United States Forest Service
Improving Management of Forest Cover
Modeling Study 100189521

Douglas Fraser

May 4, 2017

Contents

Executive Summary

The study described in this report has been conducted to investigate several data mining techniques to further serve the USFS's mission of managing America's wilderness resources. Data mining may be a popular buzzword, but it is not a brand new technology. It is merely a process involving the sophisticated use of statistics and models to discern patterns in raw data. From this, knowledge and information can be derived or inferred in order to provide better understanding for better decision making. Thus data mining can aid the USFS in more effective and efficient use of its resources, especially in a time of widespread government cutbacks for U.S. scientific agencies.

There were two main objectives to this study, which examined USFS data on forest cover (Region 2, the Rocky Mountain Region). The first involved developing models for predicting forest cover; each type of forest cover requires different management strategies and thus different allocations of resources. The second objective was to analyze specific areas where cottonwood and willow trees (Forest Cover Type 4) predominate for insights on how they might be managed better. This is needed due to the specialized requirements for this type of forest cover. Sections 2 and 3 describe in detail how these objectives were met, while section 1 is an overview of the study and how the forest cover data was first prepared for analysis.

To summarize, a decision tree model was developed that successfully predicted the type of forest cover that was present based on only 5 features: Elevation, H_Dist_Road, H_Dist_FP, Soil_Type and a combination of Shade_9 times V_Dist_Hyd. The accuracy of this decision tree model was from 91% to 93% for four of the cover types (CT1, CT2, CT3, and CT7); for the other three, it ranged from 81% to 84%. The overall accuracy on the validation set was 92.9%. These are far better results than the 70% accuracy achieved in a previous study.

The other type of classification algorithm investigated was clustering, but this was not as successful. Only the elevation of and area of wilderness a forest cell was in were shown to be useful with this type of model. The failure was due to the statistical properties of the other available data; there is a lot of commonality between most of the cover types along various axes, and clustering is not an ideal method under these conditions.

The second objective, that of determining what might be unique or notable about Forest Cover Type 4, was also successfully accomplished. First, the data supported theories about cottonwood and willow trees requiring a close water source and more sun than other types of cover. The other insights were that all the cells are fairly close to existing roads and to firepoints as well, which has logistical implications, e.g. the management of forest fires.

Based on the results of this study, the recommendation is that the USFS adopt data mining to use in fulfilling various its various responsibilities. Not only can the scientific research the USFS conducts be enhanced, the insights data mining offers are applicable to the various USFS programs for managing America's forest resources. The only recommendation going forward if this data mining program is to continue is that more data must be gathered for cells of cover types 3 to 7 and in the Neota and Cache la Poudre wilderness areas. This is because there is a surfeit of data related to cover types 1 and 2 and the other wilderness areas which impacts the data mining effort.

1 Preliminary Work

For this study, the Knowledge Discovery and Data Mining methodology was followed; the tools used were IBM's SPSS Modeler, RStudio, and Python software scripts as necessary. This section details the initial steps taken in any data mining project. First, the problem or questions were clarified, data initially analyzed, and a feasibility assessment made. As the project was deemed feasible and no other requirements were needed, the data preparation phase was then undertaken. Each step is explained below in more detail.

1.1 Problem Clarification

According to the RFP, the USFS would like to understand two things:

1. how existing data on forest cover can be analyzed to predict what type will arise in a specific area
2. if there are any patterns specifically related to Cover Type 4 (CT4) applicable for management strategies

These questions can be addressed by two data mining approaches: prediction and description. Prediction involves training machine learning classifiers on available data, and then evaluating how well these models can predict variables of interest. Section 2 describes in more detail how question 1 was investigated and the results. Question 2 required a description type of analysis to uncover information within the raw data about CT4; this phase is covered in section 3.

1.2 Data Analysis

A thorough understanding of the data is necessary before beginning any data mining effort. This assessment is to characterize the data in various ways and to gain an understanding of its statistical properties. By doing so, problems with the data, how to deal with them, and what other data might be required can be uncovered. Appendix A contains the data dictionary resulting from the analysis of the supplied forest cover data. This section highlights some conclusions about some fields in the data. The term 'field' is used more in the software development industry, with respect to databases. But in the data mining community, 'feature' is preferred and therefore is used in this report.

The preliminary analysis revealed that two sets of features (the four Wilderness_Area and the 40 Soil_Type ones) were an attempt to use a bitfield type of scheme to denote the value of something that is better expressed as a single categorical feature. This may be a carryover from the original study which investigated using neural networks for classification purposes. Details on how these features were reduced to two categoricals can be found in Appendix A. Normally this type of step would be done in the data processing phase, but immediately doing this addressed certain issues related to SPSS Modeler.

The analysis also revealed most records (85.22%) are associated with CT1 and CT2, while only 0.473% for CT4. This is a clear sign that the dataset must be balanced before dividing it into training, test, and validation. Otherwise, the likelihood of no training data related to CT4 being present would be high.

Tables A.2 and A.3 are the results of analyzing how the cover type relates to the different features, specifically the minimum and maximum for the continuous values and exactly which soil types are correlated with each cover type. This was to investigate if there were any obvious patterns or something noticeable in the data. The Python script 'breakdownData.py' (B.2) was written for this.

1.3 Feasibility Assessment

After the analysis had finished, the data was deemed fit for purpose and a determination made that no extra resources (e.g. more data, personnel, or computing resources) were required before further work could commence. This work consisted of cleaning the data and an initial pre-processing to reduce the dimensionality of the data for efficiency.

1.4 Data Cleaning

Data cleaning is a quality control process that only happens once, at the start of a data mining project, in order to regularize the data and to fix basic issues. These issues include: missing data, null values, formatting issues with the source data, and data values that are out of range or unexpected. None of these problems were detected with the Forest CoverType database, so data imputation was unnecessary and no records had to be deleted or modified.

Outliers in the data are also handled during the data cleaning phase. The benefit is that this deals with noise in the data and unusual values that might be errors or perhaps rare values that should be prevented from having an impact on the classification process. An example is clustering methods; they can be sensitive to outliers which might affect the initial starting conditions in a detrimental way. SPSS Modeler provides a function that tabulates outliers, but that display does not provide detailed information. Given the very low percentage of CT4 related records, histograms of the feature relative to a specific cover type were generated in order to understand how removing outliers might affect CT4. They also provided some insight into the utility and possible effects of removing outliers. Appendix B contains the useful charts. All of the continuous features contained outliers, while Slope, H_Dist_Hyd, and V_Dist_Hyd contained extremes. Extreme is the term SPSS uses for values greater than 5 standard deviations from the mean.

For Elevation, all of the outliers across all cover types are associated with CT7, and CT7's histogram implies they can be safely excluded. The same procedure of checking to see which CTs would be affected by removing extremes and outliers was performed for the other features. The conclusion was that except for Slope, all the extremes and outliers could be excluded without any consequences. Excluding the outliers of Slope is a bit uncertain, as a large number of records across the spectrum of cover types would be affected. So that decision was left open until it was better understood how Slope affected the modeling. Excluding some outliers of V_Dist_Hyd affected 24 records for CT4, but that was deemed acceptable. Eliminating outliers based on the IQR and not standard deviation was also examined; it generally would have resulted in more records being deleted, but would have had slightly more of an impact on a wider range of cover types. So this decision was reserved until the utility of eliminating outliers was established.

Another decision made during the project, once the problem became apparent, was to remove records with a Shade_3 value greater than 248. If Shade_3 was included in the stream in any way, its type would sometimes change to Typeless and the behavior of the software became inconsistent. The reason for this was finally traced to a limitation of SPSS which is that the maximum size for any set is 250, not 255. Setting Shade_3 to a type of continuous would solve this, but it isn't a continuous variable and using it as a nominal was desired in certain circumstances. This problem was not immediately apparent, nor where there any obvious warnings when the entire dataset was processed by reading from the file. Examining the histograms related to Shade 3 showed trimming it should not affect the data significantly.

1.5 Data Pre-processing

Data pre-processing operations (sampling, balancing, and dimensionality reduction) tend to be repeated throughout the KDD process as necessary depending on the current needs and status. For this study, sampling was investigated at the beginning of the process. But given the small amount of data and the fact that balancing was obviously necessary, there was no need to be concerned with reducing the amount of data for efficiency purposes. Section 2.1 contains the exact details on how and why balancing was done as it pertains directly to the process of building the decision tree. More advanced alternatives to balancing the dataset to deal with training the classifier properly were not investigated.

The Forest CoverType database is a small one, so the "curse of dimensionality" is not entirely applicable. But to be thorough and improve the modeling process, various ways of manipulating the features used were investigated as they still could improve accuracy. These ways include the following: feature construction, feature discretization, and feature selection. Feature construction has already been mentioned; the process of merging the 40 soil types to one feature is a classic example. Examining the other features revealed the

only other logical action to take would be to see how the three hillshade features might be merged into one; SPSS's correlation functionality returned Strong, Medium, and Weak relationships between the three. A weak correlation was unexpected, but this could be explained by variations in the terrain that affect the amount of sunlight at a certain spot. So merging Shade_3 and Shade_9 was considered, but left until further insight into the modeling process was acquired. The other two methods of reducing dimensionality, feature selection and feature discretization, are discussed in Section 2. Like balancing, they were part of the process of modeling the decision tree and investigating k-means clustering, so it is more appropriate to discuss them in context.

2 Predicting the Forest Cover

There is a wide variety of machine learning classifiers available and no definitive or absolute guidelines on how to construct them. Only two approaches were investigated in this study: decision trees and k-means clustering. The rationale for these choices, the process followed, and the details of features used and parameter settings are described in the following sections.

2.1 Decision Tree Models

Based on how there were noticeable boundaries and a few patterns in the numbers relative to each cover type for some features (tables A.2 and A.3), decision trees were a logical first choice as a classifier. The boundaries and patterns ought to aid the tree in making decisions, if the training goes well. The main benefit of decision trees is that the rules created are easily understood; trees are not black boxes like neural networks. They can also deal with different mixtures of feature types, e.g. all features don't have to be categorical ones. A disadvantage is that they can become big and complex, so there are two goals. The first is to find the right combination of as few as possible features that generates the smallest tree. The second goal is to do so without sacrificing an unacceptable amount of accuracy in its predictions.

The first decision made was how to partition the dataset into training, testing, and validation subsets. Three subsets were used instead of train plus test because the validation subset was used in the pruning process. Pruning a decision tree helps improve accuracy, combatting overfitting that is a characteristic of overly complex trees. A few different partitioning schemes were investigated (60-20-20, 70-15-15, 75-15-10), but there were no clear signs that one scheme was preferable to the others.

The next decision was to decide upon an algorithm. SPSS Modeler supports the following: CART, CHAID, and C5.0. Fortunately, SPSS also has a Auto Classifier node for investigating multiple models in a single run. However, CHAID consistently failed to build a model in these initial tests. For all, a setting of a maximum depth of 50 was used and pruning enabled, when applicable. The SPSS default pruning severity of 75% was used.

Three balancing schemes were used: oversampling (based on the minority class, CT4), undersampling (based on CT2), and a mixture. For the first two, the percentage increases suggested by SPSS for each cover type was used. This resulted in a training set composed of the same number of records for all cover types. But it was apparent that undersampling might not work well as it drastically reduced the number of CT1 and CT2 related records. This could result in an information loss impacting training - related patterns in the features might not be captured because no relevant records were selected. Therefore a mixture of under and oversampling was also used to investigate how a better, though still imbalanced, training set might affect results.

The baseline model built used all features but Cover_type; the C5.0 algorithm resulted in a model of 99.4% overall accuracy, while the CART algorithm only achieved 68.8%. But the depth of the C5.0 tree was 43 levels along with using 12 features. To see if outliers were a factor in the number of levels, a dataset with them removed was then used. The resulting tree was 40 levels deep, so outliers were not a significant factor relating to rules. But they still could be removed based on the preliminary data analysis and doing so reduced the amount of data to process to some extent.

Next, steps were taken to gauge the impact of a single feature when it was included or removed from the model. 12 data streams were built in a serial process, where one feature was removed and the previous one replaced. Other parameters were not varied during this process. Removal of Elevation, H_Dist_Road, and H_Dist_FP showed a noticeable drop in accuracy relative to the shifts for other features. The effects of removing H_Dist_Hyd and Soil_Type also were noticeable, but slightly less. Removing other features did not have an appreciable effect, or in the case of CART, improved accuracy.

The effect of different balancing schemes was then investigated. As suspected, when cover types other than CT4 were undersampled, accuracy dropped to around 90% for all variations of C5.0 and around 68% for the CART algorithm. The same general effect of removing each feature one by one was also seen in subsequent tests, but the drops were more significant than when the oversampling balancing configuration was used. The third balancing configuration was a more complex one designed to undersample just CT1 and CT2 while oversampling the other classes in order to achieve a more equitable balance. Overall, the results were much the same as for the first series of tests - accuracy dropped a little from the baseline, to 98% which is better than for balancing run 2, and removal of the same features showed the same effects. The number of levels of the tree tended to drop as well, compared to the first set of runs. This indicates information was being lost when the CT1 and CT2 classes were too undersampled, so a balancing scheme that oversampled CT4 was used from this point forward.

The next decision was to use a forward selection process to determine the smallest set of features that resulted in acceptable accuracy. Starting with a base set of just Elevation, other features were added one at a time to the current set to determine how accuracy, on the test set, might improve. The feature resulting in the largest improvement was then added and the loop started over. The final feature set was Elevation, H_Dist_Road, H_Dist_FP and Soil_Type. To add another feature beyond this point would only result in an improvement in testing of less than 2%, which wouldn't be worth the cost of making the model more complex.

Whether to discretize the continuous features was the next decision. After some initial experimentation, there was a realization that it would not be generally beneficial. This is based on an interpretation of the histograms and boxplots. The extent of values for each cover type, for each feature, overlap noticeably. If the method of discretization was not chosen carefully, then the maximum or minimum value for a feature for a specific cover type could be placed in the same 'block' as values for another cover type. Because the decision tree can not divide this block any further, classification errors for some of the records would then be certain. Discretization in essence reduces the resolution; this will prevent the decision tree from making a split that it normally could on continuous data. If discretization was essential, then customized schemes or methods based on the statistical properties of the feature values would be required. SPSS does not provide such a feature and more extensive data analysis would be needed as well.

