Assessment and predictive model for brook trout (*Salvelinus fontinalis*) population status in the eastern United States

Teresa Marie Thieling

A thesis submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

for the degree of

Master of Science

Biology

May 2006

Acknowledgements

I would like to thank Mr. Mark Hudy for giving me the opportunity to work on this project. His academic and professional guidance and advice and his encouragement in public speaking have been invaluable and will forever be beneficial in my career. Mark always made sure that I had the funding and the equipment to complete my research. I would also like to thank Mark for ensuring that I occasionally got away from my computer to get some stream time.

I would like to thank Dr. Helmut Kraenzle, Dr. Reid Harris, and Dr. Eric Smith for serving on my graduate committee. Dr. Kraenzle's willingness to act as my co-chair and his experience and insight in the GIS profession contributed to the success of my project. Dr. Reid's knowledge of population ecology and the consequences of anthropogenic influences on the environment helped me to appreciate the significance of conservation. Dr. Smith's assistance with the statistical aspects of my project was invaluable and played a vital role in the quality of my project. I would also like to thank the entire Biology and Geography Departments for their support during my time here at James Madison University.

I would like to thank all the people that I worked with in the lab. Keith Whalen and Seth Coffman's help with everything including presentations, seminar paper discussions, and aquatic ecology was essential in my transition not only to graduate school but into the world of fisheries management. I would also like to thank them for instilling in me just how much it is possible for people to talk about fishing. For their friendship I will always be grateful. I would also like to thank Kyle Spencer and Chris

McGough for all their hard work and willingness to handle all the dirty jobs (tedious box plots).

I would like to thank all my fellow graduate students for their support and for creating a friendly atmosphere here at JMU. I would especially like to thank Billy Flint and Beth Cheever for the discussion breaks that made the long hours spent in the lab bearable and Emma Kirby-Glatkowski and Rachel Comer for helping me keep everything in perspective and providing time out from the world of GIS and brook trout. Most of all I would like to thank my family for their unwavering encouragement and support of my decision to pursue a graduate degree.

Table of Contents

Abstract

Over the last 200 years, brook trout (*Salvelinus fontinalis*) have been subjected to numerous anthropogenic physical, chemical, and biological perturbations that threaten the long term viability of brook trout throughout their historic native range. The historic and current decline in brook trout populations and the threat of further habitat degradation have led to a desire to develop a large scale conservation strategy to protect and rehabilitate brook trout populations and habitat. Understanding both the current distribution of brook trout and the relationships between the brook trout population status and perturbations is essential to developing meaningful conservation strategies and tactics. My study area included the historic native range of brook trout in the eastern United States, covering 17 states stretching from Maine to northern Georgia. I developed numerous predictive models using known brook trout subwatershed population status (Extirpated/Reduced/Intact) and subwatershed metrics derived from GIS data. The purpose of the models was to predict subwatershed status for the subwatersheds where the status was either unknown or only qualitative data were available and to determine metric thresholds to aid brook trout conservation efforts. I compared the correct classification rates of multiple and single variable logistic regression, discriminant analysis (linear, quadratic, and nearest neighbor), and classification trees. I chose the classification tree as my main model for predicting brook trout subwatershed status. Based on known subwatersheds and predictive models my results show that brook trout are Intact in 1,612 subwatersheds (32%), Reduced in 1,938 subwatersheds (39%) and are Extirpated from 1,451 subwatersheds (29%) from their potential (historic) range within the study area. Six core subwatershed and subwatershed water corridor metrics (percentage of forested land, combined sulfate and nitrate deposition, percentage of mixed forest in the water corridor, percentage of agriculture, road density, and latitude) were useful as predictors of brook trout distribution and status. A total of 94% of the Intact populations of brook trout occur in subwatersheds where the percentage of forested lands is greater than 68%. The brook trout subwatershed status distribution and threshold metric values can be useful for a risk assessment and for prioritizing conservation efforts.

Introduction

Roughly 25 percent of the historic native range of the eastern brook trout (*Salvelinus fontinalis*) is located in the eastern United States. This range extends from the headwater tributaries of the Mississippi River in Minnesota, to the Atlantic coastal drainages of the Northeast (Maine to Virginia), and south along the Appalachian mountains to the headwaters of the Chattahoochee River in northern Georgia (MacCrimmon and Cambell 1969; Benhnke 2002).

Over the last 200 years, brook trout have been subjected to numerous anthropogenic physical, chemical, and biological perturbations that threaten the long term viability of brook trout throughout their historic native range in the eastern United States (Marschall and Crowder 1996; Galbreath el al. 2001). These perturbations have caused declines in brook trout populations and brook trout have already been extirpated from many areas of this range (MacCrimmon and Cambell, 1969). These perturbations include: sedimentation, acid deposition, increased water temperature, loss of riparian vegetation, non-native species, and habitat fragmentation (Brasch et al. 1958; Kelly et al. 1980; Johnson and Jones 2000; Driscoll et al. 2001; Curry and MacNeill 2004).

Although these perturbations obviously affect brook trout in their stream or lake habitat, they may stem from sources not directly adjacent to aquatic ecosystems. Because of the downhill flow of water, freshwater species are affected by activities taking place anywhere upstream or uphill in the watershed (Master et al. 1998). These watershed activities, or characteristics, include: road density, land use such as agriculture, logging, and residential areas, and human population density and growth. These watershed scale characteristics can directly or indirectly cause or affect the aforementioned list of perturbations.

According to a recent assessment, these perturbations have caused a decline in self-sustaining brook trout populations in 59% of the subwatersheds in the eastern United States (Hudy et al. 2006). The study area for this assessment consisted of approximately 70% of the United States' historic native range of brook trout and summarized existing brook trout population knowledge to classify the brook trout population status for 6th level hydrologic units, or subwatersheds. Only 5% of the subwatersheds were still intact at pre-European settlement levels, 22% of the subwatersheds no longer supported brook trout populations, and almost 35% of the subwatersheds did not have enough quantitative data available to classify the population status. This study not only highlighted where brook trout populations are in decline, but also indicated areas of population information gaps. Filling in these information gaps would help to give a complete picture of the status of brook trout populations.

Because of the historic and current decline in brook trout populations and the threat of further habitat degradation, many biologists, land managers, outdoor enthusiasts, and policy makers are concerned about the future viability of brook trout. This concern has led to a desire to develop a large scale conservation strategy to protect and rehabilitate brook trout populations and habitat throughout the historic range in the eastern United States. In June of 2004, over 50 state and federal agencies, nongovernmental organizations, and academicians decided to form the Eastern Brook Trout Joint Venture (EBTJV). The purpose of this Joint Venture is to develop meaningful strategies and tactics for the conservation of brook trout and to establish multi-organization collaborative efforts to implement and maintain these strategies. To date, there has never been a large scale assessment evaluating the span of conditions of brook trout perturbations throughout the eastern United States.

Understanding the relationship between the brook trout population status and the perturbations within watersheds is essential to developing meaningful strategies and tactics. Evaluations of the integrity of native brook trout watersheds over their native range are useful to guide decision makers, managers, and publics in setting priorities for watershed level restoration, inventory, and monitoring. Large-scale assessments for many species have been useful in identifying and quantifying: problems, information gaps, restoration priorities and funding needs (Williams et al. 1993; Davis and Simon 1995; Frissell and Bayles 1996; Warren et al. 1997; Master el al. 1998, McDougal et al. 2001). Watersheds are good units on which to base assessments because they allow for the conservation of biodiversity (Moyle and Randall 1998). Also, management goals that ensure natural processes are maintained with little human interference are obtainable at the watershed scale (Moyle and Yoshiyama 1994). Compiling a multiple variable measurement, or multi-metric index, for watersheds can assist managers in their evaluations of watershed conditions by giving an indicator of overall health when many anthropogenic factors may be contributing to a problem and by assisting in identifying key limiting factors (Barbour et al. 1999; McCormick et al 2001). Determining a method for conducting a large scale assessment measuring brook trout populations and their perturbations and creating a multi-metric index based on these measurements would aid in the creation of conservation strategies and tactics.

An example of a large scale, multi-metric assessment of an aquatic species that addressed information gaps and predicted population abundance is the study conducted on bull trout (*Salvelinus confluentus*) populations in the Northwest (Rieman et al. 1997). This study used existing knowledge of the distribution and status of bull trout and its association with landscape characteristics to predict the probability of occurrence in subwatersheds where the bull trout population status was unknown. This assessment summarized information from biologists to classify bull trout population status within subwatersheds for the Columbia River Basin east of the Cascade Mountain crest and the portion of the Klamath River basin in Oregon. They then used associations of the bull trout population status with landscape metric values of the subwatersheds to predict the probability of occurrence in subwatersheds where the population status was unknown.

To develop the prediction model Rieman et al. (1997) calculated 28 landscape variables with potential influence on aquatic ecosystems using a Geographic Information System (GIS). These variables represented vegetation, climate, geophysical properties, land use, and included metrics such as: percent vegetation cover, mean air temperature, slope, and road density. They used existing GIS databases to calculate the landscape metrics within the subwatersheds classified by the population status assessment. They then used classification trees, a type of decision tree model, to determine the association of the metrics with the bull trout population status. The classification trees were use to predict bull trout presence in subwatersheds classified as unknown and to predict the population status of subwatersheds where only the presence of bull trout was known. This analysis concluded that bull trout were still widely distributed across its potential range, but had suffered strong declines in numbers.

Some aspects of this project by Rieman et al. (1997) can be applied to brook trout populations in the eastern United States. The population status of brook trout compiled by Hudy et al. (2006) was determined using methods similar to those used by the Rieman group. However, one of the drawbacks to the study assessing bull trout populations was that they could not quantitatively measure declines in the population status because they did not know how much of the potential range was historically occupied by bull trout. In contrast, the brook trout population status classifications are in relation to the historic distribution and distinctions were made between subwatersheds that never contained brook trout and subwatersheds where populations had been extirpated. Additional methods used for bull trout could be applied to brook trout subwatersheds to fill in the gaps where the population status is unknown. Like the Rieman project, metrics developed from calculating subwatershed landscape and anthropogenic features within the brook trout subwatersheds could be used to construct a model to predict population status. This prediction model or models could also be used to determine subwatershed landscape thresholds where shifts in the population status occur. Such models and thresholds could provide a valuable tool for the conservation effort to protect and restore brook trout populations and habitat.

The objective of this study is to (1) calculate multi-metric subwatershed characteristics for a brook trout risk assessment, (2) develop a model to predict brook trout classification status by subwatershed where data are missing, and (3) specify brook trout classification thresholds for landscape metrics to aid natural resource managers.

Methods

Study Area

The study area includes the historic native range of brook trout in the eastern United States, covering 17 states stretching from Maine to northern Georgia (Figure 1). This area encompasses 25% of the brook trout's native range and about 70% of the historic native range within the United States (MacCrimmon and Cambell 1969). I chose to use 6th level Hydrologic Unit Code (HUC) watersheds, from here on out referred to as subwatersheds, (mean size 8,633 ha, SD 7,384) as the analysis resolution for this study. The subwatersheds were delineated by the Natural Resource Conservation Service (NRCS) and the United States Geological Survey (USGS) (Seaber et al. 1987; McDougal et al. 2001; EPA 2002; USGS 2002). The subwatershed data were obtained and analyzed as both ArcGIS polygon shapefiles and coverages. I chose the subwatershed level because 1) it is currently the smallest nationally delineated watershed size available, 2) it is a size useful to biologists for developing management plans (Moyle and Yoshiyama 1994; Master et al. 1998), 3) it is the same resolution used by the Eastern Brook Trout Joint Venture (EBTJV) to assess brook trout status, and 4) local populations or discrete groups of brook trout are best approximated by subwatersheds (Hudy et al. 2006).

The NRCS and the USGS are currently delineating standardized 6th level watershed coverages for the entire United States (NRCS 2005). Many of the state subwatersheds were drafts and there were no 6th level HUCs available for the state of New York. In New York I used 5th level HUCs, referred to as watersheds, (mean size 20,476 ha, SD 16,390). The analysis consisted of a total of 5,287 subwatersheds or watersheds (New York only).

Metric development

I developed a set of candidate metrics comprised of numerous anthropogenic and landscape variables having potential influences on brook trout population status and distribution. These metrics were developed and analyzed within a Geographic Information System (GIS) using ArcGIS 8.3 (Environmental Systems Research Institute, Redlands, California, USA). I took the least common denominator approach and only included variables that were available at the same resolution and had common definitions throughout the study area. This approach allowed for development of metrics that could be compared across the study area. The variables are potential surrogates for sedimentation, fragmentation, vegetation, and human land use. I calculated metrics at the subwatershed level and the subwatershed water corridor. The water corridor was defined as 100 meters on both sides of a stream or surrounding a water body that was represented by the 1:100,000 scale National Hydrography Dataset (NHD) (USGS 2004). Metrics were calculated either per area or as a percentage of the watershed to account for the variation in subwatershed size. A full list of the candidate metrics is located in Table 3.

Projection and Datum

All GIS data were converted into a common projection. I used an Albers Conical Equal Area projection adjusted for the eastern United States (parameters: Standard Parallel: 27.283806, 44.08275; Longitude of Central Meridian: -96.613972; Latitude of Projection Origin: 35.683278; False Easting: 0.0; False Northing: 0.0). I used the Albers projection because it maintains (does not distort) area, so that one can compare per area metrics from Maine to those in Georgia (Lo and Yeung 2002). I used the North American Datum developed in 1983 (NAD83) because it is the most commonly used datum for the United States. The NAD83 datum uses the 1980 Geodedic Reference System (GRS 80) ellipsoid.

Independent variables

Dams

I calculated the number of dams per square kilometer for each subwatershed. Dam information and location was obtained from the National Inventory of Dams (NID) created and maintained by the United States Army Corps of Engineers (1998) and the Federal Emergency Management Agency. This database includes dams that have a height no less than six feet and have a minimum storage capacity of 15 acre feet. I used the 2002 updated dataset which contains 9,728 dams within the study area.

Roads

I calculated road density in kilometers of road per square kilometers of land area at both the subwatershed and the water corridor levels. The road dataset used was improved Topological Integrated Geographic Encoding and Referencing system (TIGER) data enhanced by Navtech (2001).

I also calculated a roads/streams crossings metric by determining the spatial intersections of the Navtech road data and the 1:100,000 National Hydrography Dataset (NHD) stream layer (USGS 2004a). Stream crossings were summarized for each subwatershed as number of crossings per stream kilometer.

Land Use

Land use information was obtained from the 1992 National Land Cover Data (NLCD) dataset developed for the contiguous United States by the USGS (USGS 2004b). The 1992 NLCD was completed for the study area in 1998 and is the most current dataset spanning the entire study area. The NLCD was derived from LANDSAT Thematic Mapper imagery augmented by ancillary datasets and consists of 21 thematic classes stored as a grid coverage with 30 meter cell resolution (USGS 2004b). The 21 thematic classes closely resemble the Anderson land use/cover classification system (Anderson et al. 1976) and are listed in Table 1. Land use was summarized as the percentage of the subwatershed and subwatershed water corridor classified for each individual land cover class (Table 1) and the following derived metrics: total human use (sum of : low and high intensity residential, quarry/mines, commercial/industrial/transportation, transitional, pasture/hay, row crops, fallow, small grains, orchards/vineyards, urban recreation), total agriculture (sum of: pasture/hay, row crops, fallow, small grains, orchards/vineyards), total forest (sum of: deciduous, evergreen, mixed) and residential use in the water corridor (sum of: low and high intensity residential) (Table 3).

The thematic accuracy of the NLCD was assessed by the USGS as the project mapping regions were completed (USGS 2005). The mapping regions that overlap the study area are regions 1-4: New England, New York/New Jersey, the Mid-Atlantic, and the Southeast respectively. They first randomly picked a sample cell. They then determined the probability that the most common land use class within a 3x3 cell block centered on the sample cell actually matches a photo-interpreted land cover class of the same area.

