CLASSIFICATION OF STREAM BIOLOGICAL CONDITION WITHIN THE CHESAPEAKE BAY WATERSHED

by
Michelle Lee Talal

Dr. Dean L. Urban, Masters Project Advisor
Dr. Kelly O. Maloney and Dr. Donald E. Weller, Summer Internship Advisors

Masters project submitted in partial fulfillment of the
requirements for the Master of Environmental Management degree in
the Nicholas School of the Environment of
Duke University
April 2009

Abstract

Human land use activities at the landscape scale are increasingly the largest threat to the biological condition of watershed and stream ecosystems. The Chesapeake Bay watershed (CBW), a particularly valuable watershed within the United States, has undergone considerable land use change over the past 400 years and faces many restoration challenges. Using fish indicators of biological integrity (IBIs), and data for land use, land cover, and environmental attributes, five empirical models (CART model, Random Forest, Conditional Tree, Conditional Forest, and ordinal logistic regression) were used to predict the biological condition of 1st-3rd order streams within the CBW. After the models were evaluated using resubstitution and 10-fold cross validation, the highest performing model was identified (Random Forest) and extrapolated to 71,182 stream sites within the CBW using geographic information software (GIS). Of these sites, 49% (35,006 sites) were classified as having "Good" biological condition, 24% (16,826 sites) as having "Fair" biological condition, and 27% (19,350 sites) as having "Poor" biological condition. The variable importance plot generated by the Random Forest (RF) model showed that watershed area (upslope of sampling location, km^2) was the most important variable, followed by percentage of impervious surface cover, and percentage of pasture cover. Additionally, the Random Forest's partial dependence plots showed the marginal effect of each variable on the class probability. As watershed area (km^2) increases, there is a higher probability of a "fair" or "good" classification of stream biological condition; with a threshold watershed area of approximately 20 – 25 km^2. Also, as the percentage of impervious surface cover increases, there is a greater probability of a poor classification of stream condition (threshold of ~5% impervious surface cover). The results of this study may help environmental and land use managers understand the effects of human land use and make more effective land use decisions to address watershed impairment within the CBW.

Table of Contents

INTRODUCTION……..... 1
Objectives…… **2**

METHODS……..... 3
Study Area…… **3**
Fish Indices of Biotic Integrity……. **5**
Land Use/Land Cover data……... **8**
Models…….. **9**
1) CART model……..... **9**
2) Random Forest Model……... **10**
3) Conditional Tree Model…… **11**
4) Conditional Forest Model……..... **11**
5) Ordinal Logistic Regression Model…….. **12**
Model Evaluation & Accuracy Assessment……. **13**
Model Selection & Extrapolation to CBW……... **14**

RESULTS…… 14
Performance of Models…… **14**
Random Forest Model…….. **16**
1) Random Forest Model Performance……. **16**
2) Variable Importance……..... **17**
3) Partial Dependence Plots……..... **18**
4) Random Forest Model Extrapolation to CBW…… **22**

DISCUSSION AND CONCLUSIONS……..... 24
Management Implications…… **26**
Areas of Future Research and Development…… **26**

REFERENCES…… 28

APPENDIX A: Random Forest Partial Dependence Plots…….. **31**

APPENDIX B: CART Model Results……. **38**

APPENDIX C: Conditional Tree Model Results…… **41**

APPENDIX D: Conditional Forest Model Results…….. **42**

TABLES	Page
1. Sampling sites within the Chesapeake Bay watershed	**6**
2. Predictor variables used in the analyses	**8**
3. Kappa statistics and strength of agreement	**14**
4. Accuracy measures for predictions of stream conditions within CBW	**15**
5. Resubstitution confusion matrix for the Random Forest model	**16**
6. Ten-fold cross-validation confusion matrix for the Random Forest model	**17**

FIGURES	Page
1: Ecoregions within the Chesapeake Bay watershed	**4**
2. Distribution of sampling sites within 1^{st}-3^{rd} order streams of the CBW	**7**
3. Random Forest variable importance plot for predictions of CBW biological condition	**17**
4. Partial dependence plot showing effect of the watershed area variable on probability of a "Poor" classification	**18**
5. Partial dependence plot showing effect of the watershed area variable on probability of a "Fair" classification	**19**
6. Partial dependence plot showing effect of the watershed area variable on probability of a "Good" classification	**19**
7. Partial dependence plot showing effect of the % impervious surface cover area on probability of a "Poor" classification	**20**
8. Partial dependence plot showing effect of % impervious surface cover area on probability of a "Good" classification	**20**
9. Partial dependence plot showing effect of % pasture cover on probability of a "Poor" classification	**21**
10. Partial dependence plot showing effect of % pasture cover on probability of a "Fair" classification	**21**
11. Partial dependence plot showing effect of % pasture cover on probability of a "Good" classification	**22**
12. Random Forest predictions of the biological condition of 1^{st}-3^{rd} order streams within the Chesapeake Bay watershed	**23**

Acknowledgements

I would like to thank my advisors at the Smithsonian Environmental Research Center, Dr. Kelly O. Maloney and Dr. Donald E. Weller, and Dr. Dean Urban of the Nicholas School of the Environment, for their wonderful support and guidance. I also thank my family and friends for their love and encouragement. This project would not be possible without the U.S. EPA Environmental Monitoring and Assessment Program, the Maryland DNR, and all others who provided the data for this project on the Chesapeake Bay watershed. Funding and support for was provided by the Smithsonian Environmental Research Center and the U.S. EPA National Center for Environmental Research (NCER) Science to Achieve Results (STAR) grant #R831369: A Watershed Classification System for Improved Monitoring and Restoration: Landscape Indicators of Watershed Impairment.

Introduction

Human land use activities at the landscape scale are increasingly the largest threat to the biological condition of watershed and stream ecosystems (Allan 2004, Strayer et al. 2003a). These activities may include large-scale water projects such as dams for flood control or hydroelectric power, timber harvest and deforestation, agriculture, road and bridge building, conversion of forest to pasture or grass in rural areas, over-harvesting of aquatic species, industrial waste discharges, and other land uses, which all have been shown to contribute to habitat loss and degradation (Allan & Flecker 1993, Booth et al. 2002, Klein 1979, Richards et al. 1996, Strayer et al. 2003a). Habitat loss and degradation within watershed and stream ecosystems do not only disturb native fish, macroinvertebrate, and plant species populations, but they also tend to reduce the amount of clean drinking water and compromise other natural resources valuable to humans (Allan & Flecker 1993, Dynesius & Nilsson 1994). Arguably, streams are some of the most threatened ecosystems on earth (Dynesius & Nilson 1994, Vinson & Hawkins 1998).

In order to better understand stream ecosystem impairment, it is important that environmental and land use managers be able to model and predict the effects of anthropogenic disturbance, as well as effectively manage these human land activities on various spatial and temporal scales (Strayer et al. 2003a). One of the ideal ways for studying the effects of human land use activities is to replicate long term experiments of entire watersheds with particular land use activities and then to observe changes over time (Strayer et al. 2003a). However, these types of experiments are oftentimes very costly, logistically impractical and may take several years to accumulate the data needed to extrapolate the long-term effects of human land use activities (Strayer et al. 2003a). In turn, researchers can use combine existing datasets with information on

human land use, environmental attributes, and species to create predictive empirical models of the biological condition of stream ecosystems that extend to unsurveyed locations (Strayer et al. 2003a).

These predictive models may also incorporate an Index of Biologic Integrity (IBI), a broad-based multi-parameter tool that integrates the attributes of a biological community and the effects of a variety of environmental stressors across spatial and temporal scales (Karr et al. 1991). Biological integrity is the ability to support and maintain a 'balanced, integrated, adaptive community of organisms having a species composition, diversity, and functional organization comparable to that of natural habitat of the region" (Karr 1991). IBIs have been constructed from a wide range of biological data (e.g. different organisms such as macroinvertebrates, aquatic plants, riparian vegetation, fish, etc.) and been evaluated for the regions in which they are applied (Kerans and Karr 1994, DeShon 1995, Barbour et al. 1996). IBIs of fish communities are very useful for indicating the biological condition of watersheds because they reflect the cumulative effects of human land use and disturbance, provide a relatively long-term record of environmental stress, are easy to sample and identify to species, etc. (Moyle and Leidy 1992, Simon and Lyons 1995).

Objectives

One of the main objectives of this study is to create an empirical model that accurately predicts the biological condition of stream reaches within the Chesapeake Bay watershed (CBW) of the Mid-Atlantic Region of the United States. The CBW has a diversity of physiographic provinces and biological organisms, and is economically, culturally, and environmentally valuable to its inhabitants, but has also undergone considerable land use change over the last 400 years and faces several difficult restoration challenges (Boesch et al. 2001, Goetz et al. 2004a,

Jantz et al. 2005). Many of the watersheds and wetlands within the CBW have been disturbed or destroyed by logging, agricultural and urban runoff, land conversion, etc., which have contributed to a loss of normal ecosystem functioning (i.e. absorption of nutrients, recharging of water, etc.) (Weber 2004). Given the high degree of disturbed conditions within the CBW (USEPA 2006b), in addition to the large projected human population growth in the region (Weber 2004), land use and environmental managers would benefit from an accurate and comprehensive landscape scale analysis of CBW stream conditions. In this study, multiple datasets related to land use, environmental attributes, and fish IBIs will be combined to create predictive empirical models for the biological condition of the CBW. These models will then be ranked according to their predictive accuracy. After the most accurate model(s) is selected, the second objective of this study will be to extrapolate the model to the entire CBW using geographic information software (GIS). These maps will display the areas of high, medium, and low impairment of streams within the CBW, and may help land use and environmental managers make more informed stream ecosystem management decisions related to conservation, restoration, other land development.

Methods

Study Area

The Chesapeake Bay watershed (CBW) is located in the Mid-Atlantic region of the United States and extends to portions of six states and Washington D.C (Figure 1). This watershed drains an area of approximately 168,000 km^2, and flows into the Chesapeake Bay, one of the largest and most productive bays in the world (Goetz and Jantz 2004a, King et al. 2005). The potential natural vegetation of the CBW ranges from the Northern hardwood forests of the North Appalachian Plateau and Uplands and the North Central Appalachian ecoregions, to the

mixed mesophytic forest of the Central Appalachians ecoregion, to the Appalachian oak forests of the Blue Ridge Mountains and the Northern Piedmont ecoregions, to the oak/hickory/pine and southern mixed forests of the Southeastern Plains ecoregion, and the oak/hickory/pine, pocosin, southern floodplain forest, and southern mixed forest of the Middle Atlantic Coastal Plain ecoregion (Omernik 1987). In the northern reaches of the basin, the climate is humid continental, while the southern ranges of the CBW have a more humid subtropical climate (Peel et al. 2007).

