MODELING FERAL SWINE DISTRIBUTION IN GEORGIA USING LOGISTIC AND AUTOLOGISTIC REGRESSION

by

DONG CHEN

(Under the Direction of Marguerite Madden)

ABSTRACT

Feral swine (*Sus scrofa*) is a very destructive exotic mammal in the United States that carries pathogens of several diseases, endangers the safety of human beings, and disturbs local ecosystems. Although efforts have been made to monitor the distribution of feral swine, little has been done to model and predict the future distribution of feral swine. This project aims at tackling this problem by identifying the relationship between feral swine distribution and a series of environmental and cultural factors based on logistic regression. To assess the effects of spatial autocorrelation in modeling feral swine, autologistic regression was also applied to be compared with the ordinary logistic regression. The results suggest the autologistic regression model is superior to the ordinary counterpart with better performance. In addition, it is strongly recommended that the ordinary logistic regression methods should be employed with caution when spatial autocorrelation exists because they may yield misleading results.

INDEX WORDS: Autologistic Regression, Feral Swine, Habitat Modeling, Logistic Regression, Spatial Autocorrelation, Spatial Statistics

MODELING FERAL SWINE DISTRIBUTION IN GEORGIA USING LOGISTIC AND AUTOLOGISTIC REGRESSION

by

DONG CHEN

B.S. Wuhan University, China, 2009

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2011

MODELING FERAL SWINE DISTRIBUTION IN GEORGIA USING LOGISTIC AND AUTOLOGISTIC REGRESSION

by

DONG CHEN

Major Professor:	Marguerite Madden
Committee:	Xiaobai Yao
	Thomas R. Jordan

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2011

ACKNOWLEDGEMENTS

I would like to thank Dr. Marguerite Madden for supporting my research wholeheartedly. Without her, I would not be able to work for the Center for Remote Sensing and Mapping Science (CRMS) as a research assistant, where I became interested in conducting such a study on feral swine modeling. She gave me a tremendous amount of advices on every aspect of my thesis, from both professional side and linguistic side. With her help, I truly became enthusiastic about habitat modeling and this will definitely be one of the areas that I would like to devote myself to in the future.

I would like to thank Dr. Thomas Jordan for helping me on the technique part of this study. His profound expertise in image processing is truly admirable to me. In addition, his optimistic attitude toward work and study impresses me significantly, and he made me realize that with a positive attitude, there is always a way to tackle a problem no matter how difficult it is.

I would like to thank Dr. Xiaobai Yao for leading me to spatial statistics. Her course on spatial statistics made me realize the potential of this technique and helped me be prepared for this study. Her advice on statistical analysis methods made sure that I was always pointed towards the right direction throughout this project.

I would like to thank everyone in CRMS. They encouraged me every time when I was frustrated by unexpected failures or undesirable results. More importantly, they helped me analyze the problems that I had and offered a myriad of suggestions based on their expertise.

Some of these suggestions indeed inspired me and guided me in the right direction. Without them, I would not be able to finish this study on time.

I would like to thank Dr. Joseph L. Corn from the Southeastern Cooperative Wildlife Disease Study (SCWDS), College of Veterinary Medicine, University of Georgia. His knowledge on feral swine biology and diseases provided me with valuable insights for interpreting some of the patterns of feral swine distribution for which I could not find appropriate explanation.

Finally, I would like to thank to my parents, Tieping Chen and Xuxu Ren. Without their support and encouragement for the past few years, I would not be here studying in one of the best geography departments in the United States and would not be able to finish this project successfully. A series of dreams of mine have come true and it is my parents that made them happen. I love you Dad and Mom!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Feral Swine: A Dangerous and Destructive Pest

"Feral swine" is a collective term referring to escaped domestic pigs, Eurasian wild boars, and the hybrids of these two (Mapston 2004; Seward, VerCauteren et al. 2004). Native to Europe and Asia (Eurasia), domestic pigs were introduced to the United States mainland in 1539 by Spanish explorer Hernando de Soto (Towne and Wentworth 1950; Simmons 2010). The domestic pigs were not strictly controlled so they were able to roam freely, and many of them became feral. Later, Eurasian wild boars were also brought into the United States by American hunters for hunting purposes. Both of them, together with their hybrids, constitute the feral swine population existing on the American continent today (Simmons 2010).

Feral swine are the most abundant, free-ranging, exotic ungulate in the United States (McKnight 1964; Decker 1978; Seward, VerCauteren et al. 2004), and they have been spreading throughout the United States quickly due to several reasons. First is the absence of natural predators of feral swine over much of the area they occupy. Although there are certain species which prey on feral swine piglets including coyote, owls and bobcats (Stevens 1996), adult feral swine do not have many natural predators in the North American continent other than humans. Historically, species including mountain lion, black bear, wolf, and panther were the main predators of adult feral swine. However, they either do not currently exist in the United States, or their populations cover a much smaller area than feral swine (Stevens 1996; Seward, VerCauteren et al. 2004).

The second reason feral swine are spreading in the United States is their high reproductive ability. Characteristics of feral swine including large litter sizes, short gestation periods and early maturity enable them to possess very high reproductive potential. According to Seward, VerCauteren et al. (2004) and Mapston (2004), feral swine possess the greatest reproductive potential of all free-ranging, large mammals in the United States. A variety of examples in the literatures suggested that feral swine can produce two litters per year with average litter size ranging from four to ten (Choquenot, McIlroy et al. 1996; Stevens 1996; Taylor, Hellgren et al. 1998; Seward, VerCauteren et al. 2004; Simmons 2010). Their breeding behavior can occur any time throughout the year under favorable conditions (Choquenot, McIlroy et al. 1996; Mapston 2004). Given conditions with adequate nutrition, a feral swine population can double in four months (Mapston 2004).

The third reason for feral swine spread is their strong adaptability of feral swine. Although some literature reports that feral swine prefer moist bottomlands with adequate water supply and thick vegetation, they are also capable of quickly adapting to a wide range of habitat types (Choquenot, McIlroy et al. 1996; Stevens 1996; Mapston 2004). It was reported that feral swine had established populations in 38 states throughout the United States as of 2009 (Wyckoff, Henke et al. 2009).

Feral swine are very detrimental and destructive in terms of carrying diseases, causing huge economic losses and disturbing the environment.

Pathogens and Diseases

Feral swine can carry pathogens and virus of several diseases and thus are a tremendous threat to both humans and animals. There has not been a consensus on the exact number of diseases that can be carried and transmitted by feral swine. Simmons (2010) claims that feral swine can carry up to 13 serious diseases. However, according to Seward, VerCauteren et al. (2004) and Williams and Barker (2001), at least 30 significant viral and bacteriological diseases can be harbored by feral swine. Among the diseases hosted by feral swine, pseudorabies and swine brucellosis are of most concern. Feral swine have proved to be related to the spread of the pseudorabies virus (PRV), which is a highly contagious viral disease spreading among not only feral swine, but also domestic swine (Vanderleek, Becker et al. 1993; Choquenot, McIlroy et al. 1996; Corn, Cumbee et al. 2009). The swine will become lifetime carriers once they are infected by the pseudorabies virus (Stevens 1996). Adult feral swine seldom suffer from the infection of pseudorabies virus, serving primarily as reservoir hosts (Tozzini, Poli et al. 1982; Mengeling and Pirtle 1989; Pensaert and Kluge 1989; Vanderleek, Becker et al. 1993). The symptoms related to the infection of domestic swine include high piglet mortality, central nervous and respiratory system damages, anorexia and abortion in adults (Pensaert and Kluge 1989; Vanderleek, Becker et al. 1993). Other than domestic swine, the pseudorabies virus can also infect a large amount of secondary hosts including cats, dogs, cattle, horses, sheep, raccoons, skunks, goats, opossum and small rodents (Simmons 2010). Infection of these secondary hosts is often fatal (Fenner, Bachmann et al. 1987; Pensaert and Kluge 1989). According to Corn et al. (2004), among more than 15,000 feral swine of the United States through 1995, 28% were tested to be pseudorabies seropositive.

Swine brucellosis is another major disease that can be hosted by feral swine. As a chronic swine disease, its adverse effects include sterility and abortion in sows, orchitis in boars, and piglet mortality, and it is transmitted through reproductive discharges such as semen and afterbirth (Blood, Radostits et al. 1983; Fletcher, Creekmore et al. 1990; Stevens 1996). Compared to pseudorabies, a much more dreadful characteristic of swine brucellosis is that the latter can infect not only animal, but also humans, with possible symptoms ranging from severe flu-like symptoms to arthritis or meningitis (Stevens 1996). As of now, swine brucellosis has almost been eradicated from domestic swine in the United States, leaving feral swine the only remaining host reservoir (Fletcher, Creekmore et al. 1990; Olsen 2010). However, it is extremely difficult to eradicate swine brucellosis virus form feral swine because it is infeasible to test and remove infected feral swine individuals and there has not been an effective vaccine that can be delivered in oral baits and does not require capture of swine (Stevens 1996; Olsen 2010).

Other major diseases that can be harbored by feral swine include foot-and-mouth disease (FMD), African swine fever (ASF), and trichinosis (or trichinellosis). Some of these diseases are contagious to humans and therefore pose a threat to public health (Choquenot, McIlroy et al. 1996).

Economic Impacts

Feral swine cause serious damage to the local economy, especially agriculture. Known as formidable predators, they prey on a variety of livestock including lamb, goats, newborn cattle, etc. Although these livestock only make up a small portion of a feral swine’s diet, together they were of great importance to local economy. In addition, due to the facts that: 1) feral swine usually consume the prey thoroughly leaving little evidence; 2) they are not considered as a “common” predator by most people; and 3) the cause of predation is frequently misidentified (e.g., signs of feral swine are very similar to those of coyote), the damage of feral swine to livestock industry is often underestimated and the annual loss caused by feral swine predation is unable to estimate (Stevens 1996; Seward, VerCauteren et al. 2004). Plant et al. (1978) estimated that feral swine accounted for 32% of losses of newborn lambs in the semi-arid rangelands of Australia. Rollins (1993) reported that 1,243 sheep and goats were lost due to the predation of feral swine in Texas in 1990, with an estimated value of $63,000. And according to the National Agricultural Statistics Service (NASS) (1991), the total lost to feral swine in the United States in 1990 was estimated to be more than $1 million.

Feral swine also impose tremendous damage to various crops such as wheat, sorghum, barley, oilseeds, sugarcane, oats, and maize by direct consuming and trampling plants in order to form bedding or to gain access to the center of the crop (Choquenot, McIlroy et al. 1996). According to a conservative estimation by Pimentel (2002), feral swine cause a damage of $800 million in agricultural crops every year in the United States. In California, a total amount of $1,731,920 in crop damage caused by feral swine in 1998 was estimated (Frederick 1998).

Environmental Threats

In addition to disease and economic considerations, feral swine also are a considerable threat to the biological and physical environment. Being exotic, feral swine significantly and adversely impact the animals which are native to the United States in terms of at least two aspects: direct predation and indirect competition.

Due to their omnivorous diet composition, a large variety of resources can be utilized by feral swine for food, ranging from grasses, forbs, shoots, roots, tubers, fruit, and seeds to invertebrates including earthworms, leeches, grasshoppers, centipedes, beetles, and vertebrates including salamanders, frogs, fish and livestock. Their inefficient stomach also requires them to forage almost continuously in order to intake enough nutrients to sustain a high energy level (Engeman, Stevens et al. 2007). Ecol (1992) reported that in Australia, feral swine consume over 95% of the available worms at paired quadrat sites in lowland ephemeral swamps near Cape Tribulation during April - July 1992 (this is a crucial link in the ecosystems because worms decompose rotten plants and animals and therefore nutrients are returned back to soil). An experiment in Texas suggested that feral swine had a strong negative effect on bobwhite quail populations by accounting for 28% of predation of bobwhite quail (*Colinus virginianus*) (Tolleson, Rollins et al. 1993; Seward, VerCauteren et al. 2004).

Feral swine are aggressive competitors with native wildlife of the United States in terms of food, space and shelter (Stevens 1996). Potential competition for resources occurs between feral swine and animals including deer, turkey, waterfowl, squirrels, raccoons, opossums, foxes, bobcats, javelinas, bears, sandhill cranes, chipmunks, etc.

Other than their adverse impacts on animals, feral swine also directly disturb the environment, and this is largely attributed to their rooting behavior. Feral swine constantly root (i.e., dig up soil with their snouts) in order to find food beneath the soil surface. This behavior jeopardizes the environment in that it loosens the soil, changes soil properties, accelerates erosion, destroys vegetation, sets back plant succession, reduces earthworm activity, and exacerbates exotic plant invasion. Thus, feral swine are commonly considered as a major factor of habitat degradation (Stevens 1996; Mungall 2001; Seward, VerCauteren et al. 2004). Engeman, Smith et al. (2003) monitored the damage to native wet pine-flatwoods at three State Parks in Florida caused by feral swine, and the estimated cost of resource restoration was given. According to their estimation, the economic value of damage caused by feral swine at these three parks ranged from $5,331 to $43,257 per ha.

Efforts to Manage and Control Feral Swine

Due to the fact that feral swine cause extensive damage to the environment and their area of distribution is still spreading, a series of efforts have been taken to control the distribution of feral swine in the United States by different organization and institutions.

Shooting and trapping are two methods that have been most commonly used to control feral swine throughout the North America (Stevens 1996). In several states where feral swine are populated, various shooting regulations have been applied. According to the Georgia Department of Natural Resources, feral swine are considered an exotic species and there is no limit on shooting. They can be hunted any time of the year as long as the shooter owns a Georgia hunting license (Lang and Mengak 2007). Whereas in Texas and Oklahoma, hunters can shoot feral swine even without a hunting license if feral swine are damaging the landowner's property (Stevens 1996; Mapston 2004).

Trapping is a more effective way of removing feral swine than shooting because: 1) shooting will normally not reduce the feral swine population greatly; and 2) feral swine are so intelligent that they may adapt to the intensive hunting activities by shifting home range or becoming nocturnal (Mapston 2004). An extensive amount of trapping methods has been proposed. For example, Engeman, Constantin et al. (2001) implemented the passive tracking method to assess the abundance and distribution of feral swine in Florida as an important part of their swine control program. The indexes of pervasiveness and damage were used to assess the change of feral swine effects after a total of 25 feral swine were removed from the study area. Their study suggested that removing feral swine could indeed decrease the feral swine distribution, changing their normally clumped pattern of activity to a more isolated pattern, and reduce the amount of damage caused by feral swine. In addition, the passive tracking method proved to be an easy and relatively reliable source of gathering data about feral swine abundance and distribution.

Laws and regulations are another tool to control feral swine that is complementary to the common methods such as shooting and trapping. This is especially true in the southern states of the United States where feral swine are a considerable threat. For example, in Texas, a series of regulations on the movement of feral swine were promulgated by the Texas Animal Health Commission (TAHC). These regulations include prohibiting the movement of feral swine in Texas unless they are moved directly from the premises where they were trapped to a recognized slaughter facility, approved holding facility or authorized hunting preserve (Texas Animal Health Commission 2008).

In addition to these controlling methods, the distribution of feral swine has been actively monitored by researchers and managers. Prior to 2007, the spread of feral swine was tracked by means of compiling the hardcopy (for 1982 and 1988) or digital format (for 2004) maps of feral swine distribution reported by individual state resource managers and federal wildlife agencies (Tyler Campbell 2008; Madden, Zhao et al. 2009). However, this process was extremely time-consuming and costly. There was a great need for an efficient method of gathering information from various agencies and managers throughout the United States and maintaining an up-to-date feral swine distribution geodatabase that was accessible to the public (Tyler Campbell 2008; Madden, Zhao et al. 2009). Serving to meet these requirements, the National Feral Swine Mapping System (NFSMS) (http://128.192.20.53/nfsms/) was implemented in 2007. Maintained by SCWDS and based in the College of Veterinary Medicine at the University of Georgia, this web-based system aims at providing information on the distribution of feral swine across the United States. The data on which this system is based are provided by state and territorial natural resources agencies, U.S. Fish and Wildlife Service personnel and USDA, Animal and Plant Health Inspection Service (USDA-APHIS) (Tyler Campbell 2008; Madden, Zhao et al. 2009). Although being primarily maintained and edited by the Center for Remote Sensing and Mapping Science (CRMS), Department of Geography and the College of Veterinary Medicine, The University of Georgia , NFSMS allows the authorized users to add, edit or delete the polygons indicating feral swine distribution anywhere with internet access (Tyler Campbell 2008). This feature significantly accelerates the updating process, and therefore enables NFSMS to be able to be updated on a monthly basis. Figure 1 shows the NFSMS opening page with the current feral swine distribution in the United States as of February 11, 2011 (Madden, Zhao et al. 2009).

Objectives and Scientific Significances

Despite the fact that feral swine still cause a massive amount of economic damage to the United States today, and a huge amount of effort has been expended to monitor and control the distribution of feral swine in the United States, little has been done to understand preferences of feral swine in terms of habitat selection. Most examples in the literature point out that feral swine can live in almost any kind of habitat, with certain preferences favoring adequate water and thick vegetation. This vague understanding is not sufficient for an effective and efficient management of feral swine at a small scale (e.g., at the state or national level) especially when the objective is to predict the future expansion of their habitat.

Since there is still a practical need to further understand the relationship between habitat suitability and feral swine distribution at the broad state-level scale in order to better manage the spread of this destructive exotic species, this study attempts to approach this with five specific objectives:

1) Compile a geodatabase of environmental factors that may be significantly related to the distribution of feral swine from various datasets based on remote sensing within a geographic information system (GIS);

2) Identify and quantify the spatial relationships among these environmental factors and the established distribution of feral swine using logistic regression;

3) Incorporate spatial autocorrelation into the ordinary logistic regression model by means of autologistic regression modeling;

4) Evaluate whether the incorporation of spatial autocorrelation improves the logistic model outcome;

5) Assess the influence of scale on the performance of the habitat models.

Despite the fact that the current study is not the first attempt to link feral swine habitat preference to environmental factors, this is the first attempt to examine this species-environment relationship from a broad scale and relatively comprehensive perspective. There are few examples of similar previous studies, with the exception of the model constructed by Gaines et al. (2005). They used the logistic regression analysis method as a pivotal part of their research on an ecological risk assessment using feral swine as an indicator. Logistic regression models were built during their model development phase to determine the relationship between feral swine abundance and two environmental factors: habitat classification and landscape metrics. The abundance data of feral swine were obtained through the recorded number of kills of feral swine per area unit. Unlike the previous study by Gaines et al. (2005), which used the habitat classification dataset as their parimary data input, the data input of this study consists of various geographic datasets related to food, shelter, water and social requirements of the swine, as well as human influences on their distribution.

Moreover, this study tries to examine the relationship between feral swine distribution and environmental and cultural factors at multiple spatial scales. Scale is a major concept in landscape ecology. At different spatial scales, a single phenomenon can be interpreted in completely different ways. In order to yield the legitimate and reasonable results, a proper scale should be used (Wu and Hobbs 2007). There has been no consensus on the exact home range of feral swine (which is a key factor determining the proper scale to use when modeling feral swine distribution). According to Stevens (1996), the home range of feral swine can vary from 1 to 49 square kilometers (0.4 to 19 square miles). In addition, feral swine home range can be greatly subject to sex and the environmental conditions (e.g. food abundance) (Singer, Otto et al. 1981). Given these facts, this study explicitly incorporates multiple spatial scales into the model building process and compares their difference in terms of model outcomes.

Finally, this study takes into account the spatially dependent relationship between the feral swine distribution and the environmental factors both implicitly and explicitly. Distance to certain key environmental and cultural factors such as major streams and roads were incorporated into this study to evaluate the effect of both positive and negative proximity of these factors on the distribution of feral swine. Spatial autocorrelation was also examined and then explicitly employed to modify the ordinary modeling method in order to avoid committing errors such as assumption violation, and to yield more accurate results.

Study Area

Since the focus of this study is to investigate the distribution-habitat relationship for feral swine over a relatively broad scale, several criteria were considered when selecting the proper study area. Criteria included data availability and presence of feral swine presumed to result primarily from natural influences rather than translocation by human transportation and release. Given these criteria, the Coastal Plain Province of Georgia was selected as the study area, as shown in Figure 2. The Coastal Plain Province is one of four distinct physiographic regions of Georgia (the other three regions are the Valley and Ridge, The Blue Ridge and the Piedmont)

(Hodler and Schretter 1986; UGA Department of Geology 2011). To the south of the Fall Line, which is a boundary of bedrock geology separating the Piedmont Province and the Coastal Plain Province, the Coastal Plain Province occupies a territory of approximately 97,000 kilometers and takes up more than half of the total land area of Georgia. This region mainly consists of alternating layers of sand, clay, and limestone, with kaolin being the most economically significant mineral resource. Terrain elevation within this region ranges from sea level to 182 meters (Hodler and Schretter 1986; Georgia State Climate Office 1998; UGA Department of Geology 2011).

The Coastal Plain region is rich in water supply. One of the geological features of the Coastal Plain region compared to the other three regions of Georgia is the abundance in groundwater due to the water-rich Floridian aquifer system lying beneath this region. The Floridian aquifer provides approximately 50% of the groundwater for domestic consumption, industry and agricultural irrigation in Georgia. Major rivers in this region include the Altamaha, Flint, Ocmulgee, Oconee, Ogeechee, and Savannah Rivers (Hodler and Schretter 1986; UGA Department of Geology 2011).

In terms of climate, the Coastal Plain Province is mostly similar to the rest of Georgia with higher temperatures and fewer days below freezing. Temperatures are warm throughout this region, with the southern portion slightly higher than the northern part. Savannah, one of the major cities in this region, only experiences an average of 26 days of low temperatures of 0 degrees (Celsius) or below per year (Georgia State Climate Office 1998). Another city, Brunswick, which is south of Savannah, experiences only no more than 11 days of low temperatures of 0 degrees or below per year (Georgia State Climate Office 1998). The average precipitation of the Coastal Plain Province is approximately 1,143 millimeters (45 inches).

Similar to the temperature pattern, the southern part of the region has slightly more precipitation than the northern, with the number of rainy days being relatively constant throughout the entire region (121 days on average) (Georgia State Climate Office 1998).

The longleaf/slash pine forest is the dominant forest type for most of the area in the Coastal Plain region. Close to the Fall Line, loblolly/shortleaf pine forest extended from the Piedmont Province is dominant over the longleaf/slash pine forest. Oak/gum/cypress forest is mostly found dominant over other forest types in areas with abundant water supply such as river corridors and swamps. On the coast line, salt marsh is the dominant vegetation type with dense stands of salt tolerant herbs, grasses and shrubs. Common tree species in this region include white oak (*Quercus alba*), turkey oak (*Quercus laevis*), live oak (*Quercus virginiana*), loblolly pine (*Pinus taeda*), longleaf pine (*Pinus palustris*), shortleaf pine (*Pinus echinata*), slash pine (*Pinus elliottii*), bald cypress (*Taxodium distichum*), southern magnolia (*Magnolia grandiflora*) and dwarf palmetto (*Sabal minor*) (Hodler and Schretter 1986).

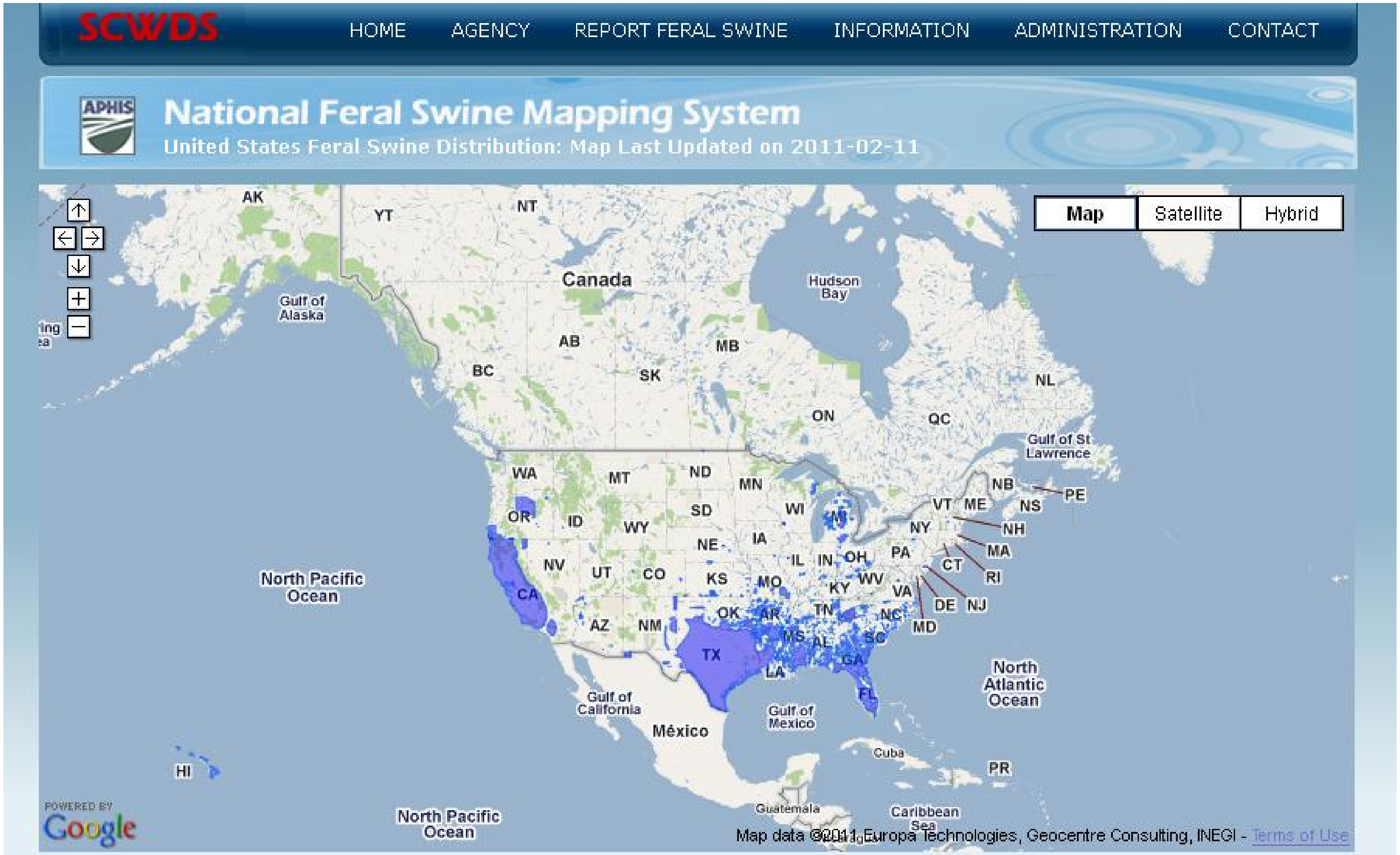


Figure 1: National Feral Swine Mapping System (NFSMS) opening page showing the current feral swine distribution in the United States (http://128.192.20.53/nfsms/). Blue areas indicate feral swine distribution.