Examining the statistics of the results in detail revealed a noticeable variation in the accuracy levels per cover type in testing. The percentage ranged from 72% (CT4) to 93% (CT7), so misclassification costs were enabled for three cover types to see if they could improve results. Weights of 5 to 10 were set, iteratively, for misclassification of CT4, CT5, and CT6, but this did not result in much improvement. CT4 was persistently misclassified as CT3 or CT6; the boxplots are evidence how the features values associated with CT4 are frequently a subset of CT3 or CT6 values. Of the features left, adding Shade_9 resulted in the greatest increase in accuracy (78%).

But given the baseline model's accuracy for classifying CT4 was 92%, this lead to the conclusion a composite feature might be the solution to improving accuracy as needed for specific cover types. By encapsulating the information of several features within one, the cost of adding more features would be reduced while increasing accuracy. Various mathematical formulas were used to combine Shade_9, Aspect, Wilderness_Area, V_Dist_Hyd, and H_Dist_Hyd. The simplest feature combination was Shade_9 times V_Dist_Hyd, which improved the accuracy rate for the three cover types to around 84%. The partitioning scheme was then changed from 60-20-20 to 70-15-15, too see if more training might be useful. The overall accuracy not only increased to 95.8%, but the accuracy rate for individual cover types all improved such that they were from 92% to 97%. It was decided at this point the model had been sufficiently tuned to acceptable levels.

10-fold cross validation was now enabled so that a variety of training, test and validation sets could be used to evaluate the generalization error of the model. Unfortunately the previous results did not persist; the accuracy rate for the three cover types in question dropped back to around 81% to 84%. The accuracy for the four other cover types stayed at around 93% to 96%. These results were deemed more than sufficient however, so work on the next approach began. Throughout the entire process, the number of levels of each decision tree built was monitored. It ranged from 38 to 43; there was no specific settings change that greatly or consistently improved this measure. As for more detailed information about the tree, every attempt to view it in SPSS and the information on how branches were created failed. SPSS Modeler would typically become unresponsive. This may be due in part to the fact SPSS Modeler was running on Windows 10 in a virtual machine, but there was sufficient memory and processing power that this should not have been a concern.

2.2 Clustering

The other approach taken to investigate predicting forest cover types was clustering; in general, the algorithm works by grouping data points and taking measures of how far apart they are from others in the same cluster. More cohesive clusters imply the data points in that cluster are more similar to each other than to other clusters'. On what basis the cluster was created (the features used) then provides some insight into the data.

K-means clustering is a version that requires just one main parameter: k, the number of clusters. The downside of clustering is that it is a random process, so many attempts to find a suitable set of clusters may be necessary. It is also not guaranteed to return useful results. Eliminating outliers in the data can be beneficial in preventing issues; their presence will affect results if they are used for the initial starting points of the clustering process. After outliers were eliminated, the next pre-processing step taken was to discretize the continuous features such as Elevation. The techniques for doing so that were used were equal width, binning by number of standard deviations, and binning based on the target feature (the cover type). Different techniques were used for different features as they have different distributions relative to the cover type. Discretization provides more of an advantage in clustering than with decision trees; the "resolution" of the data is lessened and so theoretically a decision boundary would be easier to determine. Accuracy in determining the cover type for each data point is not a goal, but finding how features best map to cover type is and so precision is not as important.

As clustering involves repeated iterations, the next decision made was to reduce the amount of data examined by using only 10% of it. The motivation was to hopefully find a set of features that resulted in good performance, i.e. clusters were found that closely corresponded to one or more of the forest cover types. The assumption was this feature set should still result in good performance when the entire dataset was used. Another decision was that since the k-means algorithm is affected by class imbalance (it assumes clusters are the roughly the same size), there might be a need to balance the dataset. So both the unbalanced dataset and datasets balanced on each cover type were used.

The final variable was that there is no definitive algorithm for determining the best value of k. Instead, there are heuristic methods that always involve repeated iterations while evaluating some measure of quality. SPSS Modeler reports the silhouette of the clustering, so this was used as the metric for evaluating performance. It is a measure of how closely a data point matches to data within the same cluster and not to data not in the cluster. Higher values closer to 1 imply more data has been placed in the correct clusters. However, the question was is there any way to create a cluster (i.e. a set of features) that fully describe a cover type versus the other types? This is different than trying to find the best set of clusters for all the data. So the silhouette measure was just used as feedback.

A baseline was established with these settings: 10% of the data, unbalanced, continuous features discretized by equal width, and all features used. The values of k varied from 2 to 15. Examining the graphics SPSS provides revealed that of the input features ranked in importance, only Elevation and the Wilderness Area had a distinct delineation in how values were associated with a cluster; the spread of the other features was across the entire spectrum for both clusters.

This lead to examining the histograms and boxplots in Appendix B for possible insights. If each feature is metaphorically thought of as a 'dimension', then the graph of how a particular CT value is spread across the range of that dimension could be used to gauge what features would actually be useful in clustering. For instance, if a particular CT is only within a certain band of values, then there is a clear boundary line between 'that' and 'not that'. A CT that is spread across the entire range means that dimension is not going to be useful for discriminating between that CT and another CT. This can be easily seen in examining the graphs for Aspect; all CTs are spread across the entire range of Aspect. Therefore it will be of no value in clustering when trying to understand the relationships between CT and the other features. This insight explains why Wilderness Area and Elevation were useful. Figure A.13 shows how CT4 is associated with only one Area (WA4) and CTs 5, 6, and 7 are in just two Areas. Examining the Elevation graphs (Figures B.1 and B.6) shows a lot of overlap in terms of the spread of a CTs, but there are some definite boundaries: CT4 is only between 2000 and 2500 while CT5 is from 2500 to 3000. So clustering that uses Elevation would or could possibly put CT4 and CT5 into two distinct clusters.

Based on this intuition, the graphs in Appendix B were analyzed further. Elevation, Wilderness Area, Soil Type, and H_Dist_Road were chosen as the subset of features to use in the next iteration. The SPSS stream was reconfigured; no other settings were changed. The silhouette measure of the best model jumped from 0.224 to 0.619 (4 clusters were found). Examining the details revealed the only features used were Elevation and Wilderness Area. However, the reduction in dimensionality from 13 features to 2 should be considered a factor as well as the selection of useful features. Fewer dimensions allows the boundary of a cluster to be more definitive; there are less opportunities for exceptions in the feature values to throw off the assignment of points to a cluster.

SPSS has facilities for graphing the results from a model in different ways; using this functionality revealed that all the data could be grouped into four clusters - Wilderness Areas 1 and 3 were associated with the middle set of Elevation values while Wilderness Areas 4 and 2 were associated with the lower and higher Elevations respectively (see Figures B.16 to B.23). But the Cover Types were still located in multiple clusters (Figure B.24) except for CT4 which is exclusively in Wilderness Area 4.

Also at this point, the importance of a balanced dataset and the use of more data became apparent. A balanced dataset would change the percentage of records associated with a specific cover type and so affect the distribution or density of points that ideally would be placed in the same cluster. Because the data is predominantly CT1 and CT2 data, a cluster that defines CT3 well, for instance, could get overlapped by a larger cluster. Hierarchical clustering may be a way to investigate this hypothesis. As for more data, it might improve the likelihood of data points with the right combination of feature values being found. These data points would better establish a cluster boundary in the multidimensional feature space being examined.

But more data may cause issues as well. This can be seen in one run that used a balanced dataset based on CT4. Figure B.25 shows how several CTs were entirely within one cluster. 9 features were used. But when the entire dataset was used, the CTs were spread across the different clusters. Only the unbalanced dataset was used thereafter as the goal was to find a set of features that could reliably describe all data, not just a subset of it.

After many iterations, the conclusion was only CT4 could be associated with a single cluster when Wilderness Area was used as a feature. CT5 and CT6 frequently ended up in just two clusters; Figure B.24 is an example of typical results. This can be explained by examining table A.3 which shows how Wilderness Area can be a strong determinate of how records are assigned to a cluster. Elevation, as explained, also was a strong determinate of how CTs might be associated with a cluster. If Wilderness Area and Elevation were left out from the feature set, then the models generated may have silhouette measures up to 0.6 but the distribution of cover types amongst the clusters showed no clear biases. These is because apparently no combination of other features can mark a clear boundary around the entire set of records for a particular cover type. To test this theory further, CT1 and CT2 records were removed from the dataset and a clustering analysis done to see if any set of features would result in the other cover types clustering in the desired fashion. But again, they were always split amongst the various clusters except for CT4.

3 Analysis of Cottonwood & Willow Forest Cover

Areas with Cottonwood and Willow Forest Cover (CT4) require specialized management, so another goal of this study was to determine if there were any patterns in the data specific to CT4. If so, these patterns could inform the USFS about ways to improve management strategies or allocation of resources.

The first approach taken was to examine the histograms in Appendix B for any obvious patterns relative to CT4. It was evident that CT4 is exclusively associated with Wilderness Area 4 (WA4, Cache la Poudre). So any changes in management should concern itself with the department that manages that area. Another obvious conclusion is that more data for CT4 cells should be gathered, if possible; more data on WA4 would help with future analysis as points for WA1 and WA2 dominate the dataset.

CT4 was also predominately associated with a H_Dist_Hyd value under 50 feet, and there was also a strong bias for values of V_Dist_Hyd to be under 25 feet. These imply cottonwoods and willows require or prefer easy access to water. A Google search for information on these trees confirmed this hypothesis. It also revealed cottonwoods and willows prefer more sun, which can be somewhat inferred from the histograms for the Shade features for CT4. The Aspect histogram also reveals a bias for CT4 cells to be oriented in an easterly direction (from 75 to 150 degrees). This meshes with the theory that CT4 cells receive more sunlight. Finally, CT4 cells are very predominantly under 1500 feet from a road and CT4 cells are the closest in terms of the maximum distance from a firepoint which should be noted. These inferences were verified by examining the web diagrams relating Cover Type to other features. Figure B.6 is an example that shows how CT4 is predominantly related to Aspect Bin 2 and 3 (bin width was 75 degrees).

The next approach taken to find any more complex patterns was the use of association rules. Association rules work with only categorical features, so the continuous features were binned as they had been for the clustering analysis. Despite numerous runs, no association rules were found using the Apriori functionality offered by SPSS. Upon analysis, the reason for this became evident. The rules being looked for are of the "if X, then CT is 4" variety. They may exist, but if the chances X leads to other cover types relatively as much, then the confidence level of that rule will be impacted. The support for rules pertaining to CT4 will also be fairly small given how little data there is for CT4 compared to CT1 and CT2.

The third approach taken was to use a decision tree to deduce rules specific to CT4. The initial conditions used were: continuous features were binned, partition of 80% training, 20% test, dataset balanced by oversampling CT4, all features selected. The difference between this phase and the first is that a decision tree that can classify all cover types reliably was not the goal; instead, accurate rules that relate features to CT4 were desired. So a decision tree that misclassifies other cover types is acceptable as long as the accuracy of classifying CT4 cells correctly appreciates. Therefore, the misclassification costs for incorrectly classifying CT4 were increased to 5.0. This initial baseline resulted in a model that had 266 rules, and was 80.9% accurate in classifying CT4. Only CT3 and CT6 cells were misclassified.

The first sets of tests investigated the impact of binning and the relative importance of the continuous features. Replacing each feature with its unbinned version, one by one, revealed the model was most sensitive to a change in Elevation; the CT4 specific accuracy improved from 80.9% to 81.8%. But unbinned version of the other features did not improve accuracy substantially or decreased it slightly. For every model, only CT3 and CT6 cells were misclassified; examining the histogram again lead to the conclusion Slope and the Shade features could be the reason for the misclassification.

The next set of tests was to determine if the feature set size could be decreased. A backwards selection process was used; each loop consisted of building multiple models in parallel where only one feature was removed from the baseline. The model that had the highest performance was chosen as the baseline for the next loop and the process stopped when there were no improvements. Elevation, Wilderness Area and Soil Type were not removed in this process; they are strongly correlated to CT4 as preliminary investigations had shown.

The following is a list of the features removed (in order) and the resulting improvement in accuracy: (Shade_3, 82.4%), (Slope, 83%), and (Shade_9, 84.9%). The features left (Aspect, Shade12, H_Dist_Hyd, V_Dist_Hyd,

H_Dist_Road, and H_Dist_FP) were all marked in the initial examination as probably being relevant to CT4. So it is not surprising they are in the final set. A check was done by removing Soil_Type and Wild_Area from the feature set; the accuracies decreased as expected (79.8% and 81.4%).

The next step taken was to increase the misclassification costs to 10.0; this was to see if the persistent failure to classify some CT3 and CT6 data could be changed. The accuracy dropped to 84.63%; raising or lowering the cost did not have any appreciable positive effect. Alternative options such as varying how features were discretized were not investigated further; the relationship between CT4 and features had been established, leading to reasonable conclusions about management strategies. The exact set of rules that the decision tree created would be of no use towards this end, especially considering they would change for every new model.

4 Recommendations and Conclusions

Based on the results, decision trees are a viable method for predicting the type of forest cover and for describing patterns in the data that can inform management decisions. An accuracy of up to 93% was achieved for the separate cover types. This is a large improvement on the previous study, which used neural networks, achieving only an accuracy of 70%. The features found to be useful for this type of model are: Elevation, H_Dist_Road, H_Dist_FP, Soil_Type and a combination of Shade_9 times V_Dist_Hyd. Combining the two features was a key improvement that allowed three cover types to be classified more accurately. Further feature engineering should be investigated given how cover types overlap in terms of the range of values for different features.

As for patterns related to Cover Type 4, several were found. First, cottonwood & willow forest cover is exclusively in the Cache la Poudre wilderness area. The department responsible for that area should be tasked with improving its management strategies. Cottonwood & willow trees tend to be located close to water sources and to be close to access roads, so the logistical aspects of managing these cells should also take that into account. Finally, the more complex patterns in the data associated with CT4 were a combination of the horizontal and vertical distance to a water source, the distance to a road, and the distance to the nearest firepoint. The firepoint distance turned out to be a important factor that distinguishes CT3 and CT6 from CT4; CT4 is the closest type of cell in that its maximum distance is the minimum amongst all cover types. This implies that cottonwood and willow cells are more at risk to wildfire and so appropriate measures should be taken to forestall that.

The other approach investigated, that of clustering, was not as successful; clusters consisting of just one cover type could not be created. This is due to the amount of overlap in the extent of a feature's range that each cover type has, relative to other cover types. Because of this overlap, clustering could not determine a definite decision boundary and so data for multiple cover types were placed in the same cluster. However, Wilderness Area and Elevation were shown to be somewhat useful for distinguishing one subset of cover types from the other subset. Another reason for this failure is the class imbalance, e.g. the amount of data related to CT1 and CT2 is excessive compared to the other CTs. CT1 and CT2 data also is very spread across the whole extent of several features. Until more data is gathered to reduce the class imbalance, clustering should not be considered as useful as decision trees.