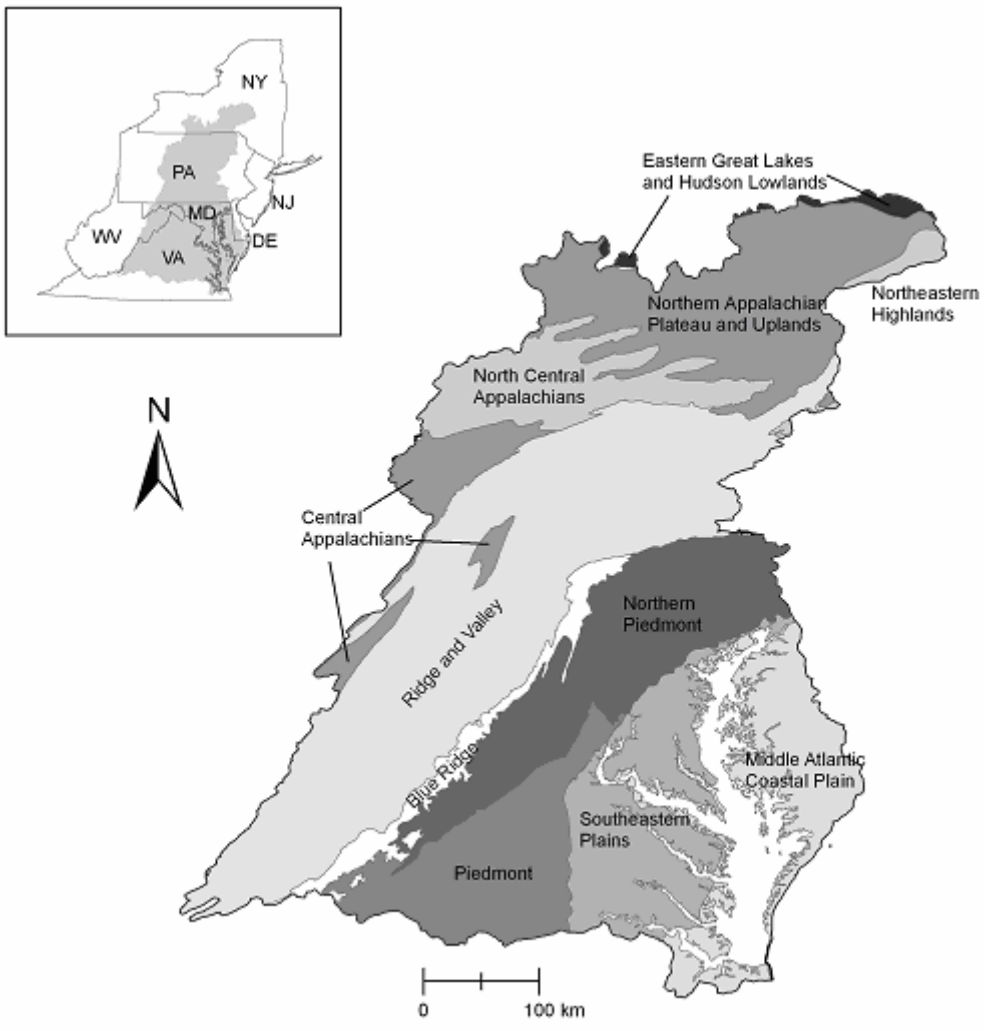


Figure 1: Ecoregions within the Chesapeake Bay watershed.

Fish Indices of Biotic Integrity

Fish data of the Chesapeake Bay watershed (CBW) were obtained from U.S. Federal and State monitoring programs including the Maryland Biological Stream Survey (MBSS, U.S. E.P.A. 2005), and Mid-Atlantic Highlands Assessment (MAHA, U.S. E.P.A. 2000b) and Mid-Atlantic Integrated Assessment (MAIA, U.S. E.P.A. 2006a) under the United States E.P.A.'s Environmental Monitoring and Assessment Program (EMAP). For the MAHA (sampled from 1993 to 1997) and MAIA (sampled from 1997 – 1998), fish were sampled from river reaches of 40 channel widths long using a stratified random sampling design. Samples were separated by habitat type (pool or riffle); only riffle data was used in these analyses. For the Maryland Biological Stream Survey (MBSS, U.S. E.P.A. 2005), fish data were collected during 1994 – 2004 using a probability-based design on 1^{st} – 4^{th} order streams. The fish were sampled during the summer months using double-pass electrofishing of 75 meter stream segments. The MBSS also developed an integrated dataset that includes several site- and landscape-scale environmental variables linked to the biological data and their derived attributes (e.g. tolerance values and functional groups). The original IBIs were developed with data collected from 1994 – 1997 from a maximum of 1098 sites, divided into 732 calibration sites and 366 (33%) validation sites. The dataset for the new IBIs included all samples from 1994 – 2004, with a total of 2508 sites and 353 (14%) of these sites reserved for validation purposes. This large number of sites provides enough reference sites to create reference conditions for additional classes of Maryland stream types (MBSS, U.S. E.P.A. 2005).

In this project, only 1^{st} to 3^{rd} order streams were analyzed because the MAHA and earlier MBSS surveys focused on these streams. In cases where sites were sampled more than once, only the sampling point closest to year 2000 was used (the year of the land use data was

compiled). Although the EMAP (MAHA and MAIA) and MBSS surveys used different metrics and were scaled somewhat differently, each survey classified sites as Poor, Fair, or Good. In cases where there was also a "Very Poor" classification of biological integrity, these were combined with the "Poor" classification. The assumption was made that site condition was classified consistently regardless of survey (i.e. a site classified as "Poor" in the EMAP assessments would also be classified as having "Poor" biological condition in the MBSS assessment). In total, there were 1713 fish sites available for the analyses (MAHA = 104, MAIA = 28, MBSS = 1581) (Table 1), which are distributed throughout the basin (Figure 3). However, to lessen the degree to which the MBSS data would overwhelm the analyses, 10% of the available MBSS data was randomly subsetted. Of these sites, 62 were classified as having "Poor" biological integrity, 91 were classified as having "Fair" biological integrity, and 137 were classified as having "Good" biological integrity.

Table 1. Sampling sites within the Chesapeake Bay watershed.

Fish IBI's	Number of sites
MBSS	158
MAHA	104
MAIA	28
Total	290

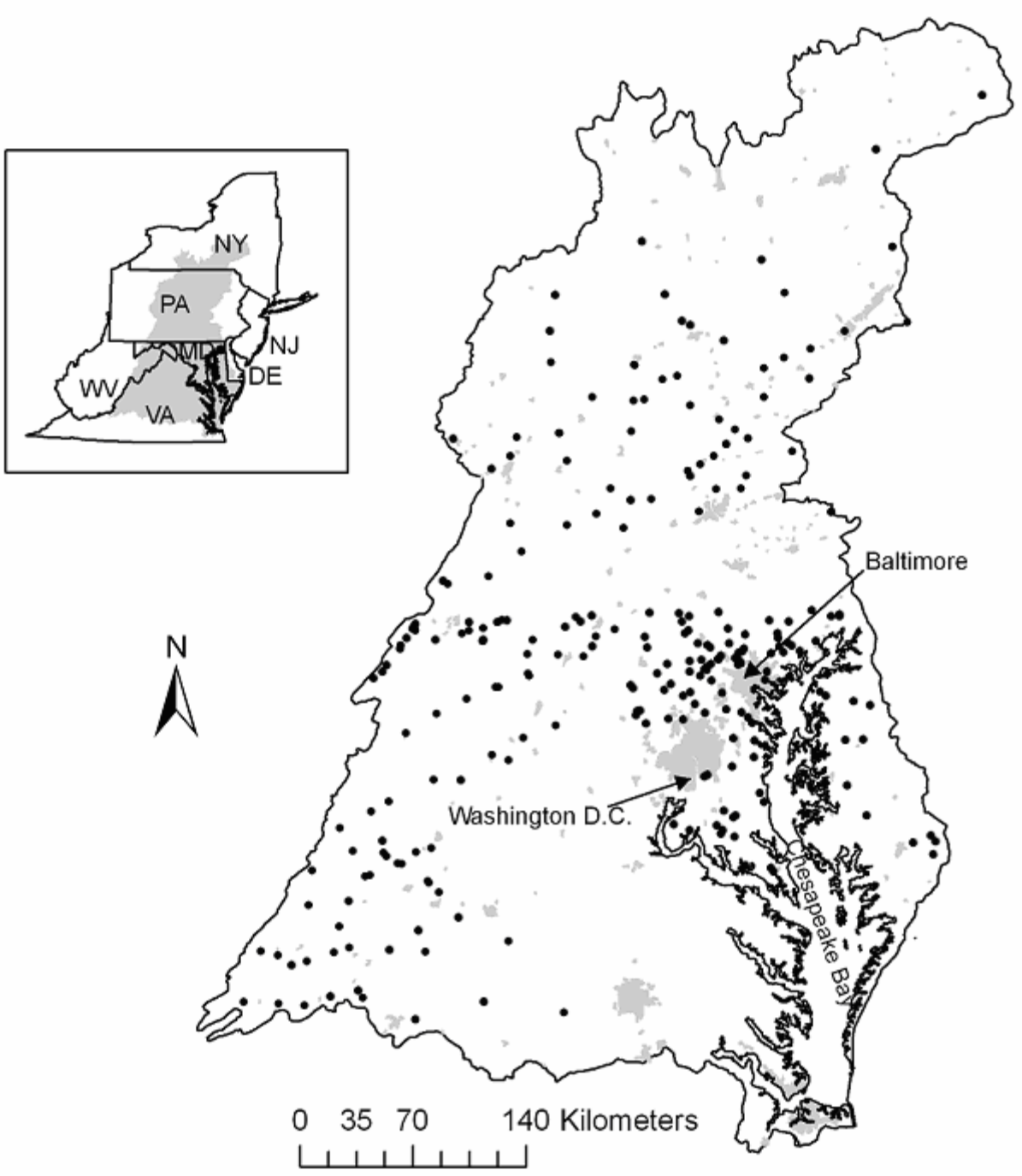


Figure 2. Distribution of sampling sites within 1st-3rd order streams of the CBW.

Land cover/land use data

Land use, biological indicator, and other environmental attribute data collected by various research institutions and centers (i.e. the University of Maryland's Regional Earth Science Application Center (RESAC, http://www.geog.umd.edu/resac), Smithsonian Environmental Research Center (SERC), etc.) were used in the analyses. All land use / land cover analyses were conducted in ArcGIS 9.3 (ESRI, Redlands, California). For each sampling site's associated upslope watershed, the land cover calculations included: percentage of the watershed as row crop agriculture, pasture, mining, impervious surface cover, and tree cover were calculated using land use/land cover data. The average catchment slope and elevation was also determined using a digital elevation model (DEM, 30 m, http://edc2.usgs.gov/geodata/index.php) (Table 2). Variable screening showed a high correlation between elevation and slope ($r = 0.80$), so slope was not considered in the development of the models. Finally, the average annual precipitation for each catchment was calculated using the Parameter-elevation Regressions on Independent Slopes Model (PRISM) (Daly et al., http://www.prism.oregonstate.edu/docs/index.phtml), a model that uses point data, a digital elevation model, and other spatial data sets to generate gridded estimates.

Table 2. Predictor variables used in the analyses.

Variable	Description
Elev	Average elevation of watershed (m)
IMP	Percentage of watershed under impervious surface cover
PRECIP	Average annual precipitation for a watershed (cm)
PerCrop	Percentage of watershed under row crop agriculture
PerPast	Percentage of watershed under pasture cover
PerTree	Percentage of watershed under tree cover
Per17	Percentage of watershed under mining cover
Slope	Average slope of watershed
WSAREA	Watershed area upslope of sampling location (km^2)

Models

Five different empirical models were used to predict the biological condition of 1st - 3rd streams within the Chesapeake Bay watershed. The models included the following:

1. Classification and regression tree (CART) model
2. Random forest (RF) model
3. Conditional Tree (cTree) model
4. Conditional Random Forest (cRF) model
5. Ordinal logistic regression (OLR) model.

1) CART Model

A classification and regression tree (CART) is a nonparametric recursive partitioning method (Breiman et al. 1984, Vayssieres et al. 2000). This method distinguishes differences among groups by examining combinations of explanatory variables and uses an algorithm to repeatedly partition the data set to find the 'purest' subgroups (Breiman et al. 1984). This process results in a tree-like structure that can explain differences between groups in terms of particular explanatory variables, show non-linear relationships, and display possibly compensatory relationships (i.e. alternative habitat within the same model) (Taverna et al. 2004). The variables that produce splits high in the tree are more influential in distinguishing general differences among the groups, while the variables that produce the last splits provide more detail, resolving idiosyncratic cases (Urban et al. 2002). However, a classification model may tend to describe the data to the point of "over-fitting" (Urban et al. 2002). In order to avoid over-fitting, "pruning" the data is useful for balancing both model accuracy and model robustness (Breiman et al. 1984). "Pruning" is a cross-validation process by which low-level splits are snipped off while higher-level splits (including the variables that best describe the differences between

groups) are retained (Landwehr et al. 2007). The tree analysis was performed using the rpart R library (Therneau & Atkinson 2007).