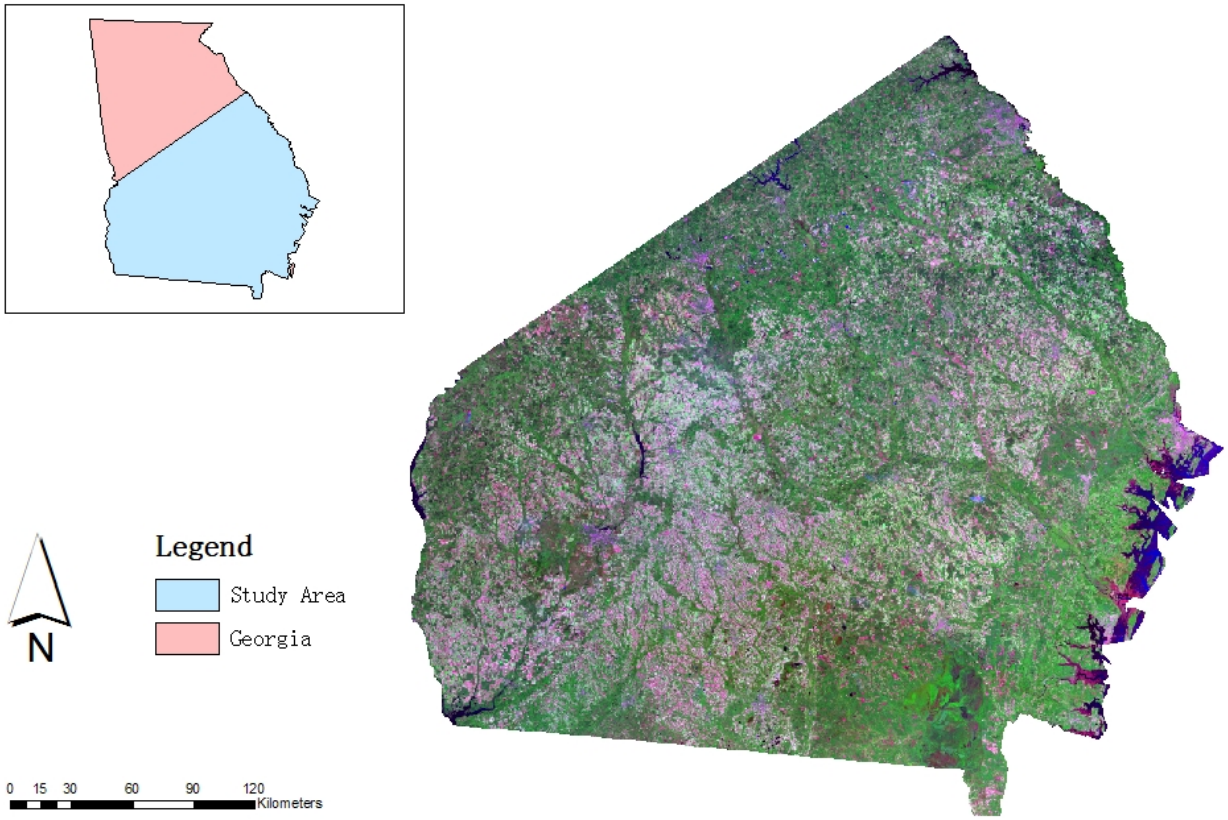


Figure 2: The study area, depicted as a mosaic of Landsat Thematic Mapper (TM) images from 1999 to 2000.

CHAPTER 2

LITERATURE REVIEW

The study presented here can be considered an application of landscape ecology theories utilizing several geospatial methods including geographic information systems, remote sensing, habitat modeling and logistic regression analysis. This chapter serves to introduce and illustrate these theories and methods.

Geographic Information System

A geographic information system (GIS) is a computer-based set of tools for capturing (collecting), storing, retrieving at will, transforming and displaying spatial data from the real world for a particular set of purposes (Burrough and McDonnell 1998). Composed of hardware, software and geospatial data, GIS is used to manipulate and operate on standard geographical primitives, such as points, lines and areas, and/or continuously varying surfaces known as grids or raster images (Bailey and Gatrell 1996; Burrough and McDonnell 1998).

The ability to process, catalog, map and analyze spatial data is the unique property of GIS which makes it different from other software programs (Wing and Bettinger 2008). Before the establishment of the first true operational GIS, spatial data were manipulated for quite a long time in hardcopy map format. Overlay analysis, which is an essential data analysis method of GIS, has been employed through manual techniques for over 200 years by printing maps at the same scale and physically overlaying them on a backlit surface (Bernhardsen 2002; Wing and Bettinger 2008). Examples of early development of GIS technology include cartographer Louis-Alexandre Berthier using overlaid maps to analyze troop movements during the American

Revolution, creation of early mapping programs in the United States such as IMGRID, CAM and SYMAP, and the establishment of the world's first GIS database called "World Data Bank" by the US Central Intelligence Agency (CIA) (Clarke 2000; Ghilani and Wolf 2002; Wing and Bettinger 2008).

In the 1960s, the first true operational GIS, Canada Geographic Information System (CGIS), was created by the Canadian government under the guidance of Roger Tomlinson in order to quantify existing and potential land uses in Canada (Bernhardsen 2002; Wing and Bettinger 2008). Other than this, the development of the Land Use and Natural Resource Inventory System (LUNR) in 1967 and the Minnesota Land Management System (MLMIS) in 1969 also are considered as the landmark events during the history of GIS development (Foresman 1998; Wing and Bettinger 2008).

With the ever-growing spatial awareness that has accompanied advances in computer technology and increased availability of geospatial data, the huge potential of GIS has been realized and utilized by people from various fields, such as health care managers, transportation authorities, forestry companies, national park authorities, etc. (Longley, Goodchild et al. 2005). Since GIS is a collection of a massive amount of techniques and software, operation of GIS requires a comprehensive understanding of a wide range of skills and knowledge, such as statistics, photogrammetry, remote sensing, programming, database design, etc. Among all the applications of GIS, manipulation of spatial data so that information hidden in the data layers can be revealed and analyzed is one of the most important one.

In this particular project, the role which GIS played was multifold. First, all the datasets that were obtained were manipulated and processed using ArcGIS (ESRI 2009) which is a prevalent GIS software. The purpose for this step was to make sure: 1) the information within each dataset could be readily extracted in the later step, and 2) all the data layers were compatible with each other in terms of projection. Then, all the layers were displayed and overlaid with the same geographic framework or GIS geodatabase. Finally, values from the various data layers spatially coincident with the specified locations of randomly selected points were extracted using a series of GIS spatial analysis methods. These data were the basis on which the statistical analysis could be carried out later.

Remote Sensing

Remote sensing is the "science and art of obtaining information about an object, area, or phenomenon through the analysis of data acquired by a device that is not in contact with the object, area, or phenomenon under investigation" (Lillesand, Kiefer et al. 2007). It is a powerful technology and has been widely used for acquisition of information (in the form of electromagnetic energy) of objects or areas on the Earth and other planetary surfaces. Different portions of the electromagnetic spectrum can be detected and utilized by various types of sensors, such as visible, infrared (IR) and microwave, each possessing unique properties. For example, the sensors that can detect the visible portion of the spectrum can create images that are convenient for manual image interpretation in that they are similar to views seen by the human eyes. The IR sensors are sensitive to infrared radiation, which is extremely helpful in vegetation monitoring because healthy green vegetation reflects strongly in the near IR spectrum. The microwave portion of the electromagnetic spectrum is used by the radio detection and ranging (radar) sensors. Due to the capability that microwave can penetrate the atmosphere under virtually all conditions, radar sensors can function well even when the ground objects of interest are covered by clouds (Lillesand, Kiefer et al. 2007).

Satellite remote sensing systems, or space remote sensing systems, refer to those systems that are mounted on satellite platforms. Although it can be dated back to 1940s, when small cameras were brought aboard by a series of rockets and satellites launched at that time, it was not until the creation of the Landsat (NASA, NOAA) program in 1967 that satellite remote sensing has shown the full potential of remote sensing in terms of data acquisition (Goward and Williams 1997). Originally named Earth Resources Technology Satellite 1 (ERTS-1), the first Landsat satellite was launched on July 23, 1972. Since then, seven Landsat satellites have been sent into space. Despite the failure of Landsat-6 to launch and the fact that the current Landsat-5 and Landsat-7 are old and suffering from failures to some degree, together they have provided a huge amount of extremely valuable data for the Earth surface. The images produced by Landsat satellites have been used in a wide variety of fields such as agricultural management, climate research, civil engineering, environmental monitoring, natural resource management, public safety, homeland security, land use planning, etc. (Behrens 2010). Table 1 shows the technical details of all the Landsat projects, and Table 2 shows the specifications of the sensors on board the Landsat satellites (Global Land Cover Facility 2010; U.S. Geological Survey 2010; NASA 2011). In addition to the Landsat program, some other major satellite remote sensing programs include MODIS (Moderate-Resolution Imaging Spectroradiometer, NASA), ASTER (Advanced Spaceborne Thermal Emission and Reflection Radiometer, NASA & ERSDAC), SPOT (Satellite Pour l'Observation de la Terre, Spot Image, France), China-Brazil Earth Resources Satellite (CBERS), etc.

Table 1: Technical details of the Landsat projects (Global Land Cover Facility 2010; U.S. Geological Survey 2010; NASA 2011).

<table>
<tr><th>Project</th><th>Launch Date</th><th>Sensors*</th><th>Altitude</th><th>Inclination</th><th>Period of Revolution</th><th>Repeat Coverage</th><th>Status</th></tr>
<tr><td>Landsat 1</td><td>July 23, 1972</td><td rowspan="3">RBV, MSS</td><td rowspan="3">nominally 900 km</td><td rowspan="3">99.2°</td><td rowspan="3">103 minutes; ~14 orbits/day</td><td rowspan="3">18 days</td><td>expired on January 6, 1978</td></tr>
<tr><td>Landsat 2</td><td>January 22, 1975</td><td>expired on February 5, 1982</td></tr>
<tr><td>Landsat 3</td><td>March 5, 1978</td><td>expired on March 31, 1983</td></tr>
<tr><td>Landsat 4</td><td>July 16, 1982</td><td rowspan="2">TM, MSS</td><td rowspan="4">705 km</td><td rowspan="4">98.2°</td><td rowspan="4">99 minutes; ~14.5 orbits/day</td><td rowspan="4">16 days</td><td>decommissioned on June 15, 2001</td></tr>
<tr><td>Landsat 5</td><td>March 1, 1984</td><td>TM still operational, MSS instrument decommissioned</td></tr>
<tr><td>Landsat 6</td><td>October 5, 1993</td><td>ETM</td><td>lost at launch</td></tr>
<tr><td>Landsat 7</td><td>April 15, 1999</td><td>ETM+</td><td>operational despite Scan Line Corrector (SLC) failure on May 31, 2003</td></tr>
</table>

Table 2: Technical specifications of the sensors on board the Landsat satellites (Global Land Cover Facility 2010; U.S. Geological Survey 2010; NASA 2011).

Sensor	Spatial Resolution	Spectral Range	Number of Bands	Image Size
RBV	80 m, 40 m (Landsat 3 only)	0.5-0.75 µm	3	185 x 185 km
MSS	57 x 79 m	0.5 - 1.1 µm	4, 5 (Landsat 3 only)	185 x 185 km
TM	30 m (120 m thermal)	0.45 - 12.5 µm	7	185 x 172 km
ETM	30 m (120 m thermal, 15 m pan)	0.45 - 12.5 µm	8	184 x 185.2 km
ETM+	30 m (120 m thermal, 15 m pan)	0.45 - 12.5 µm	8	183 x 170 km

* RBV: Return Beam Vidicon System, MSS: Multispectral Scanner System, TM: Thematic Mapper, ETM: Enhanced Thematic Mapper, ETM+: Enhanced Thematic Mapper Plus

With the development of satellite remote sensing, its ever-increasing capability has led to the fact that satellite remote sensing has been organically embedded into various environmental studies and it is a quite unlikely scenario that a study dealing with the environmental problems does not utilize any product of remote sensing as data input at all. Various applications of remote sensing have been proposed by a vast amount of researchers. For instance, Fuller, Groom et al. (1994) created the Land Cover Map of Great Britain based on the Landsat Thematic Mapper (TM) data using supervised maximum-likelihood classification method. The map consisted of 25 land cover types, including 18 types of semi-natural vegetation. Zhang, Friedl et al. (2003) proposed a method using MODIS data to monitor vegetation phenology for the New England region, United States. Based on the vegetation index (VI) portion of MODIS data, the transition dates of vegetation activity can be determined without the processes such as pre-smoothing and manual threshold identification, and the results achieved was satisfactory.

Other than the fact that a growing number of remote sensing datasets are available, another phenomenon, which is employing multiple datasets from various sources, becomes quite prevalent nowadays. A couple of reasons lead to this phenomenon, such as the data availability issue. For a certain study area, it might be difficult to obtain a certain kind of remote sensing dataset with adequate quantity and quality, under such scenario, it might be more practical to use the remote sensing images from other data sources as complementary inputs. Oliveira, Kampel et al. (2008), for example, examined the geomorphologic evolution of the coastal region of the Restinga of Marambaia, Brazil between 1975 and 2004. Multiple remote sensing images were employed as data inputs ranging from Landsat MSS, Landsat TM, Land ETM+ to CBERS-2 images, due to the fact that there was no single dataset can satisfactorily cover the study area during the period of interest. Another scenario when integration of remote sensing images from multiple sources is favorable is due to the fact that different datasets have their own limitation and strengths, thus by employing datasets of different kinds, the information which can be extracted from the images can be maximized. A striking example would be the application of an image fusion method called pan-sharpening. Pan-sharpening is a technique to merge multispectral images (high spectral resolution, low spatial resolution) with panchromatic images (high spatial resolution, low spectral resolution) to obtain images with both high spatial resolution and spectral resolution, and it is one of the commonly used techniques in image interpretation (Thomas, Ranchin et al. 2008; Guo, Chen et al. 2010).

Remote sensing technology plays a pivotal role in this project in that it provided datasets on which the final statistical analyses could be based. All the datasets that were used in this study were products of a diversity of remote sensing projects. For instance, the Normalized Difference Vegetation Index (NDVI) is a secondary product derived from the Landsat TM, ETM+ and MODIS images. And the Land Cover datasets were also developed based on the Landsat TM and ETM+ images through image interpretation and classification.

Landscape Ecology, Scale, and Habitat Modeling

Landscapes are areas that are heterogeneous in at least one factor of interest, and they are comprised of sets of ecosystems that are interacting with each other and repeated in a similar fashion in space (Forman and Godron 1986; Forman 1995; Turner, Gardner et al. 2001; Bissonette and Storch 2003; Burel and Baudry 2003). Due to its nature of heterogeneity, various patterns emerge within a single landscape or through the interaction between multiple landscapes. Landscape ecology emerged to study these patterns, and it is an individual interdisciplinary field focusing on studying landscape structure, function, and change (Liu and Taylor 2002; Burel and

Baudry 2003). As a term, landscape ecology was first introduced by German biogeographer Carl Troll in 1939, and it was derived from the traditional concepts of regional geography and vegetation in Europe and greatly influenced by impetus such as the invention of aerial photography (Turner, Gardner et al. 2001). Turner, Gardner et al. (2001) summarized two features of landscape ecology that make it unique from other ecology subdisciplines: 1) landscape ecology explicitly addresses the importance of spatial configuration for ecological processes; and 2) landscape ecology often focuses on spatial extents that are much larger than those traditionally studied in ecology. Since its recognition as a subdiscipline of ecology, a set of issues as well as topics were proposed to be considered and dealt with in landscape ecology. Table 3 shows six key issues within the domain of landscape ecology and ten key research areas to deal with these issues, identified by Wu and Hobbs (2002).

Almost all literature reviews concerning landscape ecology consider scaling as one of the key topics in this field, and it is such an fundamental concept that it is a basic concern for just about every landscape ecological study and still a question needs to be addressed (Turner, Gardner et al. 2001; Wu and Hobbs 2002; Bissonette and Storch 2003; Burel and Baudry 2003; Wu and Hobbs 2007). There is spatial pattern in the landscape at all practical scales (Bissonette and Storch 2003). A phenomenon that is significant at a particular scale may be inconsiderable at another scale, and thus totally different interpretations will be yielded. The importance of scale theory for ecological study can be understood in two aspects. Firstly, the data used for landscape ecology study is inextricably related to spatial resolutions, i.e. scale. For example, most Landsat Thematic Mapper (TM) images, a satellite imagery product commonly used for a variety of ecological studies, have 30 meter spatial resolution, and the resolution of airborne light detecting and ranging (LiDAR) data can be as small as one millimeter (Wadsworth and Treweek 1999;

Table 3: Key issues and key research areas of landscape ecology (Wu and Hobbs 2002).

Key Issues	Interdisciplinarity or transdisciplinarity
	Integration between basic research and applications
	Conceptual and theoretical development
	Education and training
	International scholarly communication and collaborations
	Outreach and communication with the public and decision makers
Key Research Areas	Ecological flows in landscape mosaics
	Causes, processes, and consequences of land use and land cover change
	Nonlinear dynamics and landscape complexity
	Scaling
	Methodological development
	Relating landscape metrics to ecological processes
	Integrating human and their activities into landscape ecology
	Optimization of landscape pattern
	Landscape conversation and sustainability
	Data acquisition and accuracy assessment

Bissonette and Storch 2003). This nature of spatial data often triggers the Modifiable Areal Unit Problem (MAUP), which results from the improper use of arbitrary areal units when data used are area-based or raster-based (Bailey and Gatrell 1995; Wu and Hobbs 2007). Secondly, a landscape phenomenon typically occurs only at a certain spatial scale; interpretation at an improper scale may produce inaccurate inference about the reality. One proper example would be the existence of habitat fragmentation. At a fine scale and small extent, this phenomenon may not be detectable (e.g. within one homogeneous patch), however, it may be quite severe at broader scales. The ecological fallacy, which is a problem occurring when unwarranted inferences about a lower level are made based on knowledge from an upper level, is a typical consequence of improper selection of scales (Wu and Hobbs 2007).

Scale is not just a topic of geography and landscape ecology, but also a typical issue to address in geography and landscape ecology study. It took researchers a long time to realize that there is no single scale that is appropriate for the study of all ecological problems (Turner, Gardner et al. 2001). Owing to the fact that scale is usually closely related to the landscape pattern that emerges, which is the key subject of landscape ecology studies, finding the proper scale at which problems should be addressed specifically becomes crucial in order to yield robust results. However, this is often difficult to achieve in practice, as numerous factors need to be taken into account, for instance, the arrangement of communities in space and how they interact with heterogeneous patterns of resources on the landscape (Turner, Gardner et al. 2001). Despite this, efforts have been made to further understand the scale issue and search for the “right” scale.

Mitchell et al. (2001) used three different scales to model the distribution of several bird species in a managed forest in South Carolina. The three scales are microhabitat scale, landscape scale and a scale that combined these two. Their study suggested that no single scale was appropriate for all the bird species, and in general, the third combined scale was found more likely to produce better fits. A study conducted by O'Neill et al. (1996) examined the sensitivity of the landscape indices (e.g. dominance, contagion, etc.) to the resolution (or grain) of the remote sensing data and calculation scale (the area unit over which the indices are calculated) based on statistical analysis. Through this study, they recommended that the resolution of the data used should be 2 to 5 times smaller than the smallest feature in the area of interest; meanwhile, the calculation scale should be 2 to 5 times larger than the largest feature in the area of interest. In addition, they also suggested that the commonly used landscape indices were not sensitive to fine scale landscape patterns and thus were limited in detecting and capturing the changes of interest at fine scales. Recently Thornton, Branch et al. (2011) examined a total of 122 studies on the response (in terms of distribution, abundance and density) of a wide range of species to the influence of various variables (i.e. taxonomic, life history, and methodological variables) at multiple scales. The variables were divided into three categories: landscape scale, patch scale and within-patch scale based on the arbitrary criteria determined by the authors. According to their review, different species differed in their sensitivity to the influence of variables at different scales. For example, mammals were more sensitive to landscape scale variables than the variables of the other two scales. This difference of sensitivity, as they pointed out, could play an important role in environment conservation.

The interdisciplinary nature of landscape ecology enables it to incorporate a vast amount of methods from the other disciplines. These methods, including approaches and tools for data collection and analysis, have significantly improved the capability of landscape ecology to deal with various empirical ecological issues (Liu and Taylor 2002). Nowadays, GIS and remote sensing have become extremely valuable for landscape ecology, with GIS being a tool for storing, analyzing and manipulating data and remote sensing being a major data provider.

Among all the applications of landscape ecology, habitat modeling is no doubt one which carries huge scientific significance. Habitat, according to Morrison et al. (1998), is "an area with the combination of resources (e.g., food, cover, water) and environmental conditions (e.g., temperatures, precipitation, presence or absence of predators and competitors) that promotes occupancy by individuals of a given species (or population) and allows those individuals to survive and to reproduce." Residing in and nourished by natural elements, species are strongly subject to the influences of the environment. Because natural landscape systems are extremely complicated, usually different aspects are dynamic and stationary at the same time, the interaction between species and the environment they inhabit is therefore complex and results in diverse distribution patterns. The purpose of habitat modeling is to capture and quantify the relationship between species and the environmental elements that they rely on through mathematical and statistical methods. This species-environment relationship can further be used for multiple research and management purposes including predicting the occurrence of species, which is a primary goal for conservation biologists, identifying the most influential predictor of a certain species, or producing habitat quality or suitability maps for certain species (Guisan and Zimmermann 2000; Zaniewski, Lehmann et al. 2002; Bissonette and Storch 2003).

Maggini et al. (2002) modeled the distribution of narrow-headed ant (*Formica exsecta*) living within a Swiss National Park based on the environmental factors including slope, vegetation, solar radiation, etc. Their results showed that with only the spatially explicit variables included, the model accounted for nearly 53% of the null deviance, or adjusted D^2, which is equivalent to the residual sum of squares in multiple linear regression (Weisberg 2005). After adding the non-spatially explicit variables, the model could account for up to nearly 74% of deviance. A study conducted by Ambrosini, Bolzern et al. (2002) investigated the relationship between the distribution and abundance of barn swallows (*Hirundo rustica*) in north-western Europe using logistic regression. The results suggested that livestock farming affected the distribution and abundance of barn swallows. Specifically, there was a positive relationship between the number of farms with livestock and the distribution and abundance of barn swallows:

breeding of barn swallows was found in more than 90% in all the farms with livestock, while only 43.9% of the farms without livestock were occupied by barn swallows. Zaniewski et al. (2002) attempted to predict the distribution of 43 native fern species of New Zealand using both generalized additive models (GAM) and ecological niche factor analysis (ENFA) models. Different results were yielded, which according to Zaniewski et al. was attributed to the fact that ENFA predicts habitat suitability while GAM produces probability of presence. Among all the methods to model species distribution, each has its own applicable requirements and limitations. For this particular project, logistic regression was chosen rather than the other methods was primarily due to the fact that the feral swine distribution data that were acquired was presence/absence data, which can be readily dealt with by logistic regression.

Logistic Regression

Species distribution data are often binomial, i.e., only occurrence or presence/absence of the species is known. Due to this characteristic, logistic regression, which is a transformed case of generalized linear model (GLM) with a logistic link function (Augustin, Mugglestone et al. 1996; Ryan 1997), is considered to be a preferred method to study the distribution data of species. Among various modeling methods, it is logistic regression that allows binomial inputs which do not have a normal distribution (Wrigley 1985; Hosmer and Lemeshow 2000). And it is for this reason that logistic regression has been extensively used to model the distribution of species by relating their distribution with the environmental factors. Gibson et al (2004a) utilized logistic regression to predict the habitat of rufous bristlebird (*Dasyornis broadbenti*) by using a series of environmental variables including elevation, distance to creek, distance to coast and sun index, and their results suggested rufous bristlebird prefer habitats with low altitude, close distance to the coastal fringe and drainage lines and less direct sunlight. They also conducted another study to predict the presence of swamp antechinus (*Antechinus minimus maritimus*) in southwestern Victoria, Australia using logistic regression (Gibson, Wilson et al. 2004b). Their results suggested a negative correlation between the swamp antechinus habitat and altitude and vertical vegetation structure complexity. Based on this relationship, the final predictive performance of the selected model was higher than 90%. Lopez-Lopez et al. (2006) performed logistic regression to model the habitat preference of Bonelli's eagle (*Hieraaetus fasciatus*) in Castellon province, east of the Iberian Peninsula at four scales. Based on automatic stepwise selection, four models were established at different scale. At the 1×1 km^2 scale, only topographic variables were included in the final model. At the 3×3 km^2 scale, the model included climate and disturbance variables. At both 5×5 km^2 and 9×9 km^2 scales, topographic, climate, disturbance and land use variables were included. Based on the study, they suggested that Bonelli's eagle preferred scrublands, agricultural areas and disperse forests.

Spatial Autocorrelation

Closely tied to the "Tobler's First Law of Geography", i.e. closer features in space tend to share more similarities than those farther apart (Tobler 1970; Wong and Lee 2005), spatial autocorrelation is defined as "the correlation among values of a single variable strictly attributable to the proximity of those values in geographic space, introducing a deviation from the independent observations assumption of classical statistics" (Griffith 2003). With an increasing awareness of spatial autocorrelation in the past decades, the drawback of the ordinary logistic regression method in modeling species distribution is revealed. In other words, the assumption on which ordinary logistic regression is based, that each individual observation is statistically independent of each other (Legendre 1993; Keitt, Bjornstad et al. 2002; De Frutos, Olea et al. 2007) is violated by the fact that both environmental factors and species distribution are often spatially autocorrelated (Keitt, Bjornstad et al. 2002; Knapp, Matthews et al. 2003). This causes the importance of environmental variables to be overestimated (De Frutos, Olea et al. 2007). It is for this reason that an increasing number of studies have employed autologistic regression modeling instead of using ordinary logistic regression modeling only, as it is believed that autologistic regression can identify and remove weak predictors from the ordinary logistic regression model and yield better fits (Boyce and McDonald 1999; Jiang, Ma et al. 2009).