The overall regional accuracy probabilities in the study area range from 0.43 to 0.66 (USGS 2005). In most cases, errors occur between related classes, for example row crops are misclassed as pasture/hay or there is confusion among the three forest classes. When aggregated into the broader groups of water, urban, barren land, forest land, shrubland, agriculture land, and wetlands, the overall regional accuracy increases and ranges from 0.74 to 0.83.

Table 1. The 21 National Land Cover Dataset thematic land classes and land class codes

Code	Land Cover Class	Code	Land Cover Class
11	Open Water	51	Shrublands
12	Perennial Ice/Snow	61	Orchards/Vineyards/Other
21	Low Intensity Residential	71	Grasslands/Herbaceous
22	High Intensity Residential	81	Pasture/Hay
23	Commercial/Industrial/Transportation	82	Row Crops
31	Bare Rock/Sand Clay	83	Small Grains
32	Quarries/Strip Mines/Gravel Pits	84	Fallow
33	Transitional	85	Urban/Recreational Grasses
41	Deciduous Forest	91	Woody Wetlands
42	Evergreen Forest	92	Emergent Herbaceous Wetlands
43	Mixed Forest		

Human Population

I used a grid coverage containing population census data in 1 km cells to calculate the population density (number of people per square kilometer) per subwatershed. The population grid coverage was developed from U.S. Census year 2000 county population data (U.S. Census Bureau 2002) divided by census blocks. The census block data were used to convert the county total population data into population per 1 square kilometer grid cells (Whalen 2004).

Acid Deposition

Acid deposition metrics were derived from the 2004 nitrate (NO_3) and sulfate (SO_4) wet deposition grid data (National Atmospheric Deposition Program 2005). The deposition grids have a 2.5 km cell resolution and contain the spatially interpolated wet deposition in kilograms per hectare. I used the Zonal Statistics function of the ArcGIS 8.3 Spatial Analyst extension to calculate the mean nitrate and sulfate deposition values for each subwatershed.

Soil Buffering Capacity

A measure of soil buffering capacity within the water corridor was determined by calculating the percentage of soil with a pH greater than or equal to 5.0. I used a database developed by Penn State, called CONUS-SOIL, which was derived from the national NRCS State Soil Geographic Database (STATSGO) dataset. The CONUS-SOIL dataset consists of multiple 1 km resolution coverages of numerous soil characteristics (Earth Systems Science Center, 2005). I derived the soil buffering metric from the pH soil coverage. The pH values represent the top 10 centimeters of soil.

Elevation

The mean, maximum, and minimum elevation in meters for each subwatershed was calculated using 30 meter Digital Elevation Models (DEM) developed by the USGS. The DEMs were obtained from the National Elevation Dataset (NED) which is a seemless elevation dataset spanning the conterminous United States (USGS 1999).

Latitude and Longitude

To develop the latitude and longitude metrics, the subwatershed polygons were first converted from the Albers projection into a Geographic, decimal degree, NAD83 datum. The latitude and longitude of the centroid of each subwatershed polygon was determined using the VBA script provided in the ArcGIS 8.3 help.

Exotic fish

I developed a metric index to measure the presence of exotic fish species within the subwatersheds. I used the professional opinion perturbation values for exotic fish species collected from the biologists in the brook trout subwatershed status classification process (Hudy et al. 2006). To create the index I gave each exotic species weighted values based on their perturbation level. I assigned Level 1 a value of 5, Level 2 a value of 3, and Level 3 a value of 1. I then summed all of the weighted values for each of the exotic species in the subwatersheds.

Dependent variable

The dependent variable in this analysis was the brook trout population status classifications by subwatershed compiled by Hudy et al. (2006) as part of the EBTJV. Subwatershed classification was based on the percentage of habitat in each subwatershed still maintaining self-sustaining populations of brook trout (Table 2). The assessment used quantitative and qualitative data from numerous biologists, databases, and other sources to make subwatershed classifications that were consistent and comparative throughout the study area. Because of small sample sizes in some of the original subwatershed classifications, I used the grouped classifications of Extirpated (n = 1,083), Reduced > 50% (n = 1,481), and Intact > 50% (n = 773) for analysis, model development, and reporting (Table 2). From here on out I will refer to these groups as Extirpated, Reduced, and Intact. The classifications of 1)Unknown: No data and 2) Present: Qualitative represent subwatersheds where there was not sufficient quantitative data to either indicate the presence of self-sustaining brook trout or specify the percentage of habitat supporting self-sustaining brook trout within the subwatershed (Table 2). These two classifications represent data gaps and will henceforth be referred to as Unknown and Present. Figure 1 illustrates the distribution of the brook trout subwatershed status for the study area.

In addition to classifying the subwatershed status of brook trout, Hudy et al. (2006) also recorded biologists' and managers' opinions of the perturbations to brook trout within each subwatershed. These opinions helped me develop the list of candidate subwatershed metrics and also to create the exotic fish metric. The perturbations were characterized as Level 1: high (life cycle component eliminated), Level 2: medium (life cycle component reduced but not eliminated), or Level 3: low (general threat, no documented loss or reduction of life cycle).

Table 2. Summary of subwatershed level brook trout population classifications used for collection and validation and final collapsed classifications used for analysis and reporting.

Collection and validation classification categories	Collapsed classifications used for analysis and reporting
Classification 1.0 Unknown: No data or not enough data to classify further	**Extirpated** Subwatersheds where all self-sustaining populations no longer exist. Same as Classification 3.0
Classification 1.1 Absent: Unknown history - Brook trout currently not in watershed; historic status unknown	**Predicted Extirpated** All self-sustaining populations are predicted extirpated. Predicted subwatersheds from Classifications 1.0 and 4.0
Classification 2.0 Never occurred - Historic self-sustaining populations never known to occur	**Reduced > 50%** Between 50% and 99% of the historic brook trout habitat no longer supports reproducing brook trout populations. Same as Classification 8.0.
Classification 3.0 Extirpated - All historic self-sustaining populations no longer exist	**Predicted Reduced > 50%** Between 50% and 99% of the historic brook trout habitat is predicted to no longer support reproducing brook trout populations. Predicted subwatersheds from Classifications 1.0 and 4.0.
Classification 4.0 Present: Qualitative - No quantitative data; qualitative data show presence	**Intact > 50%** Subwatersheds with > 50% of historic habitat occupied by self-sustaining brook trout. Formed from the collapsing of Classifications 5.0, 6.0, and 7.0.
Classification 5.0 Present: Intact large - High percentage (>90%) of historic habitat occupied by self-sustaining populations, populations greater than 5,000 individuals or 500 adults	**Predicted Intact > 50%** Subwatersheds predicted with > 50% of historic habitat occupied by self-sustaining brook trout. Predicted subwatersheds from Classifications 1.0 and 4.0.
Classification 6.0 Present: Intact small - High percentage (>90%) of historic habitat occupied by self-sustaining populations, populations less than 5,000 individuals or 500 adults	**Absent: Unknown** Brook trout currently not in watershed; historic status unknown. Same as Classification 1.1
Classification 7.0 Present: Reduced - Reduced percentage (50% - 90%) of historic habitat occupied by self-sustaining populations	
Classification 8.0 Present: Greatly reduced - Greatly reduced percentage (1%-49%) of historic habitat occupied by self-sustaining populations	

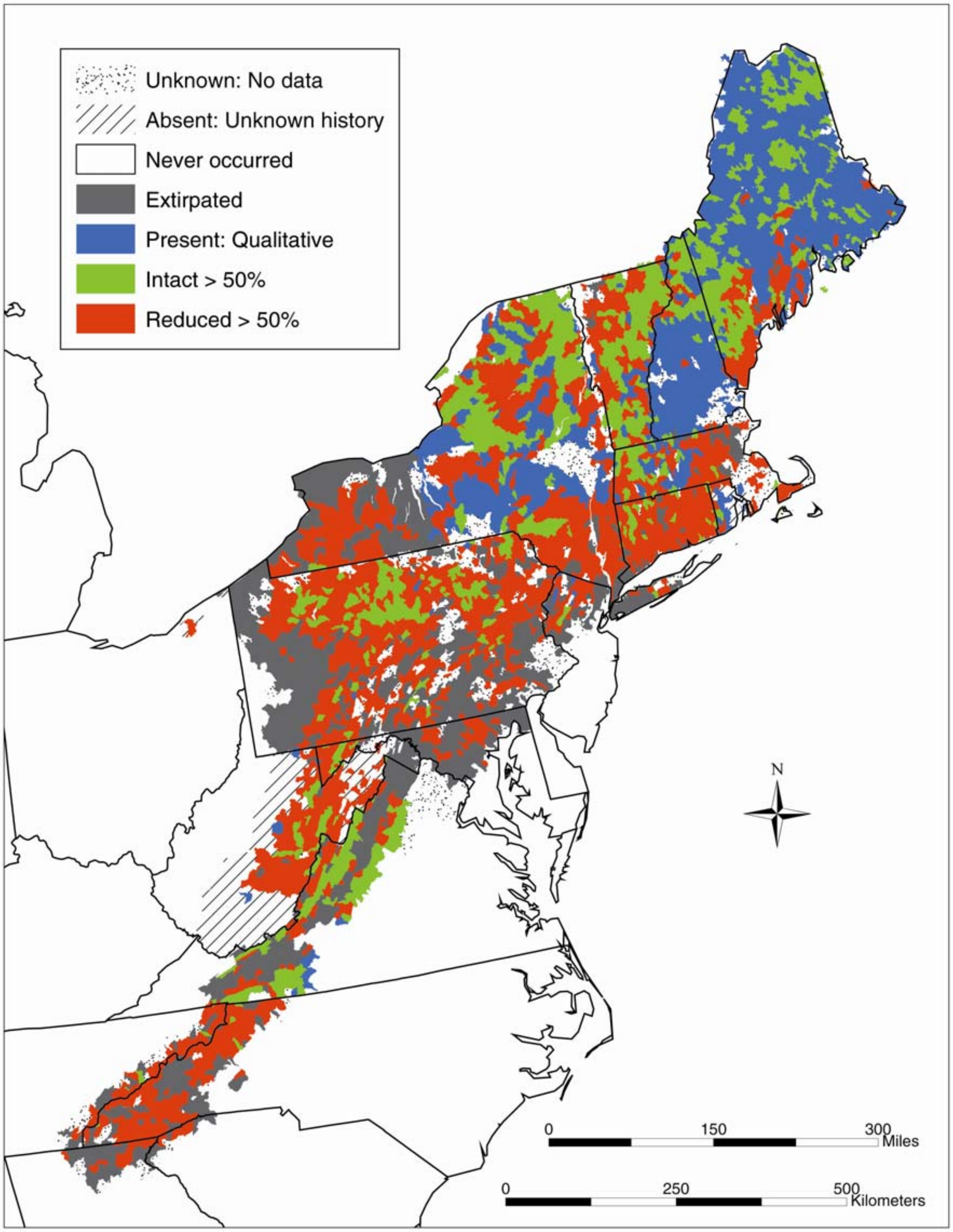

Figure 1. Distribution of brook trout subwatershed status within the study area.

Metric screening

The candidate metrics were screened according to similar methods described in Hughes et al. (1998) and McCormick et al. (2001) to help reduce the number of metrics, remove irrelevant variables, and determine which metrics are most likely to be predictive of brook trout populations (Table 3). Metrics were tested for 1) completeness, 2) range, 3) redundancy, and 4) responsiveness to the dependent variable. First, the metrics were screened for completeness to assure that measurements could be comparable throughout the study area. Metrics were excluded that did not have data available for the entire study area or did not have consistent resolution or definitions. Second, metrics with a small range of values were eliminated because they would not be useful in indicating differences in subwatershed characteristics. Next, when two metrics were highly correlated ($|r| > 0.8$) one metric was removed to eliminate redundancy. I used professional judgment to decide which single metric was retained and chose to keep the metric that would be more comprehendible, repeatable, and most useful to land managers. Finally, the responsiveness of the metrics to the brook trout subwatershed status classifications was measured using the Wald chi-square and analysis of variance tests (Sokal and Rohlf 1995; Hosmer and Lemeshow 2000). The metric screening resulted in a reduction of the original 63 metrics to a core set of six metrics: percentage forested lands (TOTAL_FOREST), percentage agriculture lands (PERCENT_AG), combined NO_3 and SO_4 deposition (DEPOSITION), road density (ROAD_DN), percentage riparian mixed forested lands in the subwatershed corridor (MIXED_FOREST2), and latitude (LATITUDE) (Table 3).

Table 3. Descriptions of subwatershed and subwatershed corridor level metrics. Screening criteria: X = eliminated for lack of range in variable; Y = eliminated for lack of response to categories; R = eliminated for redundancy with core variable; C = core variable not eliminated.

Screening	Subwatershed Metric	Description
Y	DAMS_SQKM	Number of dams per km^2
C	DEPOSITION	Derived from sum of mean SO_4 and NO_3 deposition (kg/ha)
R	NO3_Mean	Mean NO_3 deposition (kg/ha)
R	SO4_Mean	Mean SO_4 deposition (kg/ha)
Y	Pop_Density	Mean population density (# people/km^2)
Y	SOIL_GRTR5	Percentage of soils in the water corridor with a pH equal or greater than 5.0
Y	SOIL_LESS5	Percentage of soils in the water corridor with a pH less than 5.0
R	STRM_XINGS	Number of road crossings per km of stream
C	ROAD_DN	Road density (km of road per km^2 of land)
R	ROAD_DN2	Road density within the water corridor (km of road per km^2 of land)
Y	EXOTICS	Weighted number of exotic fish species within the subwatershed
C	LATITUDE	Latitude measured in decimal degrees
R	LONGITUDE	Longitude measured in decimal degrees
Y	ELEV_MEAN	Mean elevation
Y	ELEV_MIN	Minimum elevation
Y	ELEV_MAX	Maximum elevation
X	BAREROCK	Percentage bare rock in the subwatershed
X	BAREROCK2	Percentage bare rock in the water corridor
Y	DECIDUOUS	Percentage deciduous forest in the subwatershed
Y	DECIDUOUS2	Percentage deciduous forest in the water corridor
R	EVERGREEN	Percentage evergreen forest in the subwatershed
R	EVERGREEN2	Percentage evergreen forest in the water corridor
X	FALLOW	Percentage fallow fields in the watershed
X	FALLOW2	Percentage fallow fields in the water corridor
X	GRASSLAND	Percentage natural grasslands/herbaceous lands in the subwatershed
X	GRASSLAND2	Percentage natural grasslands/herbaceous lands in the water corridor
Y	HERB_WETLNDS	Percentage herbaceous wetlands in the subwatershed
Y	HERB_WTLNDS2	Percentage herbaceous wetlands in the water corridor
Y	HIGH_RES	Percentage high intensity residential lands in the subwatershed
Y	HIGH_RES2	Percentage high intensity residential lands in the water corridor
Y	INDUST_TRANS	Percentage commercial/industrial/transportation in the subwatershed
Y	INDUST_TRANS2	Percentage commercial/industrial/transportation in the water corridor
Y	LOW_RES	Percentage low intensity residential in the subwatershed
Y	lOW_RES2	Percentage low intensity residential in the water corridor
R	MIXED_FOREST	Percentage mixed forested lands in the subwatershed
C	MIXED_FOREST2	Percentage mixed forested lands in the water corridor
Y	OPEN_WTR	Percentage open water in the subwatershed
Y	OPEN_WTR2	Percentage open water in the water corridor
Y	ORCH_VINEYRD	Percentage orchards/vineyards/other in the subwatershed