2) Random Forest Model

A Random Forest (RF) is an algorithm that uses an ensemble of classification trees (Breiman 2001, Cutler et al. 2007). Each tree in the RF classification is constructed from a bootstrapped sample of the original dataset where approximately two-thirds of the original sample occurs at least once. At each split in the trees, the model uses random variable selection. In addition to bootstrapping and random variable selection in tree building, the trees remain unpruned, and the RF chooses the classification with the most votes. In this manner, the tree predictors in each tree depend on the values of a random vector sampled independently and with the same distribution for all the trees within the forest (Breiman 2001). The RF model also provides variable importance plots that are based on how much prediction error increases when out-of-bag data for that variable is permuted while other variables remain constant (Liaw & Weiner 2002, Breiman 2001). Additionally, the RF's partial dependence plots provide a graphical depiction of the marginal effect of a variable on the class probability or response, while holding all other variables constant (Cutler et al. 2007). However, these plots are not helpful for categorizing or interpreting high order interactions (Cutler et al. 2007). Another disadvantage of RF models is that their outputs may not be as easily interpretable as those of CART models.

By combining trees in this manner, the RF model provides several benefits. The RF model is typically more robust than single trees (i.e. CART models), and can cope with complex interactions and highly correlated predictor variables (Strobl et al. 2008). Additionally, and the model's bootstrapping and random variable selection result in low correlation of individual trees (Breiman et al. 2000). Finally, leaving the trees unpruned tends to produce low-bias trees

(Breiman et al. 2001). While the RF model has these favorable characteristics, is not designed specifically to handle ordinal data may not be as easily interpretable as CART models (Strobl et al. 2007). In this analysis, the RF model (# trees = 500) was created using the randomForest R library (Liaw & Wiener 2002).

3) Conditional Tree Model

A Conditional Inference Tree model (cTree) estimates a regression relationship by recursively partitioning data in a conditional inference framework (Hothorn et al. 1996). The model is applicable to many types of regression problems (e.g. nominal, ordinal, numeric, etc.), tends to not over-fit the data or have biased variable selection (Hothorn et al. 1996). The algorithm follows these steps: 1) tests the global null hypothesis of independence between any of the input variables and the response variable; if the hypothesis is not rejected, the model stops, but if the hypothesis is rejected, then the algorithm selects the input variable with the strongest association to the response variable (association measured by a p-value corresponding to a test for the partial null hypothesis), 2) a binary split is implemented on the selected input variable, 3) steps 1 and 2 are repeated recursively. The stopping criteria for the algorithm are based on hypothesis tests (e.g. $p < 0.05$), which provide a solution to over-fitting problems (Hothorn et al. 2006). However, one of the limitations of this model is that it is relatively new and experimental. In this analysis, the cTree model was created using the party R library (Hothorn et al. 2006).

4) Conditional Forest Model

A Conditional Inference Forest (cForest) model is an algorithm that uses an ensemble of classification trees in a conditional inference framework (Hothorn et al. 2007, Strobl et al. 2007).

The model is applicable to many types of regression problems (e.g. nominal, ordinal, numeric, etc.) and can produce unbiased variable importance measures (Strobl et al. 2007). The cForest model improves upon the cTree model and is typically more robust than single trees (Hothorn 2007). However, the outputs of cForest models may also not be as easily interpretable as those of cTree models. Similar to the cTree model, one of the disadvantages of cForest is that it is relatively new and experimental. In this analysis, the cForest model was created using the party R library (Hothorn et al. 2006).

5) Ordinal Logistic Regression

Logistic Regression is a type of generalized linear model (GLM), in which the response variable is binary, i.e. that is, binomial. When the response factor has more than two levels, the response is multinomial. Multinomial logits can be developed in two ways; nominal (unsorted) or ordinal (ranked). Ordinal Logistic regression (OLR) has been used successfully in classification applications in other disciplines (Worth & Cronin 2002). In this analysis, OLR was used to distinguish three levels of stream biological condition: "Poor," "Fair," and "Good". The Design R library (Harrell 2008) was used to create the OLR of stream biological condition within 1^{st}-3^{rd} order streams of the Chesapeake Bay watershed.

Model Evaluation and Accuracy Assessment

Each of the models was evaluated using resubstitution, a method in which a model is tested with the same data used to create the model (Theodoridis & Kourtroumbas 2006). However, this method typically underestimates the true error probability, or classification error of the model (Theodoridis & Kourtroumbas 2006). Therefore, models were also evaluated using 10-fold cross validation. In this type of evaluation, the data was divided into 10 subsets of equal size and the model was trained 10 times, each time leaving out one subset (Kohavi 1995a). In this manner, the mean accuracy of each model was determined by using independent data not used to create the model.

The success of each of the models was then summarized using a confusion matrix, where the aim is to maximize true classifications and minimize false classifications. In a confusion matrix, the columns correspond to the actual classifications of stream biological condition and the rows to the predicted classifications. The confusion matrix also shows the number of correctly classified and misclassified cases, which can be used to calculate the classification error and percent correctly classified (PCC) (Fielding & Bell 1997).

In addition to the confusion matrices, weighted Cohen's Kappa coefficients were employed to quantify the level of agreement between multiple ratings of categorical variables (Cohen 1968). Weighted Cohen's Kappa is the proportion of agreement corrected for chance, ranging from -1 to +1 (Cohen 1968). A coefficient of 1.0 indicates maximum possible agreement, while zero indicates exactly chance agreement, and a negative kappa coefficient indicates a less than chance agreement (Brenner & Kliebsch 1996, Cohen 1968). The following table shows somewhat arbitrary, but useful "benchmarks" for understanding the strength of agreement of weighted Cohen's Kappa coefficients (Landis and Koch 1977):

Table 3. Kappa statistics and strength of agreement.

Kappa Statistic	Strength of Agreement
<0.00	Poor
0.00 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 1.00	Almost Perfect

All of the models used in this study were completed using R-statistical software (R Development Core Team 2007). The weighted Cohen's Kappas were calculated using the Kappa function in the vcd R library (Meyer et al. 2008).

Model Selection & Extrapolation to Watershed

The model with the highest percent correctly classified (PCC) and highest weighted Cohen's Kappa coefficient under the 10-fold cross-validation was identified. This model was then extrapolated to predict the biological condition (Good, Fair, Poor) of 1^{st} – 3^{rd} order streams within the Chesapeake Bay watershed. Using the National Hydrography Dataset (NHD Plus) watersheds, predictions were made for a total of 71,182 sites within the Chesapeake Bay basin.

Results

Performance of Models

The performance of the five models under the resubstitution and 10-fold cross-validation evaluations is shown below (Table 4). For the resubstitution evaluation, the Conditional Forest (cForest) model had the highest percent correctly classified (76.2%), followed by the classification and regression tree (CART) model (63.1%), the Random Forest (RF) model (59.0%), the conditional Tree (cTree) model (53.5%), and lastly, the Ordinal Logistic Regression (OLR) model (51.7%). However, under the 10-fold cross-validation evaluation, the RF model

had the highest PCC (59.0%), followed by the cForest model (56.9%), the OLR model (49.3%), and lastly by the cTree and CART models (45.9%).

In terms of the weighted Cohen's Kappa coefficients, the cForest model performed the best during the resubstitution evaluation (0.68 – "Substantial" strength of agreement), followed by the CART model (0.49 – "Moderate" strength of agreement), the RF model (0.42 – "Moderate" strength of agreement), the cTree model (0.34 – "Fair" strength of agreement), and lastly, the OLR model (0.284 – "Fair" strength of agreement). However, under the 10-fold cross-validation evaluation, the RF model received the highest weighted Cohen's Kappa coefficient (0.40 – "Fair"), followed by cForest model (0.39 – "Fair"), OLR model (0.24 – "Fair"), cTree model (0.08 – "Slight"), and finally, the CART model (0.05 – "Slight").

Table 4. Accuracy measures for predictions of stream conditions within the Chesapeake Bay watershed. Resub = resubstitution accuracy estimates, Xval = 10-fold cross validation accuracy estimates, PCC = percentage correctly classified.

		Accuracy measure	
Classification method	Estimate	PCC	Weighted Kappa (Std. Error) & Strength of Agreement
Classification & Regression Tree (CART)	Resub	63.1	0.49 (0.065) – "Moderate"
	Xval	45.9	0.05 (0.055) – "Slight"
Random Forest (RF)	Resub	59.0	0.42 (0.065) – "Moderate"
	Xval	59.0	0.40 (0.065) – "Fair"
Conditional Tree (cTree)	Resub	53.5	0.34 (0.062) – "Fair"
	Xval	45.9	0.08 (0.057) – "Slight"
Conditional Forest (cForest)	Resub	76.2	0.68 (0.067) – "Substantial"
	Xval	56.9	0.39 (0.064) – "Fair"
Ordinal Logistic Regression (OLR)	Resub	51.7	0.28 (0.058) – "Fair"
	Xval	49.3	0.24 (0.058) – "Fair"

Under the resubstitution evaluation, the cForest model performed better than the other four models in terms of its PCC and weighted Cohen's Kappa coefficient. However, under the 10-fold cross-validation evaluation, the RF model received the highest PCC and weighted Cohen's Kappa coefficient. Since the resubstitution evaluation method typically underestimates the true error probability (Theodoridis & Kourtroumbas 2006), the best performing model under

the 10-fold cross-validation evaluation, the RF model, was selected as the highest performing model. The RF model was then extrapolated to predict the biological condition of 1st – 3rd order streams within the Chesapeake Bay watershed. All of the tree models performed better than the OLR, and so this model is not discussed further. However, the results of the other models (CART, RF, cTree, cForest) can be found in the Appendices.

Random Forest Model Performance

The resubstitution confusion matrix shows that the Random Forest (RF) model and data agree on 52% of cases classified as having "poor" biological condition, 42% of cases as having "fair" biological condition, and 74% of cases classified as having "good" biological condition (Table 5). The confusion matrix shows that model misclassified 21 "poor" sites as "good" sites, and misclassified 43 "fair" sites as "good" sites. Overall, the resubstitution evaluation shows that the random forest (RF) model accurately predicts the biological condition of 1st-3rd order streams 59% of the time.

Table 5. Resubstitution confusion matrix for the Random Forest model.

		Data			
		Poor	Fair	Good	Classification error
Model Predictions	Poor	32	9	21	0.48
	Fair	10	38	43	0.58
	Good	8	28	101	0.26

The 10-fold cross-validation confusion matrix shows that the Random Forest (RF) model and data agree on 61% of cases classified as having "poor" biological condition, 53% of cases as having "fair" biological condition, and 61% of cases classified as having "good" biological condition (Table 6). The confusion matrix shows that the model misclassified 25 "fair" sites as "good" sites and misclassified 43 "good" sites as "fair" sites. Overall, the 10-fold cross-

validation shows that the model accurately predicts the biological condition of 1st-3rd order streams within the Chesapeake Bay watershed 59% of the time.

Table 6. Ten-fold cross-validation confusion matrix for the Random Forest model.

		Data			
		Poor	Fair	Good	Classification error
Model Predictions	Poor	31	10	10	0.39
	Fair	9	38	25	0.47
	Good	22	43	102	0.39

Variable Importance

The variable importance plot generated by the Random Forest (RF) model showed that watershed area (upslope of sampling location, km^2) was the most important variable, followed by percentage of impervious surface cover, and percentage of pasture cover (Figure 3). The least valuable variables were precipitation (average annual precipitation for a watershed, cm) and % mining cover.

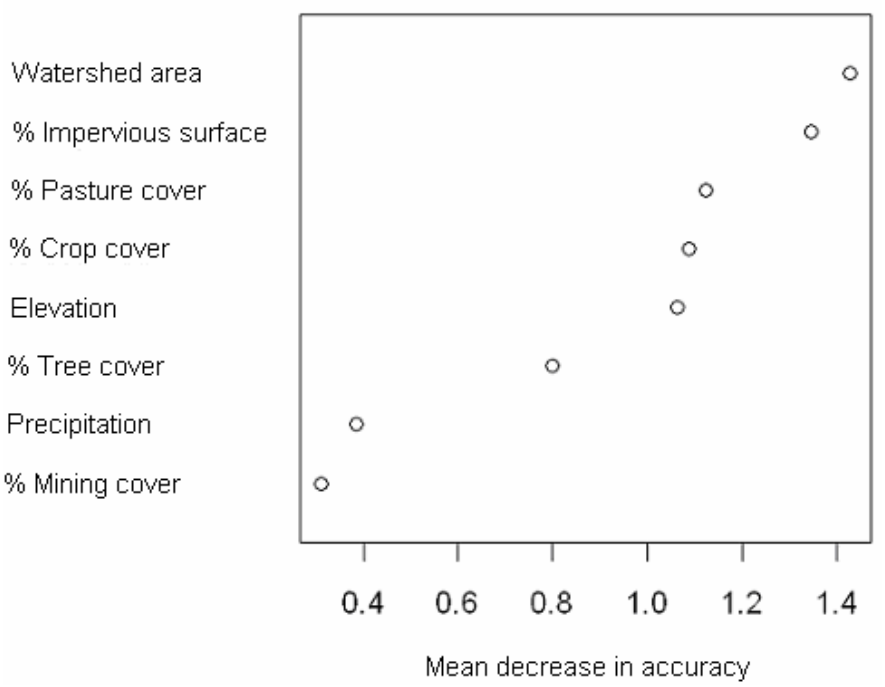


Figure 3. Random Forest variable importance plot for predictions of CBW biological condition. The horizontal axis presents the importance measure whereas the vertical axis denotes the variables.