Augustin, Mugglestone et al. (1996), for example, compared the autologistic regression model with the ordinary logistic model by using red deer census data of the East and West Grampians, UK. The comparison suggested that autologistic regression models had better predictive capacity in that they had lower mean number of misclassified patches of each type and they were especially accurate in predicting absence. Based on these results, Augustin and Mugglestone et al. came to the conclusion that the autologistic regression model was superior for estimating the spatial distribution of the deer and should be used when mapping the spatial distribution of a species. Jiang, Ma et al. (2009) tried both logistic regression and autologistic regression to model moose resource selection in a study area in northeastern China at two different scales, patch scale and landscape scale. The results suggested that autologistic regression had improved overall true skill statistic (TSS), which is a measure of performance of the species distribution models recommended by Allouche et al. (2006), lower Akaike Information Criterion (AIC) and maximum data variability accounted for.

While research conducted to compare the performance of autologistic regression models and the ordinary logistic regression models support the conclusion that the former have a generally better performance than the later in terms of species distribution modeling when spatial autocorrelation does exist, there are also some researchers indicating that this is not the case all the time. For example, a comparison between autologistic regression model and ordinary logistic regression model was described by Betts, Diamond et al. (2006). In their study, the presence/absence data of a series of bird species were modeled using both methods and no significant difference in model outputs with or without taking into account spatial autocorrelation was observed.

Given the fact that there has been no definitive conclusion in terms of which of the two logistic regression modeling methods, ordinary logistic regression and autologistic regression, is superior in modeling species-environment relationship, this study will perform both of these modeling methods using existing GIS and remote sensing geospatial data. In this way, the optimal method for assessing the relationship between feral swine distribution in the Coastal Plain Region of Georgia and the environmental factors will be determined.

CHAPTER 3

DATASETS

The presence/absence data of feral swine used as the dependent variable in this study were based on the 2004 distribution map of feral swine in Georgia. Although at the time of this study three distribution maps of feral swine in Georgia of different dates were available, the 2004 dataset was selected because all the variables used in the modeling should be close in time so that the output models are legitimate. Most available corresponding datasets which were used as explanatory environmental and cultural variables in this research documented conditions around 2004. This presence/absence of feral swine dataset was compiled by the USDA Southeastern Cooperative Wildlife Disease Study (SCWDS), College of Veterinary Medicine, University of Georgia with cooperation by the Center for Remote Sensing and Mapping Science (CRMS), Department of Geography, University of Georgia to create a digital ArcGIS geodatabase. Feral swine distribution data were independently collected for each state by state and federal nature resource agencies. Now, updated on a monthly basis via the web-based National Feral Swine Mapping System (NFSMS), this dataset is the only natural dataset of feral swine distribution at the national level (Madden, Zhao et al. 2009). Figure 3 shows the feral swine distribution in the study area.

Due to the purpose of this study, which is to find the relationship between feral swine distribution and environmental factors, several important environmental and cultural factors were derived as the independent variables using datasets from various sources. These datasets were selected based on: 1) data availability; and 2) possible association of feral swine distribution to the resulting variables. According to these two criteria, a total of eight datasets were used for this study. All of them were clipped to the outline of the study area.

1) Land Cover of Georgia in 2005

The land cover referred to in this study is a collective concept including both land cover and land use. Land cover refers to the feature type existing on the Earth surface, and land use relates to the human activity associated with a specific piece of land (Lillesand, Kiefer et al. 2007). Being the most prominent characteristic of the Earth surface, land cover greatly influences various processes and activities occurring on Earth, including most ecological processes (Li 2008). With the advent of the Landsat remote sensing satellite program, and the invention of a standardized land cover/land use classification system (i.e. Anderson Classification System), it became possible to create consistent and comprehensive datasets such as the National Land Cover Dataset (NLCD). The land cover datasets have become one of the most commonly used datasets that have been used in ecological studies (Anderson, Hardy et al. 1976; Li 2008).

The 2005 land cover of Georgia dataset (Figure 4) was used in this study and it was created as one of the products of the Georgia Land Use Trends Project (GLUT) which is conducted and managed by the Natural Resources Spatial Analysis Laboratory (NARSAL, http://narsal.ecology.uga.edu/), College of Agricultural and Environmental Sciences, University of Georgia and the Georgia Cooperative Fish and Wildlife Research Unit at the Warnell School of Forest Resources, University of Georgia (Natural Resources Spatial Analysis Lab 2011). This land cover dataset has a spatial resolution of 30 meters, and was derived from either Landsat TM or Landsat ETM+ (National Aeronautics and Space Administration, NASA) images. The original land cover dataset has 13 classes, as shown in Table 4, and was downloaded from the Georgia

GIS Clearinghouse (http://data.georgiaspatial.org/) (Natural Resources Spatial Analysis Lab 2011).

2) Digital Elevation Model

The Digital Elevation Model (DEM) dataset used in this study was produced by the U.S. Geological Survey (USGS) as part of the National Elevation Dataset (NED) (Gesch 2002; Gesch 2007) with 30-meter spatial resolution (Figure 5). The NED is a seamless dataset integrating the best available DEM obtained from multiple sources across the United States and is updated bimonthly (USGS Seamless Data Warehouse 2010). The DEM used in this project was downloaded from the USGS Seamless Data Warehouse (http://seamless.usgs.gov). While there are other DEM datasets with higher spatial resolution, such as the DEMs with 10-meter or 3-meter resolution, this 30-meter DEM is appropriate for this study because the minimum distance between the sample points which provided observation values in the later modeling stage is considerably larger than 30 meters, and therefore higher resolution is not necessary in this case.

3) Normalized Difference Vegetation Index

The Normalized Difference Vegetation Index (NDVI), developed by Rouse, Haas et al. (1974), is an index commonly used for assessing the growing condition of green vegetation. It is an important environmental factor that might be correlated with the feral swine habitat because vegetation is the most important part of the living environment of feral swine providing both food and shelter. The NDVI can be derived from optimal satellite image data such as Landsat

Table 4: Classification system of the 2005 Georgia Land Cover Dataset.

Code	Land Cover Type	Description
7	Beach/Dune/Mud	Open sand, sandbars, sand dunes, mud - natural environmentals as well as exposed sand from dredging and other activities.
11	Open Water	Lakes, rivers, ponds, ocean, industrial water, aquaculture which contained water at the time of image acquisition.
22	Low Intensity Urban	Single family dwellings, recreation, cemeteries, playing fields, campus-like institutions, parks, schools.
24	High Intensity Urban	Multi-family dwellings, commercial/industrial, prisons, speedways, junkyards, confined animal operations. Transportation, roads, railroads, airports and runways. Utility swaths.
31	Clearcut/Sparse	Recent clearcuts, sparse vegetation, and other early successional areas.
34	Quarries/Strip Mines/Rock Outcrop	Exposed rock and soil from industrial uses, gravel pits, landfills. Rock outcrops, mountain tops, barren land.
41	Deciduous Forest	Forest composed of at least 75% deciduous trees in the canopy, deciduous woodland.
42	Evergreen Forest	Evergreen forest, at least 75% evergreen trees, managed pine plantations, evergreen woodland.
43	Mixed Forest	Mixed deciduous/coniferous canopies, mixed woodland, natural vegetation within the fall line and coastal plain ecoregions, mixed shrub/scrub vegetation.
81	Row Crop/Pasture	Row crops, orchards, vineyards, groves, horticultural businesses. Pasture, non-tilled grasses.
91	Forested Wetland	Cypress gum, evergreen wetland, deciduous wetland, depressional wetlands, and shrub wetlands.
92	Non-Forested Salt/Brackish Wetland	Salt marsh, brackish.
93	Non-Forested Freshwater Wetland	Freshwater marsh.

TM, ETM+ or MODIS (NASA) images by taking the normalized ratio of near-infrared (NIR) and red bands using based on (Huete, Didan et al. 2002).

$$\mathrm{NDVI} = \frac{\rho_{\mathrm{NIR}} - \rho_{\mathrm{Red}}}{\rho_{\mathrm{NIR}} + \rho_{\mathrm{Red}}}$$

where ρ_{NIR} and ρ_{Red} are the surface bidirectional reflectance from the near-infrared and red bands, respectively. Higher NDVI values indicate healthier vegetation (Huete, Didan et al. 2002).

The NDVI dataset used in this study was downloaded from the Global Land Cover Facility (GLCF, http://www.landcover.org/) (Carroll, DiMiceli et al. 2004), and it was developed based on band 1 and 2 of the MODIS data acquired by the sensor onboard NASA's Terra satellite between July 27, 2004 and August 11, 2004 (Figure 6). The cell size for this dataset is 212 meters.

4) Major Streams in Georgia

Despite the fact that feral swine are capable of living in almost any habitat type, it is reported they prefer moist bottomlands and other areas with adequate water (Stevens 1996). Therefore sufficient water supply is a primary factor influencing the distribution of feral swine, and it is reasonable to take into account the distance to major streams in this study. This dataset was derived based on the map of major streams in Georgia (Figure 7), which was compiled by the USGS and can be downloaded from National Atlas of the United States (http://nationalatlas.gov/).

5) National Overview Road Metric Euclidean Distance

In order to obtain sufficient food, feral swine may change their feeding place often (Choquenot, McIlroy et al. 1996), and roads may serve as barriers blocking the movement of feral swine. In addition, the noise brought by the vehicles on roads may be a negative environmental impact that feral swine tend to avoid. Thus, the distance to major roads was taken into account in this project. The National Overview Road Metric Euclidean Distance (NORM ED) dataset was used to represent this environmental factor. Developed by USGS as one of the products of The Road Indicator Project (TRIP), NORM ED dataset describes the extent and configuration of the spaces between roads in the United States (USGS 2010). The metric of NORM ED is Euclidean distance to the nearest road, and the NORM ED value at any point estimates the largest radius of a circle, centered at that point, that contains no roads (USGS 2010). The original dataset was downloaded from the TRIP website (http://rmgsc.cr.usgs.gov/trip/) and has 30-meter spatial resolution (Figure 8).

6) Presence of Oak-gum-cypress Forest

Feral swine are opportunistic predators (Simmons 2010). Despite their omnivorous nature, a variety of studies suggest feral swine distribution is closely related to availability of acorns especially in winters, which are produced by oaks (Stevens 1996; Linzey 2008). Although Oak-hickory forests are prevalent in the Piedmont area of Georgia, oaks in the Coastal Plain Region of Georgia are concentrated in the Oak-gum-cypress forests in the bottomland hardwood forests in river floodplains.

A Presence of Oak-gum-cypress Forest dataset (Figure 9) was developed based on the Oak-gum-cypress class in the 2002 Forest Cover Types of Georgia dataset produced by the USDA Forest Service and USGS. This dataset has 25 classes and was interpreted from the Advanced Very High Resolution Radiometer (AVHRR) and Landsat TM images with 1000-meter spatial resolution. Field observation data and Digital elevation models were used to refine the final product (USDA Forest Service and USGS 2002). It is distributed through the National Atlas of the United States (http://nationalatlas.gov/) in GeoTIFF format.

7) Impervious Surface Cover in 2005

Impervious surfaces found in developed landscapes are anthropogenic materials including rooftops, driveways, sidewalks and other materials that prevent water from infiltrating into the ground (Slonecker, Jennings et al. 2001; Mountrakis and Luo 2010), thus they are good indicators of urban areas. Due to various reasons (e.g. disturbance of human activities, loss of vegetation, highly fragmented landscape, etc.), feral swine may tend to avoid living too close to the urban area. The Impervious Surface Cover of Georgia in 2005 dataset was used in this project and it was produced by the Natural Resources Spatial Analysis Laboratory (NARSAL), College of Agricultural and Environmental Sciences, University of Georgia and the Georgia Cooperative Fish and Wildlife Research Unit at the Warnell School of Forestry and Natural Resources, University of Georgia. It represents the percentage of each cell that is made up of impervious surface with twenty classes in 5% increments (Natural Resources Spatial Analysis Lab 2011). Similar to the 2005 Land Cover of Georgia, the 2005 Impervious Surface Cover of Georgia dataset was also one product of GLUT. Figure 10 shows the impervious surface cover map of Georgia in 2005.

8) Canopy Cover in 2005

A variety of studies report feral swine generally prefer areas with thick vegetation as these areas provide protection and sufficient food (Stevens 1996; Lang and Mengak 2007). The Canopy Cover of Georgia in 2005 dataset was used and it was produced by NARSAL. These data represent the percentage of each 30-meter cell that is made up of tree canopy with twenty classes of 5% increments, as shown in Figure 11.

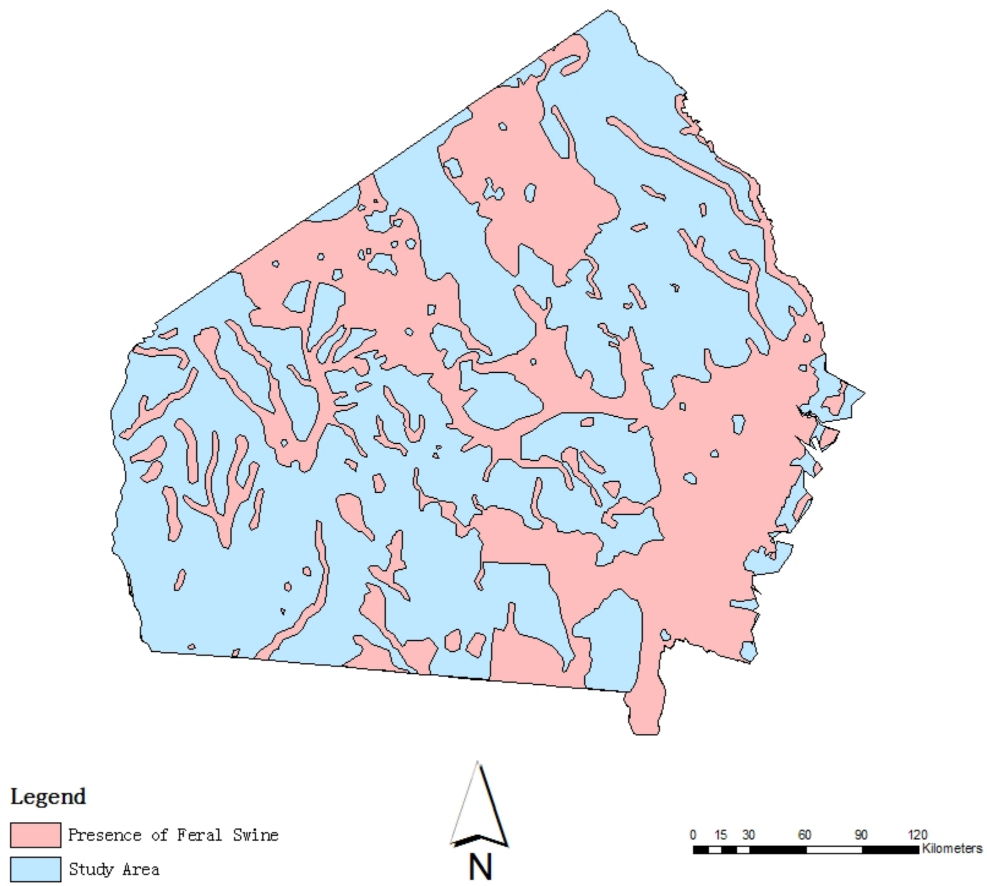


Figure 3: Feral swine distribution in the study area in 2004. Red indicates areas where feral swine were distributed.

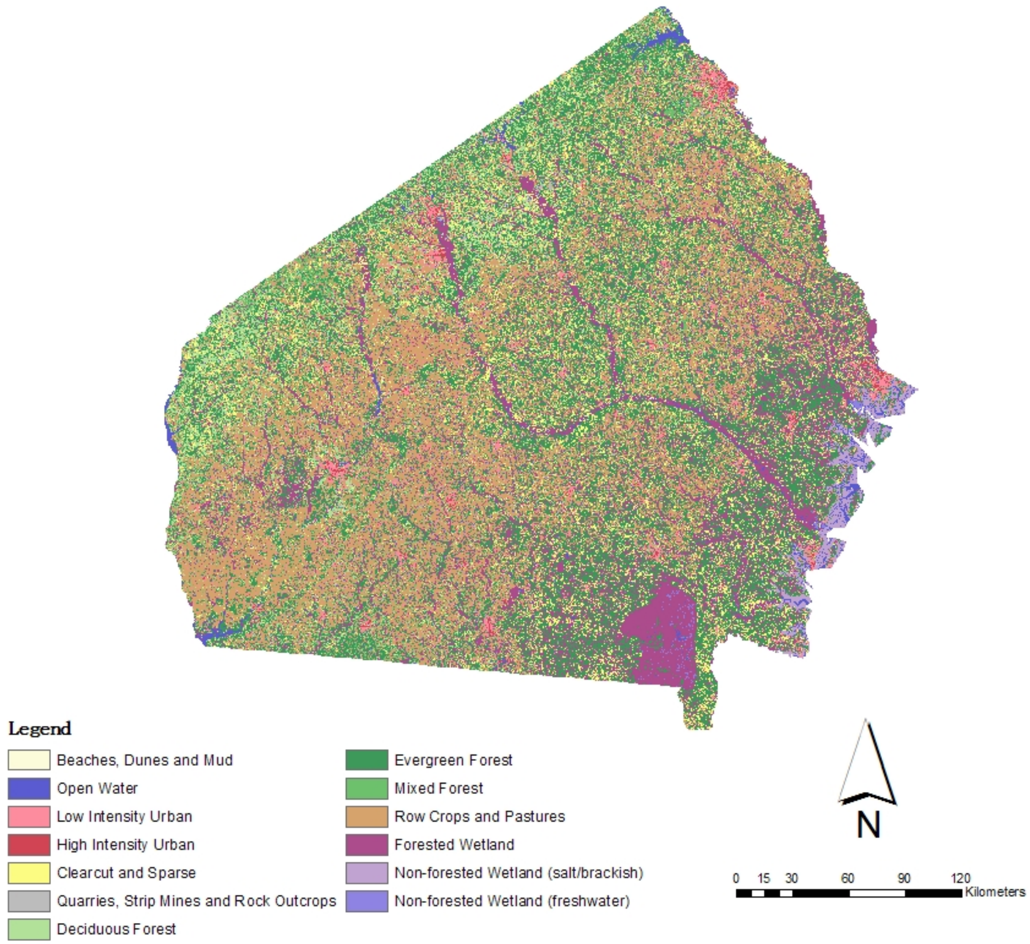


Figure 4: Land cover of the study area in 2005.

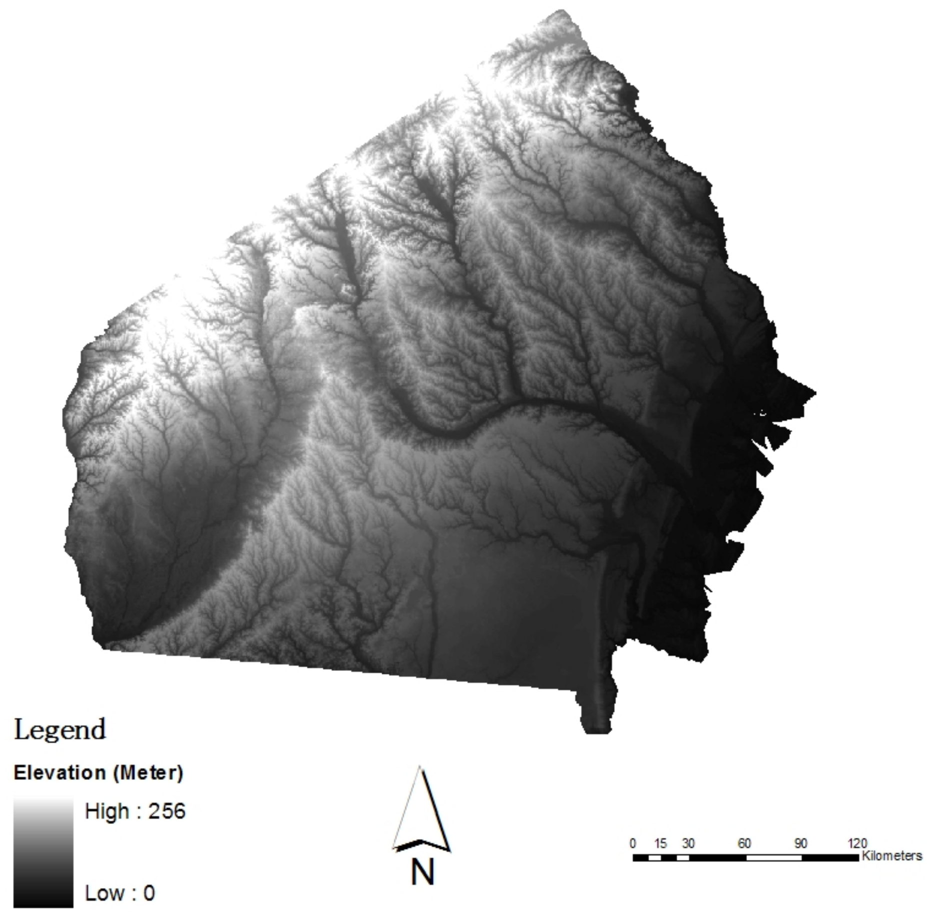


Figure 5: Digital Elevation Model of the study area.

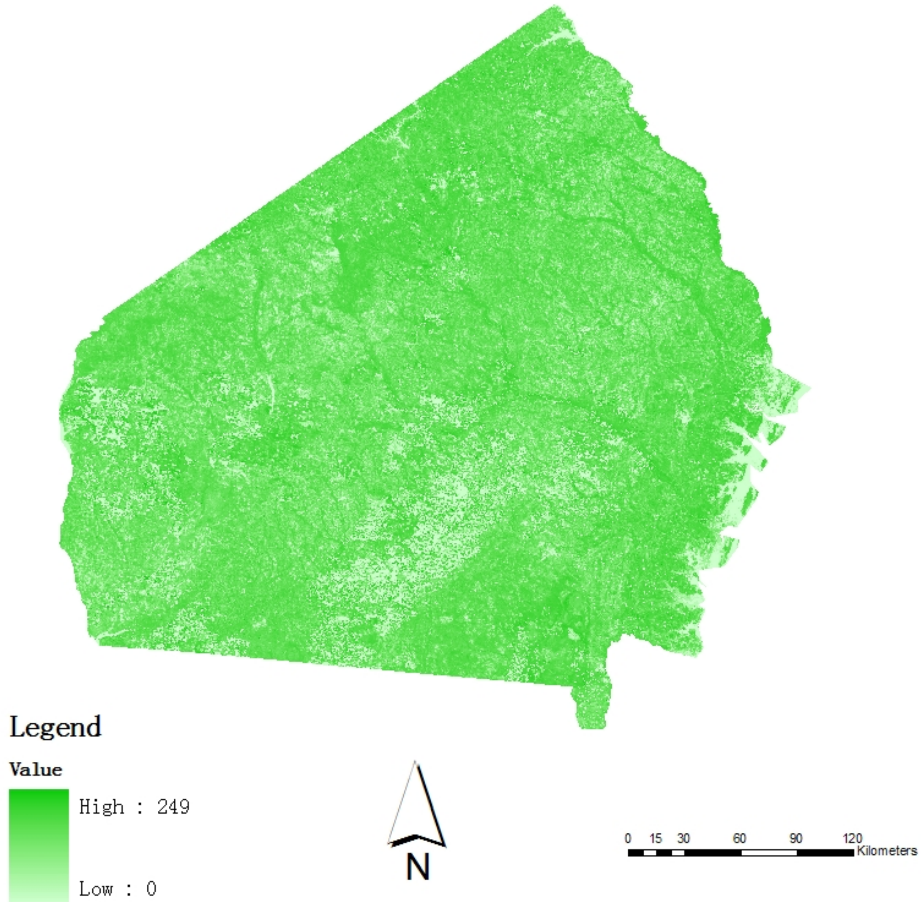


Figure 6: Normalized Difference Vegetation Index (NDVI) dataset of the study area during summer in 2004.

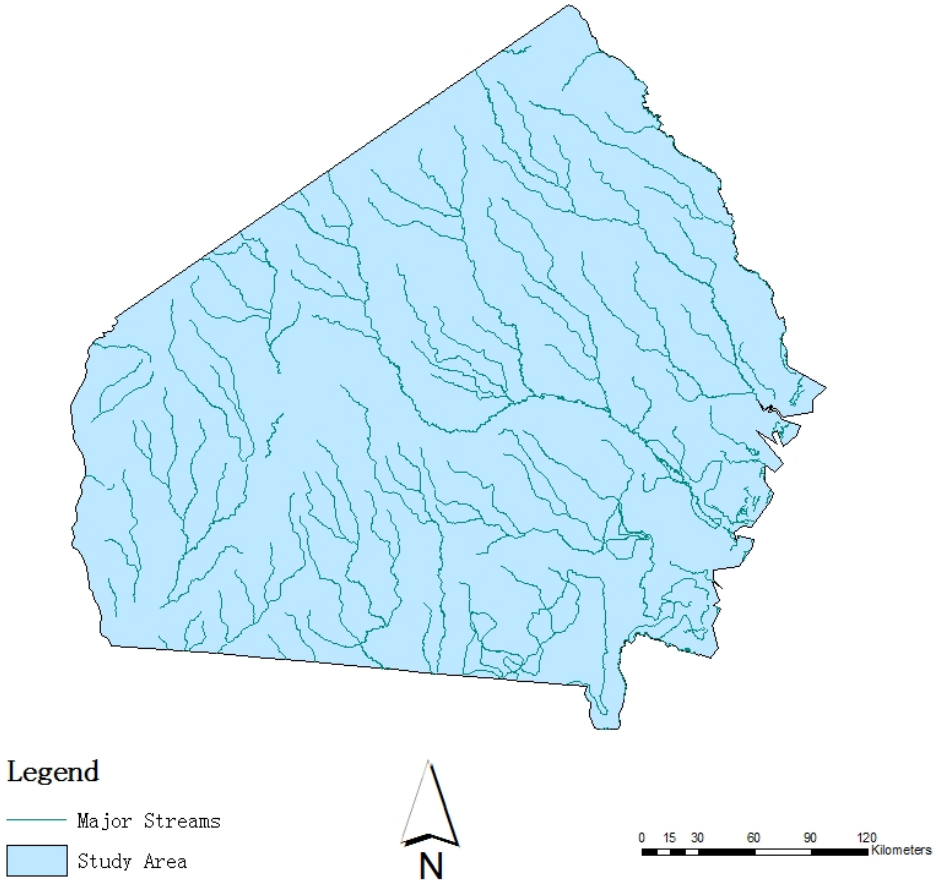


Figure 7: Major streams in the study area.