The first recommendation for future research is that more data should be collected on cover types other than CT1 and CT2. The quality of the existing dataset is more than adequate as it is very clean, with no problems, and with a low percentage of outliers. But a more equitable class balance in the dataset will only assist in future studies. The second is that other possible features relevant to forest cover type should be researched. If a feature is found that more clearly can delineate one cover type from the others, or a subset, then future data mining efforts will be even more precise and accurate.

A Data Dictionary

Two files were provided for this study: 'covtype.info' and 'covtype.csv'. 'covtype.info' describes the contents of the 'covtype.csv' database and is summarized here for the reader; specific details on the original study that used this data can be found there. 'covtype.info' contains most of a data dictionary, but the details were double checked and expanded on to be thorough in this investigation.

Name of Database: Forest CoverType data

Copyright Information

Original owners of database

Remote Sensing and GIS Program
Department of Forest Sciences
College of Natural Resources
Colorado State University
Fort Collins, CO 80523

Contact
Jock A. Blackard <jblackard@fs.fed.us> or
Dr. Denis J. Dean <denis.dean@utdallas.edu>
for further information

Database publicly released August 1998 by

Jock A. Blackard <jblackard@fs.fed.us>
GIS Coordinator
USFS - Forest Inventory & Analysis
Rocky Mountain Research Station
507 25th Street
Ogden, UT 84401

Dr. Denis J. Dean <denis.dean@utdallas.edu>
Professor
Program in Geography and Geospatial Sciences
School of Economic, Political and Policy Sciences
800 West Campbell Rd
Richardson, TX 75080-3021

Dr. Charles W. Anderson <anderson@cs.colostate.edu>
Associate Professor
Department of Computer Science
Colorado State University
Fort Collins, CO 80523 USA

This database contains data about four wilderness areas located in the Roosevelt National Forest of northern Colorado. The forest cover type for the 30m by 30m cells was obtained from the Region 2 RIS. Other data was obtained from US Geological Survey and USFS databases.

There were 581,012 records in the database.

1 Summary of Features

Total Number of Features: 55

Table A.1: Features of the Forest CoverType Database

Feature	Type	Units	Description
Elevation	continuous	meters	ground elevation
Aspect	continuous	azimuth	aspect in degrees
Slope	continuous	degrees	slope of surface
H_Dist_Hyd	continuous	meters	horizontal distance to nearest surface water features
V_Dist_Hyd	continuous	meters	vertical distance to nearest surface water features
H_Dist_Road	continuous	meters	horizontal distance to nearest roadway
Shade_9	ordinal	0 to 254	hillshade index at 9am, summer solstice
Shade_12	ordinal	0 to 254	hillshade index at noon, summer solstice
Shade_3	ordinal	0 to 254	hillshade index at 3pm, summer solstice
H_Dist_FP	continuous	meters	horizontal distance to nearest wildfire ignition points
Wild_A_1	ordinal	0 or 1	
Wild_A_2	ordinal	0 or 1	binary variables indicating which of the 4 wilderness areas
Wild_A_3	ordinal	0 or 1	the cell is present in - only one should be set per record
Wild_A_4	ordinal	0 or 1	
ST_1			
through	ordinal	0 or 1	40 binary variables indicating the presence of a particular soil type
ST_40			
Cover_type	categorical	1 to 7	forest cover type designation (the feature to be predicted)

2 Issues

The original dataset is a somewhat large one, about 75 MB, and IBM SPSS Modeler v15 has software bugs rendering it incapable of handling this amount. Once it was determined the four Wilderness_Area and 40 Soil_Type features were a type of bitfield encoding, steps were taken to reduce these to just two categorical features (Wilderness_Areas and Soil_Types). A short Python program (B.1) was written to process the original CSV file and verify that for each record in the database, only 1 of the four Wilderness_Area fields was set, as well as only 1 for the 40 Soil_Type fields. No instances of all zeros were detected either or records with null or invalid values. The Python program then transformed the records to reduce these features into the two new categoricals (where the value was the index number of the original field that was set to 1). The resulting dataset is about 26 MB.

After the derived dataset was loaded into SPSS Modeler, its functionality was used to quickly determine if there were null or missing data, problematic records, or unusual or invalid values present. No other basic quality related problems were found.

3 Feature Information

3.1 Cover_type

The type of forest cover the cell has been mapped to.

Type: integer, categorical

Category	Tree Type	Occurrences	% of total
1	Spruce/Fir	211840	36.46
2	Lodgepole Pine	283301	48.76
3	Ponderosa Pine	35754	6.15
4	Cottonwood/Willow	2747	0.473
5	Aspen	9493	1.63
6	Douglas-fir	17367	2.99
7	Krummholz	20510	3.53

An important thing to note is the fact most of the data is associated with CT1 and CT2, while CT4 is very underrepresented. This is a clear sign of a need to balance the data properly in the training set, otherwise random sampling procedures may not select enough data associated with CT4 for adequate training.

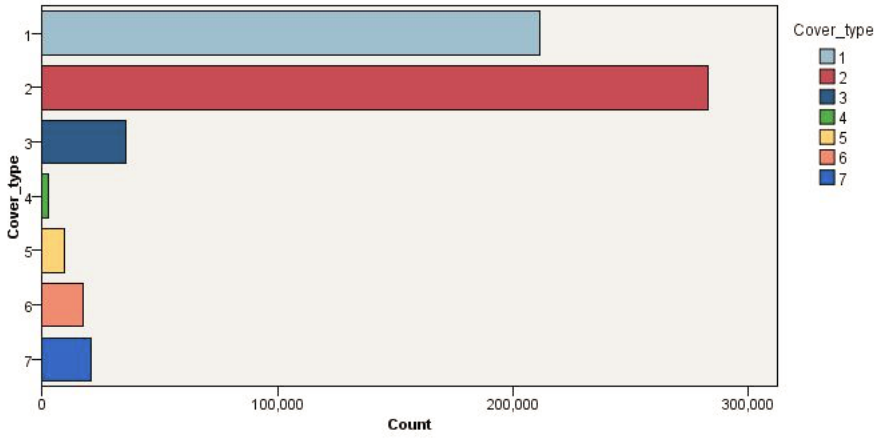

Figure A.1: Bar chart of Cover Type

Table A.2: Soil Types associated with Cover Types

Cover Type	Soil Types Present
1	4,8,9,10,11,12,13,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31
2	2,3,4,6,7,8,9,10,11,12,13,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31
3	1,2,3,4,5,6,10,11,13,14,16,17,20
4	1,2,3,4,5,6,10,11,14,16,17
5	2,4,10,11,13,16,17,18,19,20,23,24,26,28,29,30,31
6	1,2,3,4,5,6,10,11,13,14,15,16,17,20,23,24,31
7	4,13,19,21,22,23,24,27,29,30,31

Table A.3: Minimum and Maximum Values of Features per Cover Type

Feature	CT1	CT2	CT3	CT4	CT5	CT6	CT7
Elevation	(2466,3686)	(2142,3433)	(1859,2899)	(1988,2526)	(2482,3011)	(1863,2900)	(2868,3858)
Aspect	(0,360)	(0,360)	(0,360)	(0,359)	(0,359)	(0,360)	(0,360)
Slope	(0,56)	(0,66)	(0,50)	(0,46)	(0,51)	(0,54)	(0,51)
H_Dist_Hyd	(0,1200)	(0,1397)	(0,726)	(0,551)	(0,1100)	(0,644)	(0,1323)
V_Dist_Hyd	(-156,431)	(-173,601)	(-134,312)	(-25,270)	(-134,265)	(-126,288)	(-84,412)
H_Dist_Road	(0,6632)	(0,7117)	(0,3436)	(67,1702)	(30,5206)	(0,3092)	(451,5463)
Shade_9	(0,254)	(0,254)	(46,254)	(127,254)	(126,254)	(0,254)	(80,254)
Shade_12	(74,254)	(0,254)	(93,254)	(137,254)	(95,254)	(90,254)	(98,254)
Shade_3	(0,254)	(0,254)	(0,251)	(0,232)	(0,236)	(0,248)	(0,229)
H_Dist_FP	(0,7118)	(0,7173)	(0,2888)	(0,1921)	(42,6321)	(0,2940)	(0,4589)
Wild_A_1	(0,1)	(0,1)	(0,0)	(0,0)	(0,1)	(0,0)	(0,1)
Wild_A_2	(0,1)	(0,1)	(0,0)	(0,0)	(0,0)	(0,0)	(0,1)
Wild_A_3	(0,1)	(0,1)	(0,1)	(0,0)	(0,1)	(0,1)	(0,1)
Wild_A_4	(0,0)	(0,1)	(0,1)	(1,1)	(0,0)	(0,1)	(0,0)

3.2 Elevation

Elevation is the evaluation of the ground in meters, sea level assumed to be 0. Given the initial histogram, the extent of the elevation range for each cover type was determined, revealing definite limits for types 3 to 7 (see Figure B.1). This implies elevation will be useful as a feature.

Type: integer, continuous

Statistic	Value
Minimum	1859
1st Quintile	2809
Mean	2959.36
Median	2996
Mode	2968
3rd Quintile	3163
Maximum	3858
Std Deviation	279.98

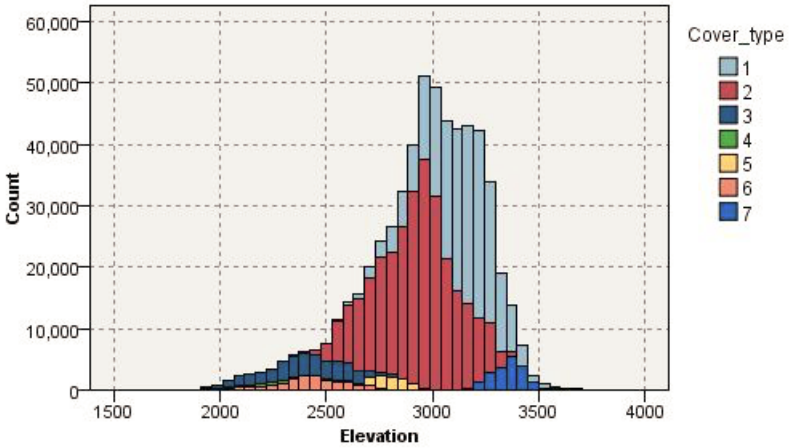

Figure A.2: Histogram of Elevation Per Cover Type

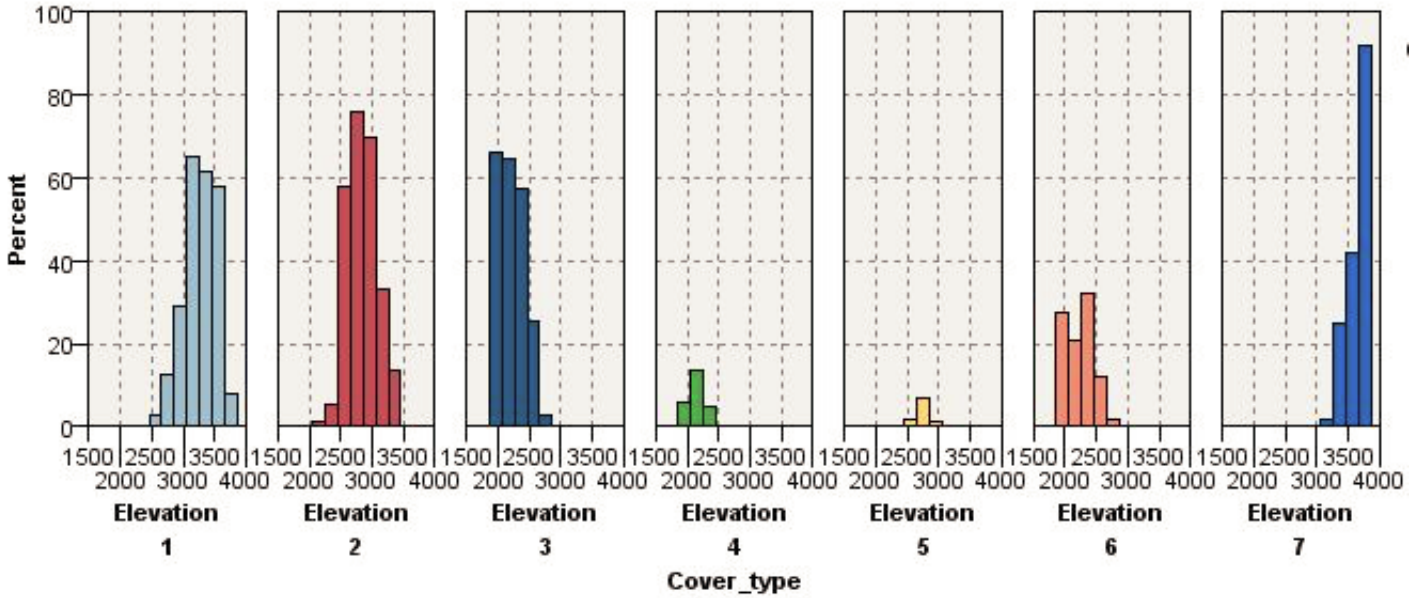

Figure A.3: Elevation Extents Per Cover Type

3.3 Aspect

Aspect is the compass direction that the ground faces, thus measured in degrees azimuth. The aspect can have a strong influence on ground temperature and thus the microclimate of that area.

Type: integer, continuous

Statistic	Value
Minimum	0
1st Quintile	58
Mean	155.66
Median	127
Mode	45
3rd Quintile	260
Maximum	360
Std Deviation	111.91

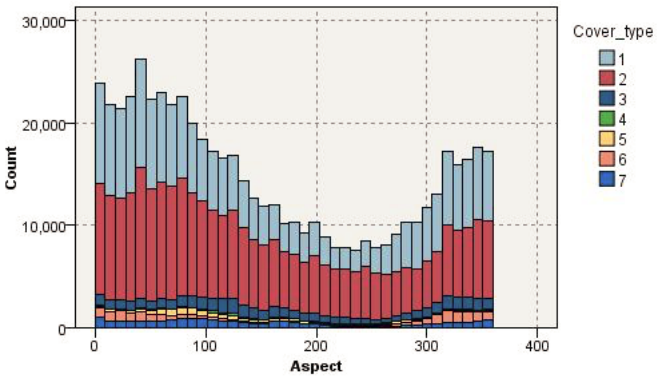

Figure A.4: Histogram of Aspect Per Cover Type

3.4 Slope

The ground slope in degrees. The appearance of periodic gaps in the data should be noted.