Partial Dependence Plots

The Random Forest (RF) generated partial dependence plots that show the marginal effect of each variable on the class probability, while holding all other variables constant. As watershed area (km^2) increases, there is a lower probability of a "poor" classification of biological condition within the Chesapeake Bay watershed (Figure 4).

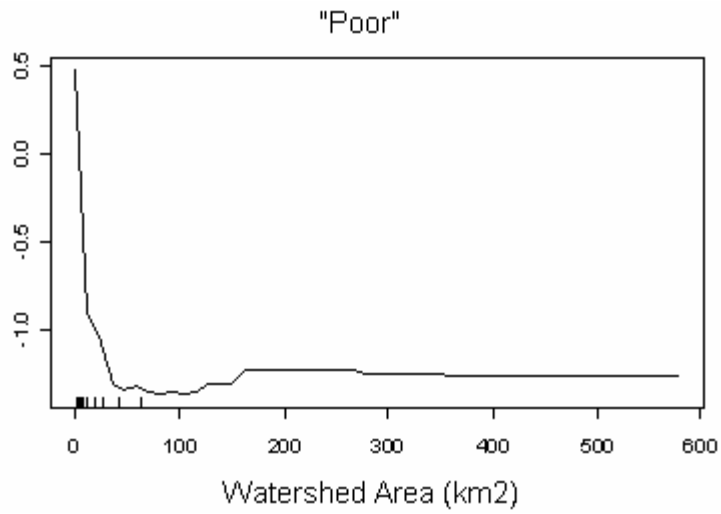


Figure 4. Partial dependence plot showing the effect of the watershed area variable on the probability of a "Poor" classification.

However, as watershed area (km^2) increases, there is a higher probability of a "fair" or "good" classification of stream biological condition (Figure 5, Figure 6). The threshold watershed area is approximately 20 – 25 km^2.

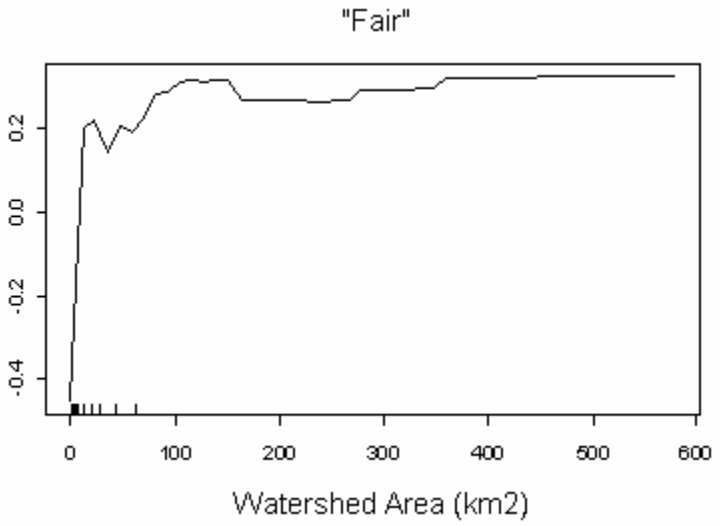


Figure 5. Partial dependence plot showing the effect of the watershed area variable on the probability of a "Fair" classification.

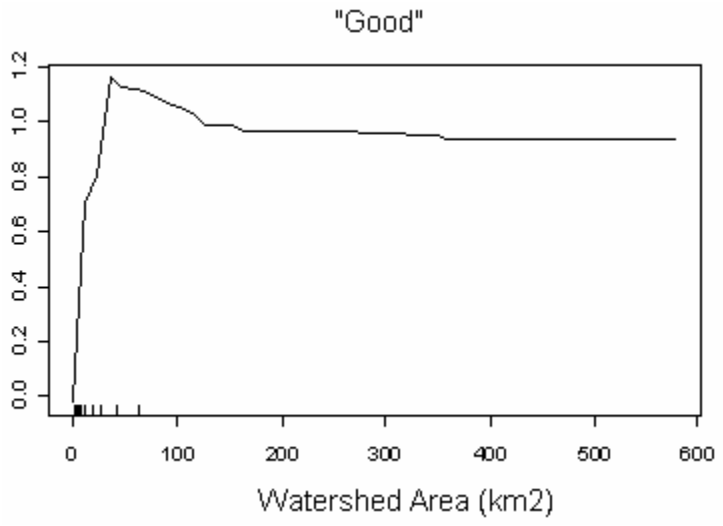


Figure 6. Partial dependence plot showing the effect of the watershed area variable on the probability of a "Good" classification.

The following partial dependence plot shows that as percentage of impervious surface cover increases, there is a greater probability of a poor classification of stream condition (Figure 7). The threshold of impervious surface cover is approximately 5%.

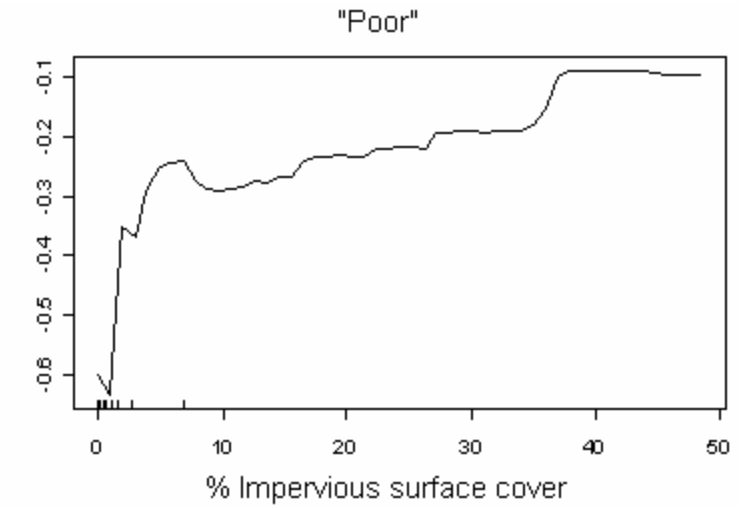


Figure 7. Partial dependence plot showing the effect of percentage of impervious surface cover area on the probability of a "Poor" classification.

Conversely, as the percentage of impervious surface cover increases, there is a decrease in the probability of "Good" classification of stream biological condition.

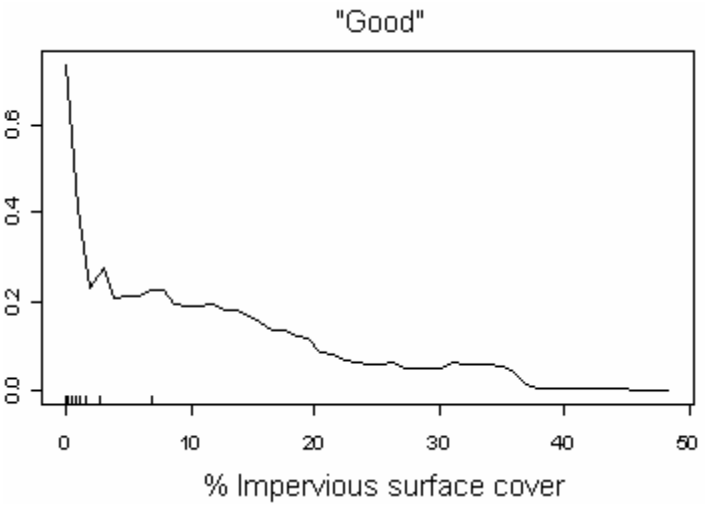


Figure 8. Partial dependence plot showing the effect of percentage of impervious surface cover area on the probability of a "Good" classification.

The following partial dependence plot shows that as percentage of pasture cover increases, there is a decrease in the probability of a poor classification of stream biological condition (Figure 9).

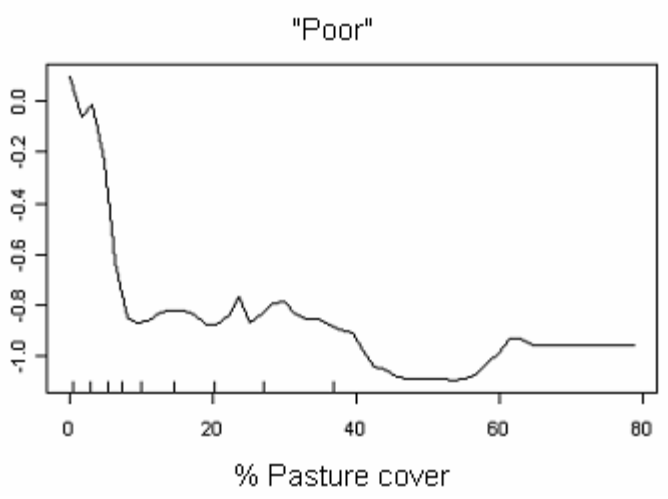


Figure 9. Partial dependence plot showing the effect of percentage of pasture cover on the probability of a "Poor" classification.

However, as percentage of pasture cover increases, there is an increase in the probability of a "Fair" stream classification (Figure 10).

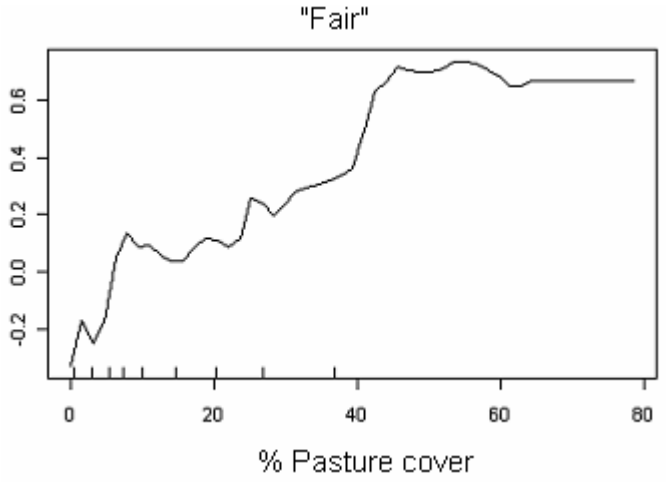


Figure 10. Partial dependence plot showing the effect of percentage of pasture cover on the probability of a "Fair" classification.

As pasture cover increases from approximately 0 – 10 %, there is also an increase in the probability of a "Good" classification of stream biological condition (Figure 11). However, when pasture cover increases beyond 10 -15%, the probability of a "Good" classification of stream condition decreases. Additional partial dependence plots for the Random Forest model can be found in Appendix C.

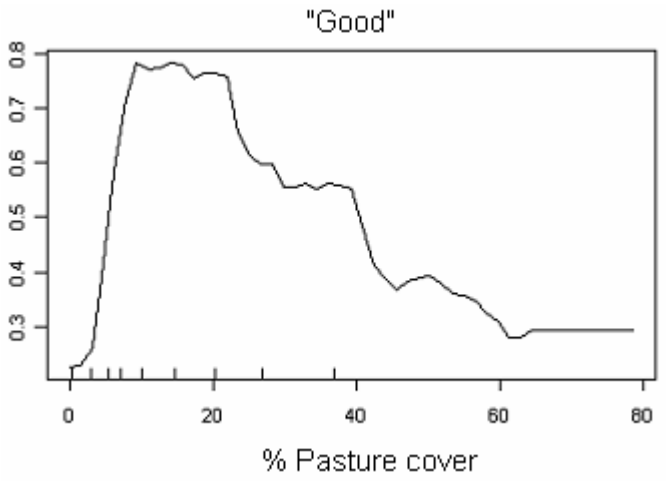


Figure 11. Partial dependence plot showing the effect of percentage of pasture cover on the probability of a "Good" classification.

