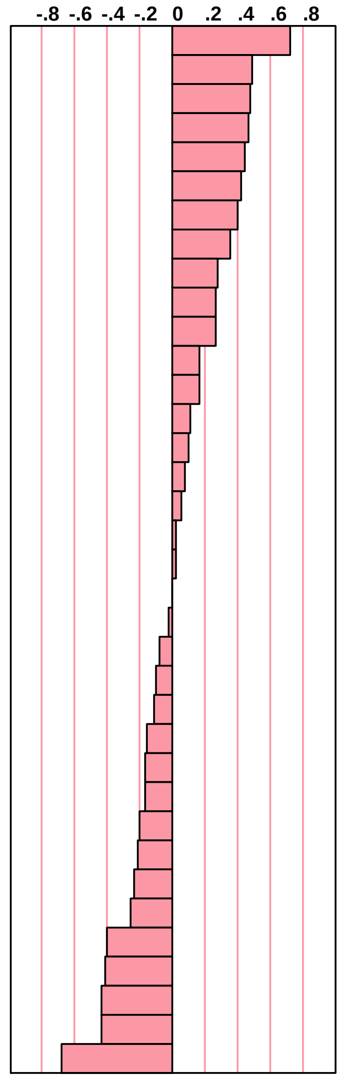
Table - Comparison of JMP® and SAS® Quadratic Discriminant Analysis Misclassification Rates with Cotton Properties and Spinning Temperature and Humidity Predictor Variables

Crop Year	JMP % Misclassified	SAS % Misclassified
2005	75.2	50.0
2005 (without classer leaf)	21.6	30.9

Appendix N: Pairwise Correlations of % DP 555 and Cotton Classification Averages by Year by State

Pairwise Correlations

Variable	by Variable	Correlation	Count	Signif Prob
Average of ClasserLeaf	Average of TrashArea	0.7314	41	<.0001*
Average of Rd	Average of Mic	0.4958	41	0.0010*
Average of Uniformity	Average of Length	0.4761	41	0.0017*
Average of ClasserLeaf	Average of Strength	0.4701	41	0.0019*
Average of Strength	Average of Length	0.4492	41	0.0032*
Average of TrashArea	Average of Length	0.4296	41	0.0051*
Average of ClasserLeaf	Average of PlusB	0.4024	41	0.0091*
Average of TrashArea	Average of Strength	0.3560	41	0.0223*
Average of PlusB	Average of Strength	0.2805	41	0.0757
Average of ClasserLeaf	Average of Length	0.2689	41	0.0891
Average of Uniformity	Average of Strength	0.2646	41	0.0946
Average of Uniformity	Average of TrashArea	0.1711	41	0.2847
Average of Length	% DP 555	0.1692	37	0.3169
Average of TrashArea	Average of PlusB	0.1087	41	0.4986
Average of Rd	% DP 555	0.1015	37	0.5500
Average of Mic	% DP 555	0.0815	37	0.6314
Average of Uniformity	Average of PlusB	0.0532	41	0.7411
Average of PlusB	% DP 555	0.0234	37	0.8905
Average of ClasserLeaf	Average of Uniformity	0.0179	41	0.9117
Average of Strength	% DP 555	0.0046	37	0.9783
Average of Length	Average of Mic	-0.0234	41	0.8843
Average of Rd	Average of Strength	-0.0813	41	0.6134
Average of ClasserLeaf	% DP 555	-0.0990	37	0.5597
Average of TrashArea	Average of Mic	-0.1171	41	0.4661
Average of Uniformity	Average of Rd	-0.1594	41	0.3195
Average of Uniformity	Average of Mic	-0.1680	41	0.2938
Average of Rd	Average of Length	-0.1709	41	0.2852
Average of TrashArea	% DP 555	-0.1980	37	0.2402
Average of Strength	Average of Mic	-0.2079	41	0.1922
Average of ClasserLeaf	Average of Rd	-0.2310	41	0.1461
Average of PlusB	Average of Length	-0.2617	41	0.0983
Average of ClasserLeaf	Average of Mic	-0.4019	41	0.0092*
Average of Uniformity	% DP 555	-0.4131	37	0.0110*
Average of TrashArea	Average of Rd	-0.4334	41	0.0046*
Average of Rd	Average of PlusB	-0.4378	41	0.0042*
Average of PlusB	Average of Mic	-0.6858	41	<.0001*