Type: integer, continuous

Statistic	Value
Minimum	0
1st Quintile	9
Mean	14.10
Median	13
Mode	11
3rd Quintile	18
Maximum	66
Std Deviation	7.49

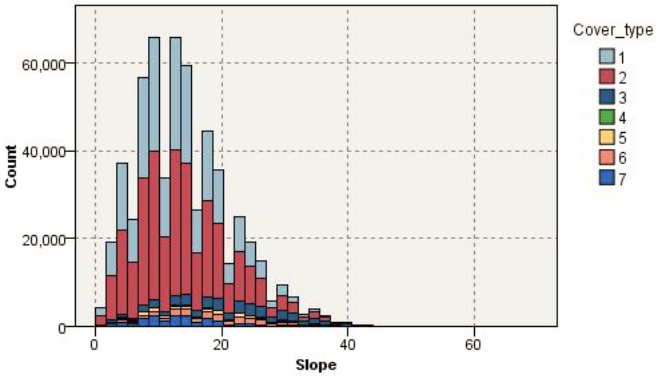

Figure A.5: Histogram of Slope Per Cover Type

3.5 Horizontal_Distance_To_Hydrology (H_Dist_Hyd)

The horizontal distance measured, in meters, to the nearest surface water features. It is not stated, but was assumed to be from the center of the cell.

Type: integer, continuous

Statistic	Value
Minimum	0
1st Quintile	108
Mean	269.43
Median	218
Mode	30
3rd Quintile	384
Maximum	1397
Std Deviation	212.55

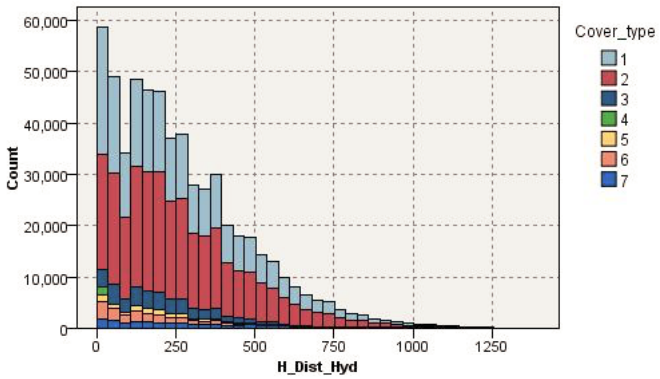

Figure A.6: Histogram of Horizontal Distance to Hydrology Per Cover Type

3.6 Vertical_Distance_To_Hydrology (V_Dist_Hyd)

The vertical distance measured, in meters, to the nearest surface water features. It is not stated, but was assumed to be from the center of the cell.

Type: integer, continuous

Statistic	Value
Minimum	-173
1st Quintile	7
Mean	46.42
Median	30
Mode	0
3rd Quintile	69
Maximum	601
Std Deviation	58.3

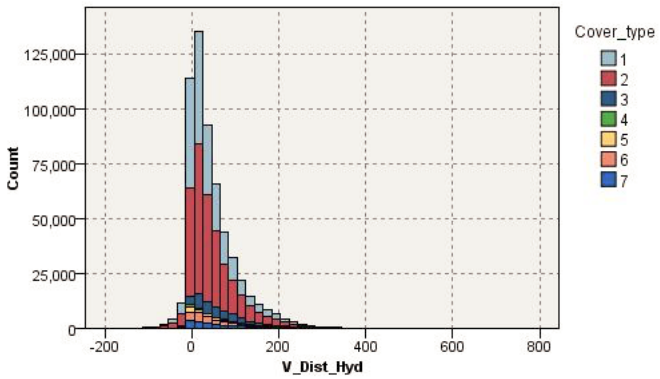

Figure A.7: Histogram of Vertical Distance to Hydrology Per Cover Type

3.7 Horizontal_Distance_To_Roadways (H_Dist_Road)

The horizontal distance measured, in meters, to the nearest roadway. It is not stated, but was assumed to be from the center of the cell.

Type: integer, continuous

Statistic	Value
Minimum	0
1st Quintile	1106
Mean	2350.15
Median	1997
Mode	150
3rd Quintile	3328
Maximum	7117
Std Deviation	1559.26

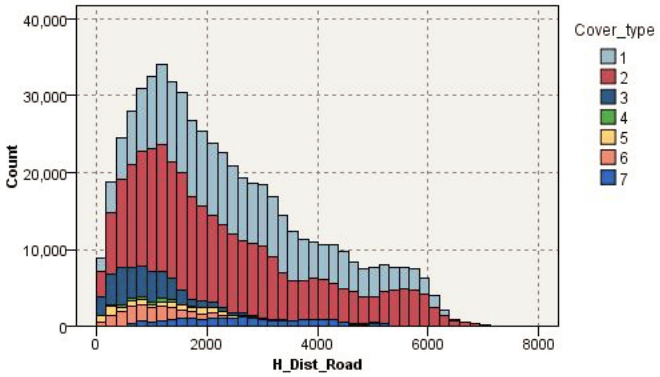

Figure A.8: Histogram of Horizontal Distance to Road Per Cover Type

3.8 Hillshade_9am (Shade_9)

An index from 0 to 255 reflecting the amount of shade at 9am on the summer solstice, i.e. the amount of illumination. It is not stated how this abstract number relates to the actual amount of sunlight, but presumably higher values correlate to more sunlight.

Type: integer, ordinal

Statistic	Value
Minimum	0
1st Quintile	198
Mean	212.15
Median	218
Mode	226
3rd Quintile	231
Maximum	254
Std Deviation	26.77

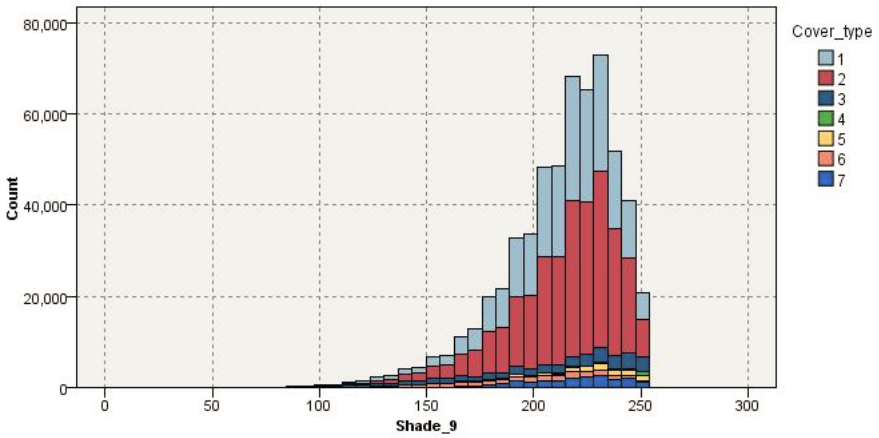

Figure A.9: Histogram of 9AM Shade Index Per Cover Type

3.9 Hillshade_Noon (Shade_12)

An index from 0 to 255 reflecting the amount of shade at noon on the summer solstice, i.e. the amount of illumination. It is not stated how this abstract number relates to the actual amount of sunlight, but presumably higher values correlate to more sunlight.

Type: integer, ordinal

Statistic	Value
Minimum	0
1st Quintile	213
Mean	223.32
Median	226
Mode	228
3rd Quintile	237
Maximum	254
Std Deviation	19.77

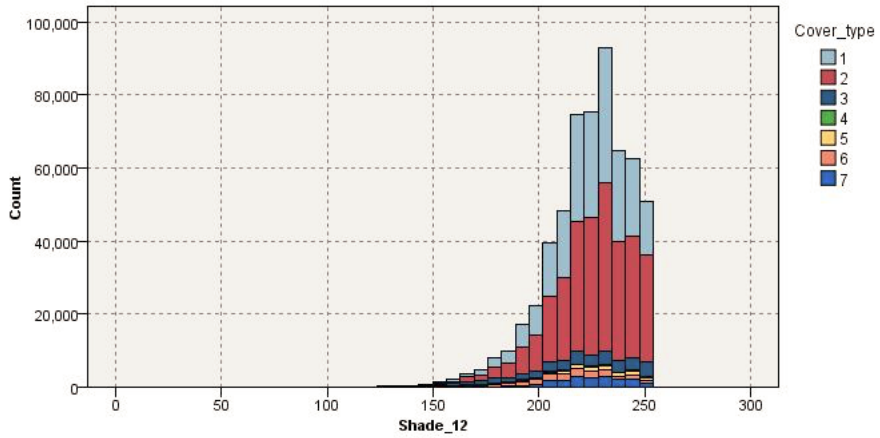

Figure A.10: Histogram of Noon Shade Index Per Cover Type

3.10 Hillshade_3pm (Shade_3)

An index from 0 to 255 reflecting the amount of shade at 3pm on the summer solstice, i.e. the amount of illumination. It is not stated how this abstract number relates to the actual amount of sunlight, but presumably higher values correlate to more sunlight. Unlike the other two Shade features, this shows a normal type

distribution. Given all 3 Shade features are related by some formula involving the angle of the sun and the ground aspect, they are a candidate for feature reduction in some way.

Type: integer, ordinal

Statistic	Value
Minimum	0
1st Quintile	119
Mean	142.53
Median	143
Mode	143
3rd Quintile	168
Maximum	254
Std Deviation	38.27

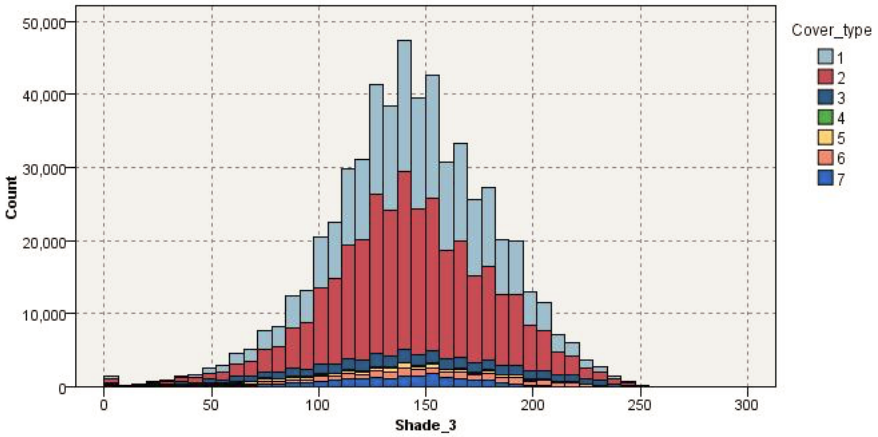

Figure A.11: Histogram of 3PM Shade Index Per Cover Type

3.11 Horizontal_Distance_To_Fire_Points (H_Dist_FP)

The horizontal distance measured, in meters, to the nearest wildfire ignition point. It is not stated, but was assumed to be from the center of the cell.

Type: integer, continuous

Statistic	Value
Minimum	0
1st Quintile	1024
Mean	1980.29
Median	1710
Mode	618
3rd Quintile	2550
Maximum	7173
Std Deviation	1324.2

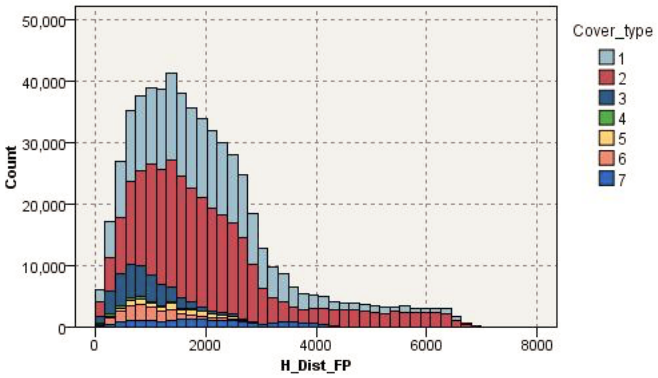

Figure A.12: Histogram of Horizontal Distance to Firepoint Per Cover Type

3.12 Wilderness_Area 1 to 4 (Wild_A_1 to Wild_A_4)

These four features categorize what wilderness area the cell is part of.

Type: binary value, 0 denoting the cell is not within the specified wilderness area, and 1 denoting that it is

There were no rows containing more than one 1 in these columns, or missing or invalid values, therefore the features were transformed into a single categorical for the purpose of data reduction. An important thing to note is that Cover Type 4 is only present within Area 4 which has implications for sampling and balancing. Cover types 5 and 7 are also associated with only 2 Areas, so this will be a useful feature.

Area Number	Description	Occurrences	% of total
1	Rawah Wilderness Area	260796	44.88
2	Neota Wilderness Area	29884	5.14
3	Comanche Peak Wilderness Area	253364	43.6
4	Cache la Poudre Wilderness Area	36968	6.36

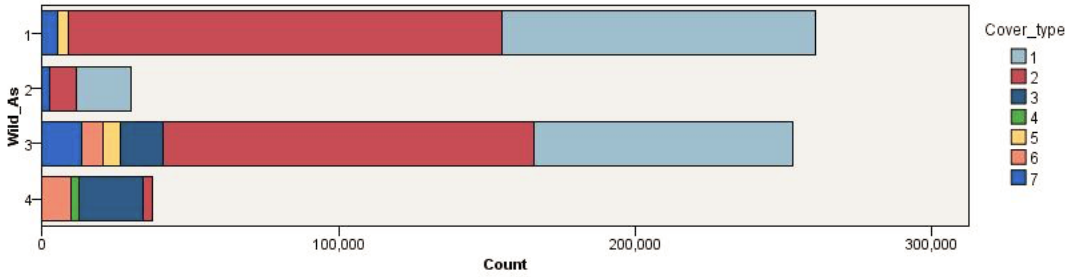

Figure A.13: Bar chart of Wilderness Area Types Per Cover Type

3.13 SoilType 1 to 40 (ST_1 to ST_40)

These 40 features indicate the type of soil (USFS Ecological Landtype Units (ELUs)). No cell had multiple soil types associated with it so these features were transformed into a single categorical. No clear patterns to the data immediately stand out.