Random Forest Model Extrapolation to CBW

The Random Forest (RF) model was extrapolated to unsurveyed stream sites within the Chesapeake Bay watershed using geographic information software (GIS). Of the 71,182 site predictions, 19,350 sites (27%) were classified as having "Poor" biological condition, 16,826 sites (24%) were classified as having "Fair" biological condition, and 35,006 (49%) were classified as having "Good" biological condition. The prediction map (Figure 12) shows a greater concentration of stream sites with "Poor" biological condition in the Southeastern Plains and Middle Atlantic Coastal Plain ecoregions that encompass large metropolitan areas such as Washington, D.C., and Baltimore, Maryland. Conversely, there is a greater concentration stream sites with "Good" or "Fair" biological condition classifications in the Northern Appalachian Plateau and Uplands, eastern portion of the Piedmont, and the Northeastern Highlands.

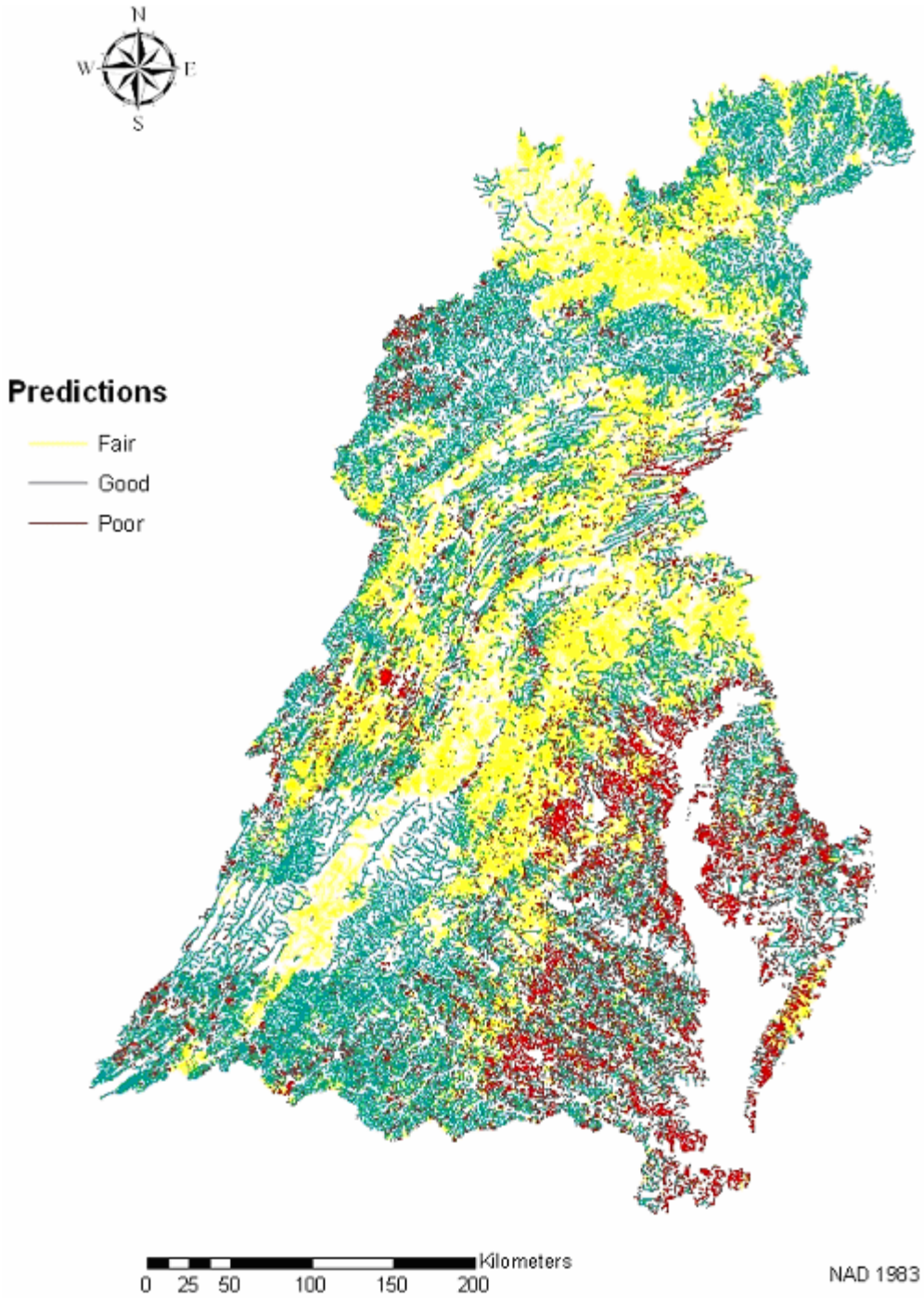


Figure 12. Random Forest predictions of the biological condition of 1st-3rd order streams within the Chesapeake Bay watershed.

Discussion and Conclusions

In this study, five empirical models were used to predict the biological condition of 1st - 3rd order streams within the Chesapeake Bay watershed. These included the CART model, Random Forest model, Conditional Tree model, Conditional Forest model, and ordinal logistic regression. Of these models, the Conditional Forest model performed the best under the resubstitution evaluation, with 76.2% correctly classified and a weighted Cohen's Kappa of 0.68 ("substantial" level of agreement). However, under the 10-fold cross-validation, the Random Forest model was the highest performing model with approximately 60% correctly classified and weighted Cohen's Kappa of 0.40 ("moderate" level of agreement). The Random Forest model was then extrapolated to 71,182 stream sites within the Chesapeake Bay watershed. Of these sites, 49% (35,006 sites) were classified as having "Good" biological condition, 24% (16,826 sites) as having "Fair" biological condition, and 27% (19,350 sites) as having "Poor" biological condition.

One of the variables that the Random Forest model highlighted as very important was watershed area (upslope of sampling location, km^2). As watershed area increases, there is a lower probability of a "poor" classification of biological condition. However, as watershed area (km^2) increases, there is a higher probability of a "fair" or "good" classification of stream biological condition; with a threshold watershed area of approximately 20 – 25 km^2. This may imply that these smaller watersheds are subject to greater levels of anthropogenic and/or environmental stress than larger watersheds within the Chesapeake Bay watershed. It may also be the case that smaller watersheds have not been the primary focus of environmental management within the Chesapeake Bay watershed (e.g. larger watersheds within the basin may have received more attention for restoration and/or conservation efforts).

Another variable that the Random Forest model highlighted as very important was percentage of impervious surface cover. As the percentage of impervious surface cover increases, there is a greater probability of a poor classification of stream condition (threshold of ~5% impervious surface cover). Conversely, as the percentage of impervious surface cover increases, there is a decrease in the probability of "Good" classification of stream biological condition. These findings are similar to previous studies that show a decline in biological condition with an increase in percentage of impervious surface cover (Booth et al. 2002, Klein 1979). However, this study's threshold of ~5% is lower than that of the other studies that report a threshold of ~10-15% (Klein 1979) or ~10-20% impervious surface cover (Schueler 1995).

Percentage of pasture cover was also an important variable within the Random Forest model. As percentage of pasture cover increases, there is a decrease in the probability of a poor classification of stream biological condition. However, as percentage of pasture cover increases, there is an increase in the probability of a "Fair" stream classification. There is also an increase in the probability of a "Good" classification as pasture cover increases from approximately 0 – 10%, but then when pasture cover increases beyond 10 -15%, the probability of a "Good" classification of stream condition decreases. This implies that to a certain extent, pasture cover may simultaneously diminish and enhance stream biological condition within the Chesapeake Bay watershed. On the one hand, human-created pasture lands replace naturally occurring vegetation and forest structure, which contributes to habitat loss and degradation. However, until pasture cover reaches approximately 10-15%, it may benefit stream biological condition by decreasing the prevalence of other human land use activities within the watersheds such as road-building or large installations of impervious surface.

Management Implications

The results of the RF model may help managers make more effective land use decisions. Since the RF model was able to correctly classify approximately 74% of the sites with "Good" biological condition under the resubstitution evaluation, the model may be useful for identifying sites for preservation within the CBW. Under the 10-fold cross-validation, the model correctly classified ~60% of sites. Managers can potentially use this model to make approximations of stream biological condition before validating this classification in the field. Additionally, managers can use the model to understand the marginal effect of particular land use and environmental variables on the stream biological condition of their areas. For example, if their area has a high percentage of impervious surface or pasture, then they can work to decrease the negative impacts of these land uses. Some potential solutions may include incentives to promote the use of permeable surfaces, improved regulations of storm-water run-off, updated land use plans, and information sharing and outreach programs to educate members of industry, agricultural, and residential sectors about the negative effects of human land use activities on streams within the CBW.

Areas of Future Research and Development

There are several areas of future research and development that can improve the ability of managers to understand stream biological condition. One of these areas could be the development of models that are more accurate at predicting biological condition and more consistent among accuracy measures. Managers should also consider standardizing monitoring and increase the number of sampling sites across the entire CBW. Finally the creation of an online database could assist managers in learning more about the effects of human land use

activities in their particular region, and the various techniques they can use to improve stream biological condition within the Chesapeake Bay watershed.

References

Allan, J.D. 2004. Landscapes and riverscapes: The influence of land use on stream ecosystems. *Annual Review of Ecology, Evolution, and Systematics* **35**:257-284.

Allan, J.D., Flecker A.S. 1993. Biodiversity conservation in running waters. *Bioscience* **43**:32-43.

Booth, D.B., Hartley D., Jackson R. 2002. Forest cover, impervious surface area, and the mitigation of stormwater impacts. *Journal of the American Water Resources* Association **38**:835-845.

Breiman L., Friedman J.H., Olshen R.A. & C.J. Stone. 1984. Classification and Regression Trees. The Wadsworth Statistics/Probability Series. Chapman & Hall Inc., New York, USA.

Breiman, L. 2001. Random forests. *Machine Learning* **45**:5-32.

Brenner, H., Kliebsch U. 1996. Dependence of weighted kappa coefficients on the number of categories. *Epidemiology* **7**:199-202.

Cohen, J. 1968. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* **70**:213.

Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A., Hess K.T., Gibson J., Lawler J.J. 2007. Random forests for classification in ecology. *Ecology* **88**(11)2783-2792.

Dynesius, M., Nilsson C. 1994. Fragmentation and flow regulation of river systems in the northern third of the world. Science **266**:753-762.

Fielding, A.H., J.F. Bell. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation* **24**(1)38-49.

Goetz, S. J., Jantz C.A., Prince S.D, Smith A.J., Wright R., Varlyguin D. 2004a. Integrated analysis of ecosystem interactions with land use change: the Chesapeake Bay watershed. Pages 263–275 in R. S. DeFries, G. P. Asner, R.A. Houghton (eds.), Ecosystems and land use.

Harrell, F.E. 2008. Design: Design package. R package version 2.1-2. http://biostat.mc.vanderbilt.edu/s/Design, http://biostat.mc.vanderbilt.edu/rms.

Hothorn, T., Hornik K., Zeilis A. 2006. Unbiased recursive partitioning: a conditional inference framework. *Journal of Computational and Graphical Statistics* **15**(3):651-674.

Hothorn, T., Zeileis A., Hornik K. 2007. Let's have a party! an open-source toolbox for recursive partytioning. Research Report Series **59**. Department of Statistics and Mathematics, Wirtschaftsuniversität Wien, Wien, Austria.

HSC. 2006. National Hydrography Dataset Plus. Horizon Systems Corporation. Accessed on January 1, 2009 at http://www.horizonsystems.com/nhdplus.

Jantz, P., Goetz S. J., Jantz C.A. 2005. Urbanization and the loss of resource lands within the Chesapeake Bay Watershed. *Environmental Management* **36**:343-360.

Klein, R. 1979. Urbanization and stream quality impairment. *Water Resources Bulletin* **15**:948-963.

Kohavi, R. 1995a. A study of cross-validation and bootstrap for accuracy estimation and model selection. In C.S. Mellish (Ed.), *Proceedings of the 14th International Joint Conference on Artificial Intelligence* (pp. 1137–1143).

Landis, R., Koch G.G. 1977. The measurement of observer agreement for categorical data. *Biometrics* **33**(1):159-174.