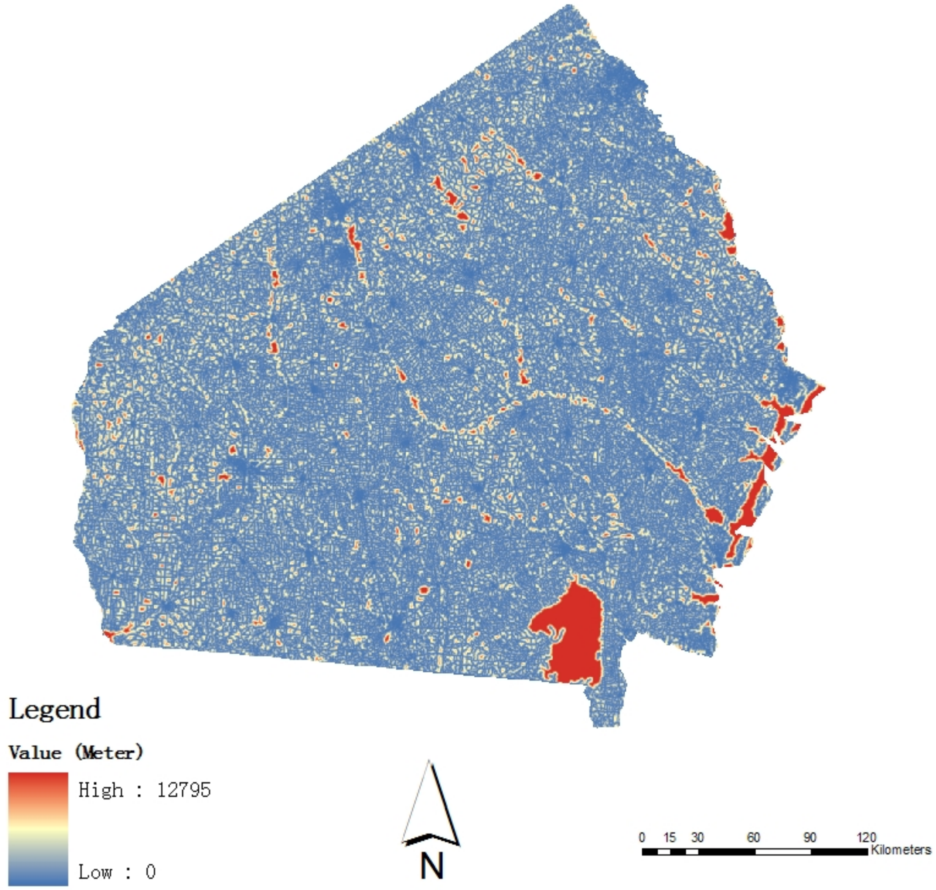


Figure 8: National Overview Road Metric Euclidean Distance (NORM ED) dataset of the study area.

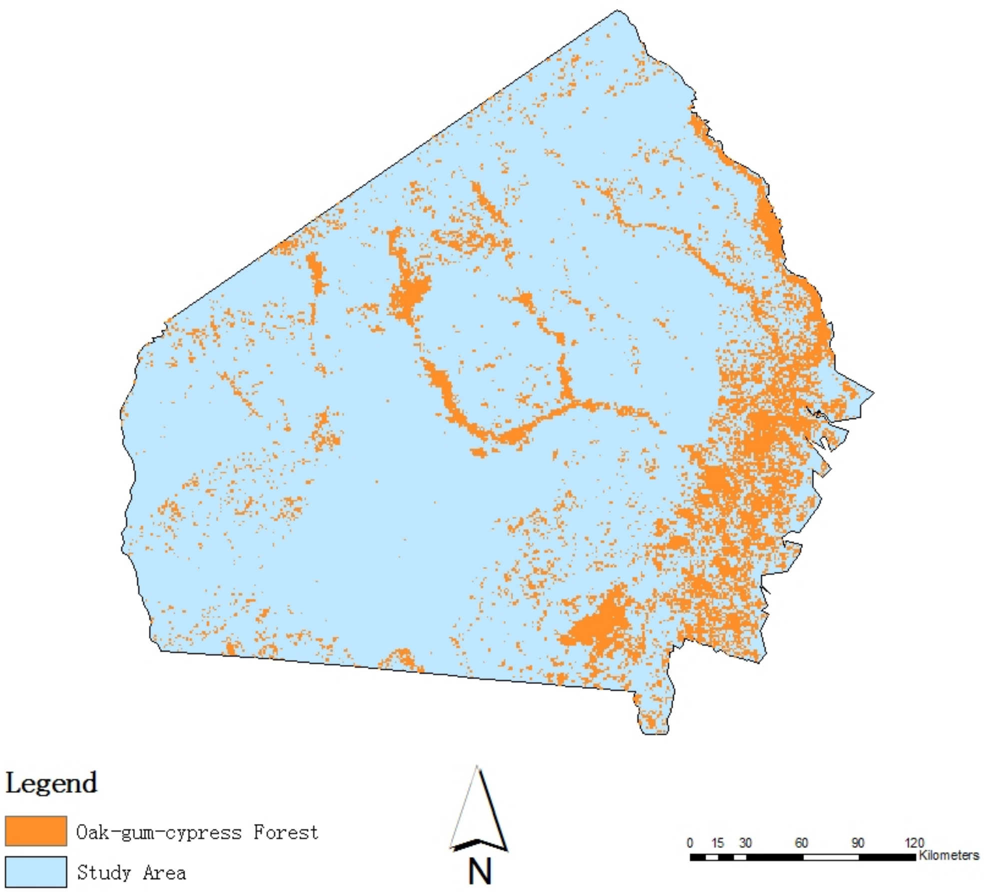


Figure 9: Presence of Oak-gum-cypress Forest in the study area in 2002. Orange indicates areas where oak-gum-cypress forests were distributed.

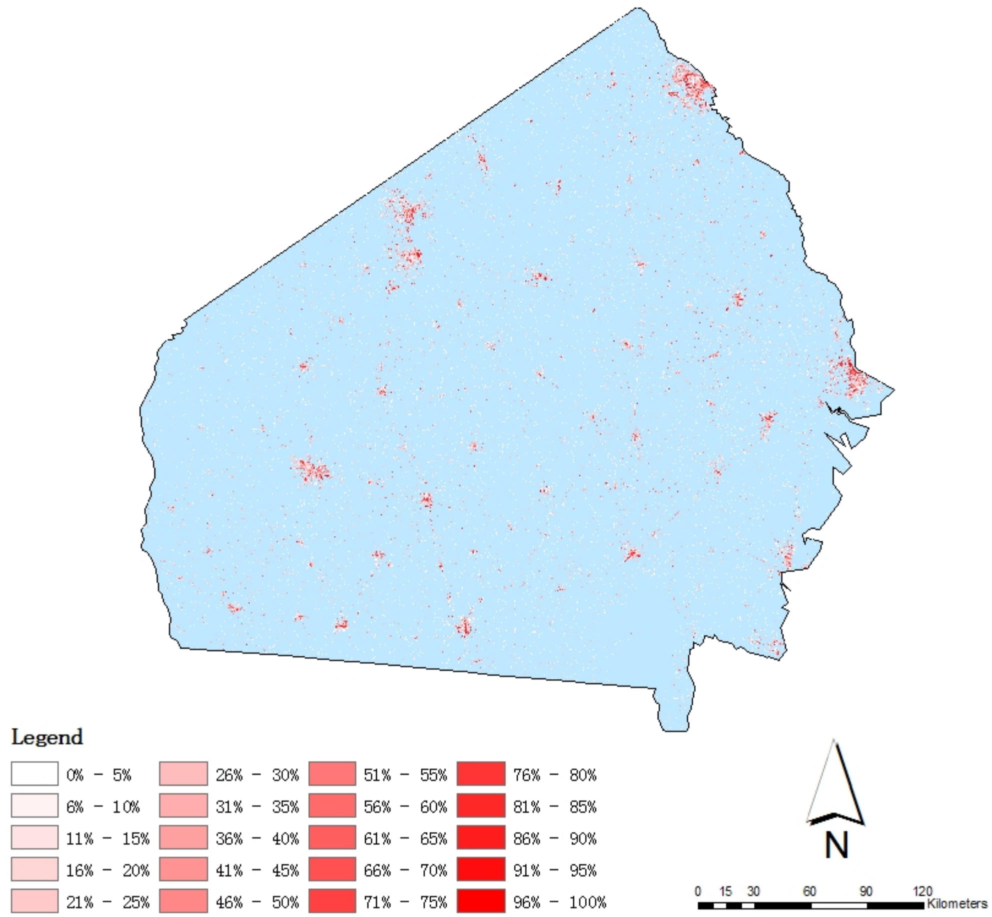


Figure 10: Impervious surface dataset of the study area in 2005. Darker red indicates higher percentage of impervious surface and lighter red indicates lower percentage of impervious surface.

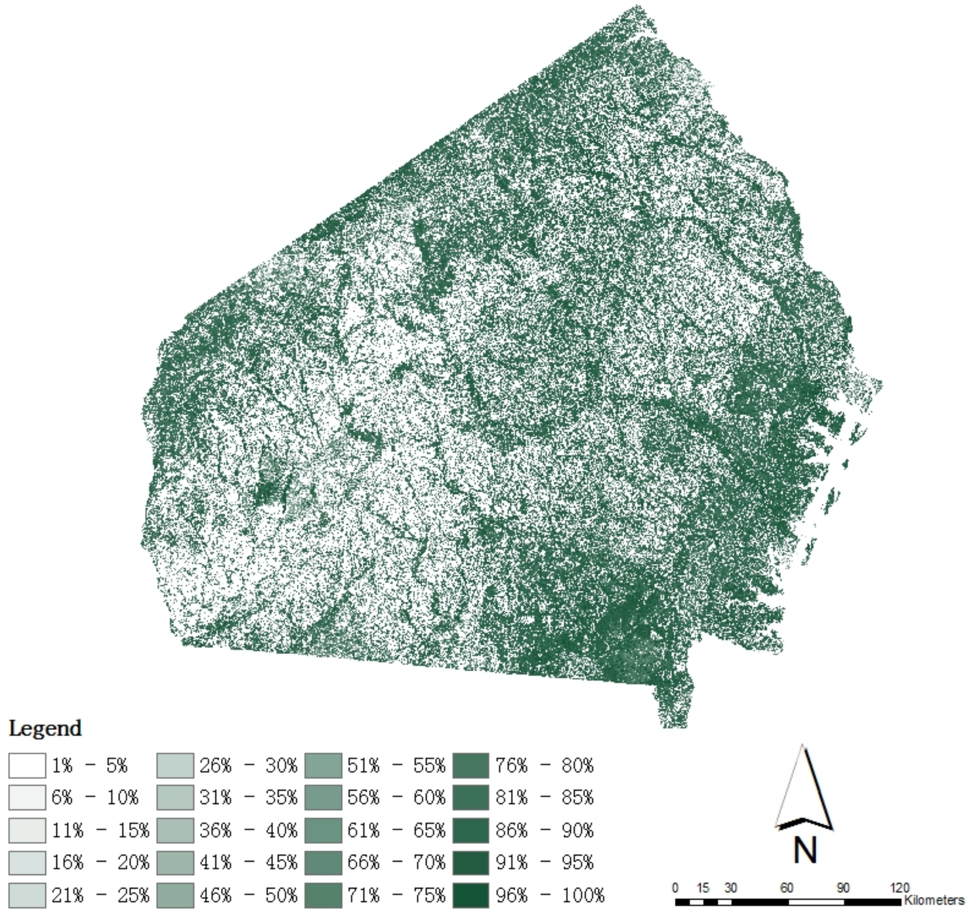


Figure 11: Canopy Cover dataset of Georgia in 2005. Darker green indicates higher percentage of canopy cover and lighter red indicates lower percentage of canopy cover.

CHAPTER 4

METHODOLOGY

Variable Manipulation

Most of the data downloaded to create a geodatabase for modeling factors spatially correlated with feral swine distribution were not ready to use directly in the modeling analysis. A variable manipulation process was therefore needed. Different methods were employed to refine different variables, and a total of nine variables were compiled for this project in Universal Transverse Mercator (UTM) ground coordinate system tied to the North American Datum of 1983 (NAD 1983).

1) Land Use and Land Cover (LULC)

In order to facilitate the modeling process, the original 13-class LULC dataset was modified and all classes whose total area was too small (i.e. less than 2% of the area of Georgia) were either grouped into an “Other” class or those classes who shared similar characteristics.

For example, the High Intensity Urban class was merged into the Low Intensity Urban and formed a new class called “Urban”, and the Beaches, Dunes and Mud class, along with the Quarries, Strip Mines and Rock Outcrops class and two Non-forested Wetland classes, was grouped into the “Other” class. The reason for eliminating those classes with small area was that they were so small that the number of the sample points that fell into those classes was statistically insignificant; in addition, smaller number of classes could facilitate the whole modeling process. After class elimination and grouping, there were nine classes in the modified classification system.

Table 5 shows the modified LULC types and their descriptions. Figure 12 shows the modified LULC map of the study area in 2005.

2) Slope and Elevation

Since the elevation information is inherently included in DEM, the elevation values were extracted directly from the DEM dataset, as shown in Figure 5. Slope was derived from the DEM using the Spatial Analyst Tool in ArcGIS and is displayed in Figure 13.

3) NDVI

The numerical values of NDVI were extracted from the original NDVI dataset described previously.

4) Distance to Major Streams

This variable was derived from the Major Streams in Georgia dataset by using Euclidean Distance Tool within Spatial Analyst Tools in ArcGIS. The output was a raster image and the value of each pixel represents the Euclidean distance from the target pixel to the nearest major streams, as shown in Figure 14.

5) National Overview Road Metric Euclidean Distance

The numerical values of the National Overview Road Metric Euclidean Distance v were extracted directly from the original National Overview Road Metric Euclidean Distance dataset.

6) Distance to Oak-gum-cypress Forest

This variable was developed based on the Presence of Oak-gum-cypress Forest dataset using Euclidean Distance Tool within Spatial Analyst Tools in ArcGIS. The output was a raster image and the value of each pixel represents the Euclidean distance from the target pixel to the nearest Oak-gum-cypress forest, as shown in Figure 15.

Table 5: Modified classification system of the 2005 Georgia LULC Dataset.

Code	LULC Type	Description
1	Open Water	Lakes, rivers, ponds, ocean, industrial water, aquaculture which contained water at the time of image acquisition.
2	Urban	Family dwellings, recreation, cemeteries, playing fields, campus-like institutions, parks, schools, commercial/industrial, prisons, speedways, junkyards, confined animal operations. Transportation, roads, railroads, airports and runways. Utility swaths.
3	Clearcut and Sparse	Recent clearcuts, sparse vegetation, and other early successional areas.
4	Deciduous Forest	Forest composed of at least 75% deciduous trees in the canopy, deciduous woodland.
5	Evergreen Forest	Evergreen forest, at least 75% evergreen trees, managed pine plantations, evergreen woodland.
6	Mixed Forest	Mixed deciduous/coniferous canopies, mixed woodland, natural vegetation within the fall line and coastal plain ecoregions, mixed shrub/scrub vegetation.
7	Row Crops and Pastures	Row crops, orchards, vineyards, groves, horticultural businesses. Pasture, non-tilled grasses.
8	Forested Wetland	Cypress gum, evergreen wetland, deciduous wetland, depressional wetlands, and shrub wetlands.
9	Others	Open sand, sandbars, sand dunes, mud - natural environmentals as well as exposed sand from dredging and other activities. Exposed rock and soil from industrial uses, gravel pits, landfills. Rock outcrops, mountain tops, barren land. Salt marsh, brackish. Freshwater marsh.

7) Distance to Impervious Surfaces

A Distance to Impervious Surfaces variable was developed based on the 2005 Impervious Surface Cover of Georgia. During the pre-processing, only the classes with more than 80% impervious surface percentage were kept in order to emphasize the highly urbanized area and to avoid the distraction from the low urbanized areas such as roads. Consequently, a raster image representing the Euclidean Distance of each pixel to the nearest urban areas was generated using the Spatial Analyst Tool in ArcGIS, as shown in Figure 16.

8) Canopy Cover

A Canopy Cover dataset was reclassified based on the 2005 Canopy Cover dataset in order to facilitate for modeling process. It contains 11 classes, with 10 classes being grouped based on the original 20-classes dataset, and one extra class representing areas with zero canopy cover values which was not contained in the original dataset (Figure 17).

Table 6 shows all variables that were used in this project as well as their abbreviations. All the datasets except for LULC and canopy cover can be input into the model directly because they are in ordinal and/or ranked numeric data format. Because the LULC and canopy cover datasets are nominal (or categorical) data, conversion to a numerical format was required before they could be used in logistic regression. This was accomplished by converting these two datasets into dummy variables, respectively. Dummy variable conversion is a prevalent approach to convert nominal variables to numerical variables, and each dummy variable represents a subgroup (Kleinbaum, Klein et al. 2010). Nine dummy variables and eleven dummy variables were generated, respectively, to replace the LULC and canopy cover in modeling (Table 7 and Table 8). LULC9 and CC11 variables were used as reference dummy variables and, in order to avoid the issue of exact collinearity, they were not entered in model building (Weisberg 2005).

Finally, all environmental and cultural variables were imported as individual layers and input to logistic regression analysis.

Table 6: List of variables and their abbreviations.

Variable Name	Abbreviation	Variable Type
Feral Swine Presence/Absence	Pre	Dependent
Land Cover	LULC	Independent
Slope	SL	Independent
Elevation	EL	Independent
Normalized Difference Vegetation Index	NDVI	Independent
Distance to Major Streams	D2S	Independent
National Overview Road Metric Euclidean Distance	NORM	Independent
Distance to Oak-gum-cypress Forest	D2O	Independent
Distance to Impervious Surfaces	D2I	Independent
Canopy Cover	CC	Independent

Logistic Regression Models

Since comparing the performance between the ordinary logistic regression and the autologistic regression is one of the objectives of this project, both of these analyses were performed. The ordinary logistic regression takes the form:

$$P_{1|i} = \frac{e^{\alpha+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_i X_i}}{1+e^{\alpha+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_i X_i}}$$

$$P_{2|i} = \frac{1}{1+e^{\alpha+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_i X_i}}$$

where $P_{1|i}$ and $P_{2|i}$ are the probability of presence and absence, respectively, α is the model constant, X_i are the independent variables and β_i are the model coefficients (Wrigley 1985).

As discussed in the previous chapter, scale plays an important role in landscape ecological modeling. This study took scale into account by establishing the models at multiple scales and selecting the optimal model to be the final scale at which the autologistic regression model was established. Based on the reported home range of feral swine (i.e., 1 to 49 square kilometers) (Stevens 1996; Mapston 2004), four different scales were considered, i.e. 1000 m, 3000 m, 5000 m and 10000 m.

The method proposed by Wiser el al. (1998) was adopted as the approach to resample and generalize all environmental variables based on four scales. Four grids were created to cover the entire study area, each consisting of square cells with cell size being 1000 m, 3000 m, 5000 m and 10000 m, respectively. All cells that intersected with the boundary of the study area were dropped to make sure that each cell is complete. All variable layers were generalized to the four cell sizes. For all the nominal data layers including presence/absence of feral swine, LULC and canopy cover, the dominant attribute (i.e., majority) within each cell was selected to represent that cell. For the rest of the ordinal data layers, the average attribute within each cell was used to represent that cell. Therefore, each variable layer was generalized and four layers for each variable were created. As an example, Figure 18 shows the distance to impervious surface data layer after it was generalized based on cell size of 10000 m.

Four sets of random points were generated across the study area in ArcGIS, each included 1000 points. These points were used as sample points for the purpose of model training. Based on the rule proposed by Pereira and Itami (1991) that the appropriate ratio of the number of points used in model training to that of points used in validation is two to one, another set of

Table 7: List of dummy variables for the LULC variable.

Code	LULC Type	LULC1	LULC2	LULC3	LULC4	LULC5	LULC6	LULC7	LULC8	LULC9
1	Open Water	1	0	0	0	0	0	0	0	0
2	Urban	0	1	0	0	0	0	0	0	0
3	Clearcut and Sparse	0	0	1	0	0	0	0	0	0
4	Deciduous Forest	0	0	0	1	0	0	0	0	0
5	Evergreen Forest	0	0	0	0	1	0	0	0	0
6	Mixed Forest	0	0	0	0	0	1	0	0	0
7	Row Crops and Pastures	0	0	0	0	0	0	1	0	0
8	Forested Wetland	0	0	0	0	0	0	0	1	0
9	Others	0	0	0	0	0	0	0	0	0

Table 8: List of dummy variables for the Canopy Cover variable.

Code	Canopy Cover	CC1	CC2	CC3	CC4	CC5	CC6	CC7	CC8	CC9	CC10	C11
0	0	1	0	0	0	0	0	0	0	0	0	0
1	1% - 10%	0	1	0	0	0	0	0	0	0	0	0
2	11% - 20%	0	0	1	0	0	0	0	0	0	0	0
3	21% - 30%	0	0	0	1	0	0	0	0	0	0	0
4	31% - 40%	0	0	0	0	1	0	0	0	0	0	0
5	41% - 50%	0	0	0	0	0	1	0	0	0	0	0
6	51% - 60%	0	0	0	0	0	0	1	0	0	0	0
7	61% - 70%	0	0	0	0	0	0	0	1	0	0	0
8	71% - 80%	0	0	0	0	0	0	0	0	1	0	0
9	81% - 90%	0	0	0	0	0	0	0	0	0	1	0
10	91% - 100%	0	0	0	0	0	0	0	0	0	0	0

500 random points was also generated across the study area to be used for the model validation. All the points were tied to the attributes of the different environmental factors extracted from the corresponding data layers in the GIS database.

Logistic regression relies on the assumption that the independent variables in the model are not correlated with each other (Chatterjee and Hadi 2006). If there is correlation between the independent variables, the regression results may be rendered ambiguous. This is commonly referred to as the problem of collinearity or multicollinearity. There are two types of collinearity. One is perfect collinearity, which is the situation when at least one independent variable is a perfect linear combination of the other variables. The other one is partial collinearity in which the degree of collinearity is less than perfect collinearity. Perfect collinearity is rare in practice, but partial collinearity is quite common. If the level of collinearity existing between variables is high enough, it becomes problematic and must be eliminated because it will lead to problems including large standard errors for regression coefficients, statistically non-significant or unreasonably high coefficients (Menard 1995; Chatterjee and Hadi 2006). In this project, Spearman's rank correlation coefficient was used to test the correlation between each pair of variables that entered the model. Spearman's rank-order correlation coefficient is a non-parametric measure of correlation between variables recommended by Borcard et al. (2011) and Cerezo et al. (2010). It is similar to Pearson's correlation coefficient, which is another correlation coefficient commonly used, in terms of the range of values and ways of interpretation. However, unlike Pearson's correlation which compares the scores of the variables, Spearman's correlation emphasizes on the difference of ranks between each pair of variables. Its expression takes the form of:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

where r_s is the Spearman's correlation coefficient, d_i is the difference in rank between pairs of scores of the variables, and n is the sample size (Clark and Hosking 1986). Partial linear regression analysis was undertaken if the correlation between any pair of variables is larger than 0.6 (Cerezo, Perelman et al. 2010). Since the nature of Spearman's correlation coefficient dictates that the variables being assessed must be in interval or ratio scale, the LULC and canopy cover variables were not considered in this step.

Each variable needs to be tested to see if their parameters are statistically significantly different from zero. The null hypothesis (H_0) is that each individual variable has zero significance, while the alternative hypothesis (H_1) is that each variable has significance. There are two tests that can be used to serve this purpose, i.e. Wald test and likelihood-ratio (LR) test. The Wald test evaluates the support for the null hypothesis by assessing the distance between the estimated coefficients and hypothesized values, while the LR test essentially compares the log-likelihood of the full model which does not suffer any constraints and a model which is under restriction of the null hypothesis. The Wald test and LR test are considered asymptotically equivalent. They usually yield very similar results and neither of them is uniformly superior (Long, Freese et al. 2006). In this project, the Wald test was used to test the significance of individual variables. For logistic regression, Wald statistics (Z score) follows a normal distribution, and in the statistical software such as Stata it associates with p-values used to judge the significance of the coefficient, and chi-squared values with certain degrees of freedom which equal the square of the Z scores (Chatterjee and Hadi 2006; Long, Freese et al. 2006). If the p-value for a coefficient is smaller than a predetermined value at a certain significance level, then the null hypothesis (which states that coefficient is equal to zero) shall be rejected at that significance level.

The stepwise regression method, which is a way of selecting and eliminating independent variables by an automatic and iterative procedure, was used to select variables to construct the optimal model among the variables that remained after the first two tests. While being criticized for a series of drawbacks including possible inflation of R^2, it is still one of the most commonly used methods to generating optimal models (Clark and Hosking 1986). According to the selecting manner, there are two forms of stepwise regression: forward selection and backward elimination (or backward selection). Forward selection starts with a model with a constant term only, then adds the variables into the model one-by-one based on whether they are statistically significant. On the contrary, backward selection starts the selection process by first considering all variables, then eliminates them one-by-one according to their significance. Since it has been suggested that backward selection is slightly superior to forward selection in that the former carries less risk to ignore a certain variable which is statistically significant only when another variable is present (or called "suppressor effect"), backward elimination was implemented in this project to determine the optimal models (Menard 1995; Ryan 1997). A threshold value of 0.05 was used to assess statistical significance. Although stepwise selection is trustworthy in most cases, when a variable was tested by the Wald Test to be statistically significant, a further scrutiny was carried out on this particular variable to insure that it was legitimate for the automatic selection process to exclude this variable.

After the optimal model was found, residuals resulting from its prediction were examined using Pregibon leverage statistics. This is a common measure to detect the influential observations, which include potential outliers. By employing such a detection analysis, all unusual observations that may be problematic can be identified, enabling further scrutiny to focus solely on these observations and determine whether they are outliers and should be removed (Christensen 1997). The Pregibon leverage statistics for each observation were obtained and plotted based on the geographic location of the respective point, and the resulting map was manually interpreted to determine whether removing influential observations was necessary, based on the spatial patterns that were exhibited. If high influential observations were highly clustered, and legitimate explanations could be proposed in accordance to this pattern, all observations in that clustered area were removed. The model fitting process was then conducted again based on the remaining observations.