Appendix O: Pairwise Correlations of % DP 555, Cotton Classification, and Ends Down Averages by Year by State

Pairwise Correlations

Variable	by Variable	Correlation	Count	Signif Prob
Average of ClasserLeaf	Average of TrashArea	0.7314	41	<.0001*
Average of Ends Down (2 day lag)	% DP 555	0.6190	37	<.0001*
Average of Rd	Average of Mic	0.4958	41	0.0010*
Average of Uniformity	Average of Length	0.4761	41	0.0017*
Average of ClasserLeaf	Average of Strength	0.4701	41	0.0019*
Average of Strength	Average of Length	0.4492	41	0.0032*
Average of TrashArea	Average of Length	0.4296	41	0.0051*
Average of ClasserLeaf	Average of PlusB	0.4024	41	0.0091*
Average of Ends Down (2 day lag)	Average of PlusB	0.3577	41	0.0217*
Average of TrashArea	Average of Strength	0.3560	41	0.0223*
Average of PlusB	Average of Strength	0.2805	41	0.0757
Average of ClasserLeaf	Average of Length	0.2689	41	0.0891
Average of Uniformity	Average of Strength	0.2646	41	0.0946
Average of Uniformity	Average of TrashArea	0.1711	41	0.2847
Average of Length	% DP 555	0.1692	37	0.3169
Average of TrashArea	Average of PlusB	0.1087	41	0.4986
Average of Rd	% DP 555	0.1015	37	0.5500
Average of Mic	% DP 555	0.0815	37	0.6314
Average of Uniformity	Average of PlusB	0.0532	41	0.7411
Average of PlusB	% DP 555	0.0234	37	0.8905
Average of ClasserLeaf	Average of Uniformity	0.0179	41	0.9117
Average of Strength	% DP 555	0.0046	37	0.9783
Average of Length	Average of Mic	-0.0234	41	0.8843
Average of Ends Down (2 day lag)	Average of Strength	-0.0365	41	0.8206
Average of Ends Down (2 day lag)	Average of ClasserLeaf	-0.0448	41	0.7811
Average of Rd	Average of Strength	-0.0813	41	0.6134
Average of ClasserLeaf	% DP 555	-0.0990	37	0.5597
Average of TrashArea	Average of Mic	-0.1171	41	0.4661
Average of Uniformity	Average of Rd	-0.1594	41	0.3195
Average of Uniformity	Average of Mic	-0.1680	41	0.2938
Average of Rd	Average of Length	-0.1709	41	0.2852
Average of Ends Down (2 day lag)	Average of Length	-0.1769	41	0.2685
Average of TrashArea	% DP 555	-0.1980	37	0.2402
Average of Strength	Average of Mic	-0.2079	41	0.1922
Average of Ends Down (2 day lag)	Average of Rd	-0.2272	41	0.1531
Average of ClasserLeaf	Average of Rd	-0.2310	41	0.1461
Average of Ends Down (2 day lag)	Average of TrashArea	-0.2481	41	0.1179
Average of PlusB	Average of Length	-0.2617	41	0.0983
Average of ClasserLeaf	Average of Mic	-0.4019	41	0.0092*
Average of Uniformity	% DP 555	-0.4131	37	0.0110*
Average of Ends Down (2 day lag)	Average of Uniformity	-0.4222	41	0.0060*
Average of TrashArea	Average of Rd	-0.4334	41	0.0046*
Average of Rd	Average of PlusB	-0.4378	41	0.0042*
Average of Ends Down (2 day lag)	Average of Mic	-0.4695	41	0.0020*
Average of PlusB	Average of Mic	-0.6858	41	<.0001*