Y	ORCH_VINYRD2	Percentage orchards/vineyards/other in the water corridor
R	PASTURE_HAY	Percentage pasture/hay in the subwatershed
R	PASTURE_HAY2	Percentage pasture/hay in the water corridor
C	PERCENT_AG	Derived from subwatershed sum of agricultural uses
R	PERCENT_AG2	Derived from water corridor sum of agricultural uses
R	PRCNT_HUMAN	Derived from subwatershed sum of percentage human uses
R	PRCNTT_HUMAN2	Derived from water corridor sum of percentage human uses
Y	PRCNT_RES2	Derived from the sum of high and low residential use in the water corridor
Y	QRY_MINE_GPIT	Percentage quarries/strip mines/gravel pits in the subwatershed
Y	QRY_MINE_GPIT2	Percentage quarries/strip mines/gravel pits
Y	ROW_CROPS	Percentage row crops in the subwatershed
Y	ROW_CROPS2	Percentage row crops
Y	SHRUBLAND	Percentage shrubland in the subwatershed
Y	SHRUBLAND2	Percentage shrubland
Y	SMALL_GRAINS	Percentage small grains in the subwatershed
Y	SMALL_GRAINS2	Percentage small grains
C	TOTAL_FOREST	Derived from subwatershed sum of forested lands
R	TOTAL_FOREST2	Derived from water corridor sum of forested lands
Y	TRANSITIONAL	Percentage transitional -areas of sparse vegetation in the subwatershed
Y	TRANSITIONAL2	Percentage transitional -areas of sparse vegetation in the water corridor
Y	URBAN_REC	Percentage urban/recreational grasses in the subwatershed
Y	URBAN_REC2	Percentage urban/recreational grasses in the water corridor
Y	WOOD_WETLNDS	Percentage wooded wetlands in the subwatershed
Y	WOOD_WTLNDS2	Percentage wooded wetlands in the water corridor

Predictive models

My main objectives were to 1) determine the relationships among the subwatershed and subwatershed corridor metrics and the brook trout subwatershed classifications, 2) use these relationships to predict the status of brook trout where the subwatershed status was unknown or only qualitative data was available, and 3) determine metric thresholds to aid land managers. The relationships among brook trout subwatershed classifications and subwatershed and subwatershed water corridor level metrics were modeled using several techniques. The modeling methods tested were: multivariable logistic regression, single variable logistic regression, discriminant analysis (linear, quadratic, and nearest neighbor), and classification trees. Although all methods were tested using the total 63 metrics, the final reported model development was completed using combinations of the final six core metrics. Each model was developed and tested to predict the brook trout subwatershed status of the subwatersheds classified as Unknown and Present (Table 2) into either a binomial or trinomial status outcome. The binomial outcome resulted in a presence/absence status classification referred to as Presence (combination of Reduced and Intact subwatersheds) and Extirpated. The trinomial outcome resulted in predicted classifications that were the same as the known status classifications of Extirpated, Reduced, and Intact. Logistic regression and discriminant analysis were run using SAS, Version 9 (SAS Institute Inc., Cary, North Carolina, USA) and CART 5.0 (Salford Systems, San Diego, California, USA) was used to fit classification trees.

Multivariable and single variable logistic regression

Multivariable logistic regression models were created to predict the brook trout subwatersheds with Unknown or Present status into subwatershed classification variables that are part of either a binary (Presence/Extirpated) or trinomial (Extirpated/Reduced/Intact) response. These classification variables were then treated as dependent variables in the logistic regression with the core metric values as the predictor variables.

In the case of a binary response variable, logistic regression analysis models p: the probability that brook trout is present in terms of one or more predictor variables. The model is nonlinear and has an “S” shape, increasing as a function of the variables (Agresti 1996). If there are k predictor variables used to model presence, the model may be written in terms of the probability of presence as

$$\Pr(\text{species present}) \ = p = \frac{\exp(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k)}$$

where $x_1, x_2, \ldots x_k$ corresponds to the k measured variables used in the model and $\beta_0, \beta_1, ..., \beta_k$ are the associated parameters (Collett 2002). The model can be transformed to a linear model using the logit transformation (Collett 2002):

$$\text{logit}(p) = \log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$

Although the transformed model is linear, the fitting process is the not the same as linear regression because the dependent variable is binary or trinomial (the response is modeled using a binomial or multinomial distribution rather than a normal distribution). The model is fitted using Proc Logistic in SAS using iterative methods of maximum likelihood. Transformations for individual predictors were evaluated using a Box-Cox transformation (Box and Cox 1964). The optimal transformation was rounded prior to application. The lack of fit of the model was evaluated using the Hosmer-Lemeshow test (Hosmer and Lemeshow 1980). Residuals and influence were checked using standard methods.

In the case of the trinomial response variable, we used methods of ordinal logistic regression that results in two S shaped curves that differ in intercept but have similar shape (Collett 2002). From these curves probabilities for each category may be computed. This model has three probabilities: $p_1, p_2,$ and p_3. Because these must sum to one, only two of the probabilities need to be modeled. A simple model to do this is to assume the same relationship with the predictors but have a different intercept i.e. (Collet 2002),

$$\text{logit}(p_i) = \beta_{0i} + \beta_1 x_1 + ... + \beta_k x_k$$

Other models, allowing for different intercepts and slopes were also evaluated. The models were summarized using prediction ability based on the ten-fold cross-validation method and resubstitution method (Breiman et al. 1984). In the ten-fold cross-validation method, a stratified random sample of ten percent (test sample) of the observations are withheld while the model is fit to the remaining 90% (learn sample). This process is repeated nine times and the error rates of the ten iterations are averaged to determine the node probabilities. The resubstitution method merely computes the classification errors using the same data as a training set that was used to build the model. The ten-fold cross-validation method results in a more realistic estimate of how well the model will fit new datasets and reduces overconfidence resulting from predicting observations using the model developed to optimize prediction of those observations (Breiman et al. 1984).

Multivariable logistic regression determines the importance of the predictor variables in the presence of other variables. It is difficult to establish overall metric value thresholds because the metric relationships vary by each subwatershed. Because of this, single variable logistic regression, using each of the six core metrics individually, was also used to predict Extirpated subwatersheds from subwatersheds where brook trout are present (binomial model). Comparing the correct classification rate to the metric value used to predict classification aided in developing metric value thresholds. Trinomial response single logistic regression models were also created for each of the core metrics.

Discriminant analysis

Discriminant analysis refers to a set of multivariable techniques that focus on the classification of an object to a group. There are several methods under the rubric of discriminant analysis including linear discriminant analysis, quadratic discriminant analysis, and nearest neighbor discriminant analysis (Huberty 1994). All these methods can be viewed as methods based on probability models for each of the groups and describe different ways to estimate the probability that an object comes from a particular group (Rencher 2002). The linear, quadratic, and nearest neighbor techniques were explored in this study.

The simplest form of discriminant analysis is linear discriminant analysis. This method uses a linear equation to predict the group that an object comes from. The genesis of the method is as follows (Rencher 2002). If each group has a multivariable normal distribution with equal variances (or variance-covariance matrices) then the only difference in the groups is in the means. To assign an object to the group we can find the group that the object is most likely to have come from. To do this with normal distributions we calculate the height of the normal curve given the characteristics of the objects. In mathematical terms, if the density of observations with characteristics, x, is $f_i(x)$ for group i then the group that the object is most likely to come from is the one with the greatest value of $f_i(x)$. So to classify, we can calculate this for each group and then assign the object. With multivariable normal distributions, there is a simpler way because the decision can be written as a linear function of x. Hence, rather than calculate the density, we simply calculate the linear function and make the assignment based on the score of the linear function. When there are just two groups, the scores can actually be estimated using a regression program (Kleinbaum et al. 1988) with the response variable being a binary variable and the independent variables being the x variables.

Quadratic discriminant analysis results when the densities are multivariable normal with unequal covariance matrices (Rencher 2002). In this case, the decision rule does not reduce to a linear function; rather it is reduced to a quadratic function.

Nearest neighbor discriminant analysis focuses on using distance rather than probability for classification (Rencher 2002). The first step in the method is to choose a distance measure. Then, a test or training set is used to define the different groups. When a new observation is to be classified, we compute the distance from the new object to the other objects in the test set. A pre-specified number of neighbors are chosen based on closeness to the given object. Then one looks at the classes associated with the neighbors. The object is assigned to the class that it is closest to in terms of the number of neighbors. Thus if the pre-specified number of neighbors is five and of the five closest to a subwatershed, three are from the Intact group and two are from the Extirpated group, the subwatershed is assigned to the Intact group (i.e. predict the brook trout population status as Intact in that subwatershed).

Multiple runs of the model were conducted, varying the number of neighbors, in order to optimize the correct classification. The final number of neighbors chosen was 11. The distance measure used for this project was the Euclidean distance. Both the ten-fold cross-validation and the resubstitution methods were used to evaluate all of the discriminant analysis models.

Classification trees

Classification trees are a type of decision tree that use input variable values to successively split data into more homogenous groups (Breiman et al. 1984; Clark and Pregibon 1992). Classification trees are similar to taxonomic keys in that they consist of a dichotomous rule set that is produced through recursive partitioning. In other words, the data are split into two groups based on a single predictor value, determined from the input variables, which produces the greatest difference in the resulting groups. Each juncture, or node, is considered in isolation without any concern for how the next node will be split (Neville 1999). These groups are then partitioned again based on a different splitting criterion and the process continues until the data can no longer be divided and result in a terminal node. In the case of this project the input variables are the core metric values and the terminal nodes represent the brook trout subwatershed status classifications. Classification trees can take the given measurements of individuals in known classifications and develop splitting criteria to predict the classifications of individuals with unknown status. Through the development of classification trees one can also determine the factors which most prominently influence or predict the terminal nodes or classifications. All classification tree modeling was done using CART 5.0 modeling program (Salford Systems, San Diego, California, USA). Both the ten-fold cross-validation and the resubstitution methods were used to evaluate the prediction errors of the classification trees.

The algorithm that CART uses to determine the splitting criteria is based on the Gini index (Breiman et al. 1984). For each node, CART determines which variable and variable value most greatly reduces the Gini index for the set of observations within the node. The Gini index is a measure of impurity with values ranging from 0 to 1, where 0 represents total purity (all observations in the node are in one class) and 1 equals total impurity (all classes are equally represented in the node). The reduction in the Gini index is measured as the impurity of a group before the split, minus the sum of the impurities of the two groups resulting from the split.

Classification trees have some advantages over linear models. First, classification trees accept multiple variable types such as numerical, ordinal, and interval (Neville 1999). Second, classification trees use surrogate variables to handle missing values (Breiman et al. 1984). Third, they are not sensitive to monotonic transformations of the variables (Statistical Sciences 1993). Also, classification trees notice relationships from the interaction of inputs, discard redundant inputs, and help determine the variable importance in prediction (Neville 1999).

Four models were created using the classification trees and the core metrics: M1 (no LATITUDE), M2, M3 (no LATITUDE), and M4. M1 and M2 sorted the subwatersheds into either a Presence or Extirpated population status classification. M3 and M4 classified the subwatersheds as Extirpated, Reduced, or Intact (Figures 2-5). I also conducted a geo-spatial analysis to map the locations of the incorrect classifications for each classification tree model.

Results

Analysis of predictive models

Numerous models were developed using the core metrics and the subwatersheds with known status classifications as a training set. The range of values of the core metrics for all the subwatersheds and for each subwatershed classification is illustrated in Figures 6-10. A summary of the models' correct classification rates is listed in Tables 4 and 5.

Binomial models

Classification trees and nearest neighbor discriminant analysis had the highest overall correct classification rates (CCR) for the binomial response models (Presence/Extirpated). Classification trees had the highest CCR when LATITUDE was included as a variable (83% with resubstitution method) and nearest neighbor discriminant analysis when LATITUDE was removed from the model (80%) (Table 4). However, when using the cross validation method, the multivariable logistic regression model had the highest CCR (79%) (Table 4). The single variable logistic regression models TOTAL_FOREST, DEPOSITION, and ROAD_DN had the highest CCR for predicting Presence subwatersheds, while PERCENT_AG, MIXED_FOREST2 and LATITUDE were better at predicting Extirpated subwatersheds (Table 4). The TOTAL_FOREST model was the best overall single metric predictor (CCR = 76%) for Presence/Extirpated (Table 4). I did not present cross validation values for the single variable models because their low resubstitution values were too low to consider them for the main prediction model.

Trinomial models

Classification trees had the highest CCR (72% with resubstitution; 65% with cross validation) of the trinomial response models (Extirpated/Reduced/Intact), followed by nearest neighbor discriminant analysis (70% with resubstitution; 64% with cross validation) (Table 5). This ranking was consistent among the models when LATITUDE was removed (Table 5). DEPOSITION had the highest CCR of the single predictor variables with 58%; however none of the variables' values fell below 51% (Table 5).

Table 4. Correct classification rates (CCR) for the binomial response models. An (L) proceeding the model name indicates that LATITUDE was included. Resubstitution values presented as percentages (cross-validation in parentheses).

	Binomial		
	Extirpated	Present	Overall
Single variable logistic regression			
TOTAL_FOREST	45	90	76
PERCENT_AG	91	41	75
DEPOSITION	3	96	66
MIXED_FOREST2	91	31	71
ROAD_DN	20	94	70
LATITUDE	89	22	68
Multivariable logistic regression (L)	59(59)	89(89)	79(79)
Multivariable logistic regression	56(56)	89(89)	79(79)
Linear discriminant analysis (L)	74(74)	80(79)	78(77)
Linear discriminant analysis	72(72)	75(75)	74(74)
Quadratic (L)	82(82)	74(74)	77(77)
Quadratic	69(69)	79(79)	76(76)
Nearest neighbor (L)	88(84)	76(74)	80(77)
Nearest neighbor	85(79)	77(74)	80(76)
Classification trees (L)	90(83)	80(77)	83(77)
Classification trees	80(76)	78(76)	79(76)

Table 5. Correct classification rates (CCR) for the trinomial response models. An (L) proceeding the model name indicates that LATITUDE was included. Resubstitution values presented as percentages (cross-validation in parentheses).

	Trinomial			
	Extirpated	Reduced	Intact	Overall
Single variable logistic regression				
TOTAL_FOREST	50	84	0	54
PERCENT_AG	55	79	11	55
DEPOSITION	4	88	34	58
MIXED_FOREST2	48	73	19	52
ROAD_DN	33	72	35	51
LATITUDE	45	48	66	51
Multivariable logistic regression (L)	64(61)	74(74)	48(46)	64(64)
Multivariable logistic regression	62(61)	73(73)	41(41)	62(61)
Linear discriminant analysis (L)	67(66)	63(63)	57(57)	63(63)
Linear discriminant analysis	66(66)	53(53)	53(53)	57(57)
Quadratic (L)	74(73)	43(43)	72(72)	60(60)
Quadratic	60(59)	45(45)	79(78)	58(57)
Nearest neighbor (L)	82(78)	54(47)	82(75)	70(64)
Nearest neighbor	78(73)	54(48)	80(71)	68(61)
Classification trees (L)	84(78)	59(53)	80(72)	72(65)
Classification trees	76(69)	64(51)	79(72)	71(62)

Using logistic regression, the best single metric models had overall correct classification rates of 51% to 58 %. Using the core variables in a multi-metric logistic regression increased the overall correct classification rate to 64%. Although the multivariable logistic regression models have a higher CCR, the varying relationships of the metrics in the multiple variable models make determining thresholds difficult. Single metric logistic regression models have a lower overall prediction rate but have the advantage of indicating specific land use metric thresholds to natural resource managers.

While all model methods showed promise, classification trees was chosen as the prediction model for this project because: 1) it had higher overall correct classifications among binomial and trinomial models; 2) there was a good balance among the correct prediction rates of each classification category; 3) there are very few assumptions and no transformations needed in any of the input data; and 4) thresholds and their interactions are easier than the other models to interpret, display, and explain to land use managers.