The goodness-of-fit of the models was examined using several methods and measures in this project, including receiver operating characteristic (ROC) plots, Akaike Information Criterion (AIC), classification accuracy and rate of successful prediction in validation (Wrigley 1985).

The ROC plot is rooted in the concept of a confusion matrix. A confusion matrix (or error matrix) tabulates the actual presence/absence pattern with the pattern predicted by a certain algorithm (or classifier), which in this case is logistic regression. There are essentially four elements in a confusion table: 1) true positive (TP) samples correctly identified as present; 2) true negative (TN) samples correctly identified as absent; 3) false positive (FP) samples incorrectly identified as present; and 4) false negative (FN) samples incorrectly identified as absent. Table 9 shows a general confusion matrix. The proportion of all TP samples of all samples that are actually positive (TP+FN) is called Sensitivity (true positive rate). Similarly, the proportion of all TN samples of all samples that are actually negative (TN+FP) is called

Specificity (true negative rate). The ROC plot is in effect a graph of x and y coordinates with the y-coordinate being Sensitivity and the x-coordinate being 1-Specificity (false positive rate). Its derivative index, the area under the ROC curve (AUC), is usually used as a measure to assess the discriminative ability of a classifier. Its value ranges from 0.5 to 1.0, with the score of 0.5 indicating the model has no discriminative ability and 1.0 suggesting that the model can discriminate two different groups perfectly (Fielding and Bell 1997; Hosmer and Lemeshow 2000; Pearce and Ferrier 2000; Reese, Wilson et al. 2005). A large number of studies employed ROC plot method in their ecological study (Guisan and Zimmermann 2000; Osborne, Alonso et al. 2001; Gibson, Wilson et al. 2004; Reese, Wilson et al. 2005; Smolik, Dullinger et al. 2010). It was chosen in this study in favor of the other counterparts due to the consideration that the ROC plot method is superior to the other measures because it is independent of the decision threshold and capable of maximizing information provided by the classifier (Fielding and Bell 1997; Pearce and Ferrier 2000).

Classification accuracy was obtained using the post-estimation command after running the logistic regression. Similar to the ROC plot, it is based on the confusion matrix produced by the classifier, with the difference being that it needs an arbitrary threshold value before calculation. The expression of classification accuracy is:

$$\mathrm{CA} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

where CA represents classification accuracy, TP, FP, FN and TN represent true positive, false positive, false negative and true negative, respectively (Wrigley 1985; Fielding and Bell 1997). The threshold value in this study was 0.5.

It is ideal to have an independent dataset to validate the model outcome (Wrigley 1985; Hosmer and Lemeshow 2000; Long, Freese et al. 2006). For this purpose, 500 random points were generated throughout the study area for validation, and the rate of successful prediction was

Table 9: A general confusion matrix. TP, FP, FN and TN represent true positive, false positive, false negative and true negative, respectively.

	Actual +	Actual -
Predicted +	TP	FP
Predicted -	FN	TN

derived based on these points. All the validating points went through the same processes as the training sample points, except that they were not used to develop the logistic regression model. Instead, their attributes were used together with the corresponding coefficients obtained by the logistic regression models based on the training data to predict the presence/absence of feral swine condition at each point. The result of prediction was then compared with the actual presence/absence of feral swine data to yield the rate of successful prediction.

All the measures of goodness-of-fit described above were calculated using Stata (StataCorp 2005) and Microsoft Excel (Microsoft 2007).

Tests for Spatial Autocorrelation

Before constructing the autologistic regression models, a series of tests were conducted to examine whether there was indeed existing spatial autocorrelation that needed to be eliminated. According to the assumptions of logistic regression, the error term should be independently distributed. Whether this assumption is violated can be determined by examining the residuals that the model yields. The basic idea behind these tests is if a significant spatial autocorrelated pattern is found within the residuals, then efforts must be taken to eliminate this effect to make sure the assumption of independency is not violated (Legendre 1993; Chatterjee and Hadi 2006).

The standardized residuals for each point were calculated in Stata based on the model with the best performance after the ordinary logistic regression was performed. Then both Global and Local Moran's I were calculated based on the standardized residuals. Moran's I is a popular measure of spatial autocorrelation originally proposed by Moran (1950) with several modified versions. In general, a Moran's I value ranges from -1, indicating complete negative spatial autocorrelation, to 1, indicating complete positive spatial autocorrelation. A Moran's I value equal to 0 means no significant spatial autocorrelated pattern is found, i.e. the distribution of the attributes is random (Moran 1950; Anselin 1995; Bailey and Gatrell 1996; Griffith 2003). The Z score and p-value associated with Moran's I indicate the significance of the calculated Moran's I and whether the null hypothesis, which is that the feature attributes are randomly distributed throughout the study area, can be rejected (ESRI 2009).

The Global Moran's I is a measure of spatial autocorrelation examining the overall pattern existing throughout the entire study area, and it is defined by ESRI (2009):

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j}(x_i - \overline{X})(x_j - \overline{X})}{\sum_{i=1}^{n}(x_i - \overline{X})^2}$$

where x_i and x_j are values for feature i and j, respectively, $\overline{X}$ is the mean of the all corresponding attributes, $w_{i,j}$ is the spatial weight between observation i and j, n is the total number of features, and S_0 is the sum of all the spatial weights shown in:

$$S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j}$$

The Z score associated with Global Moran's I is defined as:

$$Z = \frac{I - E[I]}{\sqrt{V[I]}}$$

where

$$E[I] = -\frac{1}{n-1}$$

and

$$V[I] = E[I^2] - E[I]^2$$

The Local Moran's I, unlike its global counterpart, evaluates the spatial autocorrelated pattern at the local level, and it was proposed by Anselin (1995). It is defined as:

$$I_i = \frac{x_i - \overline{X}}{S_i^2} \sum_{j=1, j \neq i}^{n} w_{i,j}(x_i - \overline{X})$$

where x_i is the value for feature i, $\overline{X}$ is the mean of the all corresponding attributes, $w_{i,j}$ is the spatial weight between observation i and j, n is the total number of features, and:

$$S_i^2 = \frac{\sum_{j=1, j \neq i}^{n} w_{i,j}}{n-1} - \overline{X}^2$$

The Z score associated with Local Moran's I is defined as:

$$Z = \frac{I - E[I]}{\sqrt{V[I]}}$$

where

$$E[I] = -\frac{\sum_{j=1, j \neq i}^{n}}{n-1}$$

and

$$V[I] = E[I^2] - E[I]^2$$

Both Global Moran's I and Local Moran's I were calculated and plotted in ArcGIS 9.3.1.

A correlogram was then generated with twofold purposes: 1) to validate the results of the previous spatial correlation measures; and 2) to examine at what spatial lag distance the spatial autocorrelated patterns are most significant. A correlogram is a distance-based diagram where the estimated spatial autocorrelation at a particular spatial lag is plotted against the lag distance (Bailey and Gatrell 1996; Rosenberg and Anderson 2011). There are a variety of forms of correlograms (e.g. Moran's I correlogram, Geary's C correlogram, etc.). Moran's I correlogram, which evaluates the relationship between Moran's I value and spatial lag distance, was constructed in this project in PASSaGE 2 (Rosenberg and Anderson 2011).

Autologistic Regression Models

Autologistic regression takes into account the spatial dependence by adding an autocorrelation (or autocovariate) term based on (Wrigley 1985; Augustin, Mugglestone et al. 1996; Hinely 2006; Santika and Hutchinson 2009):

$$P_{1|i} = \frac{e^{\alpha+\beta_1 X_1+\beta_2 X_2+...+\beta_i X_i+\beta_{i+1} \mathrm{auto\,cov}_{i+1}}}{1+e^{\alpha+\beta_1 X_1+\beta_2 X_2+...+\beta_i X_i+\beta_{i+1} \mathrm{auto\,cov}_{i+1}}}$$

$$P_{2|i} = \frac{1}{1+e^{\alpha+\beta_1 X_1+\beta_2 X_2+...+\beta_i X_i+\beta_{i+1} \mathrm{auto\,cov}_{i+1}}}$$

where $P_{1|i}$ and $P_{2|i}$ are the probability of presence and absence, respectively, α is the model constant, X_i are the independent variables, β_i are the model coefficients, and

$$\text{autocov}_i = \frac{\sum_{j=1}^{k_i} w_{ij} y_j}{\sum_{j=1}^{k_i} w_{ij}}$$

is a weighted average of the number of occupied squares amongst a set of k_i neighbors of square i where y_j is the observed value at site j surrounding i, and w_{ij} is the Euclidean distance between site i and its neighbor site j.

Three autocovariate terms were calculated based on three different levels of threshold distance in order to find out the optimal autocovariate term which enabled the final model to achieve the best performance. All points whose Euclidean distance to the point of interest was smaller than the threshold value were considered as neighbors to that particular point at that level. Threshold values were selected based on the correlograms generated in the last step. After the autocovariate terms were obtained, they were added to the corresponding equation of the identified optimal model of the ordinary logistic regression, one at a time. By doing so, three autologistic regression models were yielded, each containing an autocovariate term at different levels of distance. Finally, these three models were assessed by the same manner as the ordinary logistic regression analysis. The autocovariate term was calculated in Excel, and the performance of the autologistic regression models were analyzed in Stata.

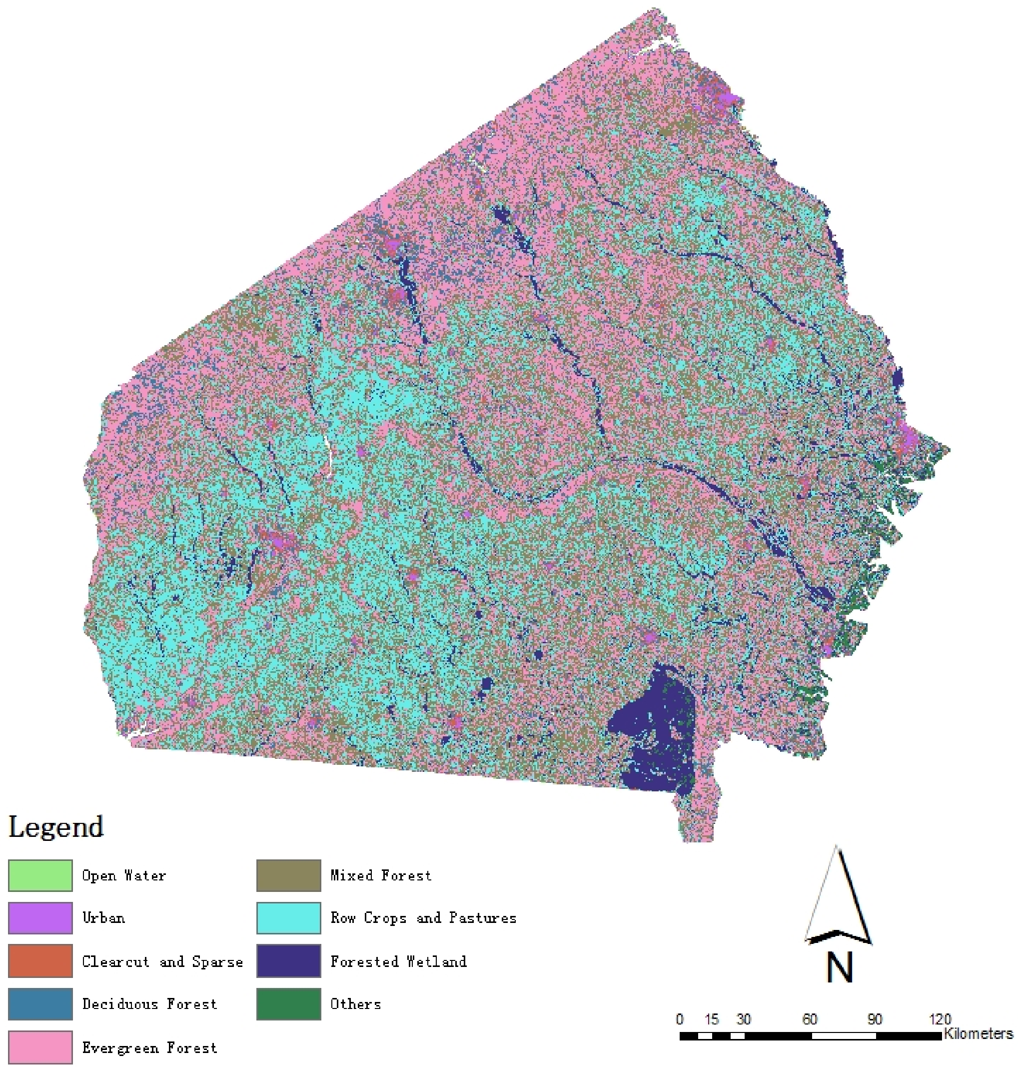


Figure 12: LULC map of the study area in 2005 based on the modified 9-class classification system.

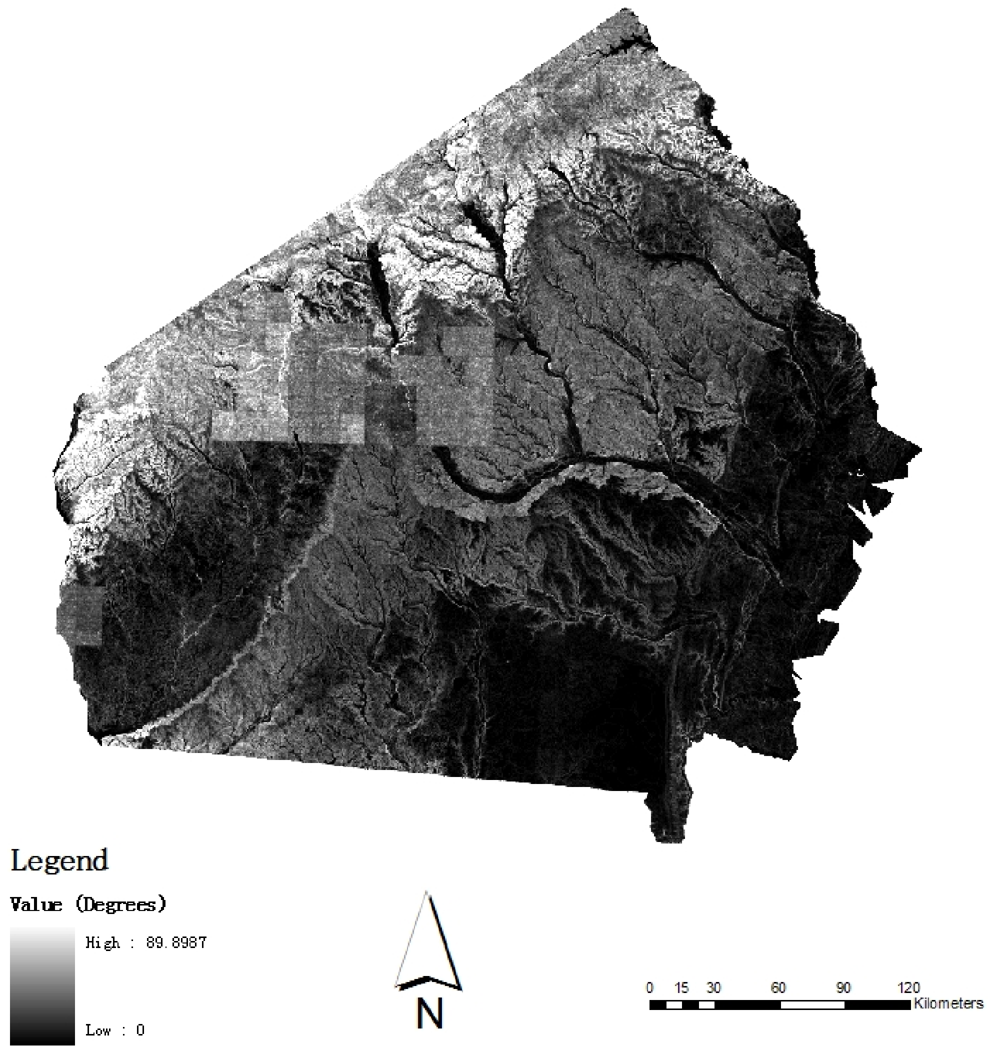


Figure 13: Slope of the study area derived from the DEM dataset.

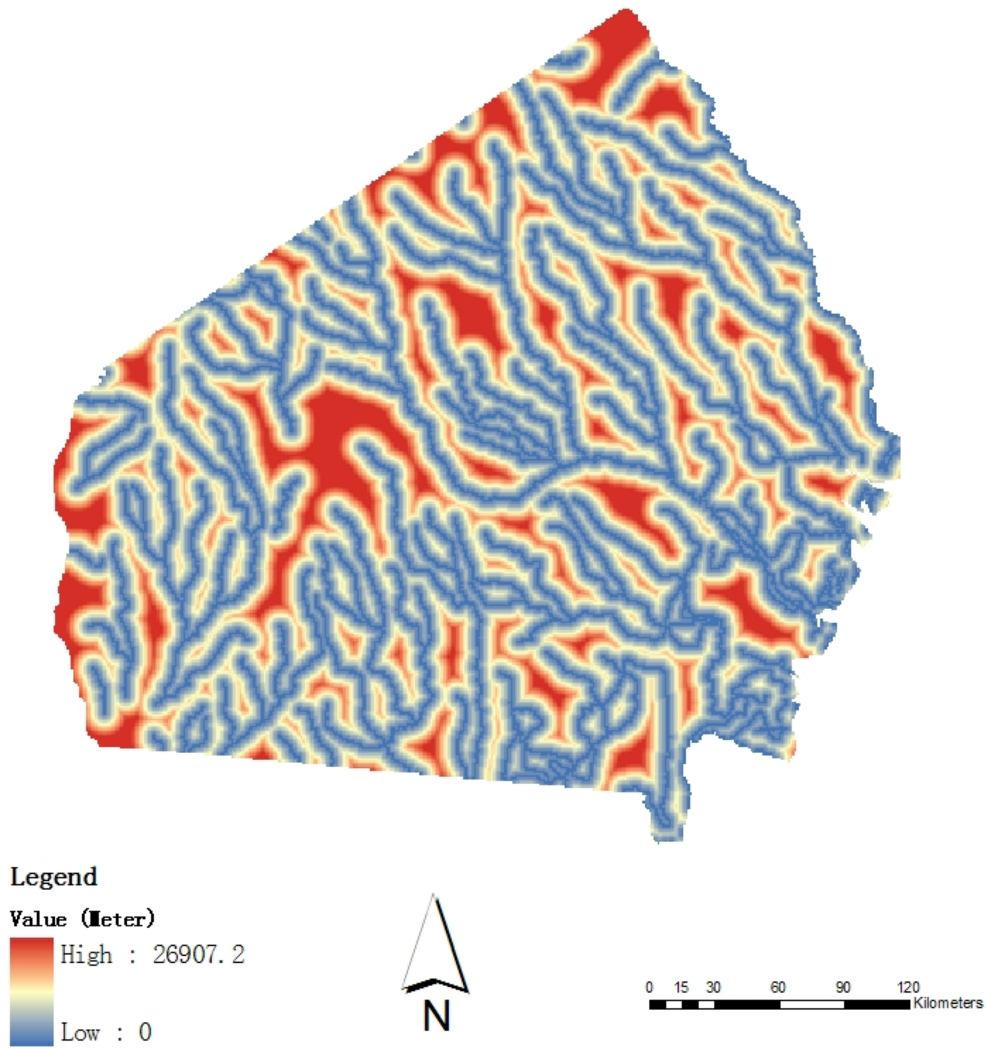


Figure 14: Map of distance to major streams.

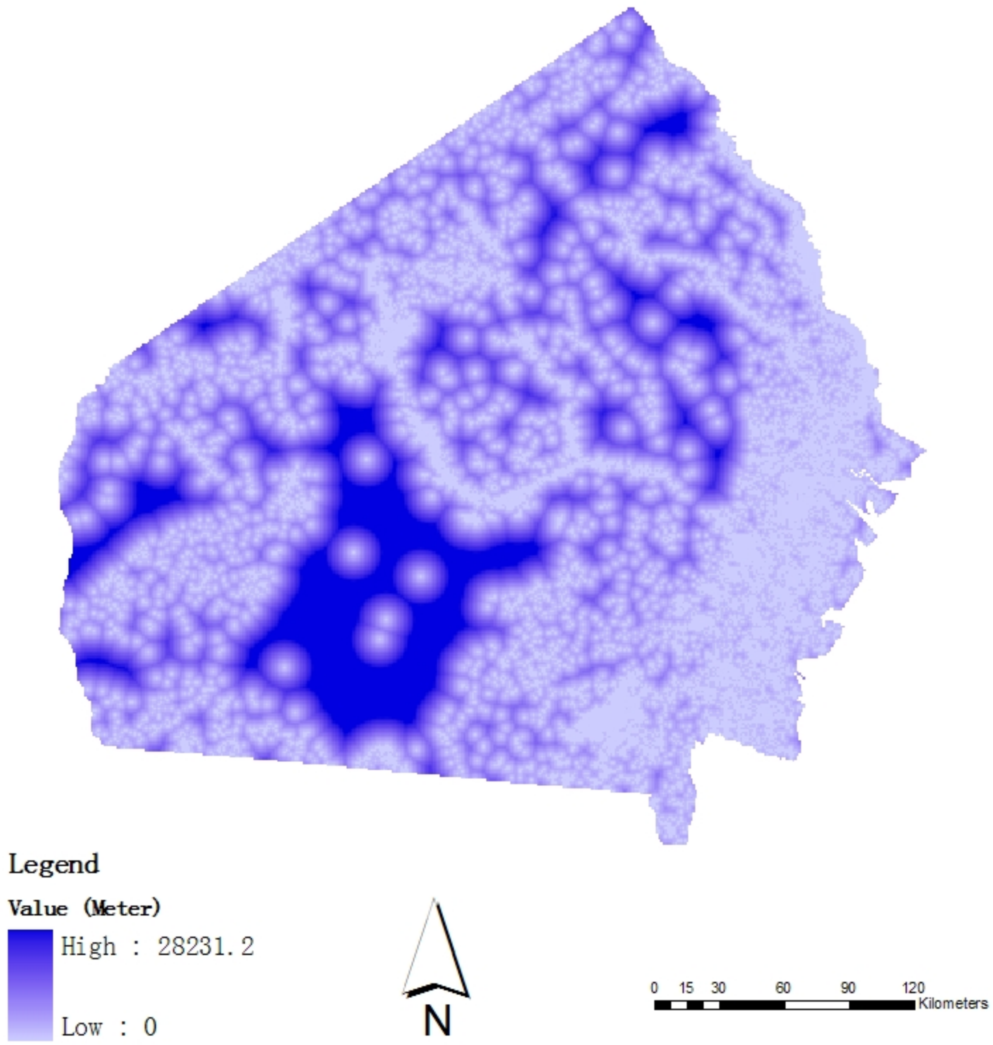


Figure 15: Map of distance to oak-gum-cypress forests.

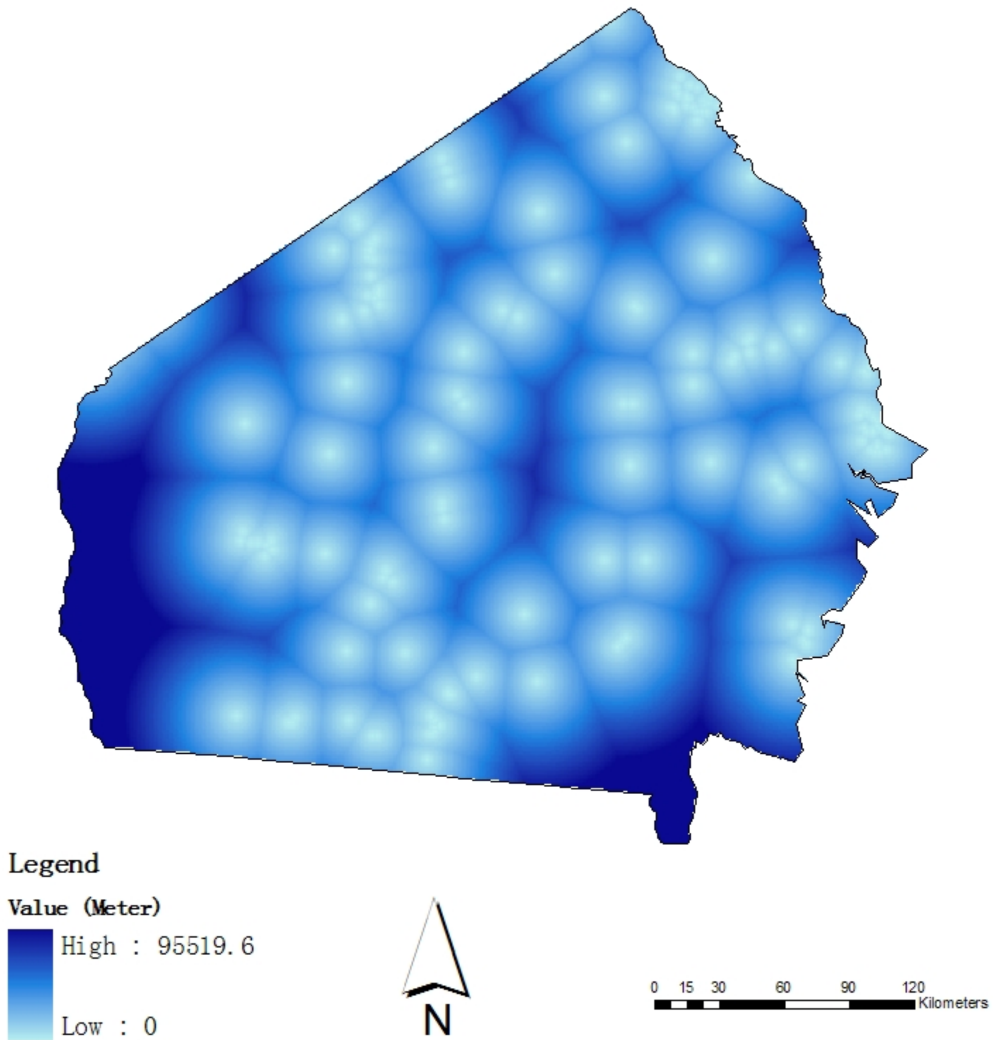


Figure 16: Map of distance to impervious surface.