Type: binary value, 0 denoting the cell is not associated with that soil type, and 1 if it is

Soil Type	Occurrences	% of total	Soil Type	Occurrences	% of total
1	3031	0.521	21	838	0.144
2	7525	1.295	22	33373	5.743
3	4823	0.830	23	57752	9.939
4	12396	2.133	24	21278	3.662
5	1597	0.274	25	474	0.081
6	6575	1.131	26	2589	0.445
7	105	0.018	27	1086	0.186
8	179	0.030	28	946	0.162
9	1147	0.197	29	115247	19.835
10	32634	5.616	30	30170	5.192
11	12410	2.135	31	25666	4.417
12	29971	5.158	32	52519	9.039
13	17431	3.000	33	45154	7.771
14	599	0.103	34	1611	0.277
15	3	0.000	35	1891	0.325
16	2845	0.489	36	119	0.020
17	3422	0.588	37	298	0.051
18	1899	0.326	38	15573	2.680
19	4021	0.692	39	13806	2.376
20	9259	1.593	40	8750	1.505

The tables describing the ELU code and the soil types have been included for easy reference.

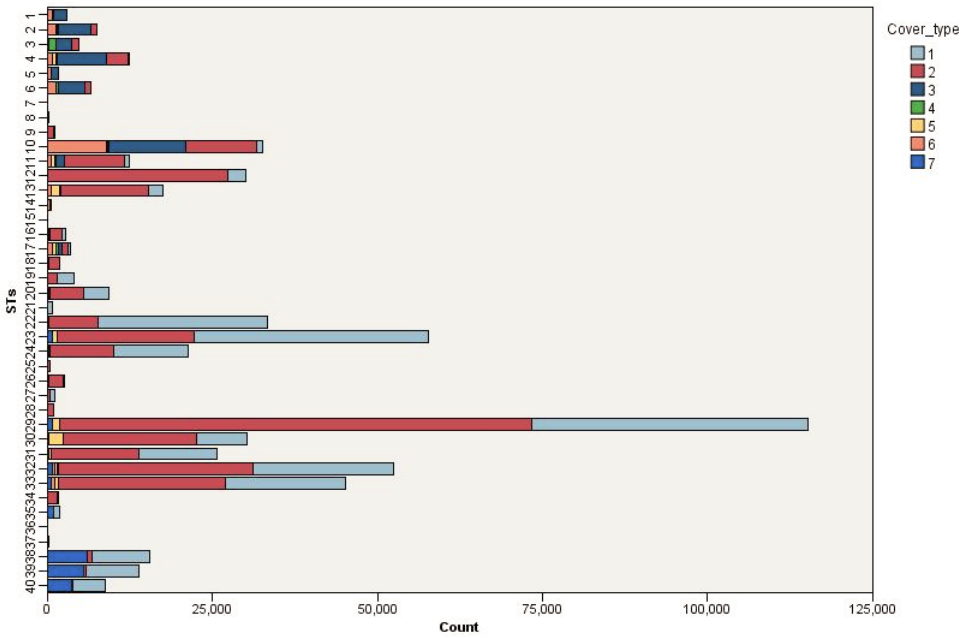

Figure A.14: Bar chart of Soil Types Per Cover Type

Soil Type	USFS ELU Code	Description
1	2702	Cathedral family - Rock outcrop complex, extremely stony
2	2703	Vanet - Ratake families complex, very stony
3	2704	Haploborolis - Rock outcrop complex, rubbly
4	2705	Ratake family - Rock outcrop complex, rubbly
5	2706	Vanet family - Rock outcrop complex complex, rubbly
6	2717	Vanet - Wetmore families - Rock outcrop complex, stony
7	3501	Gothic family
8	3502	Supervisor - Limber families complex
9	4201	Troutville family, very stony
10	4703	Bullwark - Catamount families - Rock outcrop complex, rubbly
11	4704	Bullwark - Catamount families - Rock land complex, rubbly
12	4744	Legault family - Rock land complex, stony
13	4758	Catamount family - Rock land - Bullwark family complex, rubbly
14	5101	Pachic Argiborolis - Aquolis complex
15	5151	unspecified in the USFS Soil and ELU Survey
16	6101	Cryaquolis - Cryoborolis complex
17	6102	Gateview family - Cryaquolis complex
18	6731	Rogert family, very stony
19	7101	Typic Cryaquolis - Borohemists complex
20	7102	Typic Cryaquepts - Typic Cryaquolls complex
21	7103	Typic Cryaquolls - Leighcan family, till substratum complex
22	7201	Leighcan family, till substratum, extremely bouldery
23	7202	Leighcan family, till substratum - Typic Cryaquolls complex
24	7700	Leighcan family, extremely stony
25	7701	Leighcan family, warm, extremely stony
26	7702	Granile - Catamount families complex, very stony
27	7709	Leighcan family, warm - Rock outcrop complex, extremely stony
28	7710	Leighcan family - Rock outcrop complex, extremely stony
29	7745	Como - Legault families complex, extremely stony
30	7746	Como family - Rock land - Legault family complex, extremely stony
31	7755	Leighcan - Catamount families complex, extremely stony
32	7756	Catamount family - Rock outcrop - Leighcan family complex, extremely stony
33	7757	Leighcan - Catamount families - Rock outcrop complex, extremely stony
34	7790	Cryorthents - Rock land complex, extremely stony
35	8703	Cryumbrepts - Rock outcrop - Cryaquepts complex
36	8707	Bross family - Rock land - Cryumbrepts complex, extremely stony
37	8708	Rock outcrop - Cryumbrepts - Cryorthents complex, extremely stony
38	8771	Leighcan - Moran families - Cryaquolls complex, extremely stony
39	8772	Moran family - Cryorthents - Leighcan family complex, extremely stony
40	8776	Moran family - Cryorthents - Rock land complex, extremely stony

ELU First digit: climatic zone	Second digit: geologic zones
1. lower montane dry	1. alluvium
2. lower montane	2. glacial
3. montane dry	3. shale
4. montane	4. sandstone
5. montane dry and montane	5. mixed sedimentary
6. montane and subalpine	6. unspecified
7. subalpine	7. igneous and metamorphic
8. alpine	8. volcanic

The third and fourth ELU digits are unique to the mapping unit and have no special meaning to the climatic or geologic zones.

B Miscellaneous Figures

These histograms were generated to see if there were any possibly useful or interesting aspects to each feature relative to the cover type. The data in these graphs supports preliminary ideas or intuitions about the database that arose out of the modeling process. Only the useful histograms that revealed something are included.