Binomial classification tree models: Presence/Extirpated

There were four classification tree models created. Classification tree Model 1 (M1) used five of the core metrics (no LATITUDE) and had an overall correct classification rate of 79% (resubstitution method) and 76% (cross-validation method) with a good balance in error rates between Extirpated and Presence (Table 4). The most important metrics and splitting criteria for M1 are Node 1: TOTAL_FOREST (67.9%), Node 2: DEPOSITION (22.9 kg/ha), Node 6: ROAD_DN (1.19 km/km^2), and Node 3: MIXED_FOREST2 (11.9%) (Figure 2).

The M1 classification tree model is shown in Figure 2 with predictive probabilities for each of the terminal nodes. Classification trees work much like dichotomous keys. For example, in Figure 2, at Node 1 all 3,337 subwatersheds are split on TOTAL_FOREST at a splitting criterion of 67.9%. Those subwatersheds with a percentage of forested lands less than or equal to 67.9% (n = 1,227) go to the left branch to Node 2, while those subwatersheds with a percentage of forested lands greater than 67.9% (n = 2,060) go to the right branch (Node 6). At each subsequent node the subwatersheds are split again. At Node 6 the splitting criterion is ROAD_DN with a value of 1.19 km/km^2. Subwatersheds with a road density less than 1.19 km/km^2 would follow the left branch to Terminal Node 6, while subwatersheds with road density greater than 1.19 km/km^2 would follow the right branch to Node 7. Subwatersheds proceed through the splitting criteria until they reach a terminal node where a classification is predicted with a given probability. For example, Terminal Node 6 contains all subwatersheds that have greater than 67.9 % forested lands and a road density less than 1.19 km/km^2. A total of 886 out of 3,337 subwatersheds have these two characteristics. Subwatersheds with these characteristics are predicted to have a classification of Presence at a probability of 84.9%. Out of the total 1,664 subwatersheds that were classified as either Unknown or Present, M1 predicted extirpation in 467 (28%) subwatersheds and presence in 1197 (72%) subwatersheds (Table 6).

Model 2 (M2) used all six core variables and had an overall CCR of 83% (resubstitution method) and 79% (cross-validation method) (Table 4). M2 successfully predicted extirpated subwatersheds correctly 90% (resubstitution method) and 83% (cross-validation method) (Table 4). The most important metrics and splitting criteria for M2 were Node 1: TOTAL_FOREST (67.9%), Node 2: DEPOSITION (22.9 kg/ha), Node 3: LATITUDE 37.86 decimal degrees, Node 4: LATITUDE (41.19 decimal degrees)

(Figure 3). Out of the combined total 1,664 Present and Unknown subwatersheds, M2 predicted extirpation in 380 (23%) subwatersheds and presence in 1284 (77%) subwatersheds (Table 6).

A “pruned” classification tree model is shown in Figure 3 with predictive probabilities for each of the terminal nodes. Pruning allows for easier display of the model as full classification trees can be quite large. A pruned classification tree is a tree in which the terminal and lower (near the terminal) nodes have been deleted, collapsing the tree into fewer splits. The result is that previously normal nodes become terminal nodes and the subwatersheds are sorted into less homogenous groups.

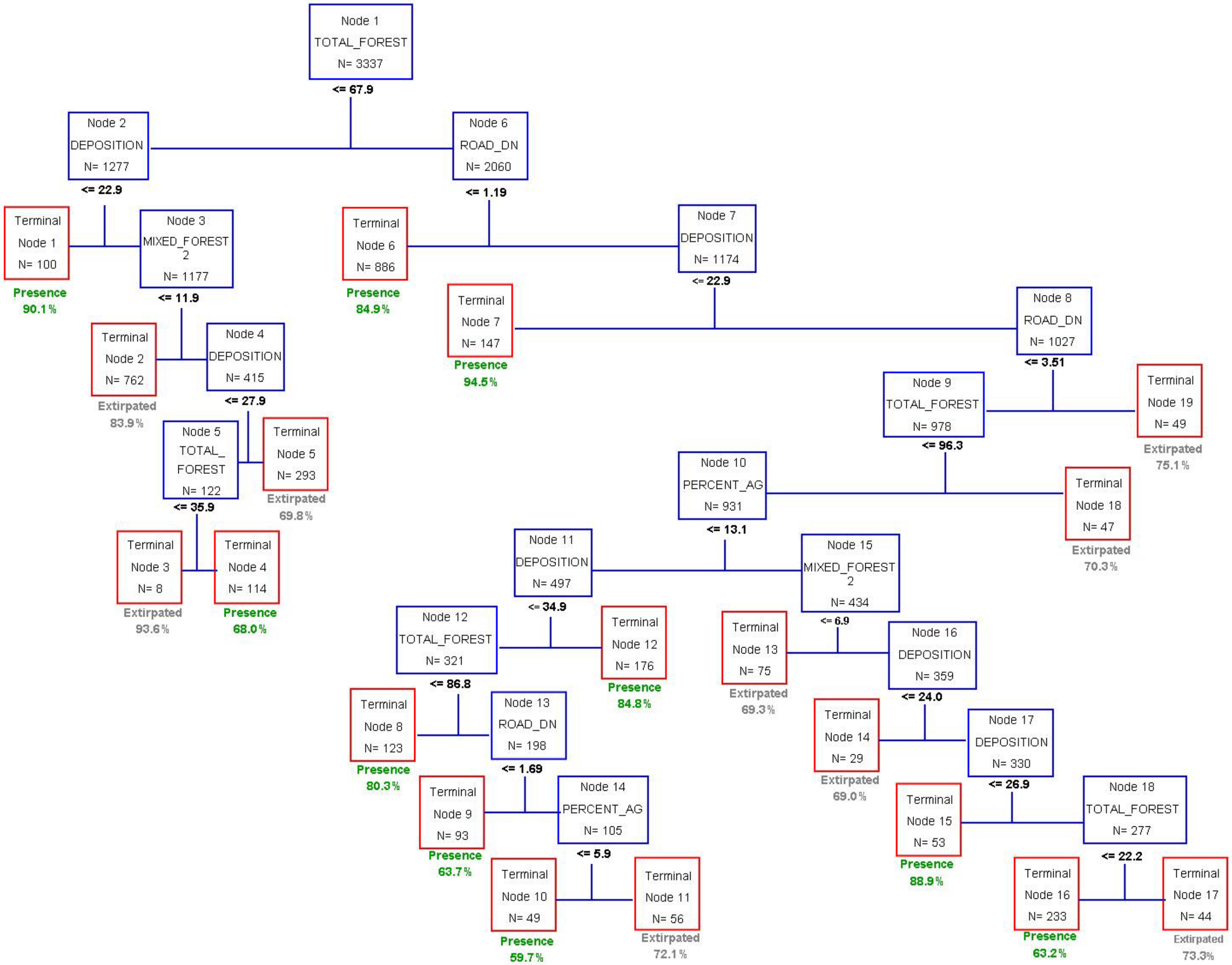

Figure 2. Complete classification tree Model 1 (M1). Terminal nodes are indicated by red boxes. Final classification and within node classification probabilities in percentages are indicated below the terminal nodes.

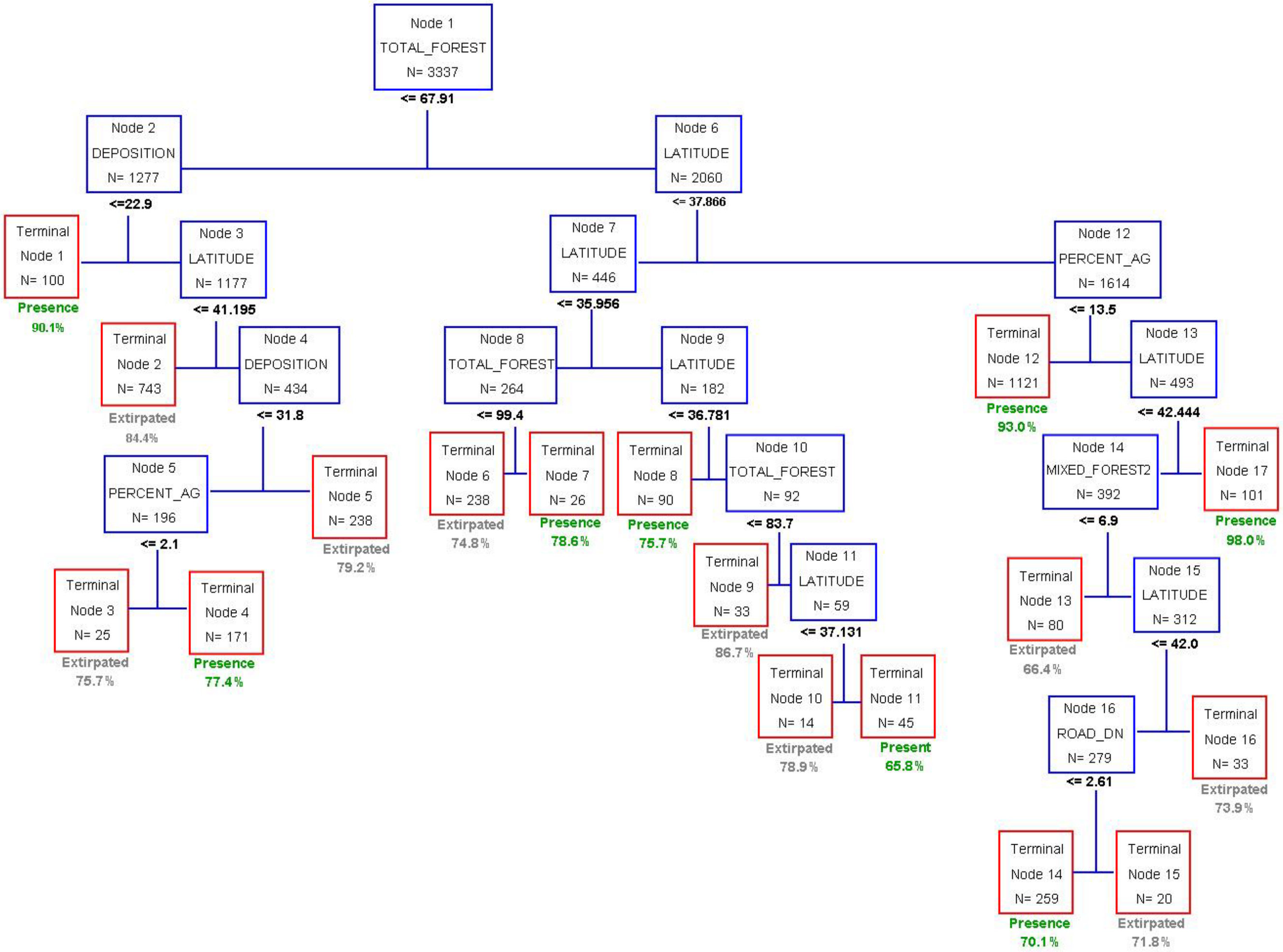

Figure 3. Pruned classification tree Model 2 (M2). Terminal nodes are indicated by red boxes. Final classification and within node classification probabilities in percentages are indicated below the terminal nodes.

Table 6. Summary of predicted classifications (number of subwatersheds) of Unknown and Present subwatersheds for classification tree models 1-4.

Model 1: Binomial - without LATITUDE

Subwatershed Classification	# of Subwatersheds	Predicted Classification
Unknown	386	Extirpated
Unknown	252	Present
Present	81	Extirpated
Present	945	Present
Unknown & Present	467	Extirpated
Unknown & Present	1197	Present

Model 2: Binomial- with LATITUDE

Subwatershed Classification	# of Subwatersheds	Predicted Classification
Unknown	340	Extirpated
Unknown	298	Present
Present	40	Extirpated
Present	986	Present
Unknown & Present	380	Extirpated
Unknown & Present	1284	Present

Model 3: Trinomial – without LATITUDE

Subwatershed Classification	# of Subwatersheds	Predicted Classification
Unknown	326	Extipated
Unknown	248	Reduced
Unknown	64	Intact
Present	42	Extipated
Present	209	Reduced
Present	775	Intact
Unknown & Present	368	Extipated
Unknown & Present	457	Reduced
Unknown & Present	839	Intact

Model 4: Trinomial - with LATITUDE

Subwatershed Classification	# of Subwatersheds	Predicted Classification
Unknown	364	Extipated
Unknown	198	Reduced
Unknown	76	Intact
Present	35	Extipated
Present	180	Reduced
Present	811	Intact
Unknown & Present	399	Extipated
Unknown & Present	378	Reduced
Unknown & Present	887	Intact

*Total predicted subwatersheds for each model is 1664

Trinomial classification tree models: Extirpated, Reduced, Intact

Classification tree Model 3 (M3) used five of the core metrics (no LATITUDE) and had an overall correct classification rate of 71% (resubstitution method) and 62% (cross-validation method) (Table 5). Correct classification rates among the three categories were Extirpated (76% resubstitution method; 69% cross-validation method); Reduced (64% resubstitution method, 51% cross-validation method); and Intact (79% resubstitution method, 72 % cross-validation method) (Table 5). The most important metrics and splitting criteria for M3 are Node 1: TOTAL_FOREST (68.1%), Node 2: DEPOSITION (27.9 kg/ha), Node 6: DEPOSITION (18.5 kg/ha), and Node 3: PERCENT_AG (27.1%)(Figure 4). The M3 pruned classification tree model is shown in Figure 4 and the predicative probabilities for each of the terminal nodes is listed in Table 7.

Table 7. Terminal node classification probabilities for Model 3 (M3).

Terminal Node	Extirpated Probability	Reduced Probability	Intact Probability
1	0.0	27.5	72.5
2	23.4	62.9	13.7
3	56.4	14.9	28.6
4	4.0	11.7	84.3
5	72.9	24.2	2.9
6	0.9	9.1	90.0
7	10.8	24.7	64.4
8	10.3	17.6	72.1
9	35.8	46.0	18.2
10	20.2	59.9	19.8
11	23.4	49.8	26.8
12	1.2	35.8	63.0
13	5.4	83.2	11.4
14	18.8	47.6	33.6
15	30.2	12.1	57.7
16	21.8	43.8	34.4
17	55.3	35.7	9.1

18	26.1	55.6	18.3

Out of the combined total 1,664 Present and Unknown subwatersheds, M3 predicted the subwatershed classification to be Extirpated in 368 (22%) subwatersheds, Reduced in 457 (28%) and Intact in 839 (50%) subwatersheds (Table 6). The spatial distribution of these predicted subwatersheds is illustrated in Figure 12.

Classification tree Model 4 (M4) uses all six core metrics and had an overall CCR of 72% (resubstitution method) and 65% (cross-validation method) (Table 5). Correct rates among the three categories were Extirpated (84% resubstitution method; 78 % cross-validation method), Reduced (59% resubstitution method, 53% cross-validation method), and Intact (80% resubstitution method, 72 % cross-validation method) (Table 5). The most important metrics and splitting criteria for M4 are Node 1: LATITUDE (43.12 decimal degrees), Node 2: TOTAL_FOREST (66.8%), Node 18: DEPOSITION (14.2 kg/ha), and Node 3: LATITUDE (41.19 decimal degrees) (Figure 5). The pruned classification tree model is shown in Figure 5 and the predictive probabilities for each of the terminal nodes are listed in Table 8.

Out of the combined total 1,664 Present and Unknown subwatersheds, M4 predicted the subwatershed classification to be Extirpated in 399 (24%) subwatersheds, Reduced in 378 (23%) and Intact in 887 (53%) subwatersheds.

Table 8. Terminal node classification probabilities for Model 4 (M4).