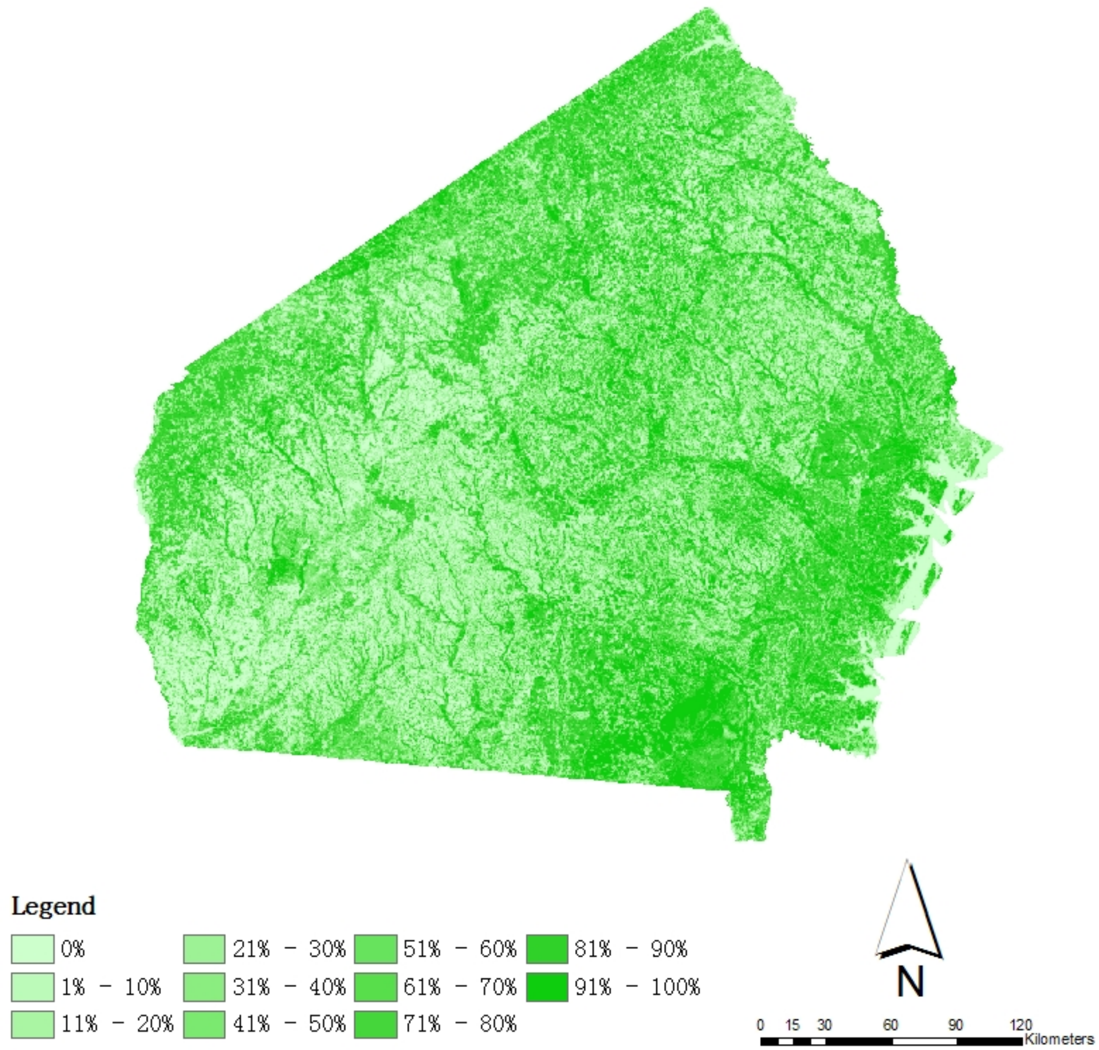


Figure 17: Map of Canopy Cover of the study area.

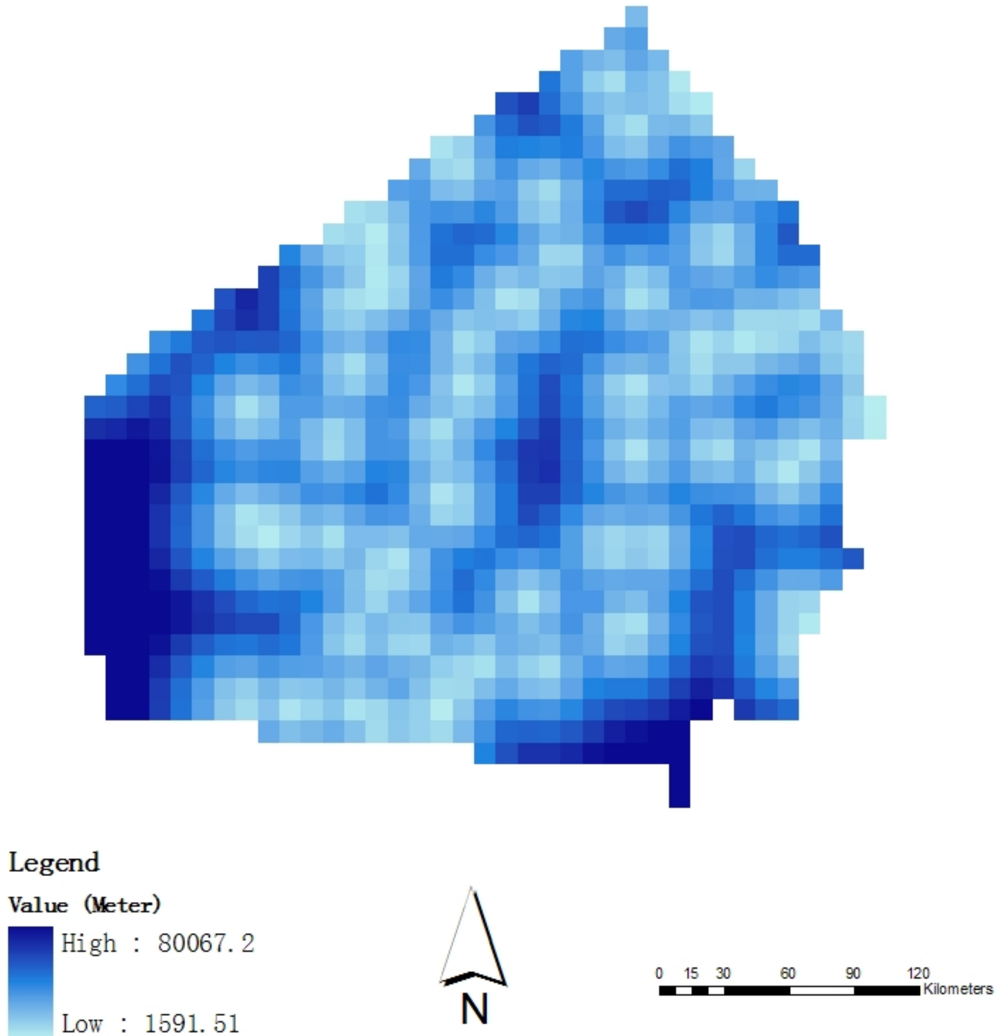


Figure 18: Distance to impervious surface data layer after it was generalized based on cell size of 10000 m.

CHAPTER 5

RESULTS

Ordinary Logistic Regression Outcomes

Table 10, 11, 12 and 13 show the Spearman's correlation coefficients for each pair of interval/ratio variables performed to assess correlation between pairs of variables intended for input to the regression models. This test was repeated for each of the scales being evaluated for optimal modeling (i.e., 1000 m, 3000 m, 5000 m and 10000 m scales). As these tables suggest, all the coefficients between each pair of variables at all scales were less than 0.5 other than the correlation between elevation and slope. This suggested that elevation or slope should not be entered into the model at the same time even if they were later tested to be significantly different from zero.

Table 10: Spearman's correlation coefficient matrix for logistic regression model generated at 1000 m scale. See Table 6 for full variable names.

	sl	el	ndvi	norm	d2s	d2o	d2i
sl	1.0000						
el	0.7256	1.0000					
ndvi	0.0084	-0.0468	1.0000				
norm	-0.0026	-0.0030	0.2317	1.0000			
d2s	0.0629	0.1611	-0.0514	-0.0543	1.0000		
d2o	0.1894	0.3292	-0.2633	-0.0886	0.0515	1.0000	
d2i	-0.0601	-0.0381	0.0900	0.1737	-0.0737	-0.1358	1.0000

Table 11: Spearman's correlation coefficient matrix for logistic regression model generated at 3000 m scale. See Table 6 for full variable names.

	sl	el	ndvi	norm	d2s	d2o	d2i
sl	1.0000						
el	0.7495	1.0000					
ndvi	0.0339	-0.0295	1.0000				
norm	0.0554	-0.0039	0.1685	1.0000			
d2s	0.1151	0.2230	-0.0706	-0.1354	1.0000		
d2o	0.2935	0.4545	-0.3105	-0.1177	0.0933	1.0000	
d2i	-0.0481	-0.0891	0.0451	0.2310	-0.1587	-0.1621	1.0000

Table 12: Spearman's correlation coefficient matrix for logistic regression model generated at 5000 m scale. See Table 6 for full variable names.

	sl	el	ndvi	norm	d2s	d2o	d2i
sl	1.0000						
el	0.7883	1.0000					
ndvi	0.0683	-0.0016	1.0000				
norm	0.0961	-0.0103	0.2455	1.0000			
d2s	0.0818	0.1780	-0.0564	-0.1201	1.0000		
d2o	0.2452	0.4463	-0.3492	-0.1991	0.1193	1.0000	
d2i	-0.0368	-0.0981	0.0571	0.2640	-0.0428	-0.1739	1.0000

The results of the Wald tests, which were used to test whether the variables were statistically significantly different from zero, for each single variable are presented in Table 14. As shown by the highlighted cells, different numbers of variables were tested to be significantly different from zero at the significance level of 0.05 at all four scales. As described previously, the stepwise selection was the primary procedure in this project to determine which variables should be included in the optimal models, the results of the Walt tests and the Spearman's correlation tests were used to provide auxiliary information required under certain circumstances, such as in the model construction by backward elimination excluding the variables which were tested to be significantly different from zero.

Table 13: Spearman's correlation coefficient matrix for logistic regression model generated at 10000 m scale. See Table 6 for full variable names.

	sl	el	ndvi	norm	d2s	d2o	d2i
sl	1.0000						
el	0.8000	1.0000					
ndvi	0.0810	-0.0175	1.0000				
norm	0.0283	-0.0722	0.1722	1.0000			
d2s	0.1071	0.1979	0.0058	-0.1699	1.0000		
d2o	0.3262	0.4769	-0.4102	-0.2470	0.1104	1.0000	
d2i	-0.0758	-0.0705	0.0890	0.3028	-0.0904	-0.1027	1.0000

Backward elimination was then implemented at each scale respectively, and the results are shown in Table 15 comparing with all the variables that were tested by the Wald tests to be statistically significant at 0.05 significance level. The highlighted cells represent the variables that were tested to be significant while excluded during backward elimination. At the 1000 m scale, all variables that were significant were included in the model constructed by backward elimination, and the automatic algorithm also included two extra land cover variables (i.e. lulc5 and lulc7) in the optimal model. Therefore the final model at this scale contained variables including ndvi, norm, d2s, d2o, lulc2, lulc4, lulc5, lulc7 and cc1. For the other three scales, there were significant variables being excluded by backward elimination. At the 3000 m scale, six variables were selected by backward elimination, with a significant variable lulc6 being excluded. Lulc6 was then added in to the model fitting along with the six pre-determined variables, and the outcome suggested that while being statistically significant alone, lulc6 was statistically insignificant when co-existing with the other variables. Thus lulc6 was confirmed to be excluded and the optimal model at this scale was finalized. Similar scrutiny was undertaken for both 5000 m scale and 10000 m scale, and after this procedure, the final models at these two scales contained eight and four variables respectively.

Table 16 shows the composition of the ordinary logistic regression models at each scale as well as the measures of goodness-of-fit associated with them, including Akaike information criterion (AIC), the area under the ROC curve (AUC), classification accuracy and prediction accuracy. The comparison of AUC of these four models is also displayed in Figure 19. There were not significant differences between AIC values of the four models, as the largest difference was only 0.268 (between Model O1 and O4). Slight differences were observed for the other three measures. In terms of AUC, Model O3 was superior to other three (0.7505), although not considerably. The AUC values of Model O1 and O2 were similar, and Model O4 had the worst AUC performance, with the value of 0.7119. However, when considering the measures of classification accuracy and prediction accuracy, Model O4 outperformed all the other three models. It correctly classified 78.40% of the total training points and correctly predicted 78.60% of the total validation points. Model O1, O2 and O3 had similar performances insofar as classification accuracy and O3 slightly outperformed O1 and O2. In terms of prediction accuracy, Model O2 performed the best among O1, O2 and O3. Model O3 and Model 1 ranked third and fourth respectively in this regard. Based on the goodness-of-fit of these models, Model O3 and Model O4 were considered to be superior to Model O1 and O2, and since the scale on which

Table 14: Results of Wald tests for each variable. Variables that no points were tied to are not included. Highlighted cells represent variables that were tested to be statistically significantly different from zero at significance level of 0.05.

1000 m scale			3000 m scale			5000 m scale			10000 m scale		
Variable	Z score	p-value	Variable	Z score	p-value	Variable	Z score	p-value	Variable	Z score	p-value
sl	1.62	0.105	sl	1.14	0.253	sl	0.89	0.375	sl	0.89	0.374
el	-1.79	0.074	el	-0.41	0.678	el	-0.77	0.442	el	-1.06	0.291
ndvi	2.26	0.024	ndvi	1.76	0.078	ndvi	1.23	0.219	ndvi	1.91	0.056
norm	-2.99	0.003	norm	-3.04	0.002	norm	-3.95	0.000	norm	-2.44	0.015
d2s	-4.73	0.000	d2s	-7.10	0.000	d2s	-7.07	0.000	d2s	-6.12	0.000
d2o	-4.54	0.000	d2o	-3.75	0.000	d2o	-3.37	0.001	d2o	-3.27	0.001
d2i	0.21	0.833	d2i	0.81	0.418	d2i	-1.63	0.104	d2i	-0.27	0.783
lulc1	-0.41	0.680	lulc1	-1.29	0.198	lulc2	-12.22	0.000	lulc2	-10.47	0.000
lulc2	-2.44	0.015	lulc2	-1.83	0.068	lulc3	-10.50	0.000	lulc4	-9.84	0.000
lulc3	-1.19	0.235	lulc3	-1.19	0.235	lulc4	-11.99	0.000	lulc5	-10.60	0.000
lulc4	-2.41	0.016	lulc4	-3.03	0.002	lulc5	-12.31	0.000	lulc7	-10.40	0.000
lulc5	-1.69	0.091	lulc5	-1.95	0.051	lulc6	-9.14	0.000	lulc8	-9.84	0.000
lulc7	-1.84	0.066	lulc6	-2.26	0.024	lulc7	-12.60	0.000	cc1	-1.05	0.294
lulc8	-1.23	0.220	lulc7	-2.25	0.024	lulc8	-10.73	0.000	cc8	-0.39	0.695
cc1	-2.93	0.003	lulc8	-1.28	0.202	cc1	-2.00	0.046			
cc7	-0.29	0.772	cc1	-1.65	0.100	cc8	-1.27	0.206			
cc8	-1.50	0.133	cc8	-0.29	0.771						

Table 15: Comparison of variables that were tested by the Wald tests to be significantly different from zero and variables that were selected by backward elimination. The highlighted cells represent the variables that were tested to be significant while excluded during backward elimination, and variables in bold font are those comprised of the base model at each scale.

1000 m Scale		3000 m Scale		5000 m Scale		10000 m Scale	
Stepwise	Wald Test	Stepwise	Wald Test	Stepwise	Wald Test	Stepwise	Wald Test
ndvi	ndvi	**norm**	norm	**norm**	norm	**norm**	norm
norm	norm	**d2s**	d2s	**d2s**	d2s	**d2s**	d2s
d2s	d2s	**d2o**	d2o	**d2o**	d2o	**d2o**	d2o
d2o	d2o	**lulc4**	lulc4	**lulc2**	lulc2	**cc1**	lulc2
lulc2	lulc2	**lulc7**	lulc6	**lulc4**	lulc3		lulc4
lulc4	lulc4	**cc1**	lulc7	**lulc5**	lulc4		lulc5
lulc5	cc1			**lulc7**	lulc5		lulc7
lulc7				**cc1**	lulc6		lulc8
cc1					lulc7		
					lulc8		
					cc1		

Model O3 was based is within the general home range of feral swine described previously (1 to 49 square kilometers), Model O3 was selected as the optimal model during the ordinary logistic regression model stage.

Pregibon leverage analysis was performed based on the fitting results of Model O3 with Stata and was plotted in ArcGIS, as shown in Figure 20. There were three areas where high Pregibon leverage values were found to be aggregated: one at the southern end of the Fall Line, one close to the mid-point of the Fall Line, and one in the south of the study area. For the former two areas, strong evidence indicated that the points within these areas were outliers. The third

Table 16: Composition of the ordinary logistic regression models at each scale associated with the measures of goodness-of-fit.

Scale	1000 m	3000 m	5000 m	10000 m
Model	O1	O2	O3	O4
Variable	ndvi	norm	norm	norm
	norm	d2s	d2s	d2s
	d2s	d2o	d2o	d2o
	d2o	lulc4	lulc2	cc1
	lulc2	lulc7	lulc4	
	lulc4	cc1	lulc5	
	lulc5		lulc7	
	lulc7		cc1	
	cc1			
AIC	1.237	1.214	1.17	0.969
AUC	0.7352	0.7452	0.7505	0.7119
Correctly Classified	68.50%	68.61%	71.10%	78.40%
Correctly Predicted	66.40%	71.20%	68.40%	78.60%

area was the most evidently clustered compared with the former two and, more importantly, it exhibited a strong pattern of all the highly influential observations in this area being located simultaneously in the Okefenokee Swamp. Given the fact that feral swine are widely reported to prefer swamp habitat and that the boundary of the Okefenokee Swamp was almost identical to the outline of the absence-of-feral-swine area at this location, this phenomenon was considered as unusual. It indicated that there might be certain factors within the Okefenokee Swamp that had stronger influence on the distribution of feral swine differently than the factors considered in this project. Since the purpose of this study was to obtain a general relationship between feral swine and their environment conditions, all the observations within the Okefenokee Swamp were considered to be outliers and removed. Then logistic regression was fitted based on specifications of Model O3 again using the remaining observations, and the fitting results are displayed in Table 17.

Tests for Spatial Autocorrelation

After the optimal model was fitted, a standardized residual was calculated for each point in the training set, and then tests for both Global Moran's I and Local Moran's I were conducted. The results of Global Moran's I test (Figure 21) suggested that a strong clustered

Table 17: Fitting results of logistic regression based on Model O3 with the remaining observations.

	Coef.	Std.	z	P>|z|	[95% Conf. Interval]	
cc1	-0.6489	0.202414	-3.21	0.001	-1.04562	-0.25217
lulc5	-0.74578	0.315227	-2.37	0.018	-1.36361	-0.12794
lulc7	-1.12578	0.322498	-3.49	0	-1.75786	-0.49369
norm	0.001522	0.000731	2.08	0.037	0.000089	0.002955
d2s	-0.00018	0.000025	-7.17	0	-0.00023	-0.00013
d2o	-6.4E-05	1.79E-05	-3.57	0	-9.9E-05	-2.9E-05
lulc4	-1.84184	0.619229	-2.97	0.003	-3.0555	-0.62817
Constant	2.39277	0.431894	5.54	0	1.546273	3.239267
	AIC	AUC	Correctly Classified		Correctly Predicted	
	1.173	0.7452	70.67%		66.87%	

pattern was detected across the study area, and that the null hypothesis should be rejected, i.e. spatial autocorrelation did exist. The results of the Local Moran's I test, as shown in Figure 22, confirmed the conclusion of the Global Moran's I test, and local clustered pattern (indicated by the red dots) was found throughout the study area.

Figure 23 shows the correlogram produced by PASSaGE 2 based on the standardized residual term of Model O3. Red dots and blue crosses represent significant and insignificant autocorrelated patterns respectively. Results of this test suggested that the autocorrelated pattern was significant ($p<0.05$) within a range of approximately 45000 meters, and within this range its significance declined with the increasing of distance. Beyond 45000 meters, the autocorrelation could be considered as insignificant, although significant spatial autocorrelation was found at certain distance which was the results of chance fluctuation (Sokal and Wartenberg 1983).

Autologistic Regression Outcomes

In order to explore the relationship between scale and its influence on the power of the autocovariate term, and given the fact that spatial autocorrelation vanished with the increasing of distance and became insignificant at the distance of approximately 45000 m, 10000 m, 30000 m, and 50000 m were selected as the distance values for the neighbor size of three autocovariate terms. These three autocovariate terms, i.e. Auto1, Auto3, and Auto5, were calculated and added separately into the Model O3, forming Model A1, A2, and A3 respectively, as shown in Table 18. Logistic regression was then conducted based on these three models, and their goodness-of-fit measures, in comparison with those of Model 3, are also displayed in Table 18. Figure 24 shows the comparison of AUC of these four models.

In order to further test whether by employing autologistic regression modeling spatial autocorrelation was effectively removed, Global Moran's I and Local Moran's I tests were again conducted based on the standardized residual produced by Model A1, which had the best performance among all three autologistic models. Results are displayed in Figure 25 and Figure 26 in comparison with the results of the same tests for Model O3.

Table 18: Comparison of the compositions and goodness-of-fit of the ordinary logistic regression model and the autologistic regression models.

Model	O3	A1	A2	A3
Variables	norm	norm	norm	norm
	d2s	d2s	d2s	d2s
	d2o	lulc7	d2o	el
	lulc4	**auto1**	lulc3	lulc4
	lulc5		lulc4	**auto5**
	lulc7		lulc5	
	cc1		lulc7	
			auto3	
AIC	1.173	0.837	1.011	1.026
AUC	0.7452	0.8869	0.8223	0.8165
Correctly Classified	70.67%	81.47%	74.54%	73.73%

Based on the coefficients for each variable after autologistic regression was fitted to A1, a map of predicted probability of feral swine presence was produced using the Kriging Interpolation Method in ArcGIS (Figure 27a). This map was compared with the one predicted by Model O3 (Figure 27b). The actual presence of feral swine was overlaid on top of these two maps to assess the accuracy of prediction made by these two models (Figure 27c and 27d). The probability map of feral swine predicted by Model A1 was then reclassified into two classes based on the cut-off probability of 0.5, as shown in Figure 28. Areas with probability larger than 0.5 indicate locations where local habitat is suitable for feral swine, and areas with probability less than 0.5 indicate locations where feral swine are unlikely to inhabit. A pie chart was also generated to compare the area of predicted favorable habitats of feral swine versus predicted unfavorable habitats.

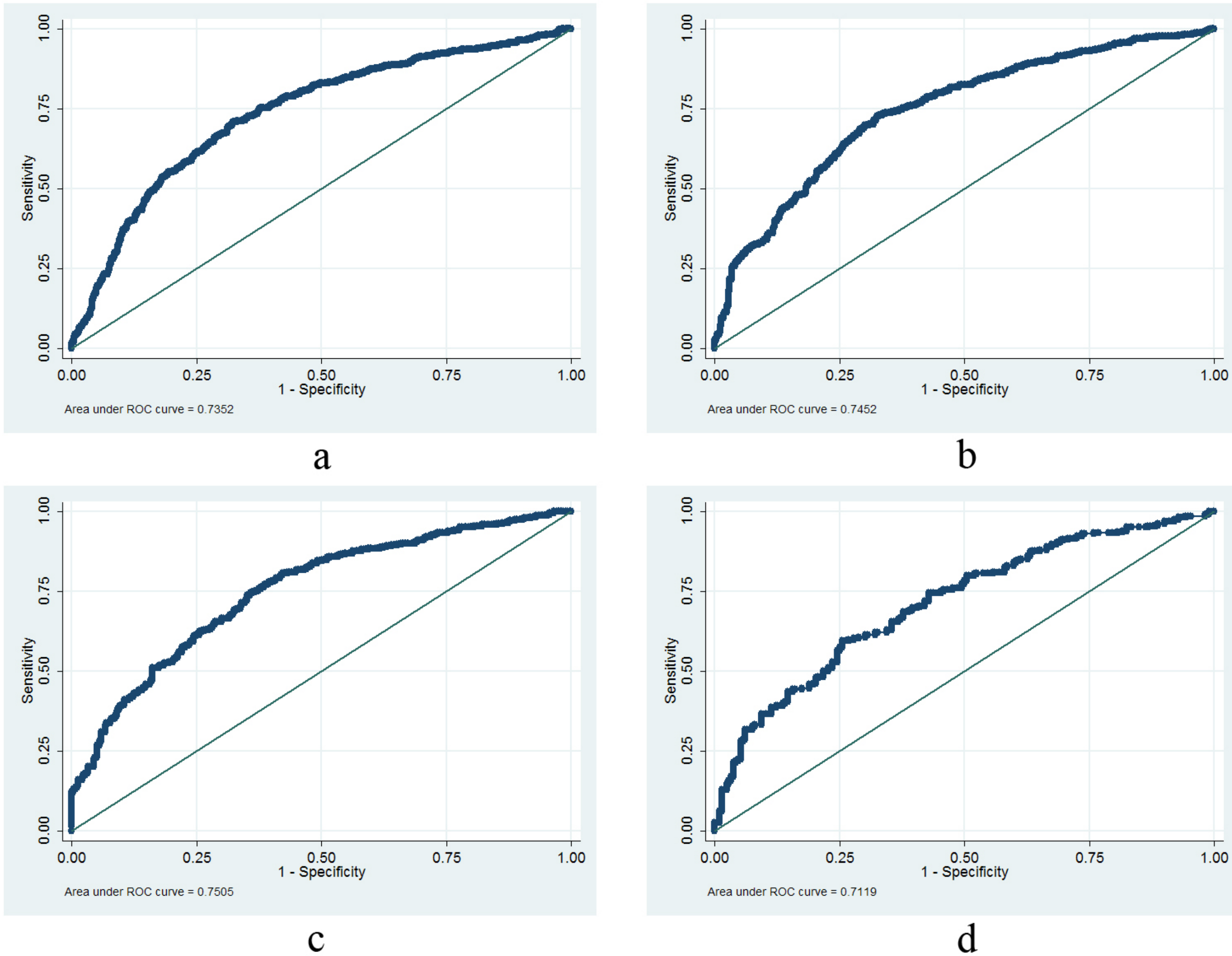


Figure 19: Comparison of ROC curves generated by four ordinary logistic models. a) Model O1, b) Model O2, c) Model O3, d) Model O4.