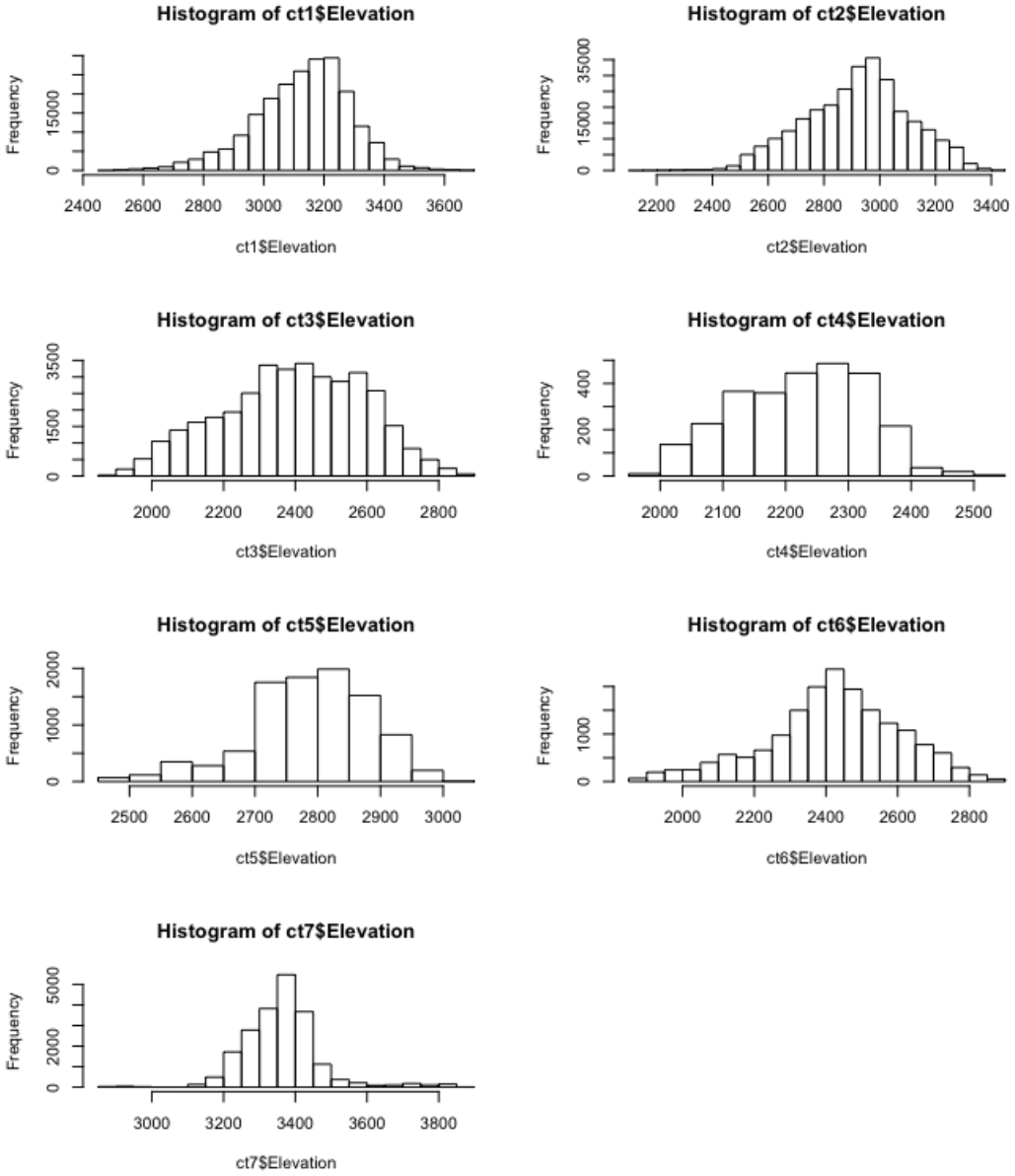

Figure B.1: Histogram of Elevations per Cover Type

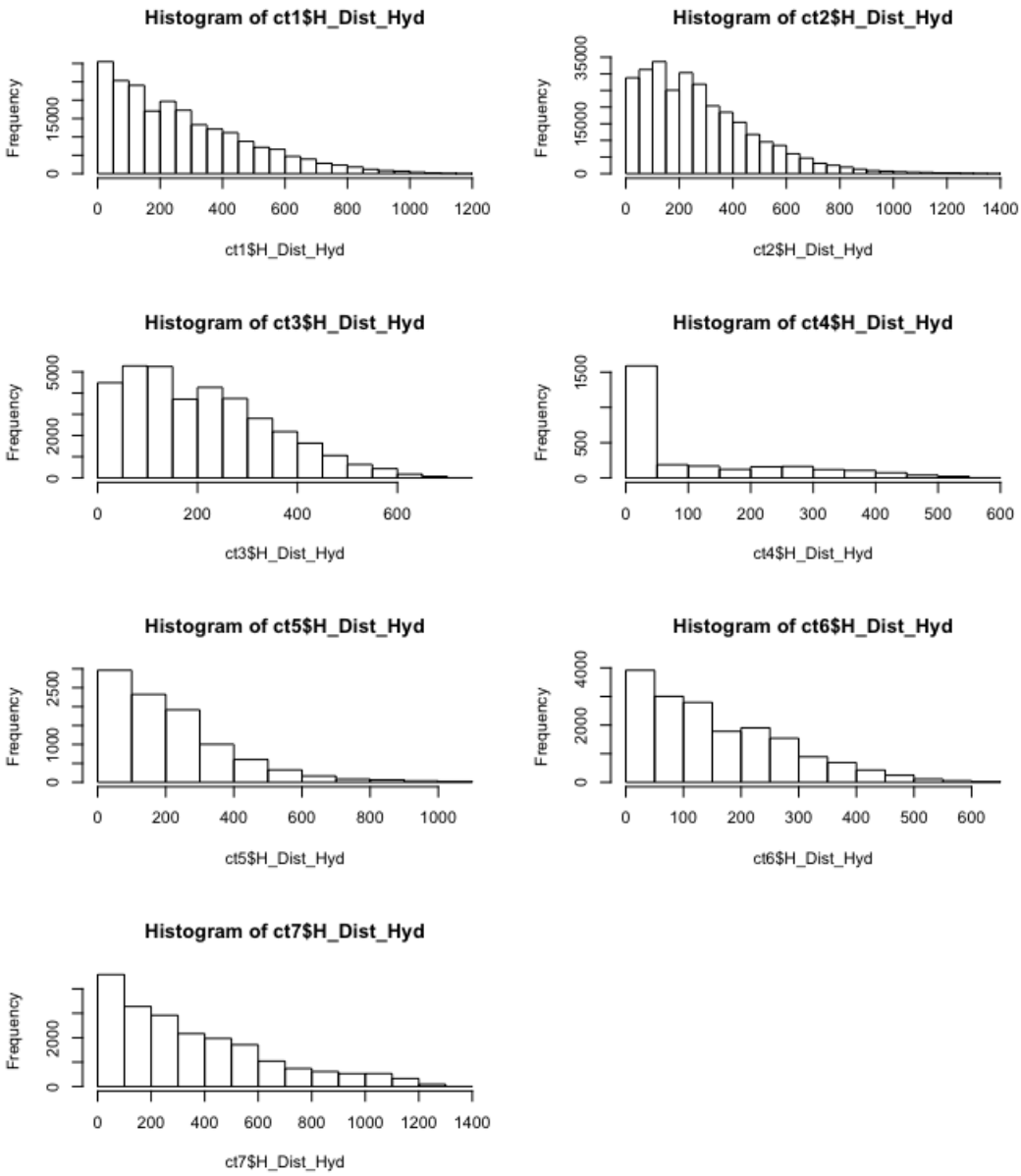

Figure B.2: Histogram of H_Dist_Hyd per Cover Type

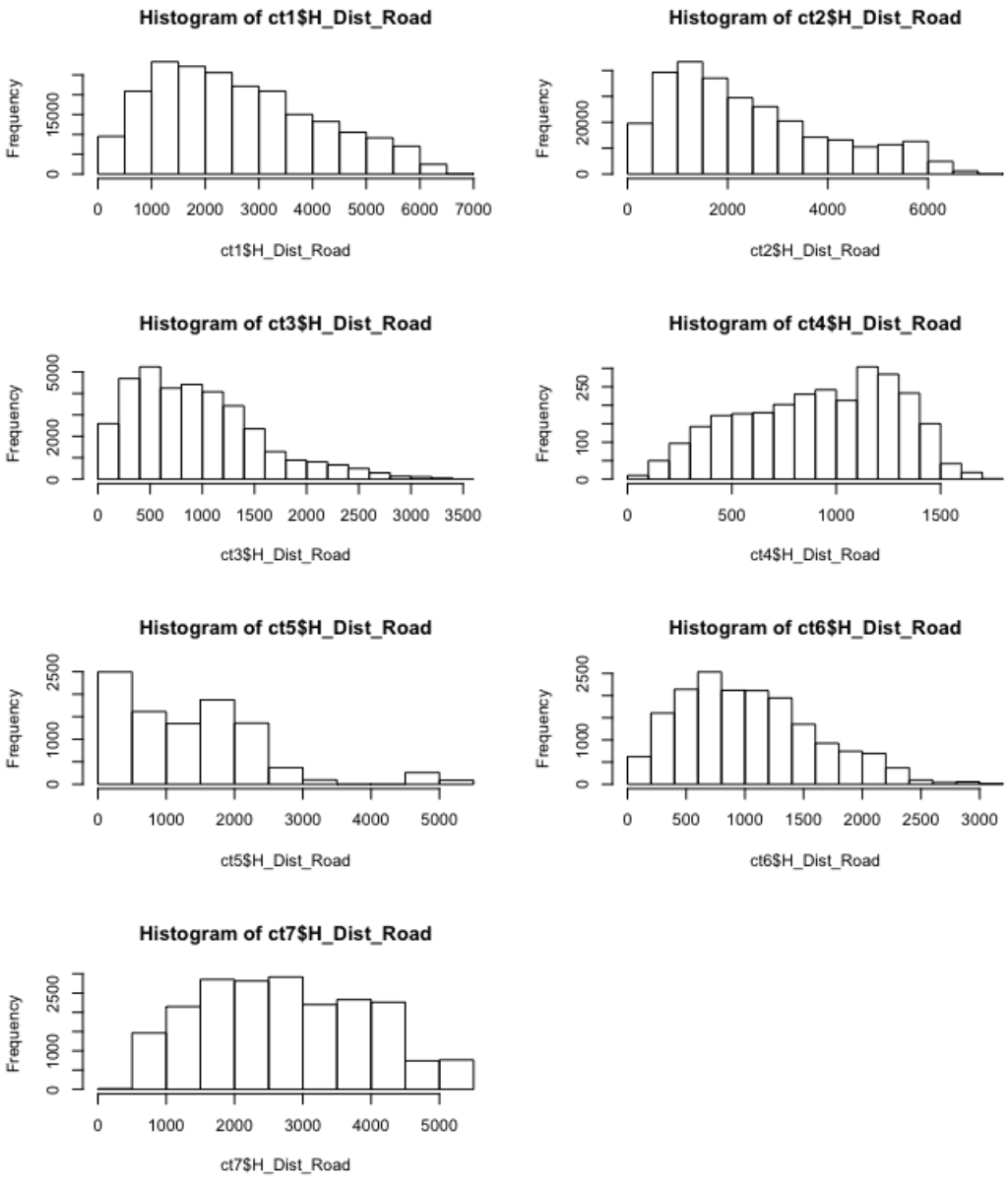

Figure B.3: Histogram of H_Dist_Road per Cover Type

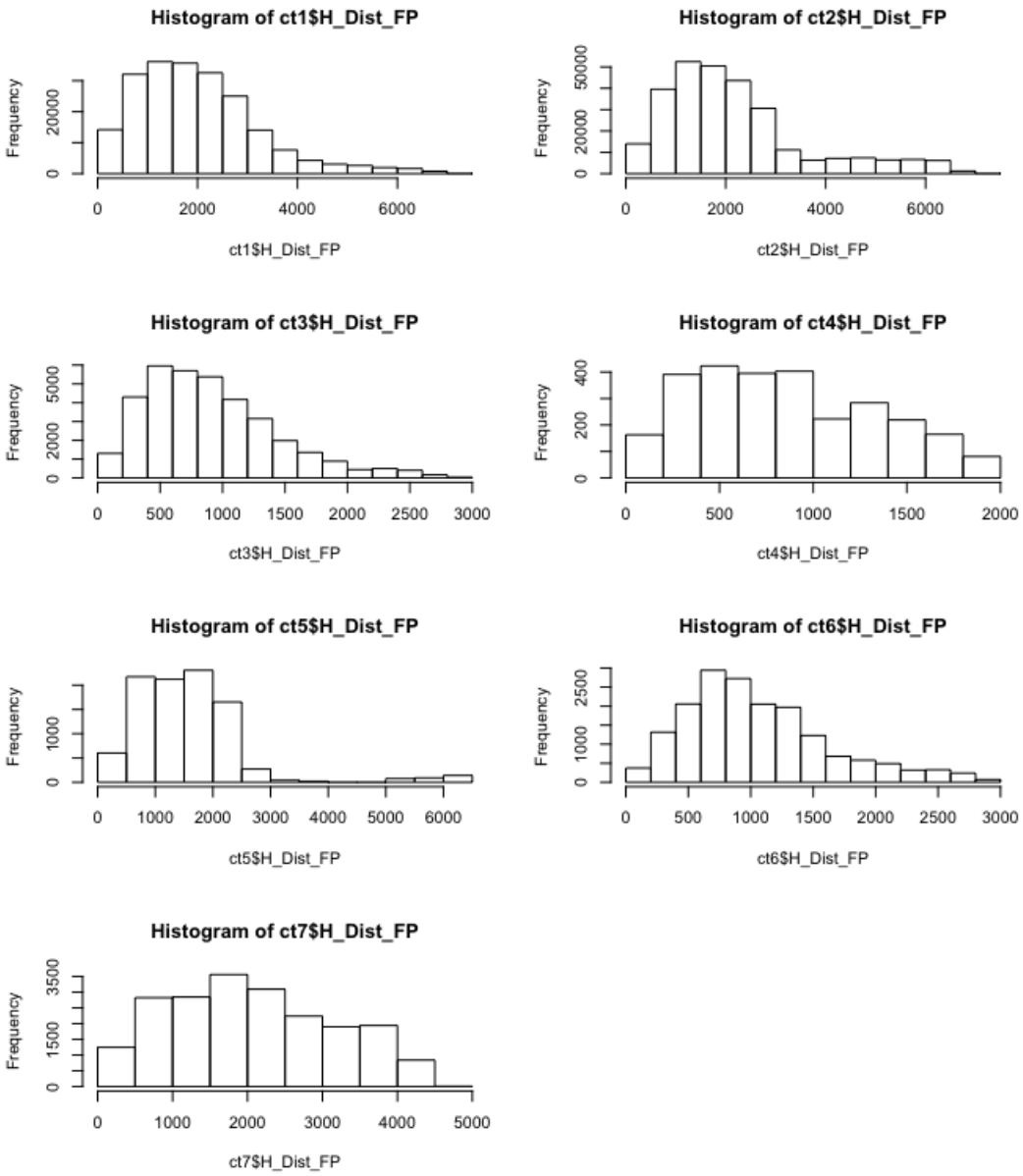

Figure B.4: Histogram of H_Dist_FP per Cover Type

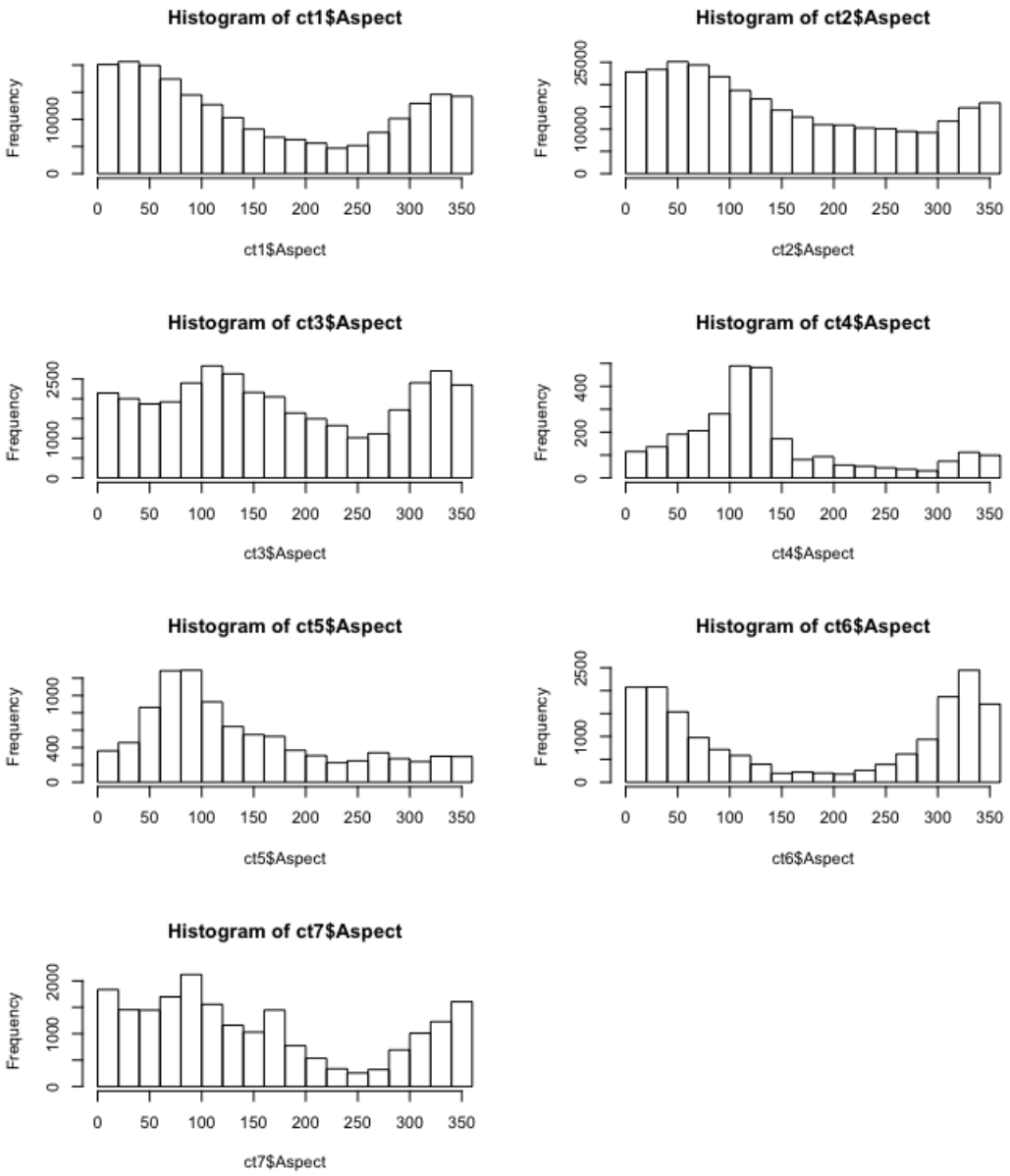

Figure B.5: Histogram of Aspect per Cover Type

These boxplots were generated to confirm ideas based on the histograms and to get a visual sense of the distribution and range of outliers.