Terminal Node	Extirpated Probability	Reduced Probability	Intact Probability
1	76.6	18.0	5.3
2	64.6	35.4	0.0
3	12.6	70.8	16.5
4	71.3	27.2	1.4
5	66.1	32.7	1.3
6	14.6	78.5	6.8
7	37.0	63.0	0.0
8	0.0	7.4	92.6
9	55.5	14.4	30.1
10	4.5	76.4	19.1
11	5.8	26.8	67.5
12	15.2	65.2	19.6
13	59.1	34.8	6.1
14	27.8	58.9	13.2
15	56.8	16.6	26.5
16	20.2	59.2	20.6
17	2.5	35.5	62.0
18	0.4	3.8	95.8
19	2.0	31.1	66.8
20	100.0	0.0	0.0

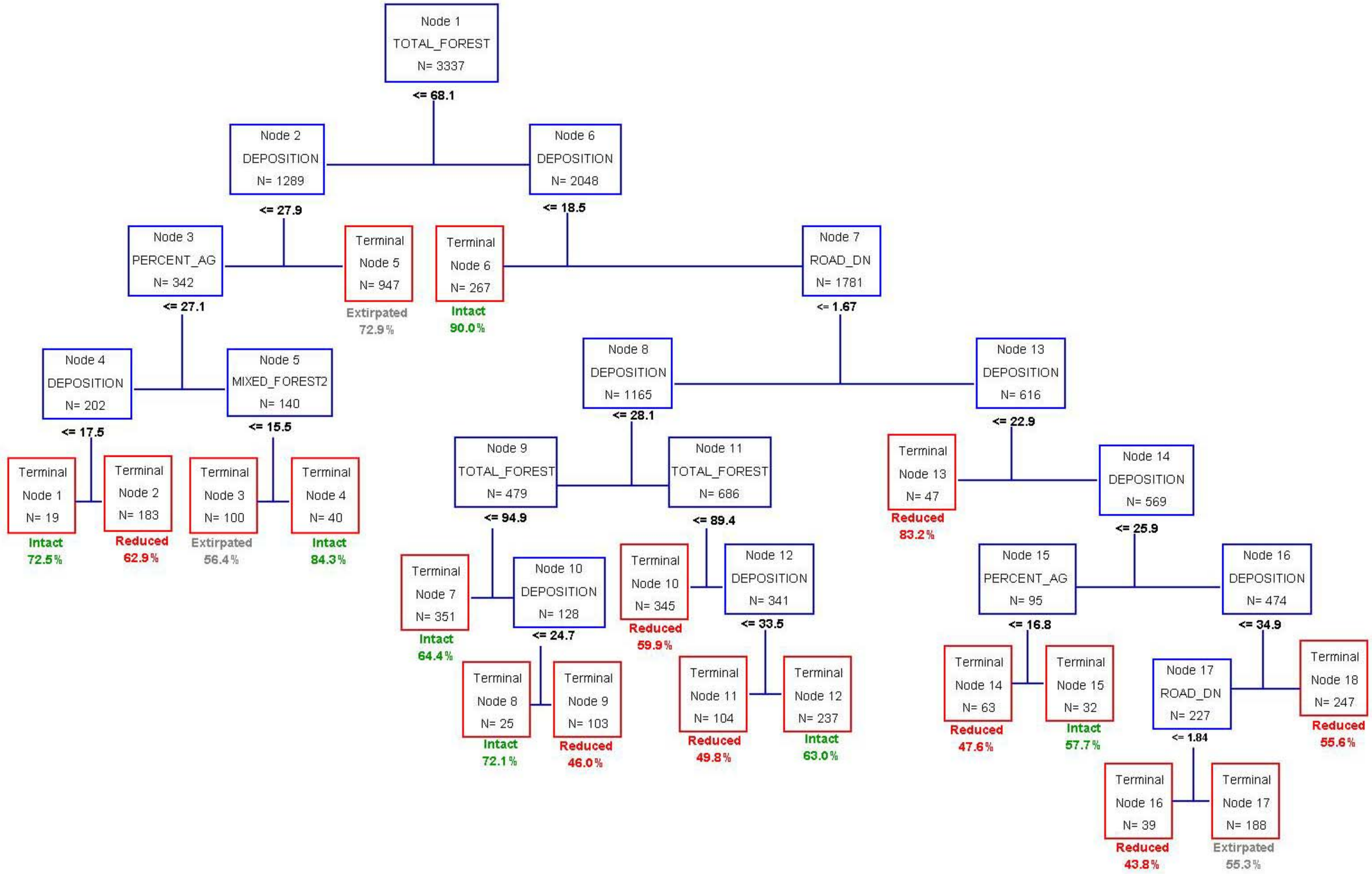

Figure 4. Pruned classification tree Model 3 (M3). Terminal nodes are indicated by red boxes. Final classification and within node classification probabilities in percentages are indicated below the terminal nodes.

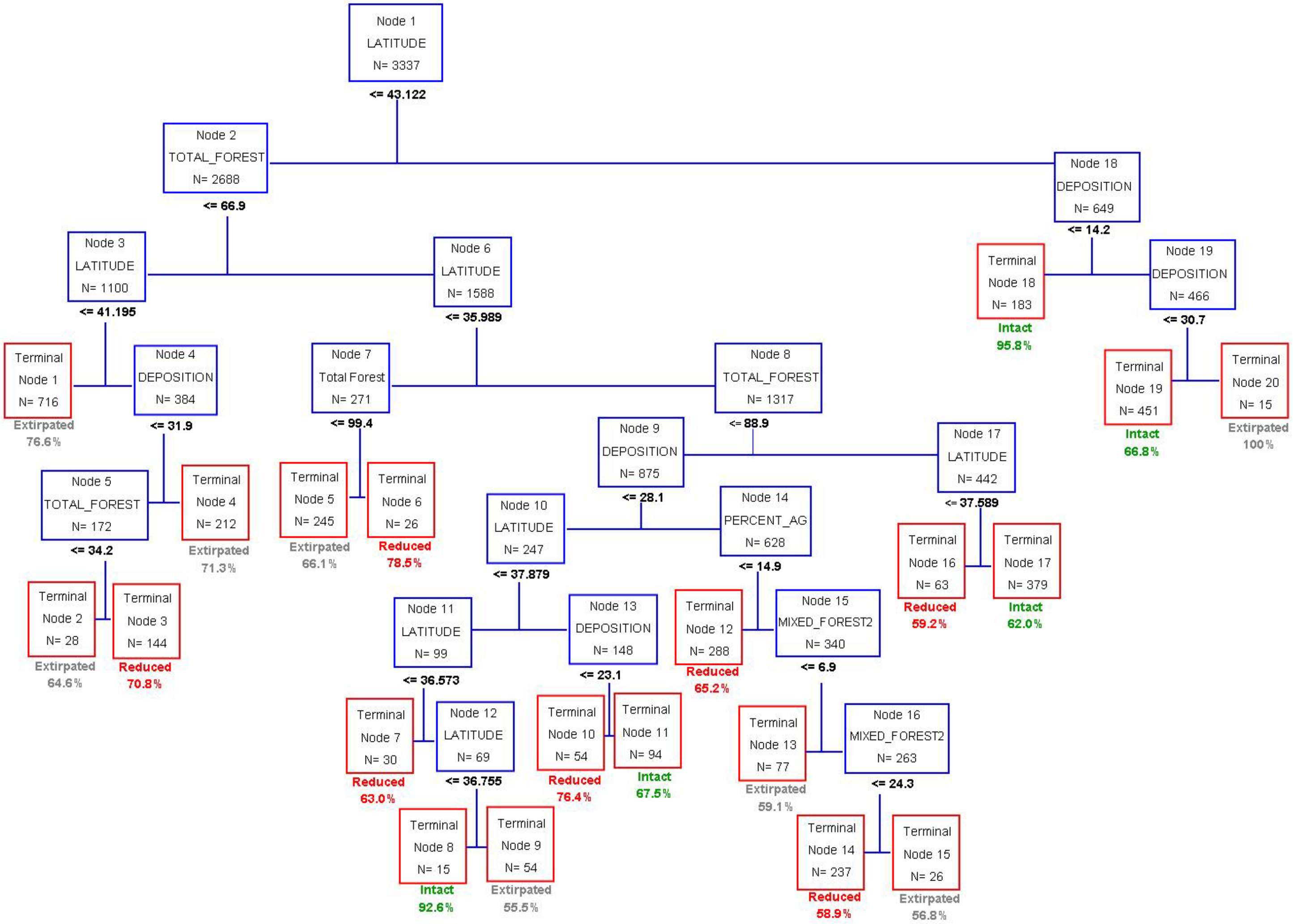

Figure 5. Pruned classification tree Model 4 (M4). Terminal nodes are indicated by red boxes. Final classification and within node classification probabilities in percentages are indicated below the terminal nodes.

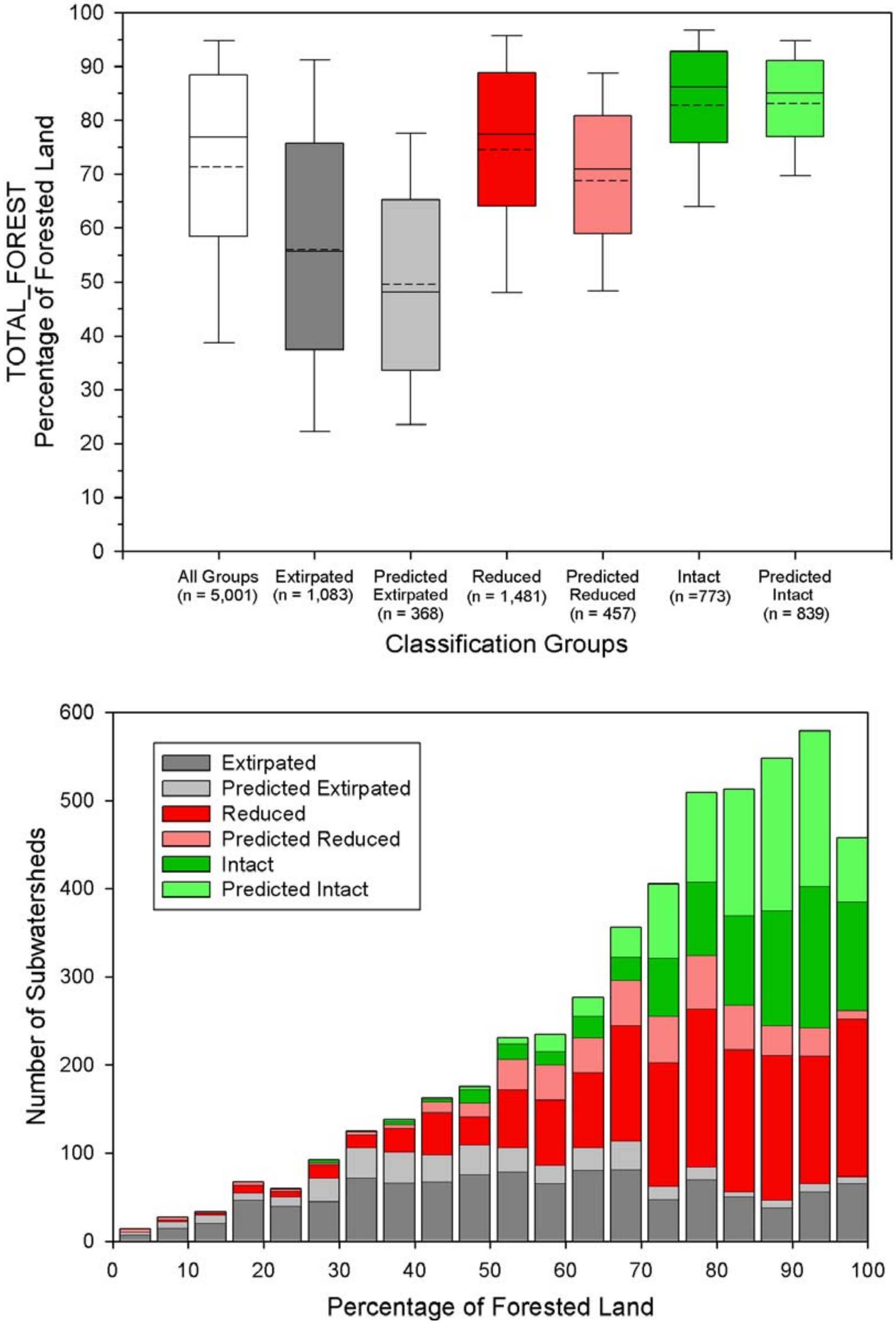

Figure 6. Box plot and histogram distribution of percentage of forested land per subwatershed for each subwatershed classification. Predicted subwatersheds based on Model 3.

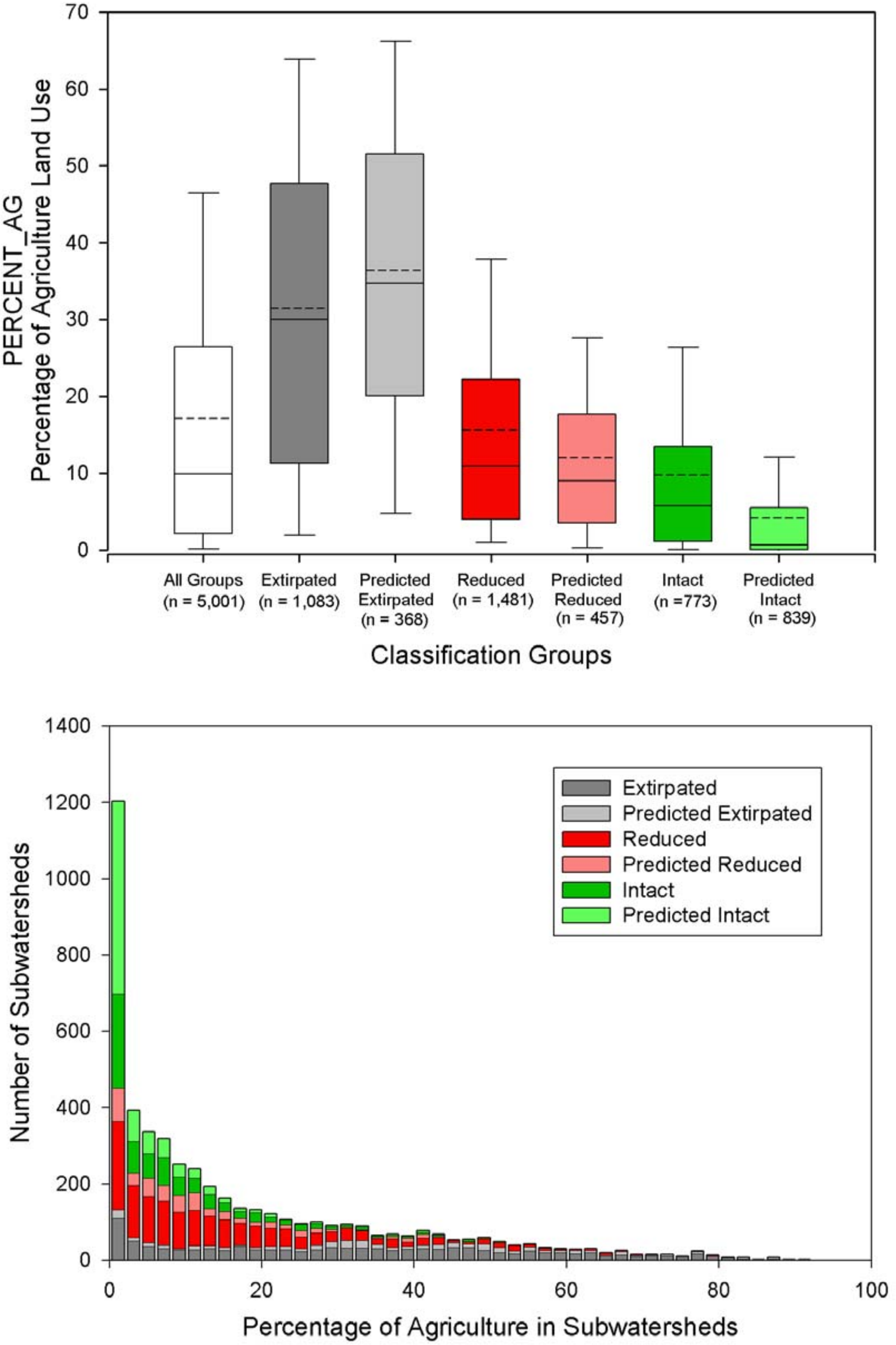

Figure 7. Box plot and histogram distribution of percentage of agriculture lands per subwatershed for each subwatershed classification. Predicted subwatersheds based on Model 3.