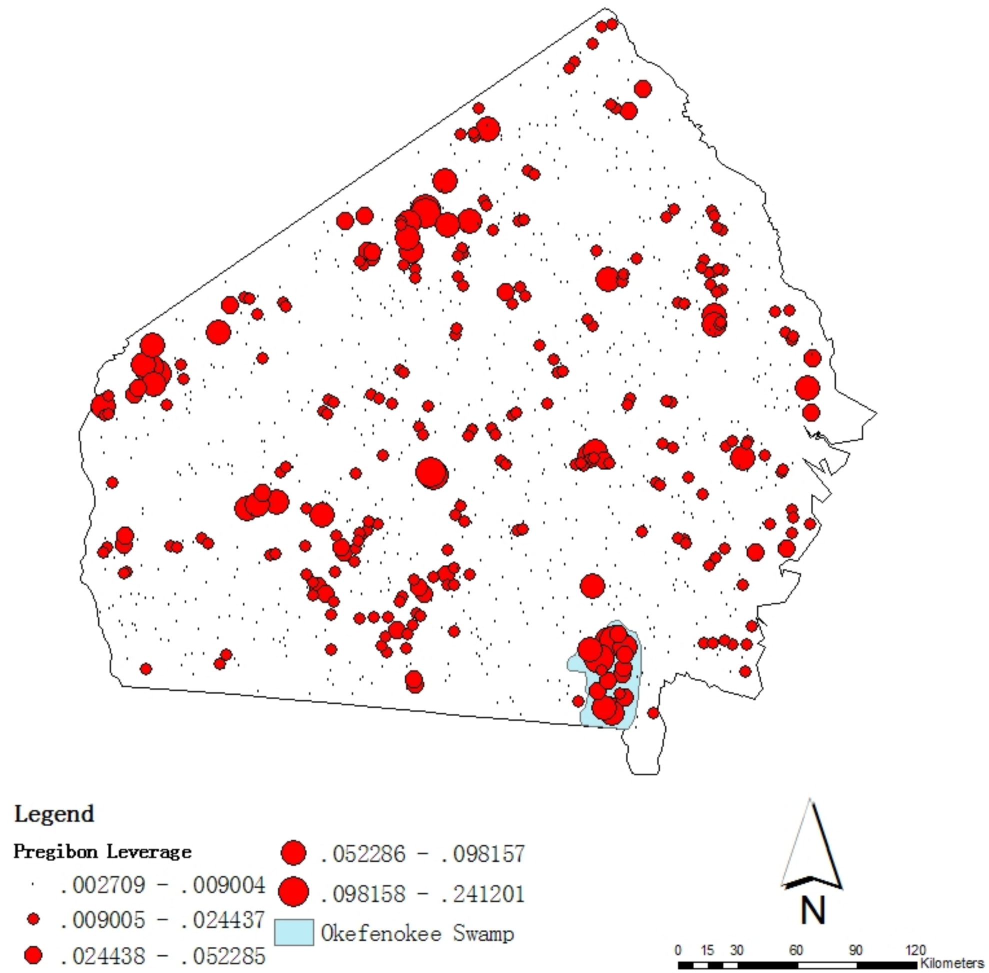


Figure 20: Distribution of the Pregibon leverage statistics based on the fitting results of Model O3. Red circles with larger area indicate higher leverage values. The blue area in the south is the Okefenokee Swamp.

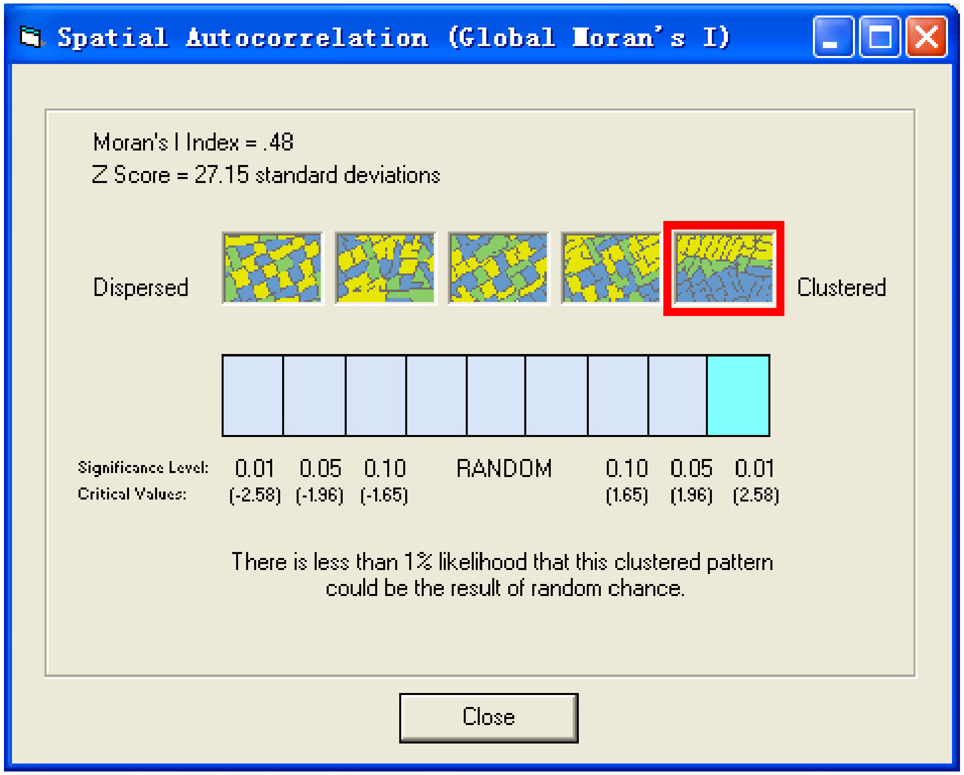


Figure 21: Graphic results of the Global Moran's I test based on the standardized residual term generated in ArcGIS.

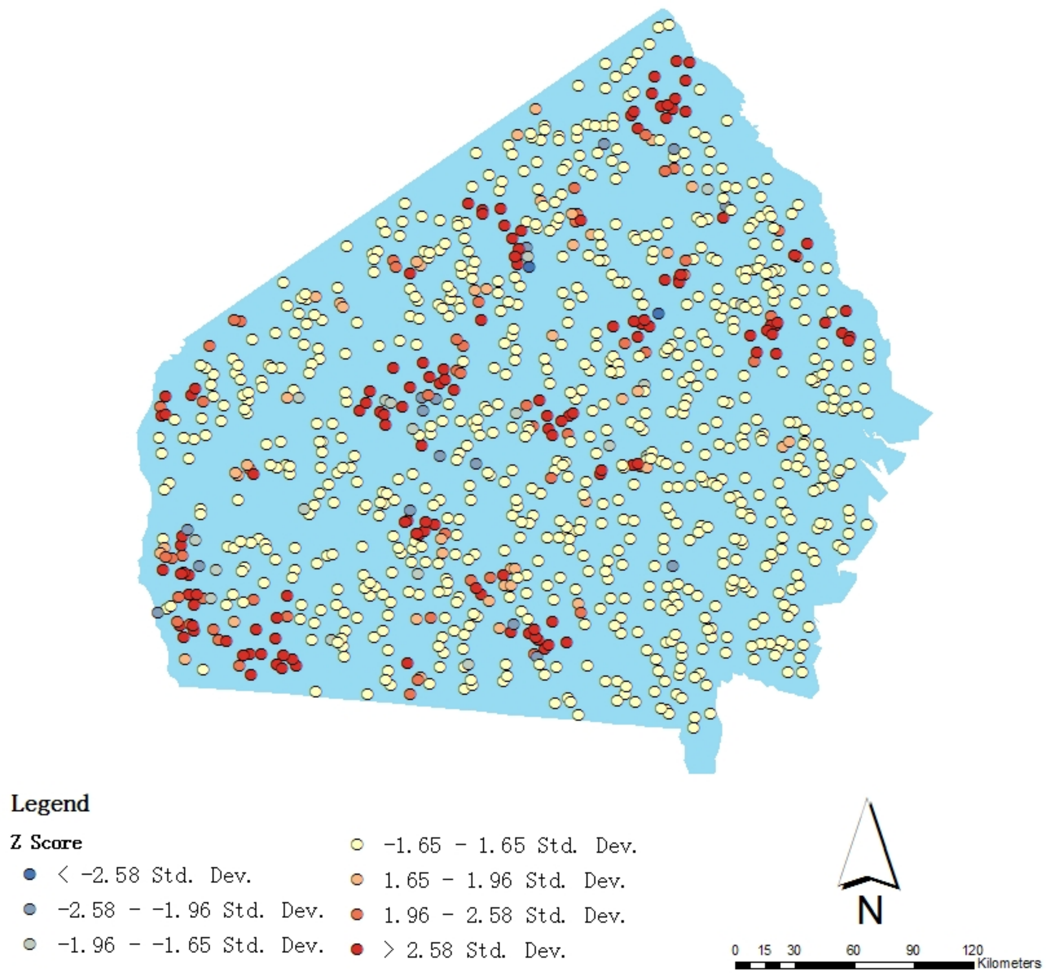


Figure 22: Graphic result of the Local Moran's I test based on the standardized residual term generated in ArcGIS. Red dots represent the strong clustered pattern and blue dots represent strong dispersed pattern.

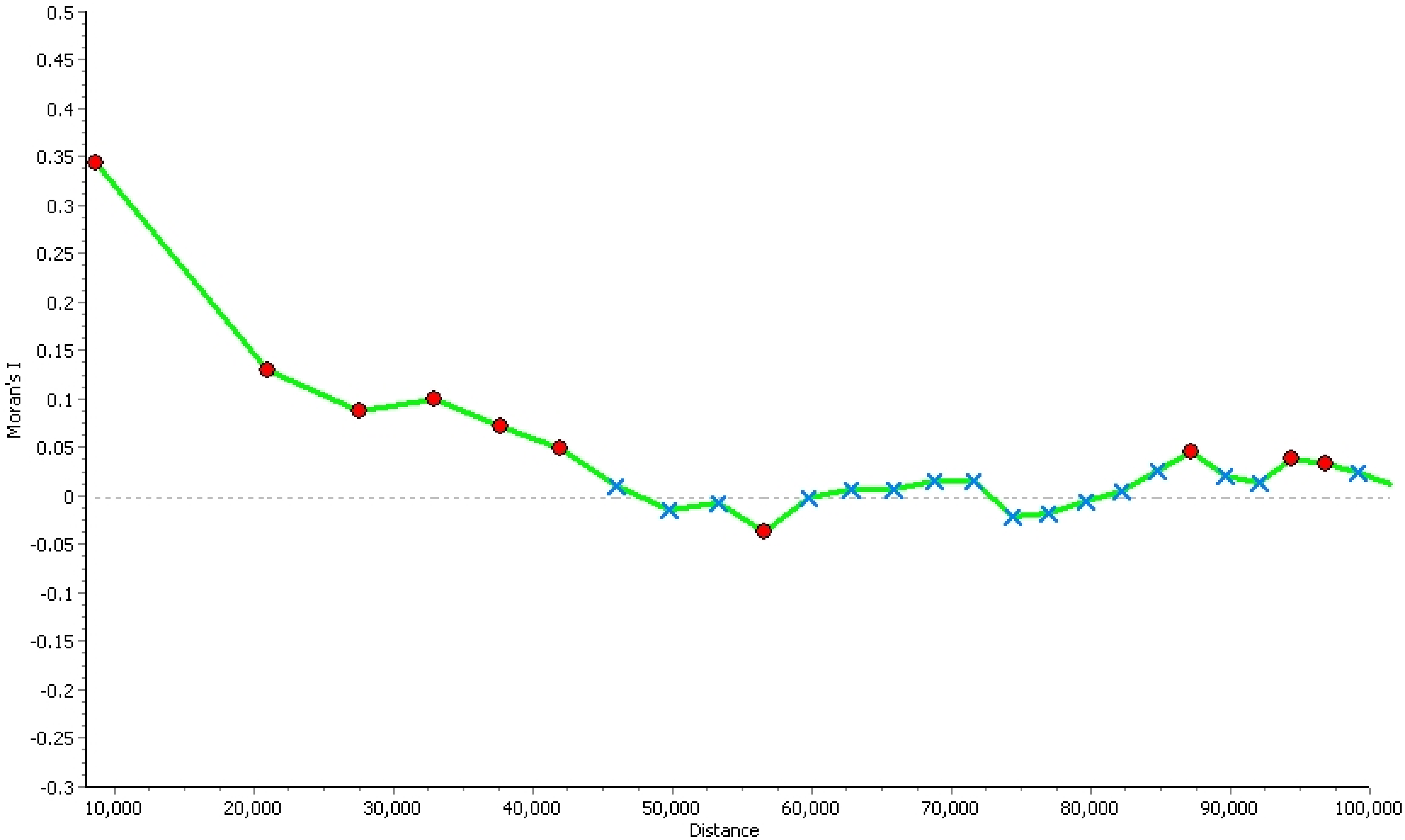


Figure 23: Correlogram created based on the standardized residual term. Red dots represent significant autocorrelated patterns, and blue crosses represent insignificant autocorrelated patterns.

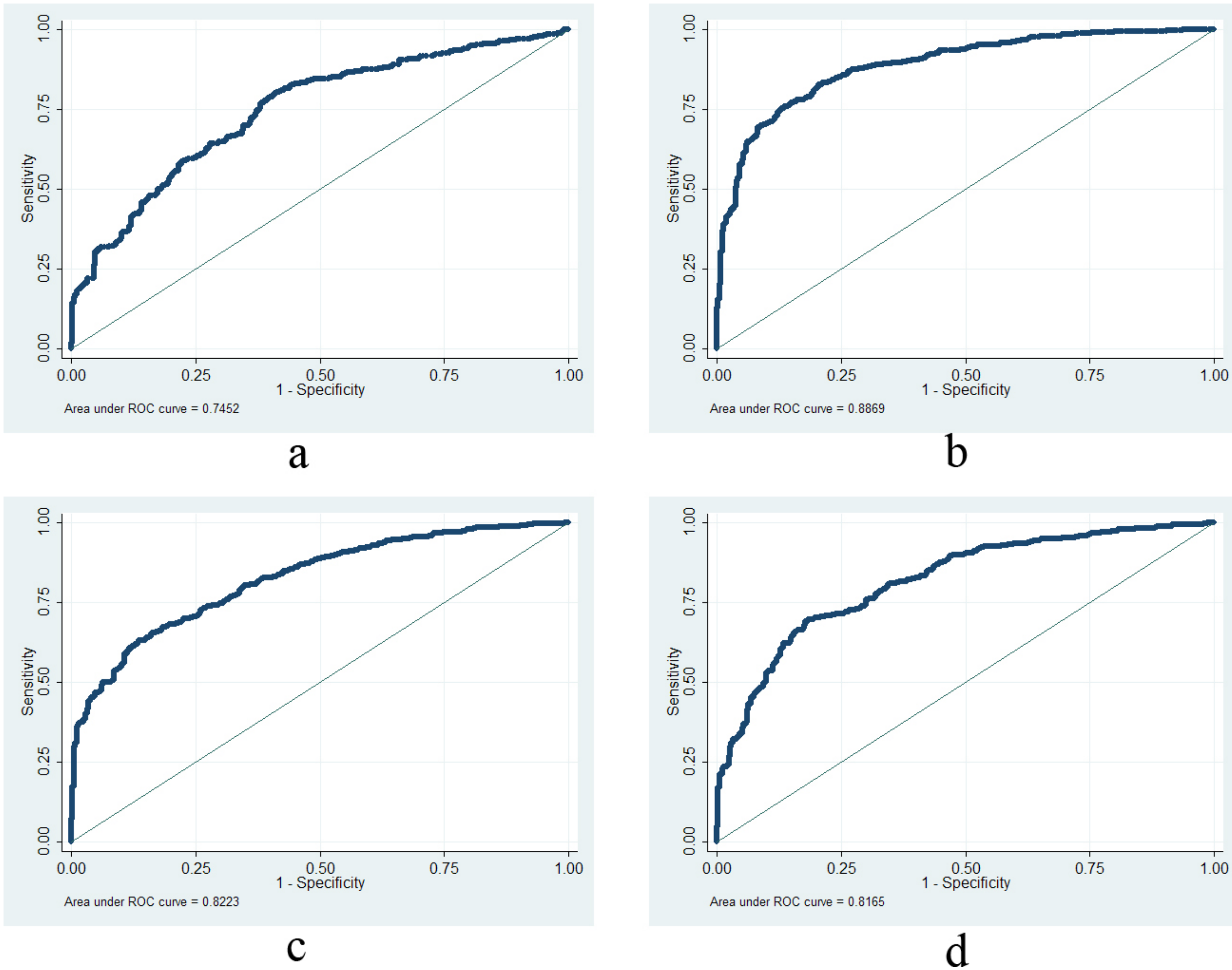
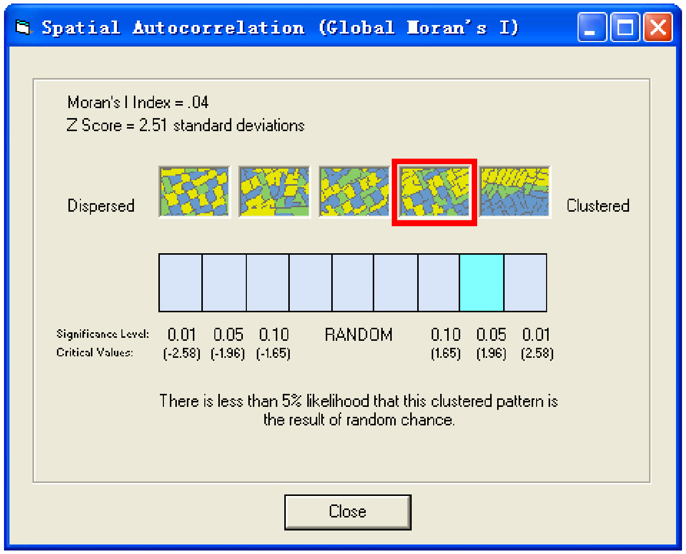
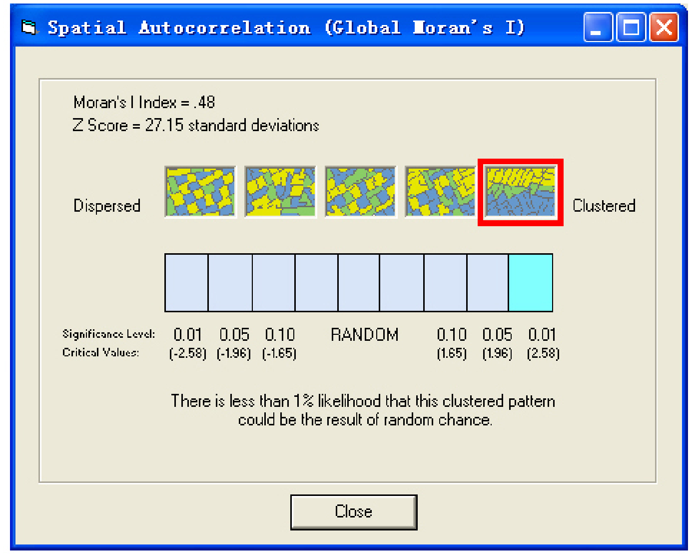


Figure 24: Comparison of ROC curves generated by both ordinary logistic and autologistic models. a) Model O3, b) Model A1, c) Model A2, d) Model A3.



a



b

Figure 25: Results of Global Moran's I tests based on standardized residual produced by a) Model A1, and b) Model O3.

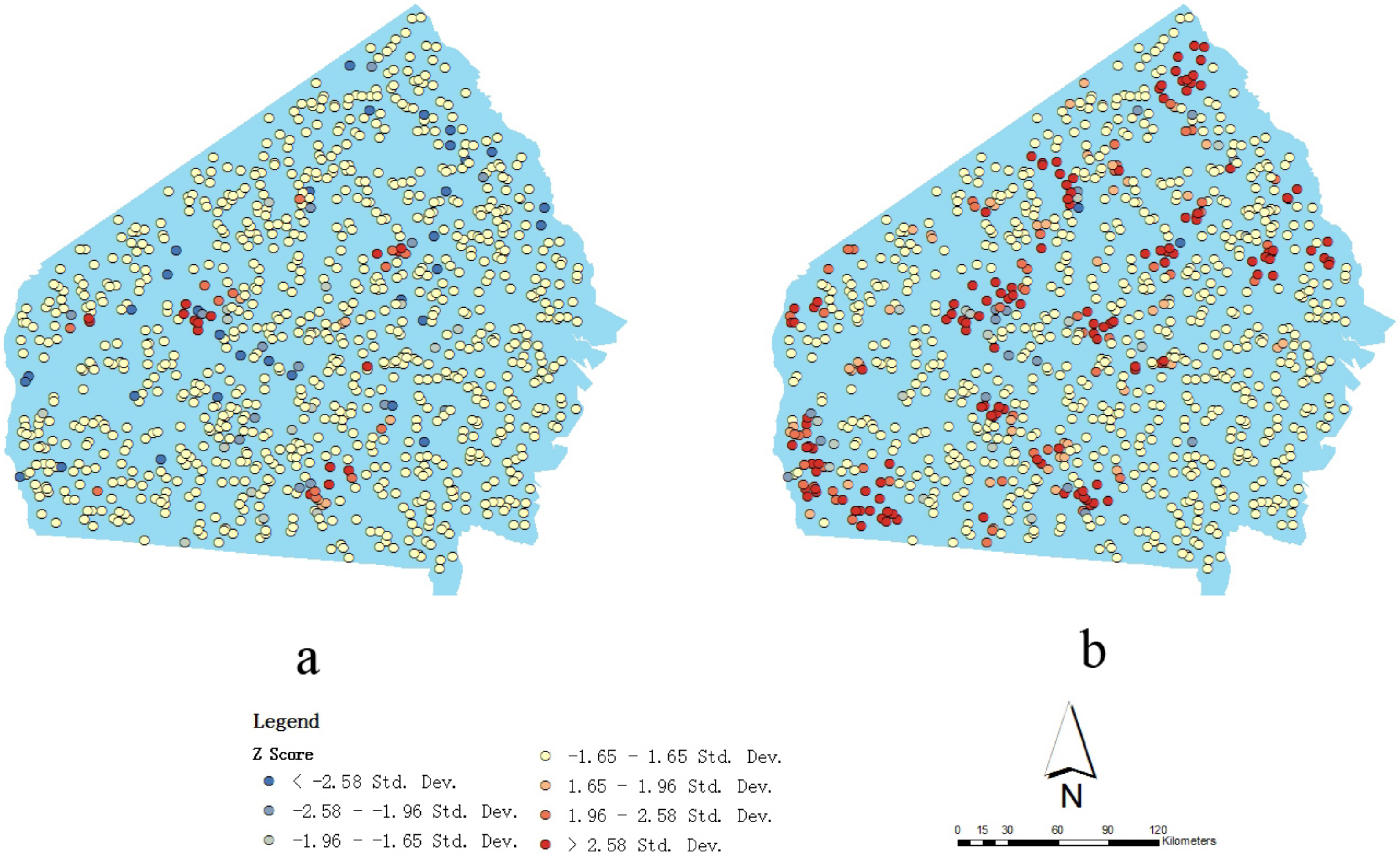


Figure 26: Results of Local Moran's I tests based on standardized residual produced by a) Model A1, and b) Model O3. Red dots represent the strong clustered pattern and blue dots represent strong dispersed pattern.

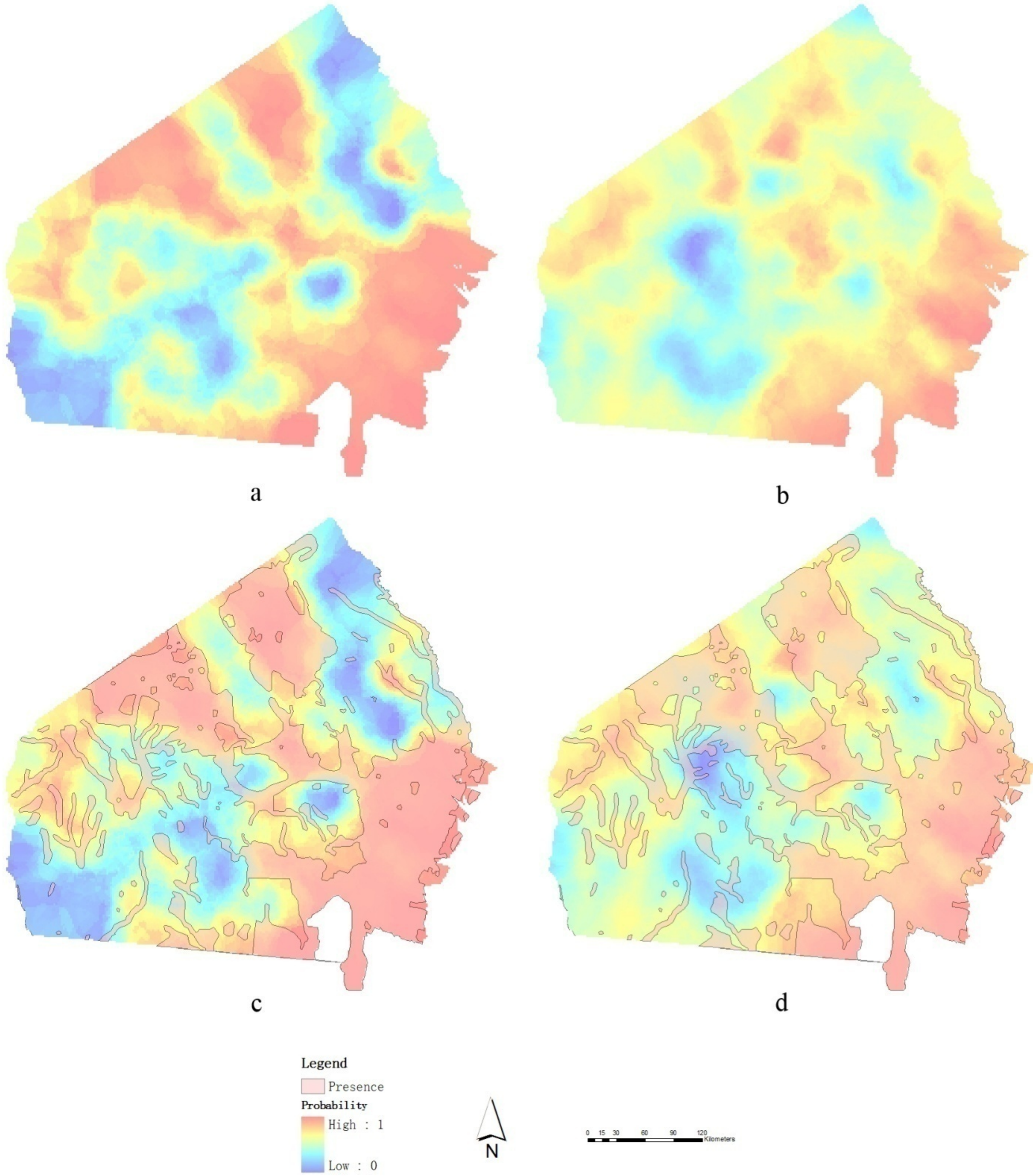


Figure 27: Maps of predicted probability of feral swine presence: a) prediction made based on Model A1, b) prediction made based on Model O3, c) prediction made based on Model A1 overlaid with actual presence of feral swine, d) prediction made based on Model O3 overlaid with actual presence of feral swine.