Landwehr, N., Hall M., Frank E. 2003. Logistic model trees. *Machine Learning*, **59**:161-205.

Liaw, A., Wiener M. 2002. Classification and regression by randomForest. *R News* **2**(3), 18-22.

Meyer, D., Zeileis A., Hornik K. 2008. Vcd: visualizing categorical data. R package version 1.0-9.

Moore, D.M., Lee B.G., Davey S.M. 1991. A new method for predicting vegetation distributions using decision tree analysis in a geographic information system. *Environmental Management* **15**:59-71.

Omernik, J. M. 1987. Ecoregions of the conterminous United States. Annals of the Association of American Geographers **77**:118-125.

Peel, M.C., Finlayson B.L., McMahon T. A. 2007. Updated world map of the Koppen-Geiger climate classification. Hydrology and Earth System Sciences **11**:1633-1644.

Schueler, T. 1995. *Site Planning for Urban Stream Protection.* Metropolitan Washington Council of Governments: Washington D.C.

Strayer D. L., Beighley R.E., Thompson L.C, Brooks A., Nilsson C., et al. 2003a. Effects of land cover on stream ecosystems: roles of empirical models and scaling issues. *Ecosystems* **6**:407–23

Strobl, C., Boulesteix, A.L., Zeileis, A., Hothorn, T. 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics* **8**:25.

Strobl, C., Boulesteix A., Kneib T., Augustin T., Zeileis A. 2008. Conditional variable importance for random forests. *BMC Bioinformatics* **9**:307.

R Development Core Team. 2007. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Taverna, K., Urban D.L, McDonald R.I. 2004. Modeling landscape vegetation pattern in response to historic land-use: a hypothesis driven approach for the North Carolina Piedmont, USA. *Landscape Ecology* **20**:689-702.

Theodoridis, S., Kourtroumbas K. 2006. Pattern Recognition, Ed. III. Academic Press.

Therneau, T.M., Atkinson B. 1997. An introduction to recursive partitioning using the rpart routine. Technical Report 61, Section of Biostatistics, Mayo Clinic, Rochester, 1997.

Urban, D.L., Goslee S., Pierce K., Lookingbill T. 2002. Extending community ecology to landscapes. *Ecoscience* **9**(2):200-202.

U.S. E.P.A. 1998. Surface waters: Field operations and methods for measuring the ecological condition of wadeable streams. EPA/620/R-94/004F, Office of Research and Development, United States Environmental Protection Agency, Washington DC.

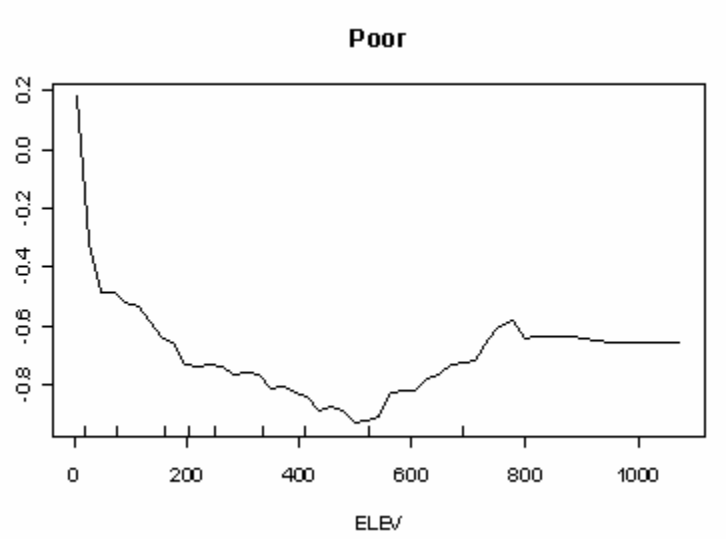
U.S. E.P.A. 2006b. Wadeable Streams Assessment: A collaborative survey of the Nation's streams. EPA 841-B-06-002, Office of Water, United States Environmental Protection Agency, Washington, DC.

Vassières, M.P., Plant R.E., Allen-Diaz B.H. 2000. Classification trees: an alternative nonparametric approach for predicting species distributions. *Journal Vegetation Science* **11**:679-694.

Weber, T. 2004. Landscape ecological assessment of the Chesapeake Bay Watershed. *Environmental Monitoring and Assessment* **94**: 39-53.

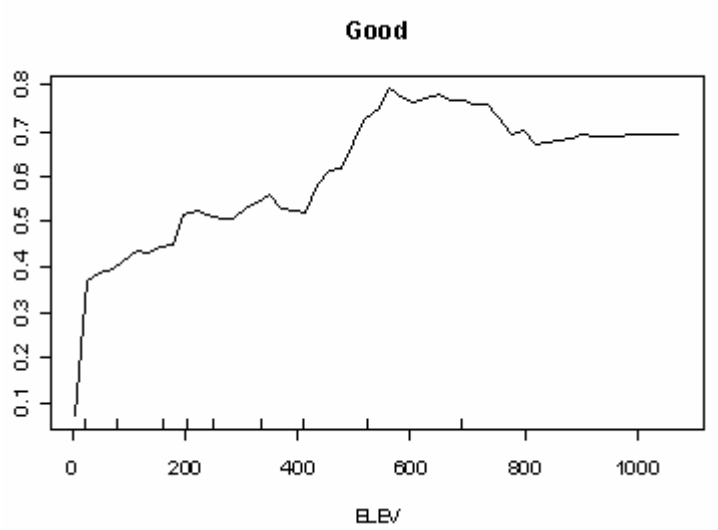
Worth, A.P., Cronin M.T.D. 2002. The use of discriminant analysis, logistic regression and classification tree analysis in the development of classification models for human health effects. *Journal of Molecular Structure: THEOCHEM* **622**(1-2):97-111.

APPENDIX A: Random Forest Partial Dependence Plots

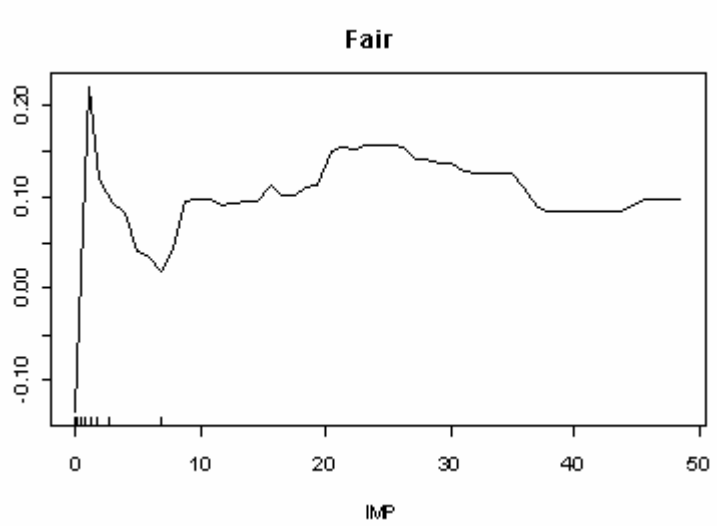


A.1 Partial dependence plot showing the effect of the elevation variable on the probability of a "Poor" classification.

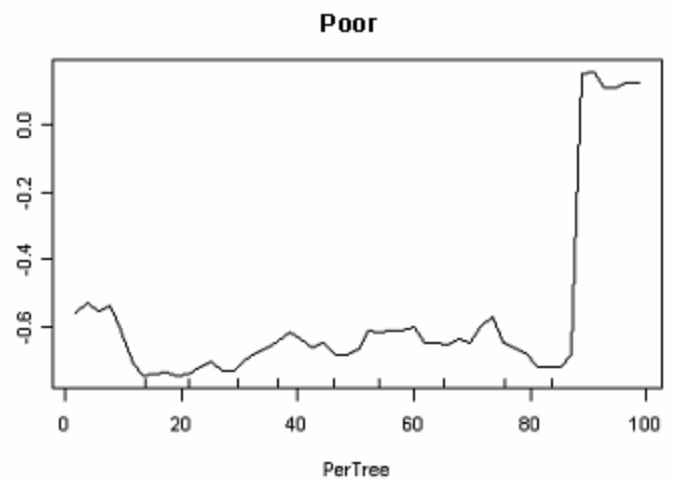
A.2 Partial dependence plot showing the effect of the elevation variable on the probability of a "Fair" classification.



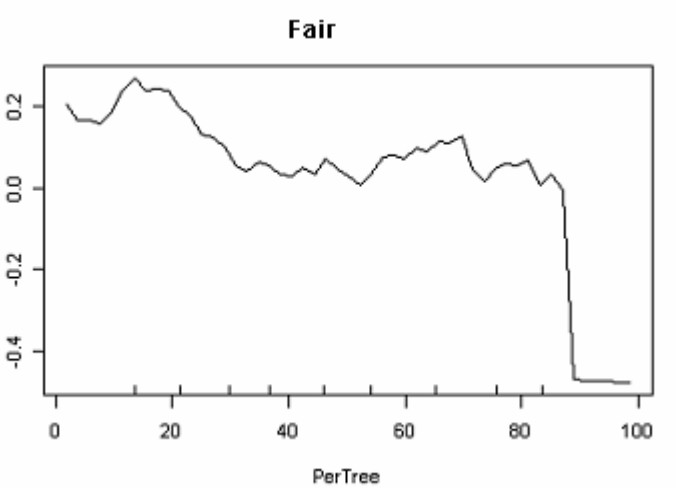
A.3 Partial dependence plot showing the effect of the elevation variable on the probability of a "Good" classification.



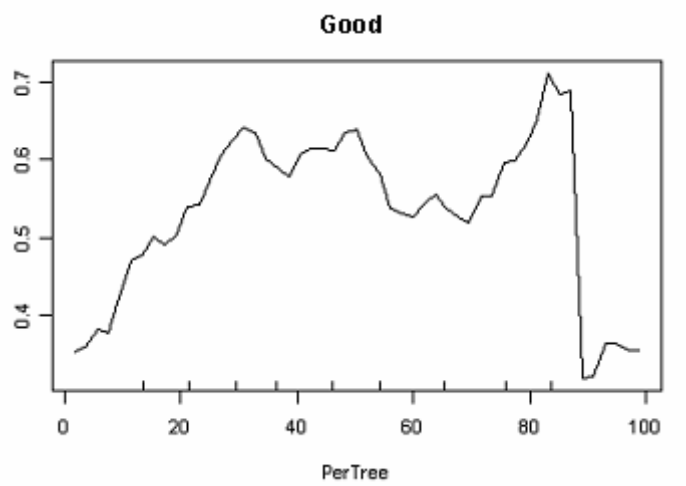
A.4 Partial dependence plot showing the effect of the percentage impervious surface variable on the probability of a "Fair" classification.



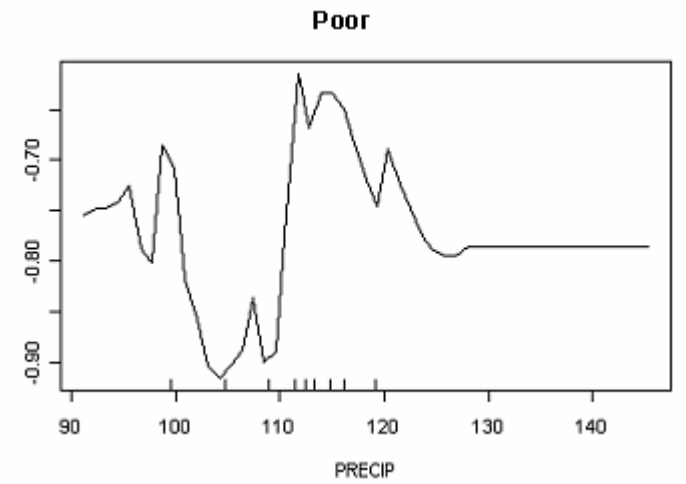
A.5 Partial dependence plot showing the effect of the percentage of tree cover on the probability of a "Poor" classification.