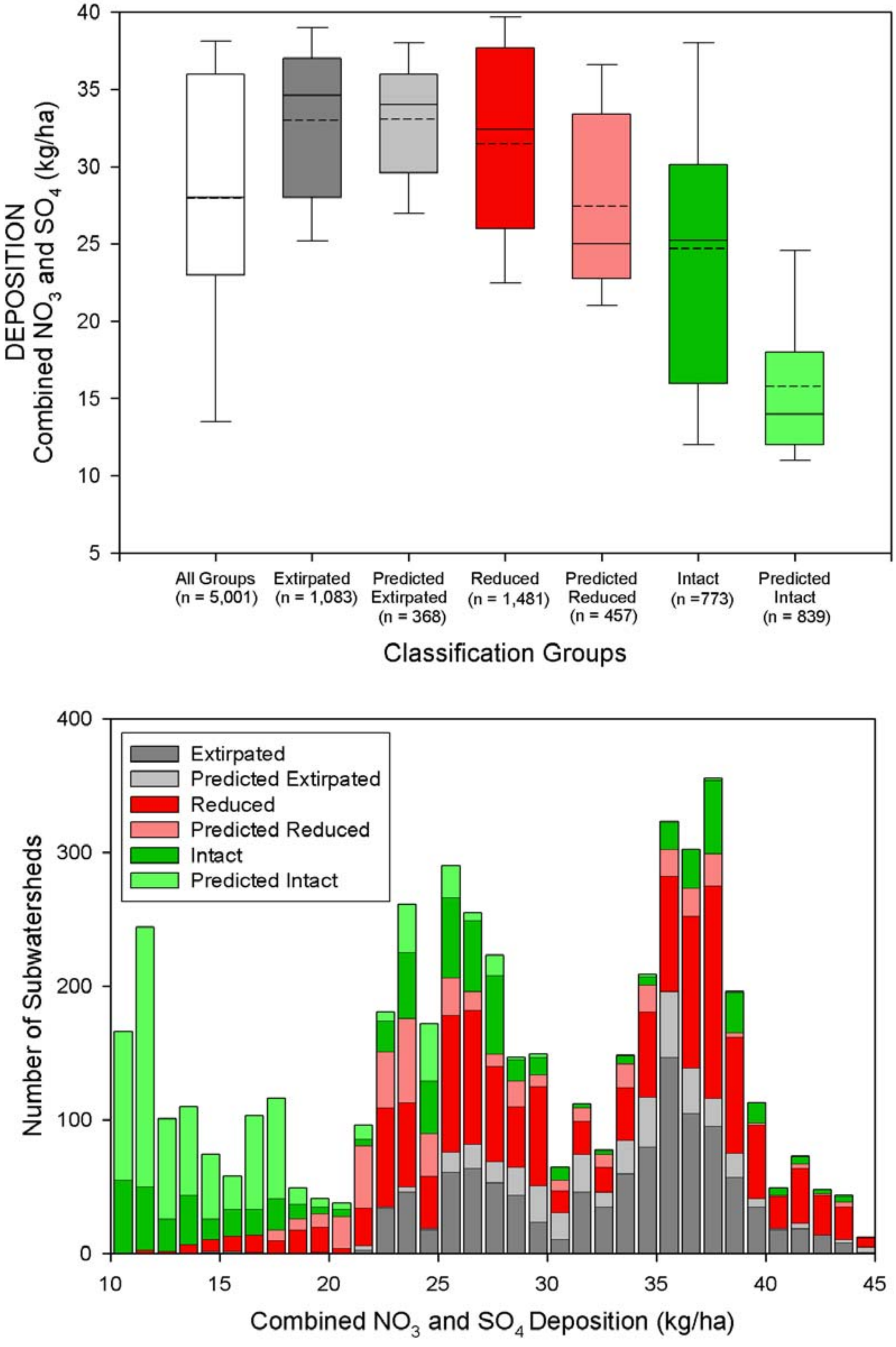

Figure 8. Box plot and histogram distribution of the combine NO_3 and SO_4 deposition per subwatershed for each subwatershed classification. Predicted subwatersheds based on Model 3.

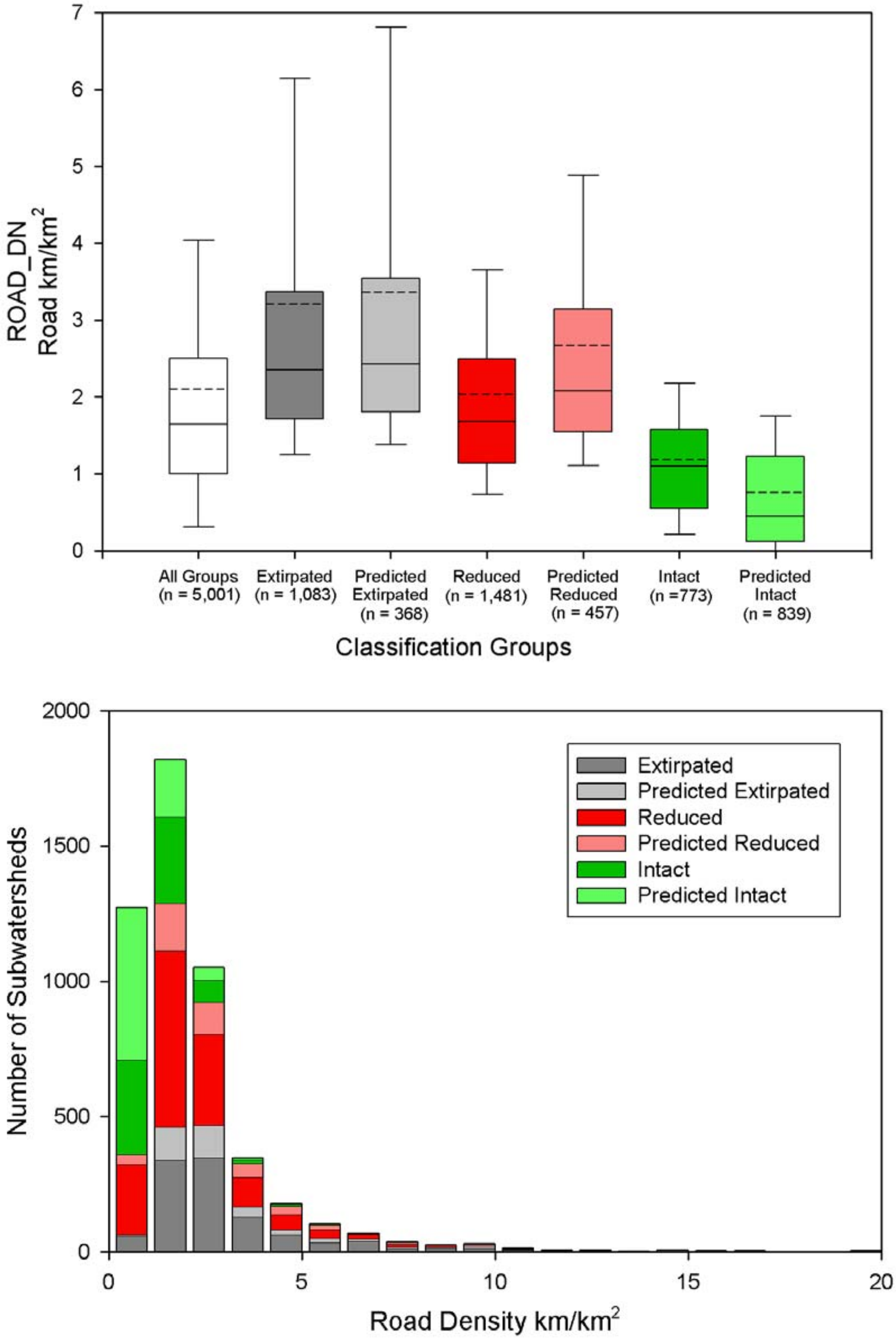

Figure 9. Box plot and histogram distribution of road density (km/km^2) per subwatershed for each subwatershed classification. Predicted subwatersheds based on Model 3.

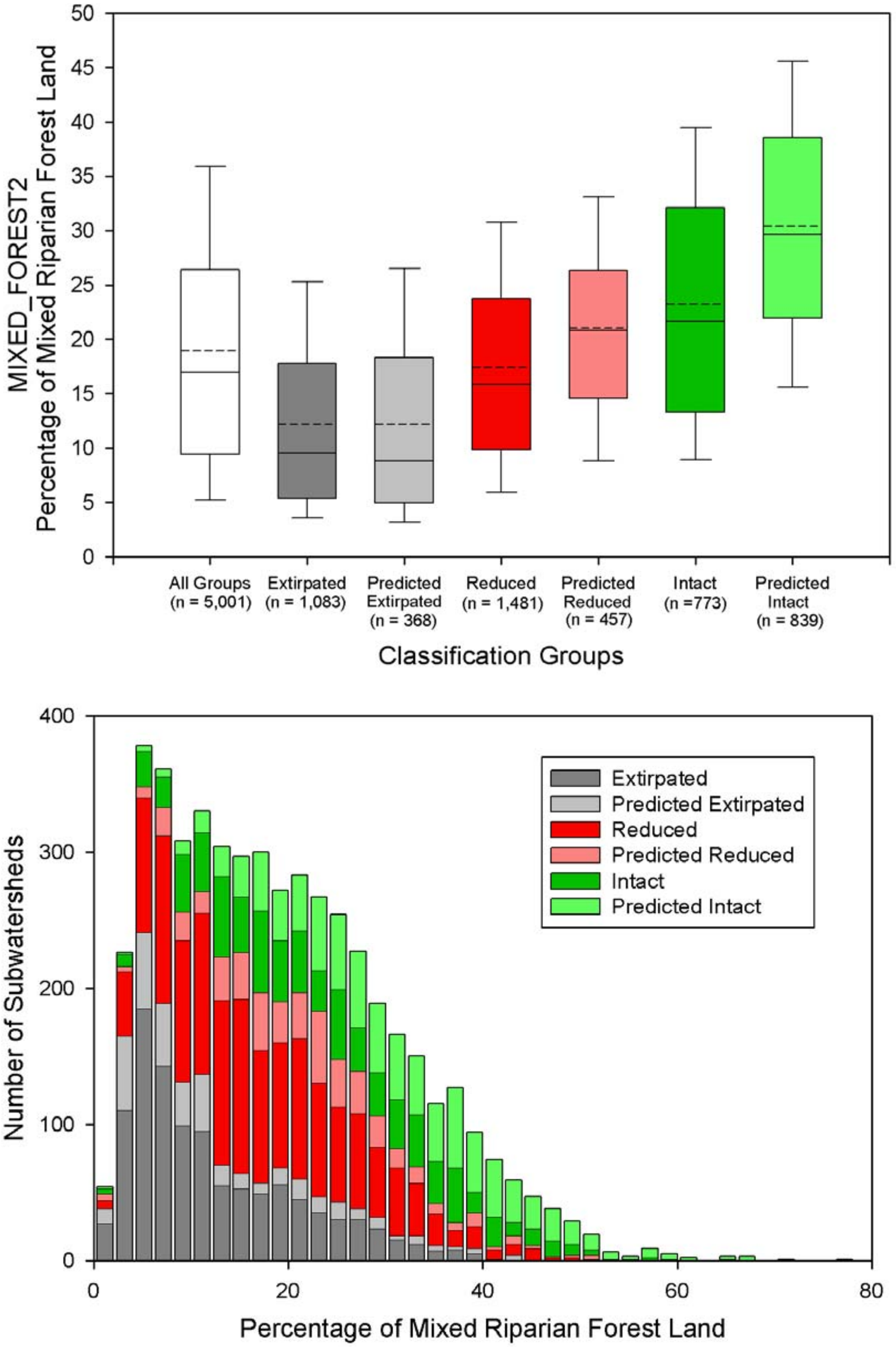

Figure 10. Box plot and histogram distribution of the percentage of mixed riparian forest land per subwatershed for each subwatershed classification. Predicted subwatersheds based on Model 3.

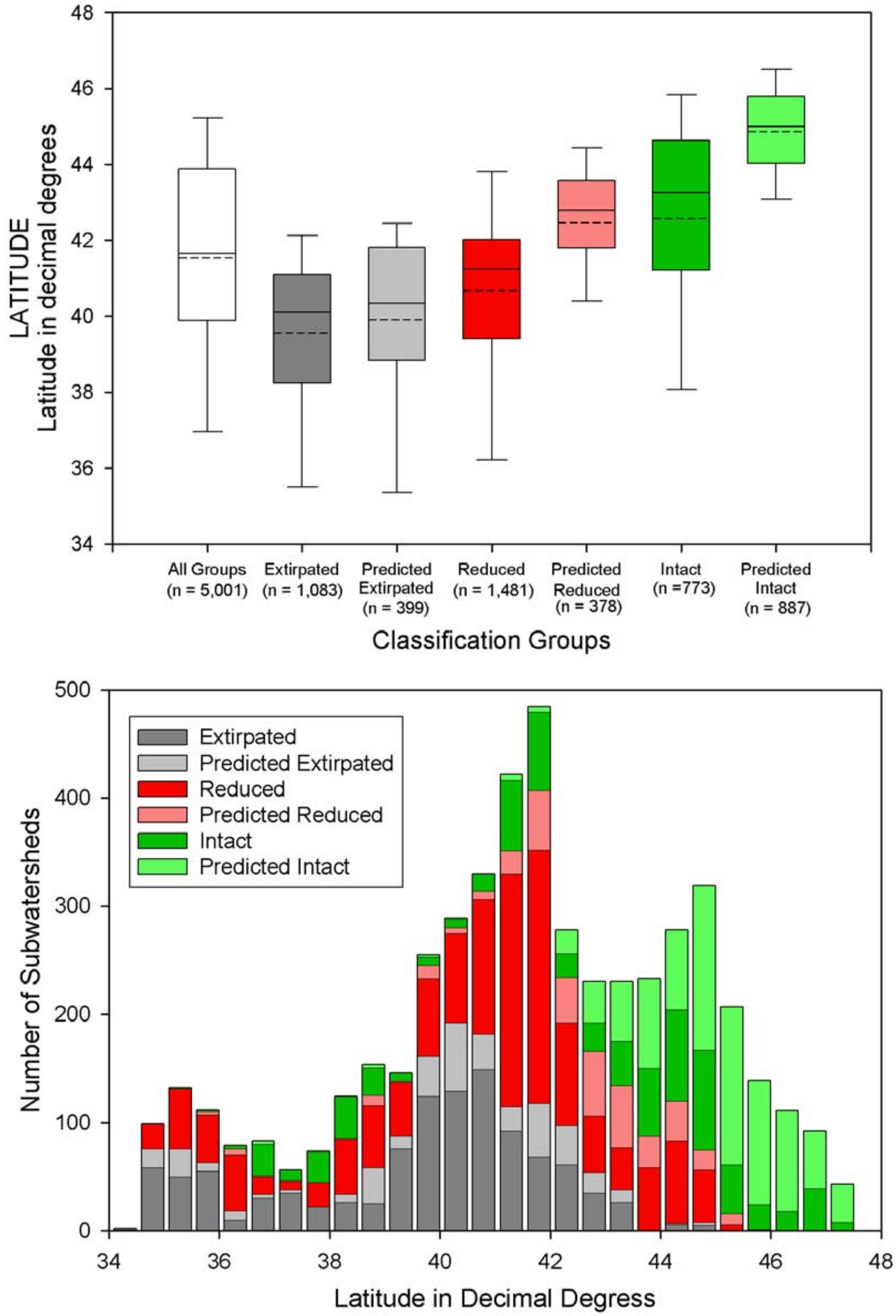

Figure 11. Box plot and histogram distribution of the percentage of mixed riparian forest land per subwatershed for each subwatershed classification. Predicted subwatersheds based on Model 4.