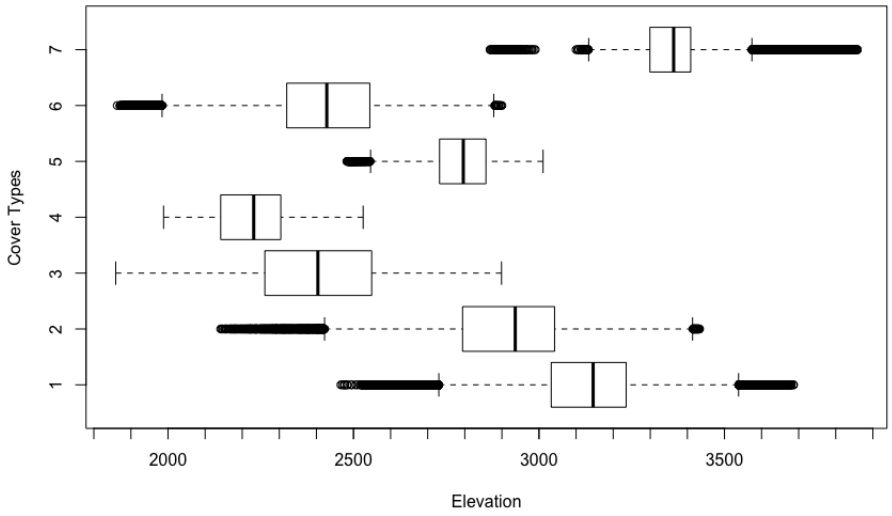

Figure B.6: Boxplots of Elevation versus Cover Type

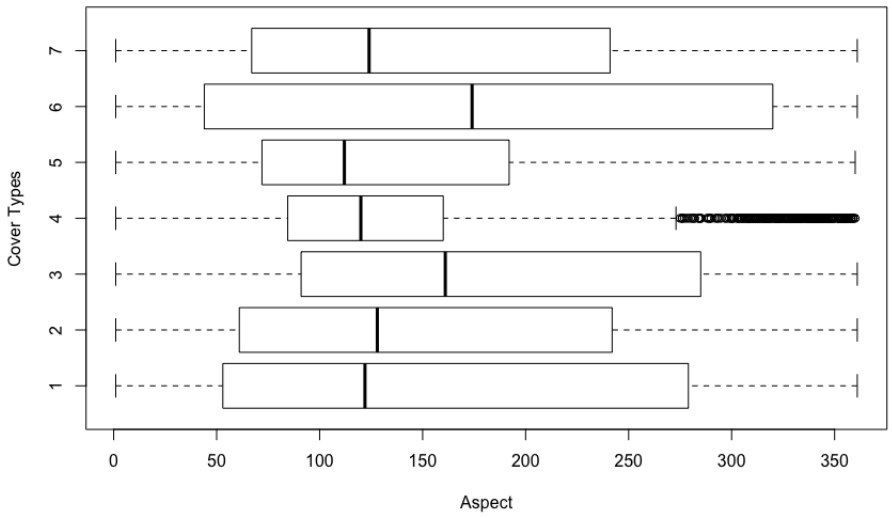

Figure B.7: Boxplots of Aspect versus Cover Type

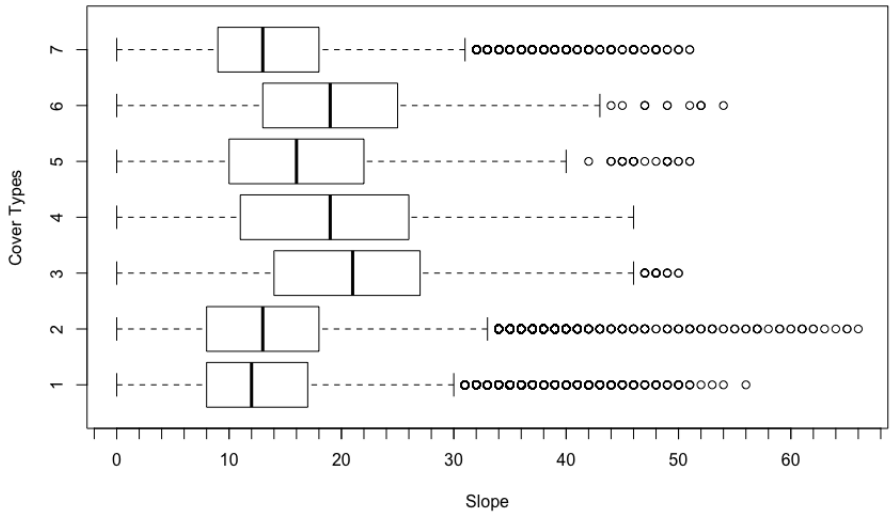

Figure B.8: Boxplots of Slope versus Cover Type

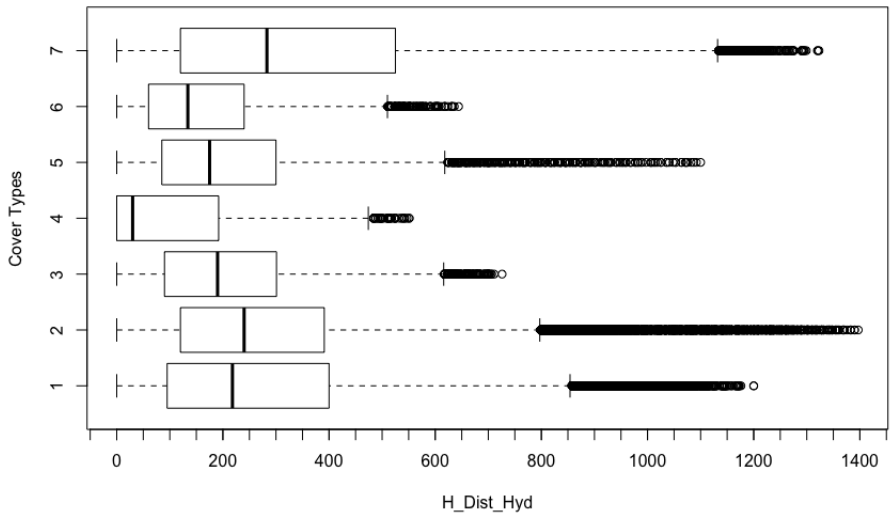

Figure B.9: Boxplots of H_Dist_Hyd versus Cover Type

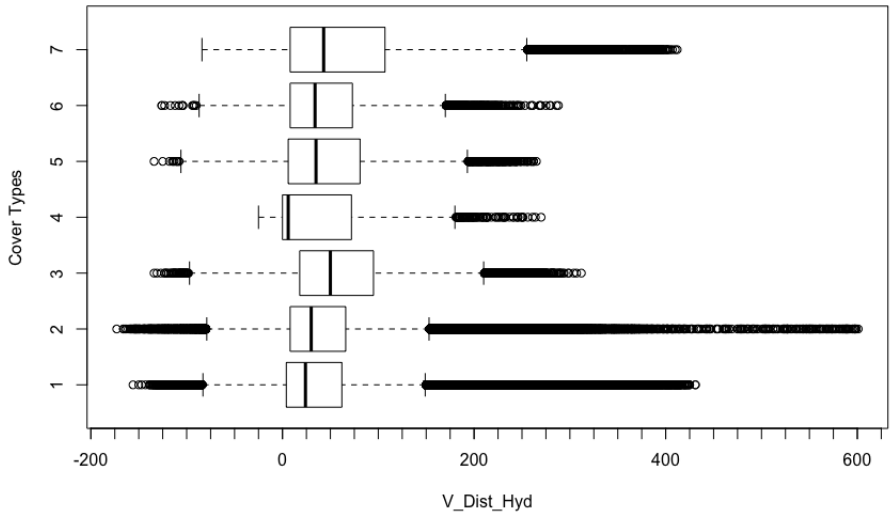

Figure B.10: Boxplots of V_Dist_Hyd versus Cover Type

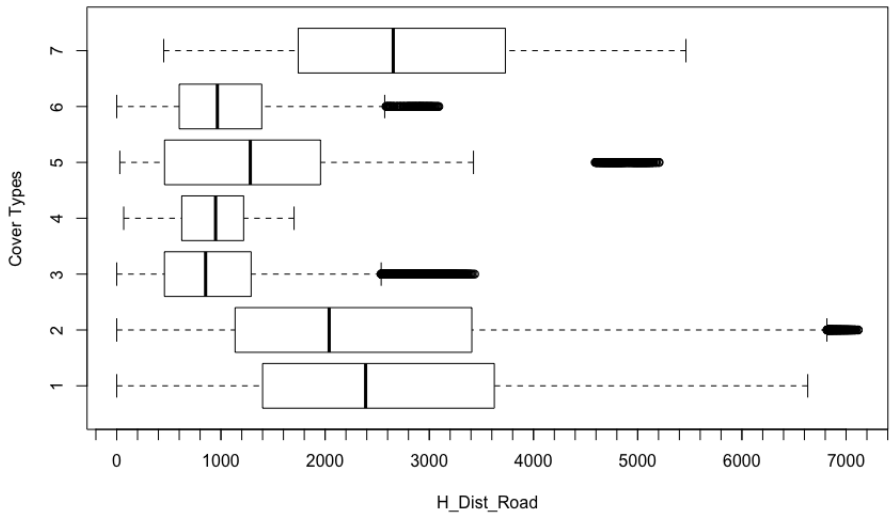

Figure B.11: Boxplots of H_Dist_Road versus Cover Type

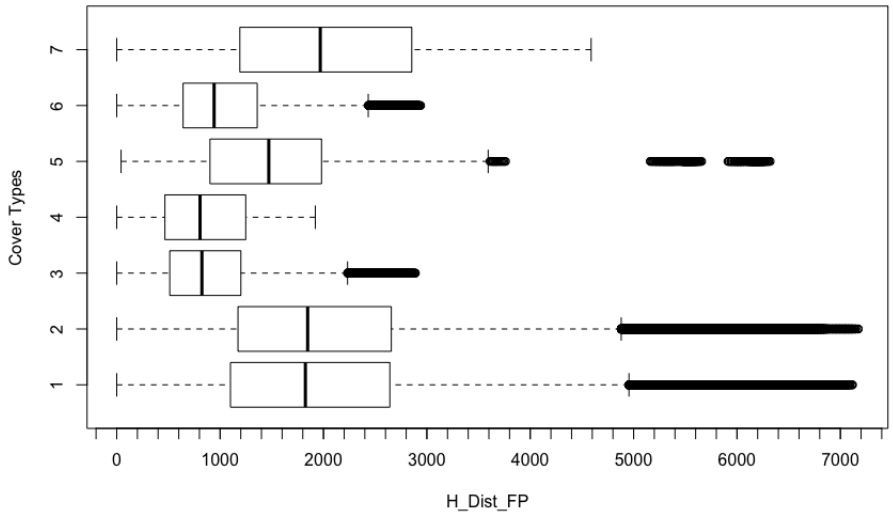

Figure B.12: Boxplots of H_Dist_FP versus Cover Type

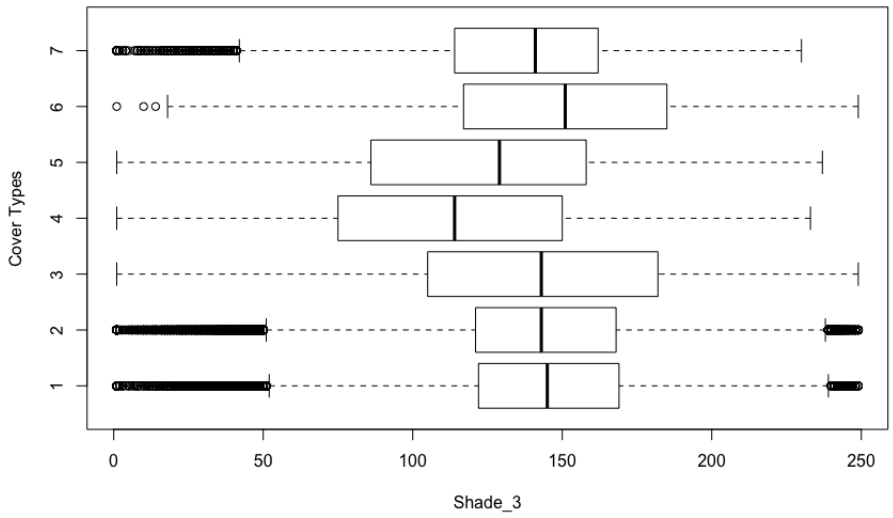

Figure B.13: Boxplots of Shade_3 versus Cover Type

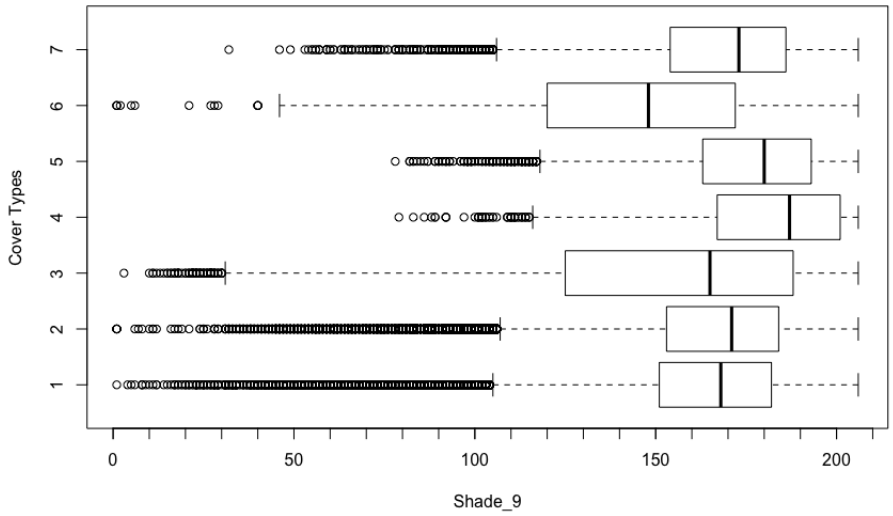

Figure B.14: Boxplots of Shade_9 versus Cover Type

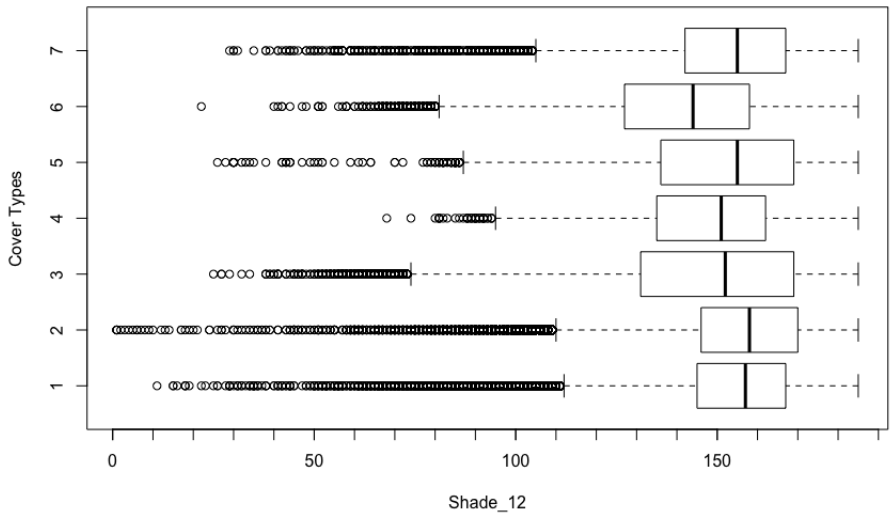

Figure B.15: Boxplots of Shade_12 versus Cover Type

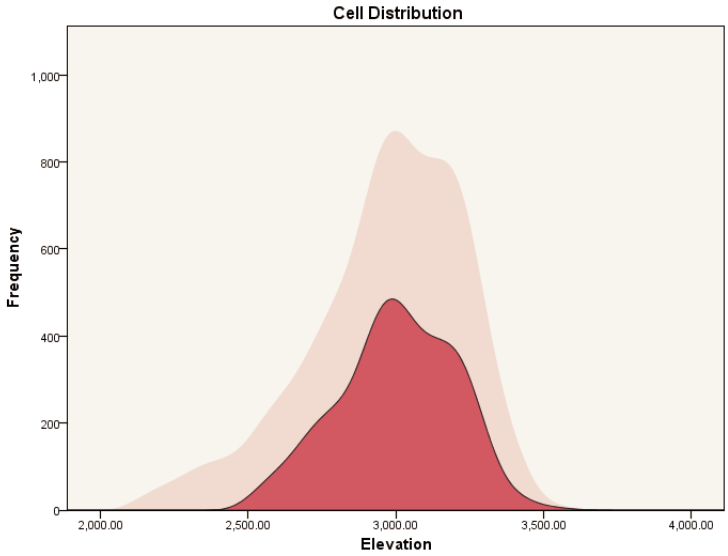

Figure B.16: Elevations in Cluster 1

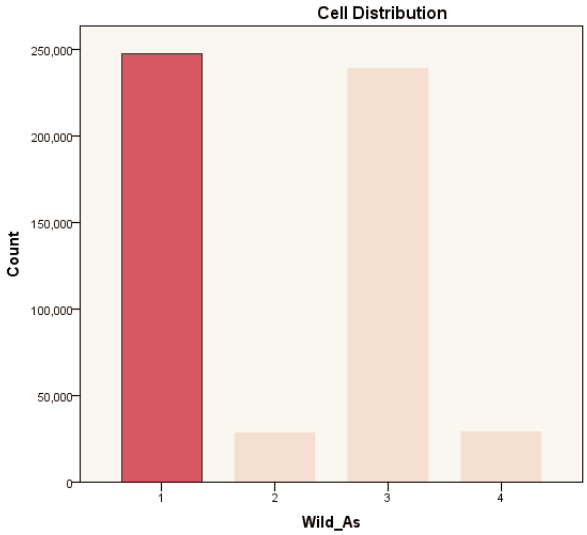

Figure B.17: Wilderness Areas in Cluster 1

These would be one cluster if it was not for Wilderness Area.