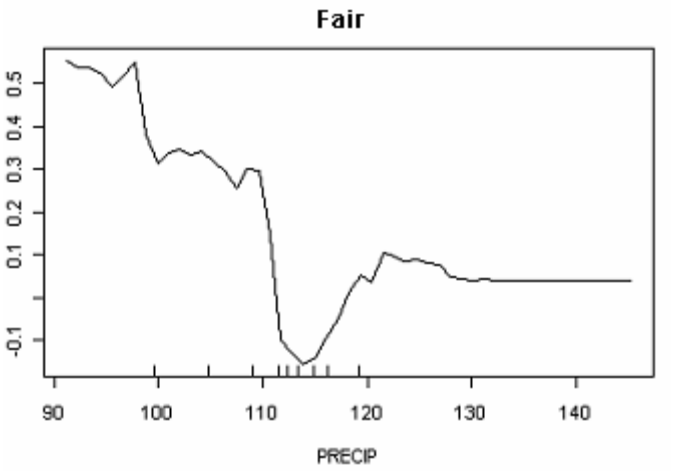
A.6 Partial dependence plot showing the effect of the percentage of tree cover on the probability of a "Fair" classification.



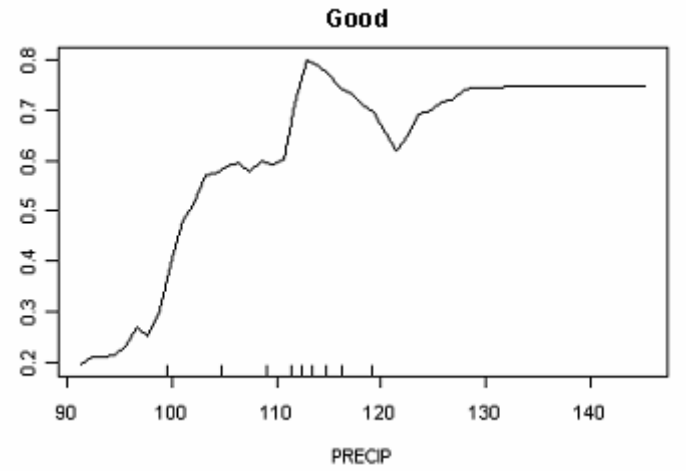
A.7 Partial dependence plot showing the effect of the percentage of tree cover on the probability of a "Good" classification.



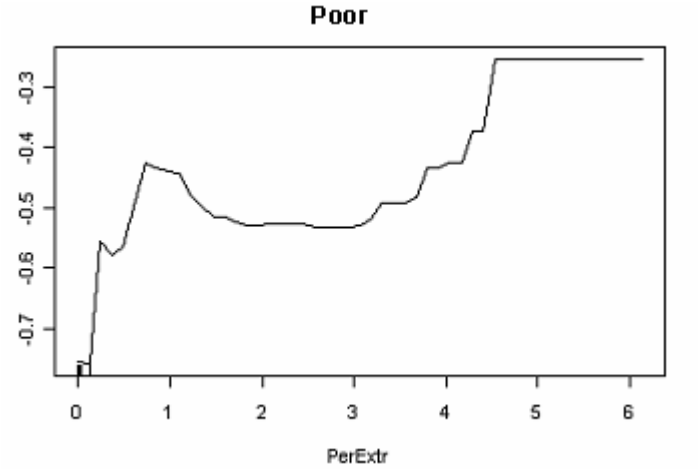
A.8 Partial dependence plot showing the effect of precipitation on the probability of a "Poor" classification.



A.9 Partial dependence plot showing the effect of precipitation on the probability of a "Fair" classification.

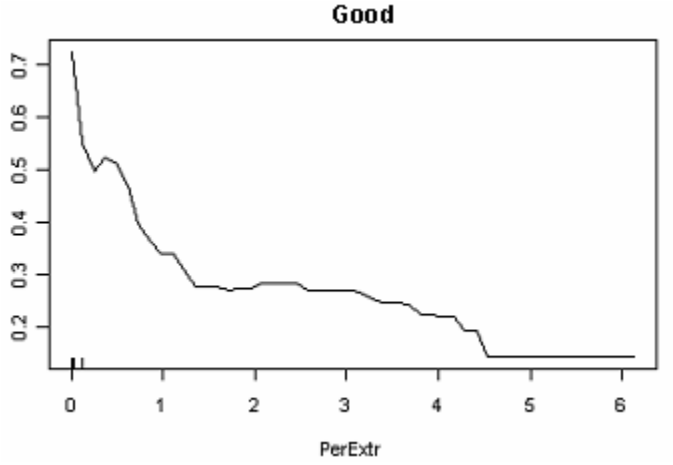


A.10 Partial dependence plot showing the effect of precipitation on the probability of a "Good" classification.

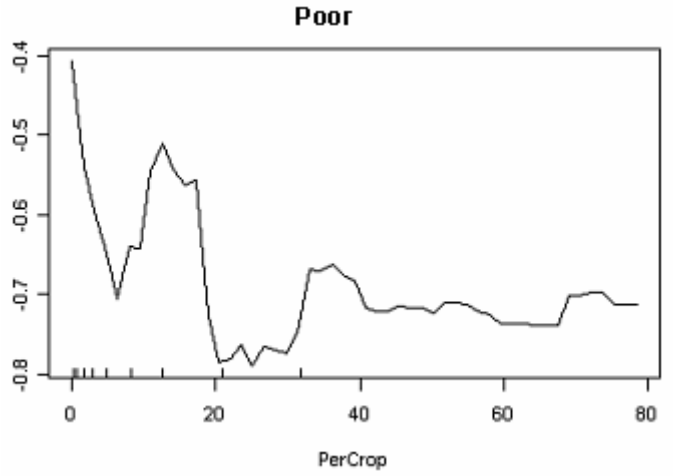


A.11 Partial dependence plot showing the effect of percentage of mining cover on the probability of a "Poor" classification.

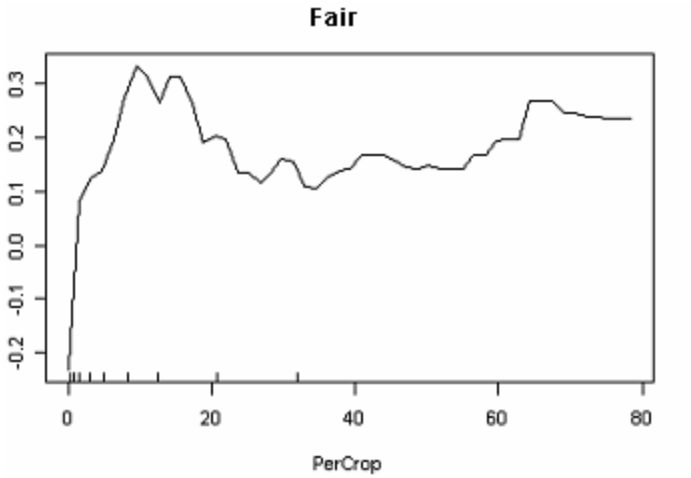
A.12 Partial dependence plot showing the effect of percentage of mining cover on the probability of a "Fair" classification.



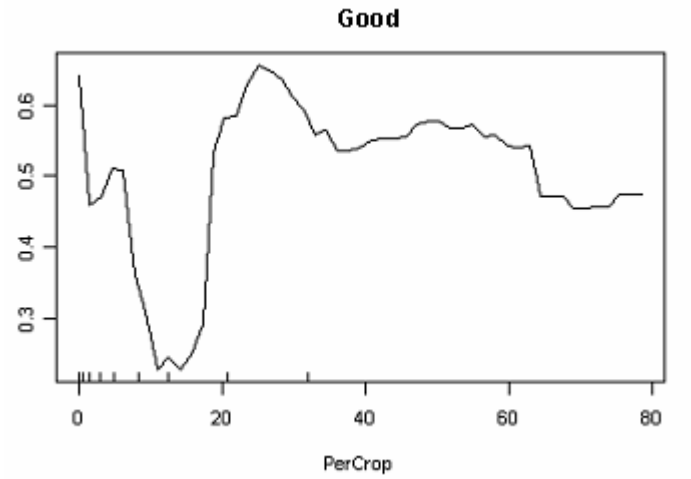
A.13 Partial dependence plot showing the effect of percentage of mining cover on the probability of a "Good" classification.



A.14 Partial dependence plot showing the effect of percentage of crop cover on the probability of a "Poor" classification.



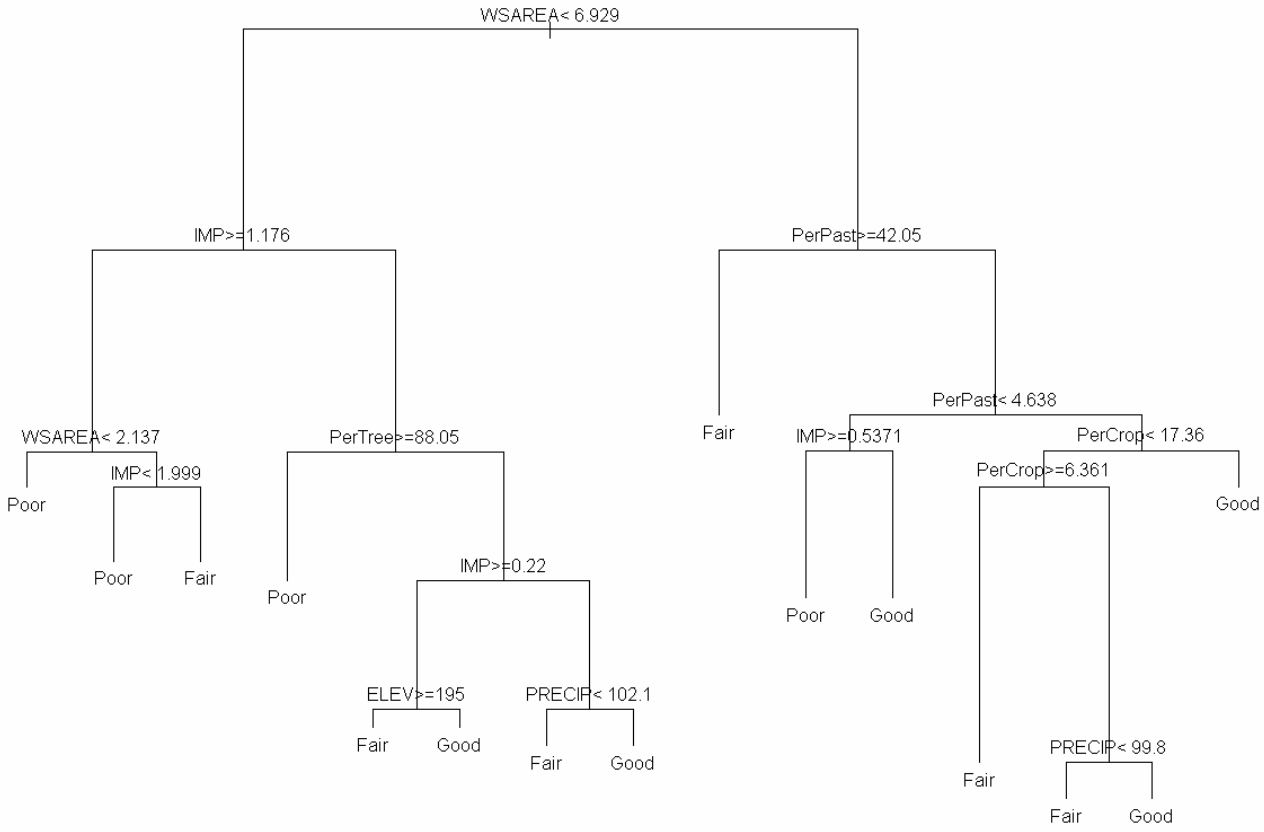
A.15 Partial dependence plot showing the effect of percentage of crop cover on the probability of a "Fair" classification.