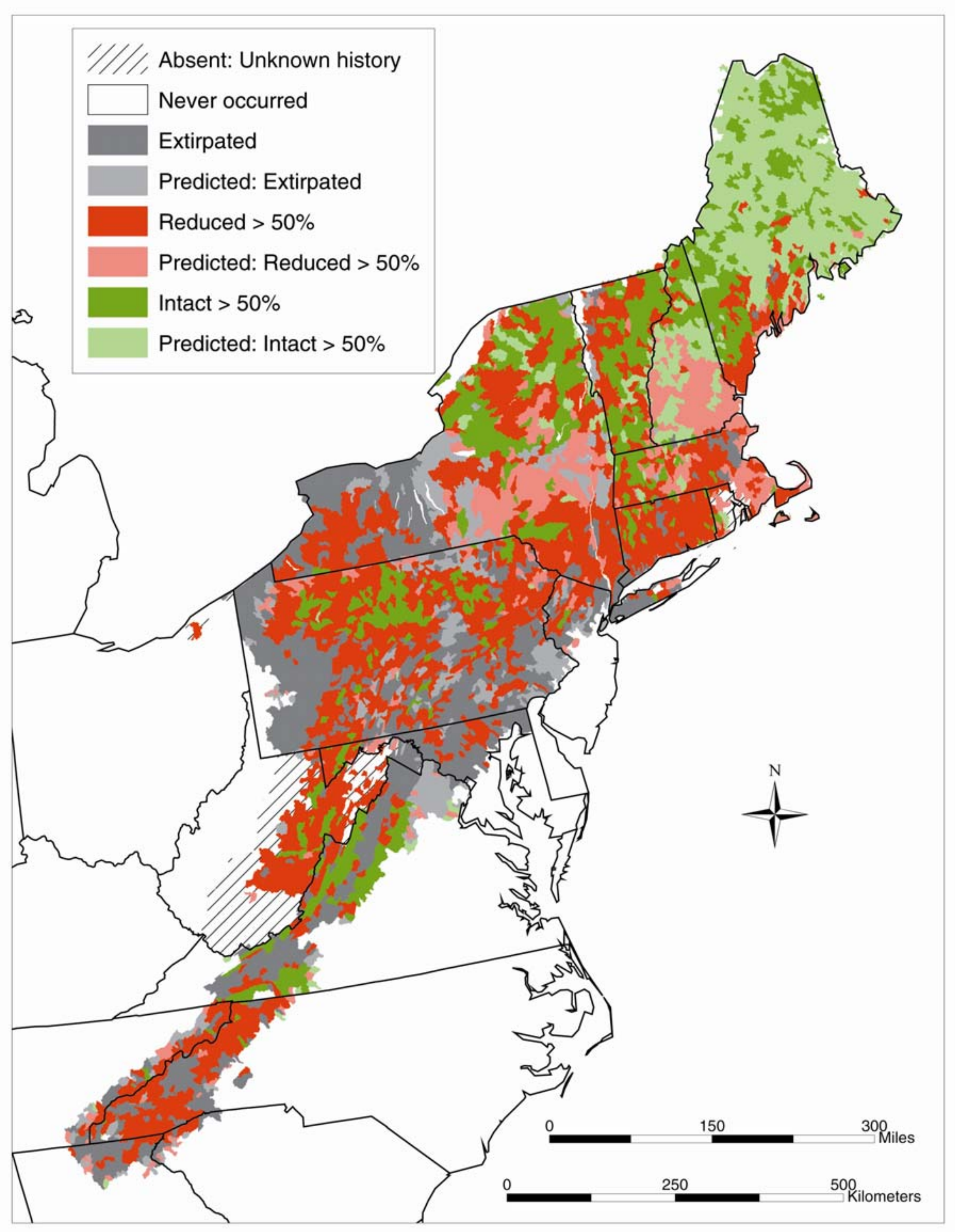

Figure 12. Distribution of known and predicted (Model 3) subwatershed classifications within the study area.

Misclassified subwatersheds

Subwatersheds that were predicted to be Intact or Reduced but in fact were Extirpated using M3 were predominately (31%) found in the southern Appalachians (Tennessee, North Carolina, Georgia, and South Carolina) (Figure 13).

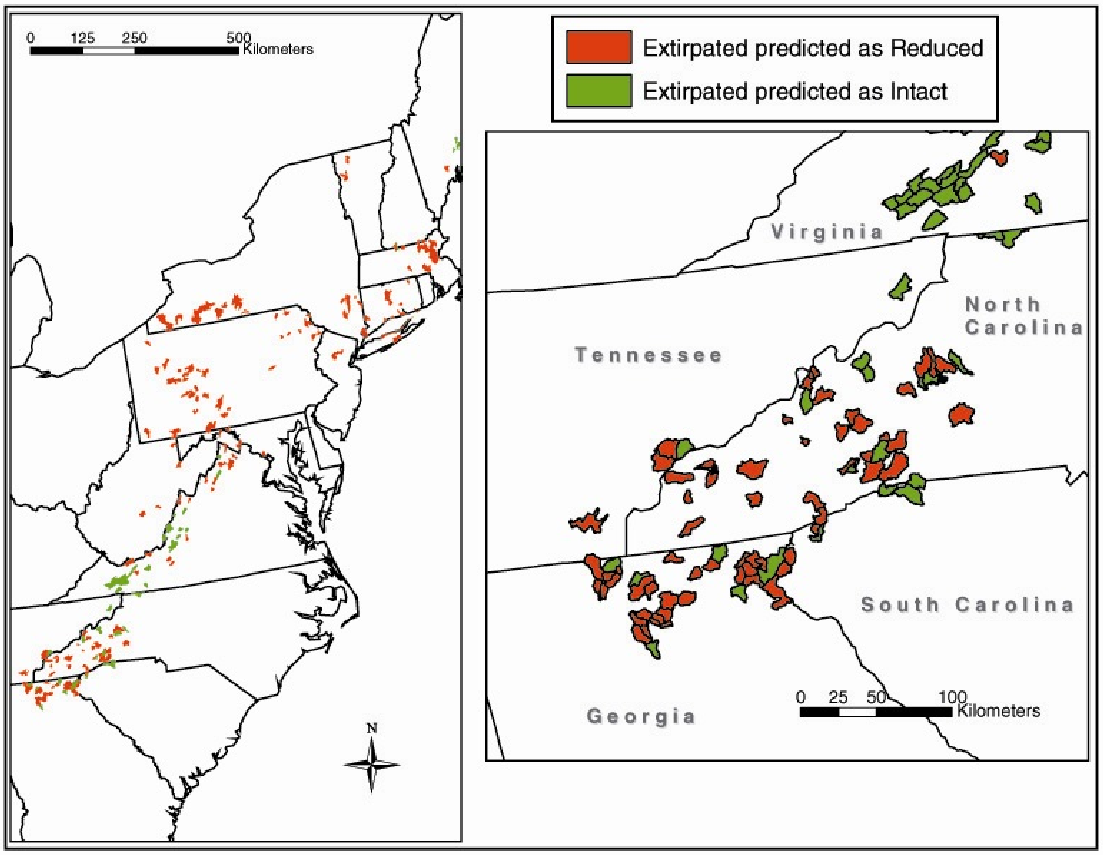

Figure 13. Extirpated subwatersheds incorrectly predicted as Reduced or Intact in the southern Appalachians based on Model 3 (M3).

The distribution of subwatersheds predicted to be Extirpated but classified as Reduced or Intact using M3 is illustrated in Figure 14. A high concentration of these misclassified subwatersheds was in Pennsylvania (45%) and New York (14%) (Figure 14).

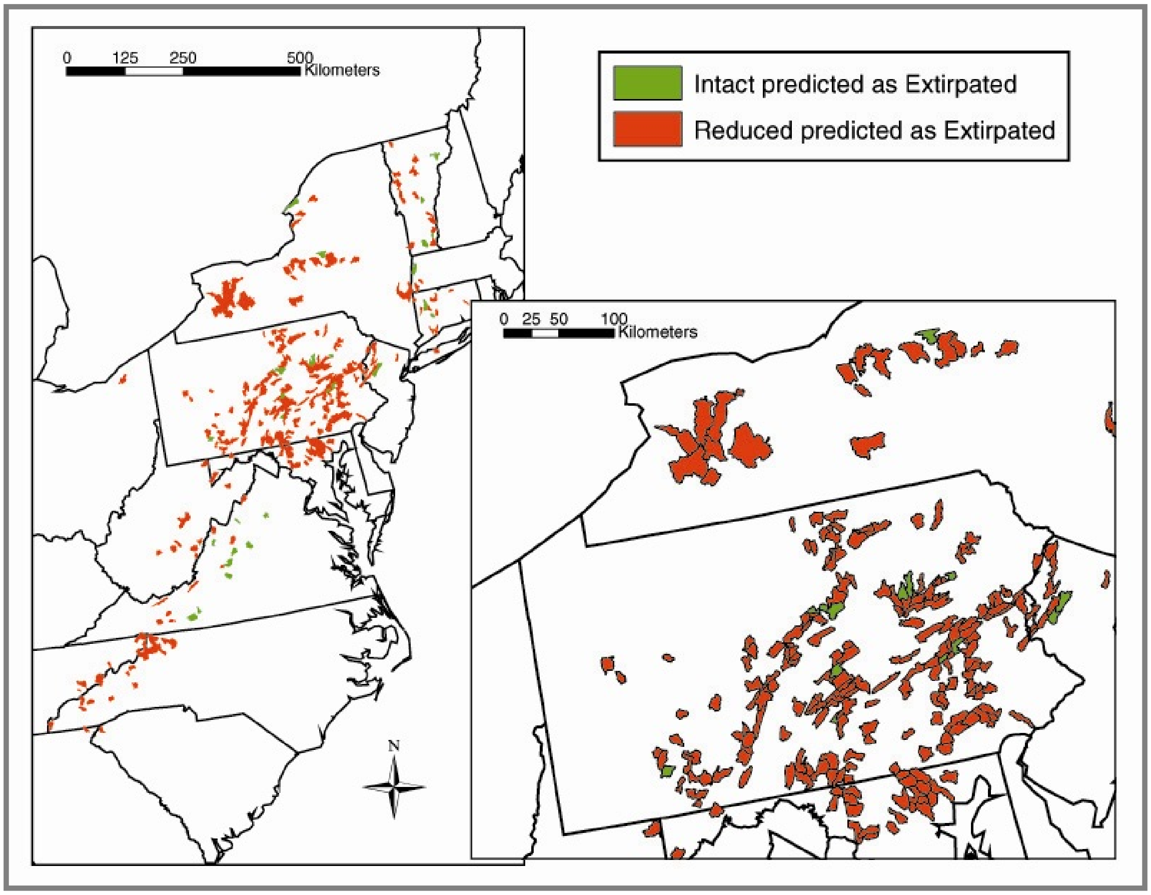
0 125 250 500 Kilometers
Intact predicted as Extirpated
Reduced predicted as Extirpated
0 25 50 100 Kilometers
N

Figure 14. Reduced and Intact subwatersheds incorrectly predicted as Extirpated based on Model 3 (M3).

Discussion

The classification tree models show that land use metrics at the subwatershed scale are effective predictors of brook trout. I chose the M3 model for the majority of final reporting and mapping because it is the most useful model for biologists and land managers. The binomial response models, M1 and M2, are useful for indicating the presence of brook trout, but do not give as much information as to the abundance or status of the populations. Although M4 (with latitude) has a slightly higher overall correct classification rate, M3 uses only the metrics that managers can influence.

Understanding the current distribution and population status of a species is one of the key tools in the conservation of that species (Williams et al. 1993; Warren et al. 1997). Although the Hudy et al. (2006) assessment produced a comprehensive large scale appraisal of the distribution of brook trout within the eastern United States, 33% of the subwatersheds within the study area did not have enough information to indicate the percentage of habitat within the subwatershed that supported self-sustaining brook trout populations. Filling in those information gaps by predicting the subwatershed status using core subwatershed metrics will be of use to natural resource managers by providing a baseline to help set goals and priorities and measure effectiveness of conservation and restoration efforts. By combining known and predicted status of brook trout subwatersheds in the study area, a complete picture is now available to land use managers. My analysis shows that only 32% of the subwatersheds are Intact for self-sustaining brook trout populations. The remaining 68% are either Reduced (39%) or Extirpated (29%). Although the brook trout is not threatened throughout its range, regional declines and local extinctions have occurred. While I was generally satisfied with the correct classification rates of the various classification tree models, as with any model, they need to be validated to confirm the accuracy of the predicted status.

In addition, to further understand the model, I conducted a geo-spatial analysis of the locations of the incorrect classifications (Figures 13 and 14). The southern Appalachians, which as a region only contains 10% of the subwatersheds in the study area, contained a high percentage (31%) of the subwatersheds that were predicted to be Intact or Reduced but in fact were Extirpated (Figure 13). The southern Appalachians is an area where exotic rainbow trout have displaced brook trout. Many of these subwatersheds currently have core metric values that would predict Intact brook trout (i.e. high percentage of forested lands, low percentage of agriculture, and low deposition). However, past land use practices and subsequent stocking of rainbow trout into restored subwatersheds have led to the establishment of rainbow trout (King 1937; King 1939; Lennon 1967; Kelly et al. 1980). Exotics were indicated as one of the top perturbations to brook trout by biologists and managers in the population assessment completed by Hudy et al. (2006). However, exotics were not one of the core metrics used in the predictive models. The exotic metric that I developed was derived from professional opinions of perceived subwatershed perturbations because consistent smaller scale (stream segment) data for exotic fishes are highly variable among states, making development of a quantitative exotic metric impossible at this time. The exotic metric was not responsive to the subwatershed classifications probably because of the complex interaction of natural and manmade barriers, stocking history, and the variability by experts in identifying exotics as a perturbation at the subwatershed level. I believe exotic fishes are an important variable for brook trout status but the impacts may not be adequately addressed at the subwatershed level over a large geographic area as was done in this study. A more localized predictive model for the southern Appalachians was not attempted because of the small numbers of subwatersheds that remain Intact (only six below the state of Virginia). Development of an accurate more quantifiable metric for exotics may improve future prediction models.

Subwatersheds predicted to be Extirpated but classified as Reduced or Intact were wider spread geo-spatially (Figure 14). The majority of these misclassified subwatersheds were located in Pennsylvania (45%) and New York (14%) (Figure 14). The model predicted these subwatershed as Extirpated because of the high road density associated with increasing urbanization (or suburbanization) of eastern Pennsylvania and the high acidic deposition. In general these misclassified subwatersheds had only one or two isolated brook trout populations and represent some of the subwatersheds most vulnerable to extirpations.

I believe the larger watershed sizes used for New York made the probabilities of the extirpation of brook trout in New York watersheds harder. When the smaller 6th level subwatersheds become available for New York I believe the model will be slightly improved for the correct classification rates of Extirpated subwatersheds.

Most of the misclassified subwatersheds were from the Reduced classification group. This suggests that the models are better at separating the two extremes of the classification status.

The misclassified subwatersheds bring up the caveat that these models are basically taking a snapshot in time. They use current subwatershed characteristics even though past land use practices may have caused brook trout extirpations from habitats and even whole subwatersheds. Even if past land use practices have been remedied, for example reforestation of agricultural and clear cut areas, it may take greater than 50 years for the stream habitat to recover to pre-impact conditions (Harding et al. 1998). Biologists, as part of the Hudy et al. (2006) assessment, indicated historical land use such as agriculture and logging as high perturbations to brook trout populations. However, much like the exotic species metric, the variability and geographic inconsistency of biologists' responses does not make for a robust metric.