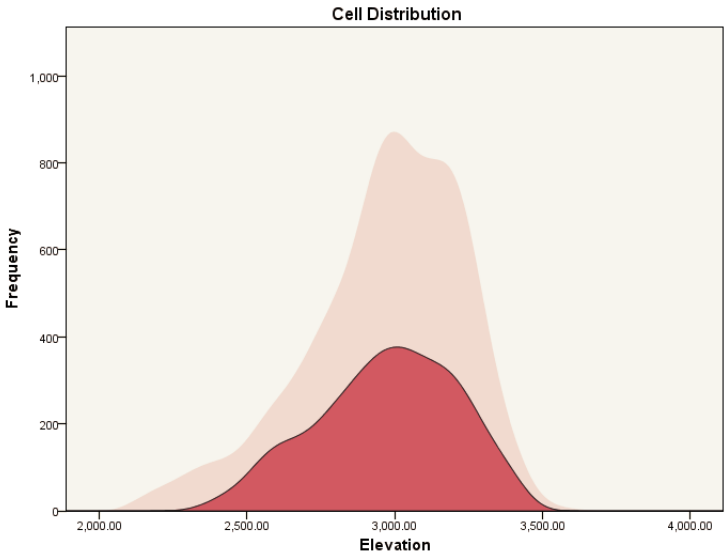

Figure B.18: Elevations in Cluster 2

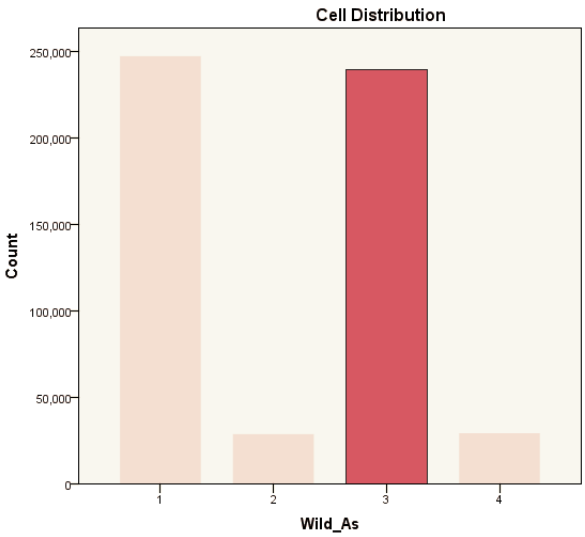

Figure B.19: Wilderness Areas in Cluster 2

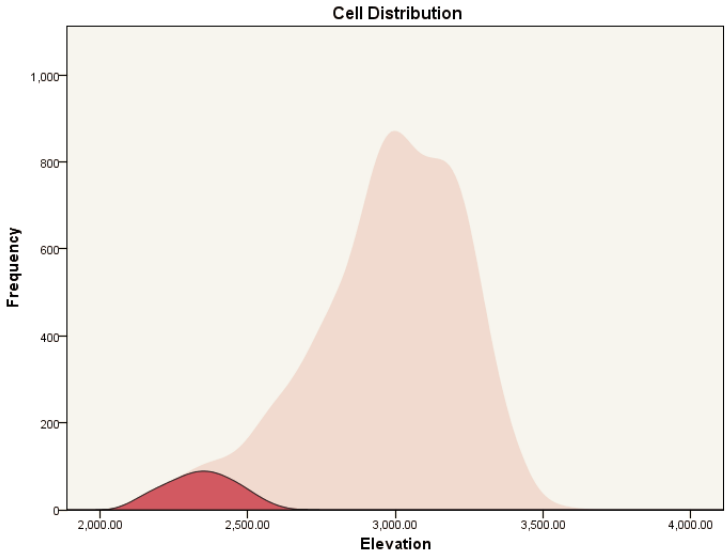

Figure B.20: Elevations in Cluster 3

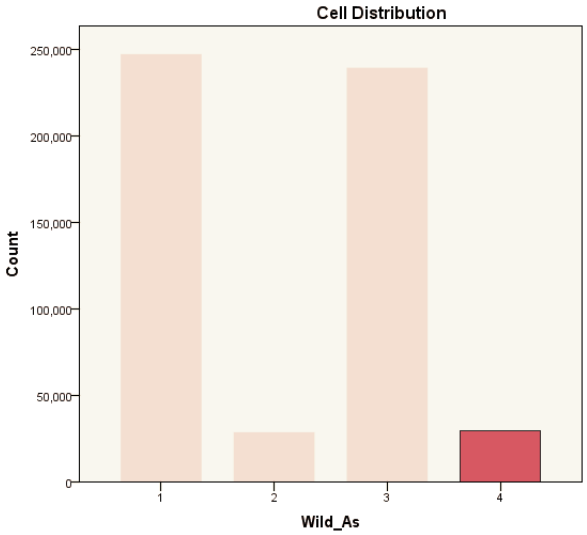

Figure B.21: Wilderness Areas in Cluster 3

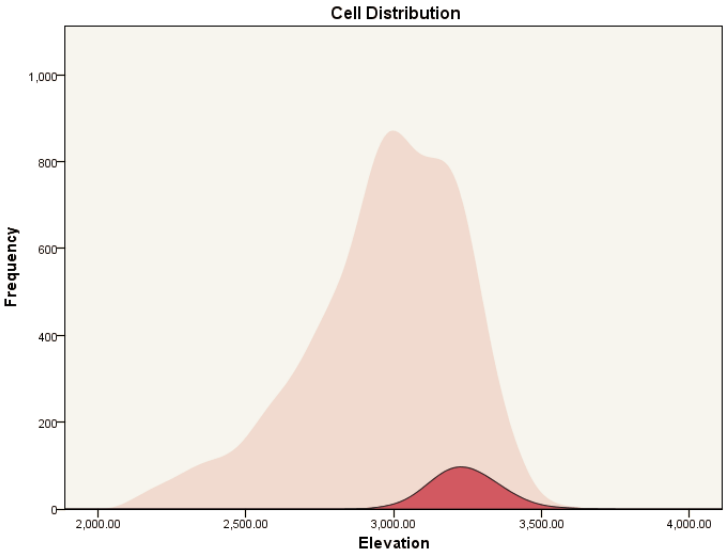

Figure B.22: Elevations in Cluster 4

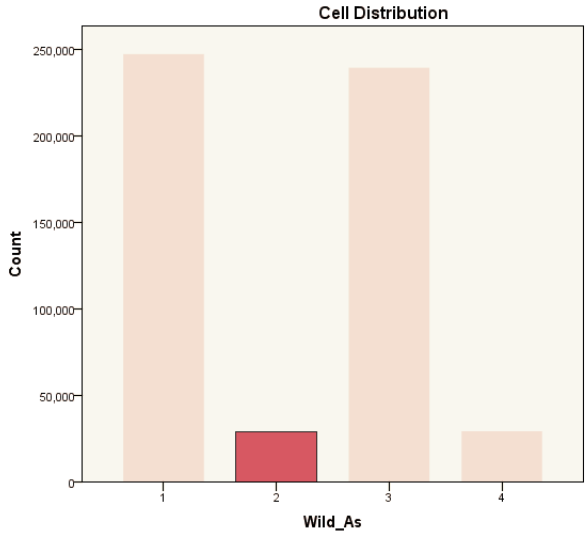

Figure B.23: Wilderness Areas in Cluster 4

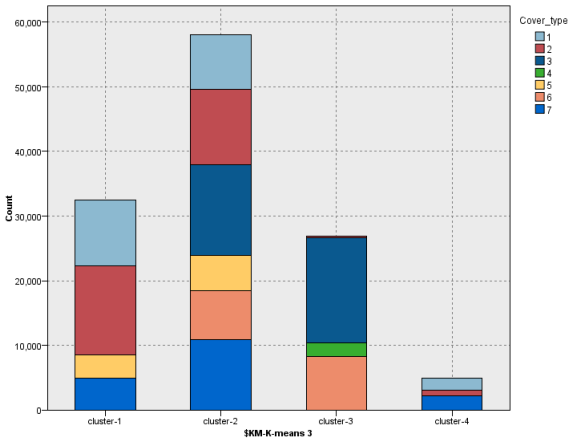

Figure B.24: Typical example of clustering results

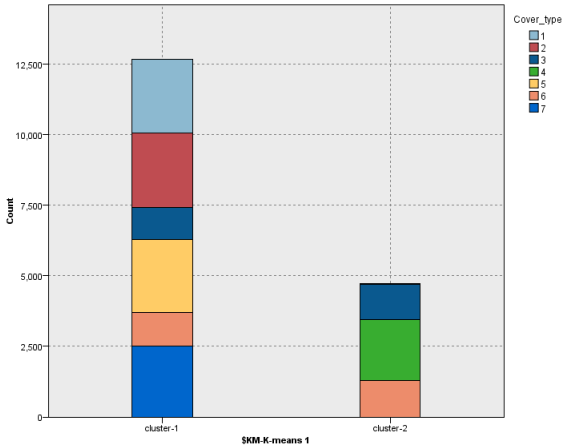

Figure B.25: Clustering of a balanced dataset

This is an example of the web diagrams used to confirm the strength of the relationship between cover types and specific levels of a feature.

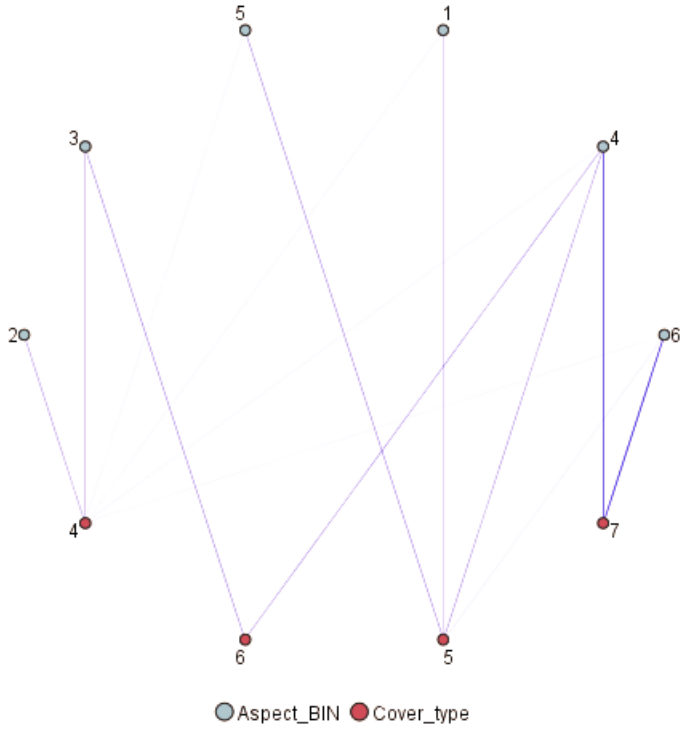

Figure B.26: Cover Type and Aspect Relationships

C Software Scripts

```
import sys
import csv
import pdb

wa_stats = [0,0,0,0]
st_stats = [0] * 40

outf = open('condensed.csv', 'w')
newCSV = csv.writer(outf)

with open(sys.argv[1], 'r') as inf:
  coverTypeCSV = csv.reader(inf)
  rowcount = 0
  for line in coverTypeCSV:
    rowcount = rowcount + 1
    if (rowcount > 1):
      line[:] = map(lambda i: int(i), line)
      newrow = line[0:10]
      newdata = [0,0]

      if (sum(line[10:14]) == 0) or (sum(line[10:14]) > 1):
        print("WA ERROR at line {rowcount}")
        newdata[0] = 100
      else:
        for i in range(10,14):
          if line[i] == 1:
            newdata[0] = i - 9
            wa_stats[i - 10] += 1

      if (sum(line[14:54]) == 0) or (sum(line[14:54]) > 1):
        print("ST ERROR at line {rowcount}")
        newdata[1] = 100
      else:
        for i in range(14,54):
          if line[i] == 1:
            newdata[1] = i - 13
            st_stats[i - 14] += 1

      newrow += newdata
      newrow += [line[54]]
    else:
      newrow = line[0:10] + ['Wild_As', 'STs'] + [line[54]]
    # filter out Shade_3s greater than 248 so SPSS will work
    if rowcount == 1 or newrow[8] < 249:
      newCSV.writerow(newrow)

print(str(wa_stats))
print(sum(wa_stats))
percs = list(map(lambda i: i / sum(wa_stats) * 100, wa_stats))
print(str(percs))

print(st_stats)
print(sum(st_stats))
percs = list(map(lambda i: i / sum(st_stats) * 100, st_stats))
print(str(percs))

outf.close()
```

Script C.1: reduceData.py

```
import sys
import csv
import pdb

min_maxs = {};
fields = ['elevation', 'aspect', 'slope', 'h_dist_hyd', 'v_dist_hyd', 'h_dist_road',
   'shade_9', 'shade_12', 'shade_3', 'h_dist_fp', 'wild_a_1', 'wild_a_2', 'wild_a_3',
   'wild_a_4']
for fld in fields:
  min_maxs[fld] = []
  for x in range(0,8):
    min_maxs[fld].append([100000,0])

cover_types = []
for x in range(0,8):
  cover_types.append([0]*41)

with open(sys.argv[1], 'r') as inf:
  coverTypeCSV = csv.reader(inf)
  rowcount = 0
  for line in coverTypeCSV:
    rowcount = rowcount + 1
    if (rowcount > 1):
      line[:] = map(lambda i: int(i), line)
      ct = line[-1]
      for i in range(0,14):
        ind = fields[i]
        if min_maxs[ind][ct][0] > line[i]:
          min_maxs[ind][ct][0] = line[i]
        if min_maxs[ind][ct][1] < line[i]:
          min_maxs[ind][ct][1] = line[i]
      for i in range(14,45):
        cover_types[ct][i-13] += line[i]


for fld in fields:
  print(fld, end='')
  for i in range(1,8):
    print(" CT{} ({},{})".format(i, min_maxs[fld][i][0], min_maxs[fld][i][1]), end='')
  print()

for i in range(1,8):
  print("Cover Type {}:".format(i), end="")
  for y in range(1,41):
    if cover_types[i][y] > 0:
      print(" ST{}".format(y), end="")
  print()
```

Script C.2: breakdownData.py