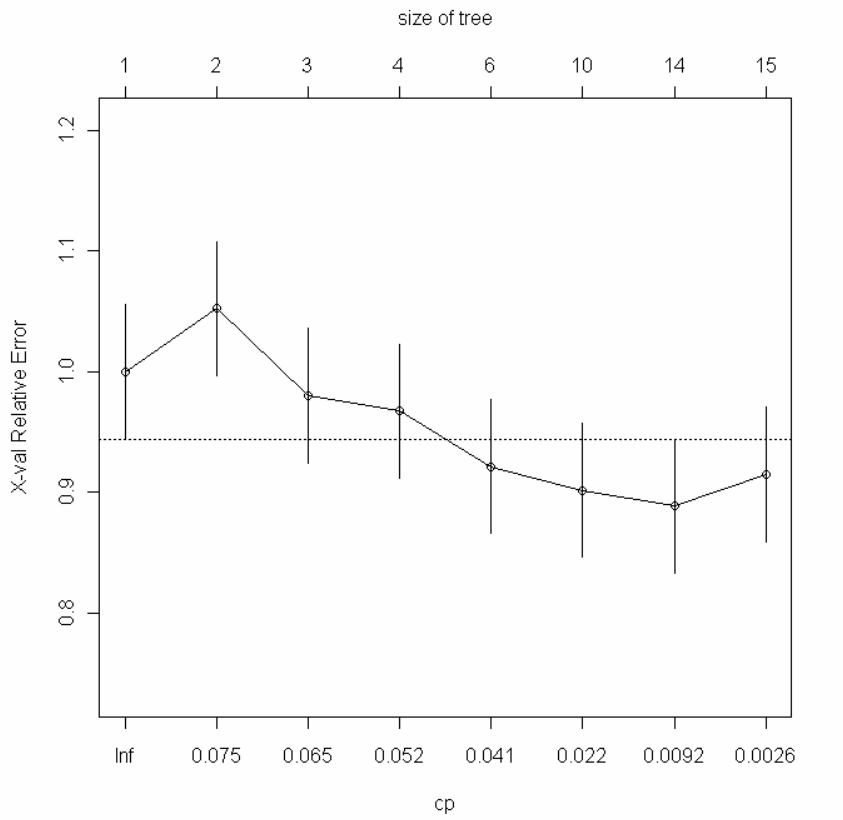
A.16 Partial dependence plot showing the effect of percentage of crop cover on the probability of a "Good" classification.

APPENDIX B: CART Model Results

The classification tree for the biological condition of 1st-3rd order streams within the Chesapeake Bay watershed has 15 nodes (i.e. branches). The tree also shows that the key explanatory variables for the key explanatory variables for determining biological condition are watershed area (<6.929 m2), percent impervious surface (>= 1.176%), and percent pasture (>= 42.05%) (B.1). The variables with the least explanatory ability are elevation and precipitation.

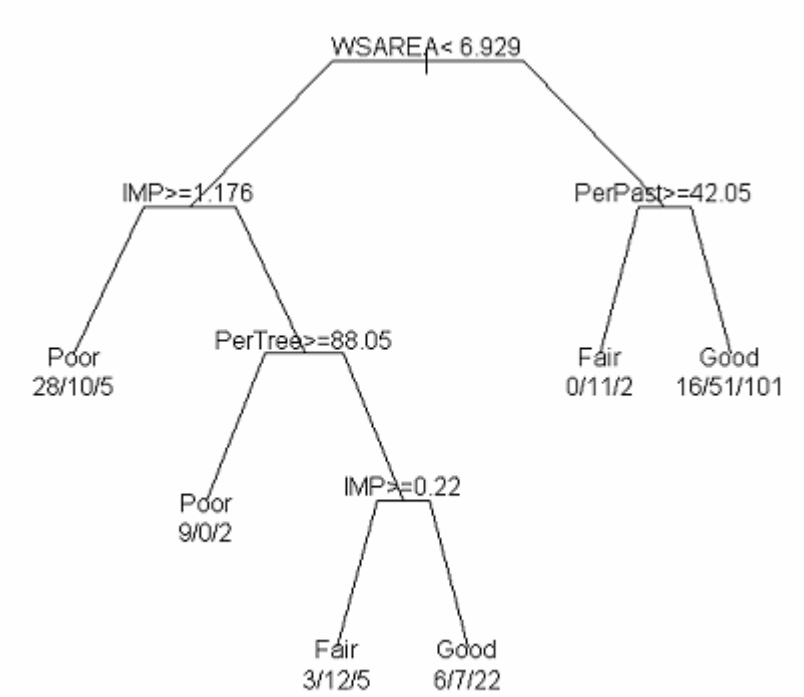


B.1 Classification tree for the biological condition of 1st-3rd order streams within the Chesapeake Bay watershed.



B.2 Cost-pruning curve for the classification tree model of 1st-3rd order streams within the Chesapeake Bay watershed.

The pruned classification tree model has six branches and shows the key explanatory variables for the biological condition of the 1st-3rd order streams within the Chesapeake Bay watershed are still watershed area (<6.929), percent impervious surface (1.176%), and percent pasture (42.05%) (B.3). However, elevation, percent crop cover, and precipitation were removed ("pruned") from the classification tree.



B.3 Pruned classification tree (using the misclassification method) for the biological condition of 1st-3rd order streams within the Chesapeake Bay watershed.

CART Model Performance

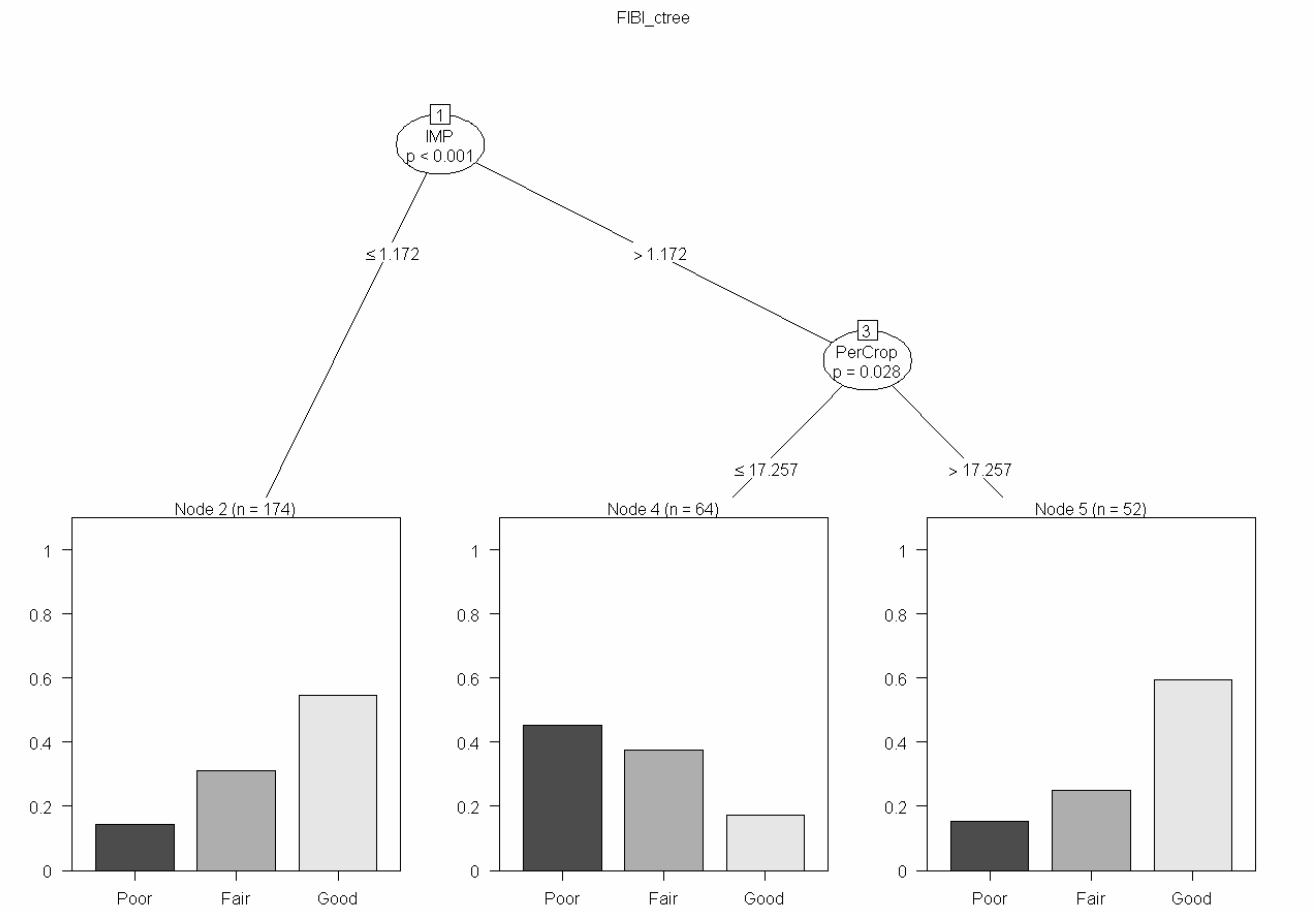
B.4 Confusion matrix (resubstitution) for the classification tree of 1st-3rd order streams within the Chesapeake Bay watershed.

		Data			
		Poor	Fair	Good	Classification error
Model Predictions	Poor	37	10	7	0.31
	Fair	3	23	7	0.30
	Good	22	58	123	0.39

B.5 Confusion matrix (10 – fold cross-validation) for the classification tree of 1st-3rd order streams within the Chesapeake Bay watershed.

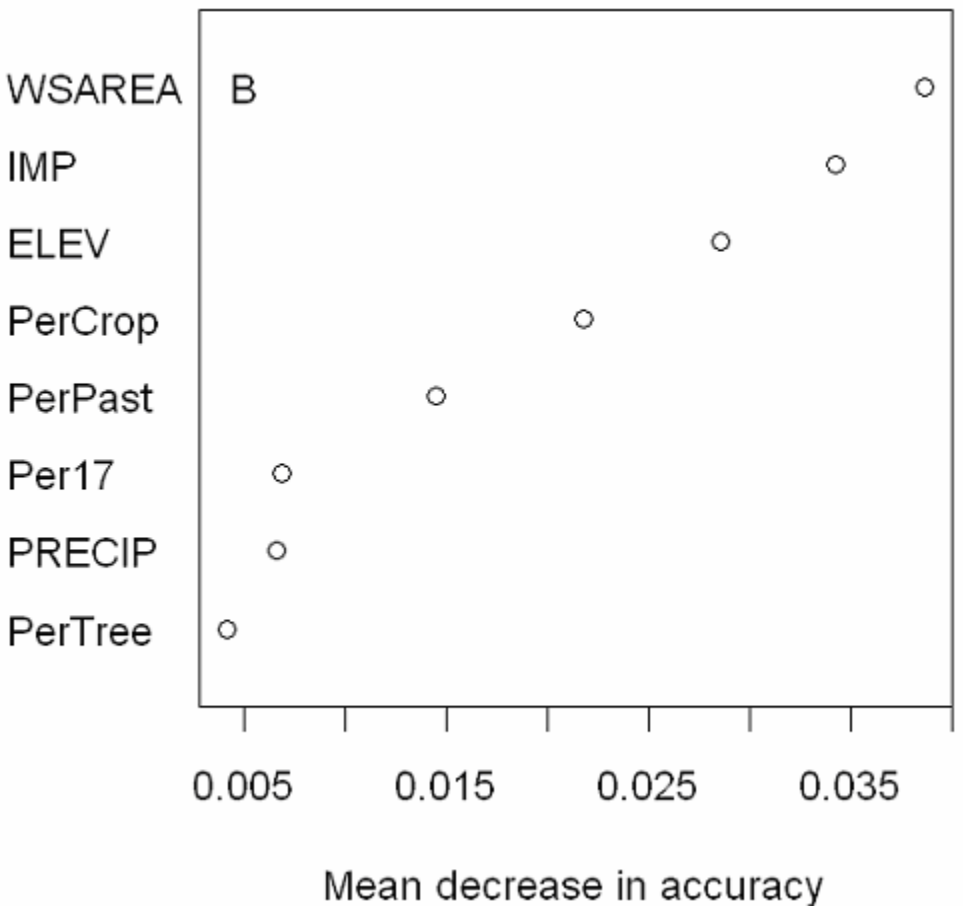
		Data			
		Poor	Fair	Good	Classification error
Model Predictions	Poor	11	4	13	0.61
	Fair	0	1	3	0.75
	Good	51	86	121	0.53

APPENDIX C: Conditional Tree Model



C.1 Conditional Tree model for predictions of stream biological condition within the CBW.

APPENDIX D: Conditional Forest Model



D.1 Conditional Forest variable importance plot for predictions of CBW biological condition. The horizontal axis presents the importance measure whereas the vertical axis denotes the variables.