The models I developed are also useful for developing metric thresholds, metric values at which a subwatershed switches from one status classification to another, to be used by biologists and land managers as warning flags or impact indicators. For example, 68% forested land appears to be an important factor in determining brook trout subwatershed status. A value of 68% forested lands was the first splitting criterion for three of the four classification tree models. In M3, only 6% of the Intact subwatersheds had a TOTAL_FOREST value of less than 68% (Figures 6). Although the value that resulted in the best overall correct classification rate of the TOTAL_FOREST single variable binomial response model was 52% forested land, the value that produced the best CCR for both classifications without sacrificing much in the overall CCR (74%) was 65% forested land. I suggest that land managers should consider subwatershed values below 65-70% forested lands as a tipping point for brook trout status.

Similar thresholds can be determined for the other metrics by maximizing the CCR among the classifications for the single variable logistic regression models and examining the histograms and classification tree splitting criteria. However, it should be noted that these thresholds, including the percentage of forested lands, are not absolute. Due to the interactions with the other metrics, the impact of the metric at these thresholds can either be compounded or mitigated. For example, the deposition value that results in the optimized CCR for the single variable logistic regression model is 33 kg/ha. However in M3 (Figure 4) it is possible to get Extirpated subwatersheds with a deposition value as low as 28 kg/ha if the percentage of forested land is below 68%. Therefore, I recommend that managers be aware when subwatersheds approach a combined NO_3 and SO_4 deposition value of greater than 24 kg/ha (Figure 8).

Managers should be concerned when the percentage of agricultural land in the subwatershed is in the 12-19% range or higher.   Only 17% of the Intact subwatersheds have a PERCENT_AG value greater than 19% and 74% of the Extirpated subwatersheds have a PERCENT_AG value greater than 12% (Figure 7). Another subwatershed threshold range that land managers should be aware of is a road density value greater than 1.8-2.0 km/km^2.   The road density value that optimized the CCR for predicting Extirpated from Presence was 2.0 km/km^2. Although 47% of the subwatersheds have a road density value greater than or equal to 1.8 km/km^2, Intact subwatersheds only constitute 8% (17% of the Intacts) of that group (Figure 9).

The remaining two core metrics do not produce thresholds that are as significant to brook trout conservation efforts. The percentage of riparian mixed forested land metric should not be confused with the percentage of riparian forested land. The MIXED_FOREST2 metric measures the percentage of land within the water corridor where both deciduous and evergreen trees are present but neither type represents over 75% of the cover. The metric measuring percentage of forested land in the water corridor was removed in the screening process due to redundancy with the subwatershed level percentage forested land metric. Mixed forested land in the water corridor helps to separate classifications in the models, but the biological reasons for this is unknown, reducing its utility as a threshold.

Latitude is another metric that is not as helpful to land managers. A high percentage (77%) of the Intact subwatersheds were above 41.0 degrees latitude, and in M4, the first node had a spitting criterion of 43.122 degrees latitude (Figure 11). Although it increases the correct classification rates of the models, latitude was not included in some of the models because it is not a metric that land managers can control. Separate models were not presented for each region (i.e. New England, Mid-Atlantic Highlands, and southern Appalachians) because each region did not contain adequate sample sizes of the subwatershed classifications to produce reliable models.

Although at the subwatershed level the core metrics effectively predicted the classification status; these metrics are not the only influences on brook trout. Simply because a metric did not make it through the screening process does not mean that it is not a biologically significant factor for brook trout populations. Also, some metrics may have greater influence on brook trout populations and are better predictors at different scales (Kocovsky and Carline 2006). For example, the U. S. Environmental Protection Agency was able to predict brook trout presence/absence in stream segments in the Mid-Atlantic Highlands with a CCR of 79% using the metrics: depth, temperature, substrate, percentage riffles, cover, and riparian vegetation (Rashleigh et al. 2005). Because of the large geographic area of this study, I used larger scale (subwatershed) metrics due to the inconsistency of smaller scale (stream segment) data.

The objective of this study was to determine if models using land use metrics at the subwatershed scale could be used to predict brook trout subwatershed status and develop metric thresholds that could be used by land managers for conservation. The land use models were successful in predicting brook trout status and filled in information gaps by classifying the Unknown and Present subwatersheds. The models also aided in indicating useful thresholds for forested lands, agricultural lands, acid deposition, and road density. I believe the models could be improved by acquiring a quantitative exotic metric and 6th level watersheds for New York and that an important next step would be to validate the model with stream inventories of a sample of the predicted subwatersheds. Overall, the brook trout subwatershed status distribution and threshold metric values can be useful for risk assessments and for prioritizing conservation efforts; whether one's conservation strategy is to protect the "best of the best" or rehabilitate the worst (Frissell 1993).

References

Agresti, A. 1996. An introduction to categorical data analysis. John Wiley and Sons Inc., New York.

Anderson, J. F., E. E. Hardy, J. T. Roach, and R. E. Witmer. 1976. A land use and land cover classification system for use with remote sensor data, U.S. Geological Survey Professional Paper 964, U.S. Geological Survey, Washington, DC.

Barbour, M. T., J. Gerritsen, B. D. Snyder, and J. B. Stribling. 1999. Rapid Bioassessment protocols for use in streams and wadeable rivers: periphyton, benthic macroinvertebrates, and fish, Second Edition. EPA 841-B-99-002. U.S. Environmental Protection Agency; Office of Water; Washington, D.C.

Behnke, R.J. 2002. Trout and Salmon of North America. Free Press, Simon and Shuster, Inc. New York, New York.

Box, G. E. P. and D. R. Cox. 1964. An Analysis of Transformations. Journal of the Royal Statistical Society. 211-252.

Brasch, J., J. McFadden, and S. Kmiotek. 1958. The eastern brook trout: its life history, ecology, and management. Wisconsin Conservation Department, Publication 226, Madison, Wisconsin.

Breiman, L., J. H. Friedman, R. Olshen, and C. J. Stone. 1984. Classification and regression trees. Wadsworth, Belmont, California.

Clark, L. A., and D. Pregibon. 1992. Tree based models. Pages 377-419 *in* J. M. Chambers and T. J. Hastie, editors. Statistical models. Wadsworth and Brooks/Cole, South Pacific Grove, California.

Collett, D. 2002. Modelling Binary Data, 2nd edition. Chapman and Hall, New York.

Curry, R. A., and W. S. MacNeill. 2004. Population-level responses to sediment during early life in brook trout. North American Benthological Society 23(1):140-150.

Davis, W. S., and T. P. Simon. 1995. Biological assessment and criteria: tools for watershed resource planning and decision making. Lewis Publishers, Washington, D.C. Doppelt, B., M. Scurlock, C. A. Frissell, and J. Karr. 1993. Entering the watershed. Island Press, Covello, California.

Discroll, C. T., G. B. Lawrence, A. J. Bulgur, T. J. Butler, C. S. Cronan, C. Eagar, K. F. Lambert, G. E. Likens, J. L. Stoddard, and K. C. Weathers. 2001. Acidic deposition in the Northeastern United States: sources and inputs, ecosystem effects, and management strategies. Bioscience 51(3):180-198.

Earth Systems Science Center. 2005. Pennsylvania State University College of Earth and Mineral Sciences. Soil Information for Environmental Modeling and Ecosystem Management. http://www.essc.psu.edu/soil_info/index.cgi?soil_data&conus&background (August, 2004)

EPA (U. S. Environmental Protection Agency). 2002. Hydrologic unit boundaries from USGS. http://www.epa.gov/region02/gis/atlas/hucs.htm. (November 2004).

Frissell, C. A. 1993. A new strategy for watershed recovery of Pacific salmon in the Pacific Northwest. Report prepared for the Pacific Rivers Council Inc., Eugene, Oregon. March 1993.

Frissell, C. A., and D. Bayles. 1996. Ecosystem management and the conservation of aquatic biodiversity and ecological integrity. Water Resources Bulletin 32:229-240.

Galbreath, P. F., N. D. Adams, S. Z. Guffey, C. J. Moore, and J. L. West. 2001. Persistance of Native Southern Appalachian Brook Trout Populations in the Pigeon River System, North Carolina. North American Journal of Fisheries Management 21:927-934.

Harding, J. S., E. F. Benfield, P. V. Bolstad, G. S. Helfman, and E. B. D. Jones III. 1998. Stream biodiversity: The ghost of land use past. Proceedings of the National Academy of Sciences 95:14843-14847.

Hosmer, D. W., and S. Lemeshow. 1980. Goodness-of-fit testing for multiple logistic regression model. Communications in Statistics. A10:1043-1069.

Hosmer, D. W., and S. Lemeshow. 2000. Applied logistic regression, 2nd edition. John Wiley and Sons Inc., New York.

Hudy, M., T. M. Thieling, N. Gillespie, and E. P. Smith. 2006. Distribution, status, and perturbations to brook trout within the eastern United States. Final report to the Eastern Brook Trout Joint Venture.

Huberty, C. J. 1994. Applied discriminant analysis. John Wiley and Sons, Inc., New York.

Hughes, R. M., P. R. Kaufmann, A. T. Herlihy, T. M. Kincaid, L. Reynolds, and D. P. Larsen. 1998. A process for developing and evaluating indices of fish assemblage integrity. Canadian Journal of Fisheries and Aquatic Sciences 55:1618-1631.

Johnson, S. L. and J. A. Jones. 2000. Stream temperature responses to forest harvest and debris flows in western Cascades, Oregon. Canadian Journal of Fisheries and Aquatic Sciences 57:1-10.

Kelly, G. A., J. S. Griffith, and R. D. Jones. 1980. Changes in distribution of trout in Great Smoky Mountains National Park, 1900-1970. U.S. Fish and Wildlife Service Technical Paper 102.

King, W. 1937. Notes on the distribution of native speckled and rainbow trout in the streams of Great Smokey Mountains National Park. Journal of the Tennessee Academy of Science 12:351-361.

King, W. 1939. A program for the management of fish resources in Great Smokey Mountains National Park. Transaction of the American Fisheries Society 68:86-95.

Kleinbaum, D.G., L. L.Kupper, and K. E. Muller. 1988. Applied Regression Analysis and Other Multivariable Methods. PWS-Kent, Boston.

Kocovski, P. M., and R. F. Carline. 2006. Influence of landscape-scale factors in limiting brook trout populations in Pennsylvania streams. Transaction of the American Fisheries Society 135:76-88.

Lennon 1967, R. E. 1967. Brook trout of Great Smokey Mountains National Park. U.S. Fish and Wildlife Service Technical Paper 15.

Lo, C.P., and A. K. W. Yeung. 2002. Concepts and Techniques of Geographic Information Systems. Prentice-Hall Inc., Upper Saddle River, New Jeresey.

MacCrimmon, H. R. and J. S. Campbell. 1969. World Distribution of Brook Trout, *Salvelinus fontinalis.* Journal of Fisheries Research Board of Canada 26(7):1699-1725.

Marschall, E. A. and L.B. Crowder. 1996. Assessing Population Responses to Multiple Anthropenic Effects: A Case Study with Brook Trout. Ecological Applications 6(1):152-167.

Master, L. L., S. R. Flack, and B. A. Stein, editors. 1998. Rivers of Life: Critical watersheds for protecting freshwater biodiversity. The Nature Conservancy, Arlington, VA.

McCormick, F. H., R. M. Hughes, P. R. Kaufmann, D. V. Peck, J. L. Stoddard, and A. T. Herlihy. 2001. Development of an Index of Biotic Integrity for the Mid-Atlantic Highlands Region. Transactions of the American Fisheries Society 130(5):857-877.

McDougal, L. A., K. M. Russell, and K. N. Leftwich, editors. 2001. A Conservation Assessment of Freshwater Fauna and Habitat in the Southern National Forests. USDA Forest Service, Southern Region, Atlanta Georgia. R8-TP 35 August 2001.

Moyle, P. B., and R. M. Yoshiyama. 1994. Protection of aquatic biodiversity in California: a five-tiered approach. Fisheries 19:6-18.

Moyle, P. B. and P. J. Randall. 1998. Evaluating the biotic integrity of watersheds in the Sierra Nevada, California. Conservation Biology 12(6):1318-1326.

National Atmospheric Deposition Program. 2005. Isopeth Grids. http://nadp.sws.uiuc.edu/isopleths/grids.asp. (July 2005).

Navtech. 2001. Navstreets: street data. CD-Rom, Version 2.5. http://www.navtech.com/data/data.html

Neville, P. G., 1999. Decision trees for predictive models. SAS Institute Inc. Cary, North Carolina.

NRCS (Natural Resource Conservation Service). 2005. Watershed Boundary Dataset (WBD). http://www.ncgc.nrcs.usda.gov/products/datasets/watershed/ (November 2005)

Rashleigh, B., R. P. Parmar, J. M. Johnston, and M. C. Barber. 2005. Predictive habitat models for the occurrence of stream fishes in the Mid-Atlantic Highlands. North American Journal of Fisheries Management 25:1353-1336.

Rencher, A. 2002. Methods of Multivariate Analysis, 2nd edition. John Wiley and Sons Inc., New York.

Rieman, B. E., D. C. Lee, and R. F. Thurow. 1997. Distribution, Status, and Likely Future Trends of Bull Trout within the Columbia River and Klamath River Basins. North American Journal of Fisheries Management 17:1111-1125.

Seaber, P. R., F. P. Kapinos, and G. L. Knapp. 1987. Hydrologic unit maps: U.S. Geological Survey Water-Supply Paper 2294.

Sokal, R.R. and F.J. Rohlf. 1995. Biometry, 3rd edition. W. H Freeman and Company, New York.

Statistical Sciences. 1993. S-PLUS guide to statistical and mathematical analysis, version 3.2. Mathsoft, Inc, Seattle, Washington.

U. S. Army Corps of Engineers. 1998. National Inventory of Dams data. http://crunch.tec.army.mil/nid/webpages/nid.cfm (November 2004).

U.S. Census Bureau. 2002. Census 2000: Summary Files. http://www.census.gov. (November 2002).

USGS (U. S. Geological Survey). 1999. National Elevation Dataset. http://edcnts12.cr.usgs.gov/ned/ned.html. (June 2004).

USGS (U. S. Geological Survey). 2002. Water resources of the United States: Hydrologic unit maps. http://water.usgs.gov/GIS/huc.html. (November 2004).

USGS (U. S. Geological Survey). 2004a. National Hydrography Dataset. http://nhd.usgs.gov. (June 2004).

USGS (U. S. Geological Survey). 2004b. National Land Cover Dataset 1992 (NLCD 1992). http://landcover.usgs.gov/natllandcover.asp. (June 2004).

U.S. Geological Survey. 2005. Accuracy Assessment of 1992 National Land Cover Data. http://landcover.usgs.gov/accuracy/index.asp. (January 2006).

Warren, M. L., Jr., P. L. Angermeier, B. M. Burr, and W. R. Haag. 1997. Decline of a diverse fish fauna: patterns of imperilment and protection in the southeastern United States. Pages 105-164 in Benz, G. W. and D. E. Collins (editors). Aquatic Fauna in Peril: The Southeastern Perspective. Special Publication 1, Southeast Aquatic Research Institute, Lenz Design and Communications, Decatur, GA.

Whalen, K. 2004. A risk assessment for crayfish conservation on National Forest lands in the eastern United States. Master's thesis. James Madison University, Harrisonburg, Virginia.

Williams, J. D., M. L. Warren, Jr., K. S. Cummings, J. L. Harris, and R. J. Neves. 1993. Conservation status of freshwater mussels of the United States and Canada. Fisheries 18(9):6-22.