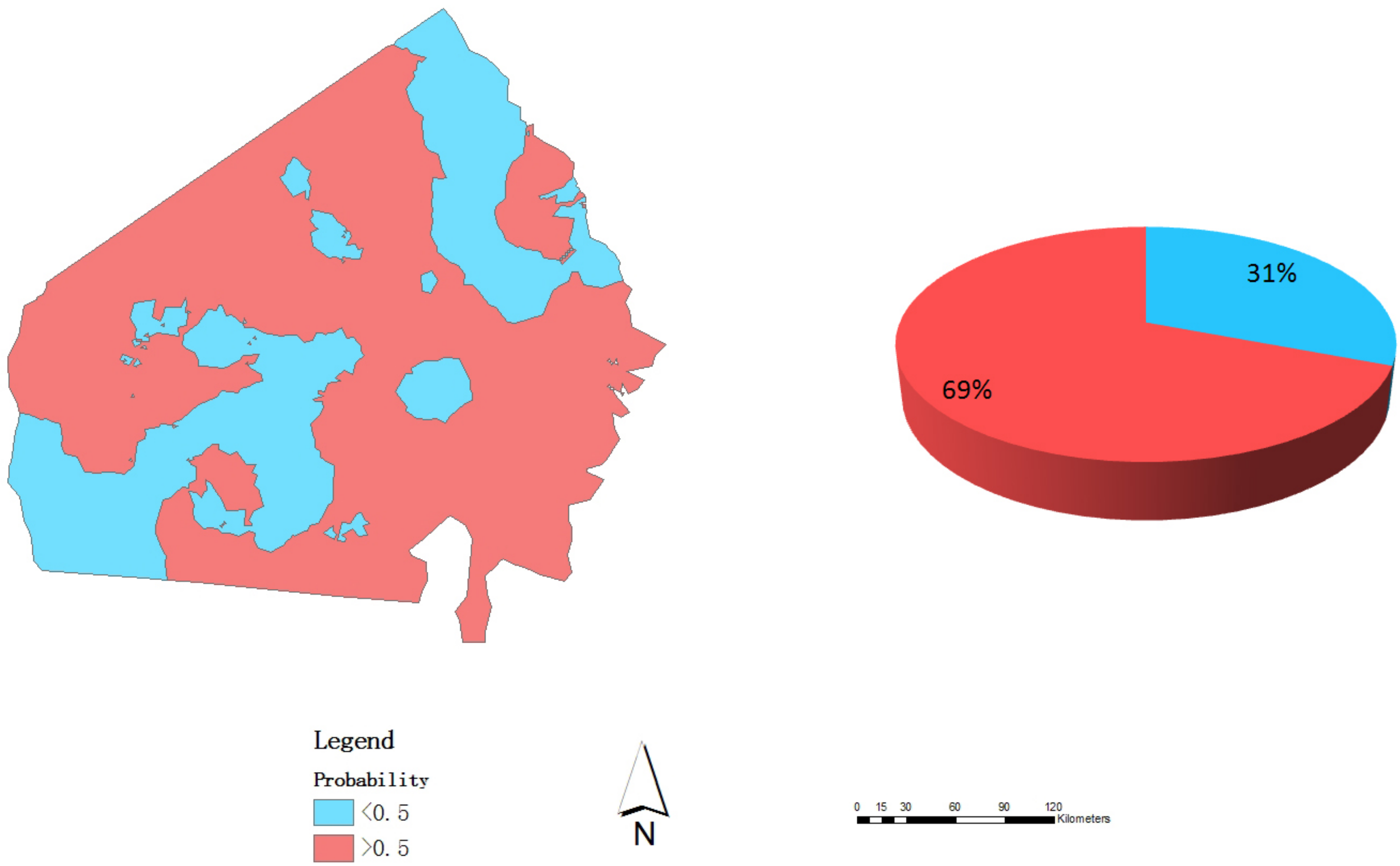


Figure 28: Map of predicted habitat suitability for feral swine in the study area based on Model A1 using 0.5 as cut-off probability value. Blue indicates areas where feral swine are likely to inhabit and red indicates areas where feral swine is unlikely to inhabit. The pie chart on the right shows the area composition of the predicted favorable habitats versus unfavorable habitats.

CHAPTER 6

DISCUSSION

Ordinary Logistic Regression v.s. Autologistic Regression

As shown in Table 18 and Figure 24, it is evident that all three autologistic models outperform the ordinary logistic model, especially in terms of AUC and classification accuracy. This accords with the assumption that autologistic regression models can produce superior results to ordinary logistic regression models when spatial autocorrelation exists since the former obeys the assumption of independence. Comparisons of the results of both Global and Local Moran's I tests before and after the autologistic regression was fitted was designed to test if spatial autocorrelation was effectively removed by autologistic regression. Slight autocorrelation still exists in the standardized residual term after the model was fitted, as shown in Figure 25a and Figure 26a (i.e. there are still several patches of red dots in the study area), and this may be due to the fact that the autocovariate term calculated in Model A1 is still incapable of capturing all spatial autocorrelation presented in the study area. Also it is entirely possible that in practice there is no single model that can remove spatial autocorrelation perfectly. In this regard, the autocorrelated pattern observed after Model A1 was fitted might simply be the results of random chance. Despite this, spatial autocorrelation has been considerably lessened by employing autologistic regression modeling, and there is a dramatic difference between the level of autocorrelated patterns presented in the study area before and after the model was fitted.

Figure 27 further compares the outcomes produced by the ordinary logistic regression model (O3) and its counterpart with an extra autocovariate term (A1), by comparing the maps of predicted probability of feral swine presence produced by them separately. The outcomes of these two models differ greatly throughout the study area. High probability of feral swine presence in the northern part of the study area is predicted by Model A1, however, Model O3 only suggests a slightly higher-than-normal probability of feral swine presence for the same area. According to Model O3, feral swine are extremely unlikely to inhabit the region slightly to the southwest of the center of the study area, while A1 reports that feral swine may live in that region because the calculated probability for that region fluctuates around 0.5. After the actual presence of feral swine was overlaid on top of these two maps, their prediction accuracy became apparent: the prediction map produced based on Model A1 coincides with the actual map quite well, with a number of small patches of feral swine presence having been captured by the model (e.g. high-probability region in the north); while the one produced by Model O3 clearly fails to correctly predict a considerable amount of the study area.

The comparison among the three models which took into account spatial autocorrelation reveals that the model performance decreases dramatically with the increase of neighboring size from 10000 meters to 30000 meters. This is in accordance with the phenomenon suggested by Figure 21 that within this range, the spatial autocorrelation experienced a significant decrease (i.e. Moran's I value decreases from more than 0.34 to 0.09). Model A1 was generated based on a smaller neighboring size, and therefore its capability of accounting for spatial autocorrelation is considerably stronger than Models A2 and A3. The performance of Model A2 was expected to be superior to Model A3, however, this is not the case. Their performances turned out to be quite similar in general, with the former being slightly superior in terms of AIC and AUC, while the latter outperforms insignificantly insofar as classification accuracy. This deviation from expectation might be due to the fact that spatial autocorrelation observed at 30000 meters is not considerably strong. After all, the observed Moran's I value at this distance is 0.09, which is not significantly greater than 0 at 50000 meters.

The various tests and comparisons conduced based on the model outcomes of ordinary logistic regression and autologistic regression in this project indicate that the existence of spatial autocorrelation jeopardizes the reliability of the traditional logistic regression modeling method. Although the influence of spatial autocorrelation on habitat modeling has been recently recognized, unfortunately a considerable amount of research of this kind is still ignoring this fact and employing the ordinary methods. While this study alone is insufficient to fully explore the impact that spatial autocorrelation has on habitat modeling and more similar studies are in great need, this study successfully discloses the drawbacks of the traditional method of habitat modeling using ordinary logistic regression. One conclusion can be drawn in this regard, that ordinary logistic regression method should be carried out in great caution in habitat modeling, as it may result in misleading results (highlighted by the comparison made by Figure 27). Habitat modeling should explicitly take into account spatial autocorrelation existing in various forms. At minimum, a test for spatial autocorrelation should be included.

Coefficient Interpretation

The influence of each variable involved in model fitting on the dependent variable is often evaluated by interpreting the raw regression coefficients, standardized coefficients or odds ratios. Raw coefficient represents the change in the dependent variable given one unit increase in the independent variable (Menard 1995). While it has been most commonly used, this measure has some limitations. For example, evaluation of the strength of the influence of the independent variables on the dependent variable using the raw coefficients is not straightforward when the independent variables are measured in different units. Therefore, standardized coefficients are proposed to assist the interpretation under circumstances like this. Standardized coefficients are modified raw coefficients which represent the change in standard deviation in the dependent variable given one standard deviation increase in the independent variable while holding all other variables constant (Menard 1995; Long, Freese et al. 2006). Odds ratio is another transformation of the raw coefficient which is friendly to interpret. It takes the form:

$$\mathrm{OR} = e^{\beta_i}$$

where OR represents odds ratio, e is the natural logarithm (2.718) and β_i is the raw coefficient of variable X_i. Odds ratio can be interpreted as the factor change in odds in the dependent variable associated with one unit increase in the independent variable while holding all other variables constant (Long, Freese et al. 2006). All three of these measures were used together to examine the relationship between the dependent variable and each independent variable in this project and the results for Model O3 and Model A1 are displayed in Table 19 and Table 20, respectively.

Table 19: Logistic regression analysis results for Model O3.

Variables	Coefficient	Standardized Coefficient	Odds Ratio	1/Odds Ratio
cc1	-0.64890	-0.1439	0.5226222	1.9134281
lulc5	-0.74578	-0.1774	0.4743663	2.1080756
lulc7	-1.12578	-0.2644	0.3243998	3.0826160
norm	0.00152	0.1383	1.0015230	0.9984793
d2s	-0.00018	-0.2620	0.9998206	1.0001794
d2o	-0.00006	-0.1325	0.9999362	1.0000638
lulc4	-1.84184	-0.1112	0.1585258	6.3081215

Table 20: Logistic regression analysis results for Model A1.

Variables	Coefficient	Standardized Coefficient	Odds Ratio	1/Odds Ratio
lulc7	-0.62550	-0.1152	0.5349917	1.8691879
d2s	-0.00017	-0.1932	0.9998312	1.0001688
norm	0.00235	0.1671	1.0023500	0.9976555
auto1	3.95028	0.5892	51.9498100	0.0192493

By examining the components of Model O3, which was selected as the optimal model among all four ordinary logistic models, several patterns can be revealed. Firstly, there are seven variables that have significantly influence on the distribution of feral swine, and among them, the NORM ED index is the only variable that is positively related to the feral swine distribution, i.e., increase in NORM ED index will increase the possibility of feral swine presence. This relationship suggests that feral swine may indeed tend to avoid areas with high road density, potentially due to the disturbing effects of traffic and other human activities associated with roads. Secondly, all other variables are negatively related to the possibility of feral swine presence. Among all canopy cover dummy variables, only cc1 (no canopy cover) was confirmed to correlated with feral swine distribution, and its negative influence is accordant with the assumption that feral swine prefer habitat with vegetation, although the further assumption, that feral swine prefer thick vegetation as their shelters was not able to be confirmed. Both distance to major streams and distance to oak-cypress-gum forests are negatively correlated with feral swine distribution, indicating major streams and forests containing oak species are indeed attractive to feral swine and they tend to live close to these two factors. The results related to the land cover dummy variables, lulc4 (deciduous forest), lulc5 (evergreen forest) and lulc7 (row crops and pastures) deviate from the previous expectation. It was expected that feral swine would prefer forest and agricultural areas, however, they were found negatively related to forest and agricultural LULC distribution. This obvious deviation may indicate the intrinsic problems (e.g., data inconsistency) of the datasets that were employed, but in order to confirm this assumption, further study with higher-quality land cover information is needed.

After the autocovariate term was added into the model, while the overall model performance increased dramatically, several variables including lulc4, lulc5, cc1 and d2o were dropped from the model composition. The remaining variables, d2s, norm, lulc7 retained their own direction of influence on feral swine distribution, i.e. NORM ED index still positively influences the distribution of feral swine while distance to major streams and the row crops and pastures land cover still possess the negative influences. With the spatial autocorrelation taken into account, the autocovariate term has become the most influential variable, and the difference in influential power between the autocovariate term and the other three terms is dramatic, i.e. the reciprocal of the odds ratio of cc1 (1.869), is considerable less than the odds ratio of auto1 (51.950). This means that whether feral swine are present at adjacent locations plays a strong role in influencing the probability of feral swine distribution at each point, and this relationship outweighs the relationship between feral swine distribution and other environmental factors. Moreover, this finding also supports the previous conclusion that spatial autocorrelation should be explicitly incorporated in habitat modeling.

Figure 27 shows the map of predicted presence/absence of feral swine based on the prediction made by Model A1 and 0.5 was used as the cut-off probability value. If 0.5 is arbitrarily selected as the threshold value for determining whether a certain location is a suitable habitat for feral swine, this map can further be interpreted as a map of suitable habitats of feral swine verse unsuitable habitats of feral swine. The red polygons, which represent the suitable habitats, take up 69% (66070 square kilometers) land area of the study area, while the blue polygons only take up 31% (29427 square kilometers). This explains to some degree why feral swine cause tremendous damage to Georgia: because Georgia has a large portion of land surfaces that are suitable for feral swine.

Scale Effects

This project attempts to address two questions that relate to the influence of scale: 1) is scale related to the performance of feral swine distribution models? And 2) is scale related to the performance of the autocovariate terms in counteracting the inherent spatial autocorrelation?

In order to address the first question, the models in this project were generated at four scales. The fact that Model O3 (5000 m scale) and Model O4 (10000 m scale) have superior performance to Model O1 (1000 m scale) and Model O2 (3000 m scale) suggests that at a broader scale, the feral swine distribution correlates better with the environmental factors. In addition, it is believed that evaluating the models generated at a scale even broader than 10000 m is unnecessary because the 10000 m scale has already exceeded the reported home range of feral swine. Thus there is no justification for using models generated at a broader scale even if they are numerically superior.

Through the comparison of these models, it is evident that habitat models are sensitive to scales. And similar to the fact that there is no single model that has the best performance for every species, there is no single optimal scale when modeling the distribution of a certain species. The question of upon which scale should the model be constructed for a species should be determined on case-by-case analyses.

The autocovariate terms were created at three scales in order to evaluate the scale effects on the explanatory power of the autocovariate terms. The results suggest that at a finer scale, the autocovariate terms can better counteract the negative influence of spatial autocorrelation. With the increase of the neighborhood size, the ability of the autocovariate term to remove spatial autocorrelation decreases. Since the purpose of this analysis is to explore whether there is a relationship between scale and power of the autocovariate terms, the current project does not attempt to determine the optimal scale at which a “best” autocovariate term should be constructed.

Research Considerations

The current study of modeling feral swine distribution has achieved satisfactory results. However, several considerations must be pointed out. First, the relationships between feral swine distribution and the environmental and cultural factors revealed by this study are not necessarily causal. Even if a particular location possesses all the desirable conditions for feral swine indicated by the habitat model, it does not mean there are definitely feral swine distributed at that location. Possible explanations include the fact that no feral swine have been introduced to that area or factors have thus far prevented their spread to that location. As indicated previously, the predicted probability of feral swine distribution can better be interpreted as the equivalent measure of habitat suitability for feral swine. Second, it is entirely possible that this study overlooked certain other environmental and cultural factors which may also play important role in influencing the distribution of feral swine. Constraints of time and resources for this study governed the number of variables that could be considered in this modeling effort. Future work, however, is definitely needed to obtain an even more thorough understanding of feral swine distribution by incorporating more variables.

Similarly, for those variables that were considered in this project, their influence on feral swine may differ from the conclusions made by this study under certain circumstances. For instance, although elevation has been reported by a variety of habitat modeling research as an influential factor directly related to the distribution of a large number of species (Wiser, Peet et al. 1998; Loyn, McNabb et al. 2001; Reese, Wilson et al. 2005), it was not confirmed by the current study. This indicates that, in general, feral swine distribution is not directly correlated with elevation, at least in the study area. However, as suggested by Ackerman, Harmon et al. (1978), feral swine may prefer high elevation areas in March and April in the Great Smoky Mountains due to the higher food availability there. Thus, when drawing insights from the current study, one should bear in mind that this study only discloses general relationships at a broad scale. Finally, there may be some special factors influencing the feral swine distribution only at a local scale that cannot be explained by the models discussed in this study. For instance, feral swine have been manually introduced to a large number of regions across the Southeastern United States mainly for recreational hunting purposes. Since this introduction phenomenon is most likely governed by different assumptions such as hunting preferences, ease of transport and opportunity for release, human-assisted spread of feral swine cannot be accounted for by the habitat models. Another example would be the effects of the Okefenokee Swamp on preventing the models from accurately predicting feral swine spread and inhabitation. A detailed examination of the geophysical conditions of the Okefenokee Swamp suggests that one potential reason that feral swine are not able to inhabit in this swamp is that unlike other regular swamps that can be occupied by feral swine, the Okefenokee Swamp is a deep-water swamp containing peat soils. The deep water pockets and the soft peat can effectively block feral swine from entering the Okefenokee Swamp (US Department of the Interior Fish and Wildlife Service 2006; Aicher 2011).

Error Sources

The distribution map of feral swine in 2004 is the pivotal dataset utilized in this project and it is assumed that this dataset is flawless. However, in reality this is impossible. Unlike the datasets that were produced from remotely sensed data and therefore possess a high level of consistency throughout the study area, the 2004 feral swine distribution map was compiled based on a number of data inputs from different sources within each state and in various formats. This may have resulted in the discrepancy in ways of interpretation of data inputs and levels of accuracy. For instance, the expert who reported the presence/absence status of feral swine in the field based their conclusion on the various traces left by feral swine, including body remains, food remains, signs of rooting, footprints, etc. Interpretation of these traces requires profound knowledge of feral swine and speculation which is often subjective. Since a large number of field experts were involved in reporting the feral swine distributions, it is inevitable that different standards were applied across the study area in determining whether particular evidence is indicative of feral swine distribution, thus causing the potential inconsistency resided in the distribution of feral swine dataset. Unfortunately, it was not possible to accomplish validation of this dataset within the range of this study.

In addition, several factors related to the other data layers may negatively affect the accuracy of the final results of this study. 1) Data accuracy. Similar to the limitations and uncertainties of the distribution of feral swine dataset, other data layers might be problematic in terms of accuracy. All datasets that were used in this project were of the best quality possible among all datasets available at the time of study. Owing to the impracticality of assessing the accuracy of each dataset, it could only be assumed that the datasets that were employed were sufficient in accuracy to yield justifiable results. Despite this, modelers must be aware of potential inaccuracies of the datasets input to predictive models. 2) Data inconsistency. The datasets in this project were produced separately by multiple organizations or agencies, and similar to the presence of feral swine dataset, they may suffer from issues of data inconsistency. One striking example in this regard would be the time discrepancies that existed among all datasets. Ideally, the condition described by all datasets that were employed should focus on the same time point, i.e. the year 2004. Since this is impractical to achieve, efforts were made to utilize those datasets that were closest to 2004 in time. For example, the NORM ED dataset in this project evaluates the road network condition in the United States in 2007 (USGS 2007), it was chosen for this project in that a) this was the only dataset available, and b) it was assumed that this particular information did not change significantly from 2004 to 2007.

Future Work

This project is the first attempt to explore the relationship between the distribution of feral swine and a comprehensive set of environmental factors at a relatively broad scale. Due to the time limit, further habitat suitability factors were not able to be incorporated in this project. Never-the-less, it certainly opens the door for future work to further understand this species-environment relationship, which may be focused on the following aspects. 1) Incorporating more environmental variables. It is desirable for habitat modeling to take into account more variables, even those which are seemingly irrelevant to the distribution of the particular species. Due to the various constraints described previously, only nine variables were selected as the environmental factors used in this project. Variables excluded by this project, such as landscape metrics, may improve the model outcomes. 2) Implementing time series analysis. Similar to spatial autocorrelation, which is a correlation of variables within a two dimensional surface, the attributes of some particular variables may be also strongly correlated in time. This gives rise to the study of time series analysis (Shumway and Stoffer 2006), and the availability of the distribution map of feral swine of multiple dates renders this analysis possible. Through time series analysis, the mechanisms behind the patterns in which feral swine distributions expanded through time with the change of environmental factors may be understood more thoroughly. 3) Expanding the study area. The current project focuses solely on the Coastal Plain Region in Georgia, which is only a small portion of the Coastal Plain areas in the Southeastern United States where feral swine are widely distributed. It might be worthwhile to study the relationship between feral swine distribution and the same set of environmental factors for some areas in other states and compare the results to see if they agree with each other. Since some of the datasets have been collected and processed by each state independently, such comparison may serve to validate the quantified relationship found in this study. 4) Integrating Agent-based Modeling techniques. Agent-based Modeling (ABM) is a newly emerging modeling method that directly focuses on individuals who are autonomous and takes action which consists of various patterns at a broader scale. Its unique nature enables it to tackle some of the complex questions that the classic modeling methods cannot (Parker, Manson et al. 2003; Brown, Aspinall et al. 2006). Because the distribution pattern of feral swine is an aggregated phenomenon based on all the actions made by individual feral swine living alone or in extended family groups known as sounders, by employing ABM and focusing on the behaviors of each individual, fair understanding of their habitat selection preference may be yielded.

CHAPTER 7

CONCLUSION

All the objectives of this project have been successfully accomplished. During the ordinary logistic regression stage, a set of nine variables were developed individually based on various datasets, including land cover, slope, elevation, NDVI, distance to major streams, NORM ED, distance to oak-gum-cypress forest, distance to impervious surfaces, and canopy cover. According to the characteristics of each dataset, different methods of data processing were applied at four different scales (i.e. 1000 m, 3000 m, 5000 m and 10000 m) in order to obtain data layers that were appropriate for the following statistical analysis. After the automatic backward elimination procedure, a few variables were dropped for models generated at each scale. Through comparison between the goodness-of-fit of all the models, Model O3 was selected as the optimal model during the ordinary logistic regression stage, and it was based on eight variables (i.e., norm, d2s, d2o, lulc2, lulc4, lulc5, lulc7 and cc1) at the scale of 5000 meters. This model had a medium performance in terms of predicting the feral swine distribution with AUC being 0.7505 and prediction accuracy of approximately 70%. A general relationship between the feral swine distribution and these environmental factors was therefore obtained. Namely, feral swine prefer habitat that has less road density, is covered by vegetation and is close to major streams and oak-cypress-gum forests. In addition, the results suggested that the existence of forest and agricultural areas negatively influence the distribution of feral swine, which contradicted with our expectation. Although it is likely that this was the result of data inaccuracies, further and more thorough analysis should be conducted to confirm this assumption.

In order to test whether spatial autocorrelation existed and whether this influenced the model performance, several tests of spatial autocorrelation were explicitly employed based on the residuals calculated using the selected optimal model. Results from multiple perspectives indicated that strong autocorrelated patterns were present. Under this scenario the assumption of independence of the ordinary logistic regression method might be jeopardized and lead to unjustifiable conclusions. Given this, autologistic regression was employed by constructing an autocovariate term which accounted for spatial autocorrelation and added the autocovariate term to the optimal model. The fitting results of the autologistic model were improved significantly compared to those of the ordinary logistic models. Its prediction accuracy increased up to more than 80%, and AUC increased even more dramatically to 0.8869.

In addition to the quantitative comparison, the difference in predictive power of these two models was also compared visually by producing the prediction maps based on each model separately and comparing them with the actual distribution of feral swine. Such a comparison revealed that the ordinary logistic model was significantly restrained and affected by the presence of spatial autocorrelation, and it was, therefore, problematic under such a circumstance. This discovery led to one of the conclusions of this study, that is, the ordinary logistic regression modeling methods may suffer from problems such as the presence of spatial autocorrelation, under which circumstance the results might be dubious and misleading. It is strongly recommended that any habitat modeling should adopt a stage of testing spatial autocorrelation as one of the routine procedures, based on which undesirable spatial autocorrelation can be detected and solutions to remove such a negative influential factor can be established.

The concept of scale was explicitly incorporated in this study and its role in habitat modeling and influence on the significance of spatial autocorrelation was analyzed. During the ordinary logistic regression stage, four models were established at four scales: 1000, 3000, 5000 and 10000 meters, respectively. Their goodness-of-fit were then compared and the model based on 5000 meters scale seemed to be superior to the other counterparts. Although this study was unable to fully explore the relationship between the scale factor and model performance, it was confirmed that the scale factors play a role influencing the performance of the corresponding models and it is something that needs to be taken into account when constructing meaningful habitat models. In addition, this study suggested that the supplementary predictive power added by incorporating an autocovariate terms differs for autocovariate terms generated at different scales. Based on the current study, it seemed the autocovariate term created at the distance where spatial autocorrelation is the most evident possesses the largest predicting power, but further and more thorough analyses should be performed to yield a thorough understanding of the relationship between scale and the power of the autocovariate terms.

This study successfully revealed the less-known relationship between a series of environmental factors and the distribution of feral swine at a broad scale, and this information might be valuable for agencies or experts who seek to monitor and control the spreading of feral swine effectively and efficiently. It is believed this obtained quantitative relationship may be able to be applied to a vast area in the Southeastern United States due to the fact that a considerable amount of land surfaces in the southeast share the same characteristics in terms of environmental conditions. Moreover, the methods proposed in this study can be applied when modeling other exotic species. The current study also leaves room for further analyses which are expected to improve the model. Future working might be focused on incorporating more environmental variables, adapting time series analysis, expanding the study area, and integrating Agent-based Modeling techniques.

REFERENCES

Ackerman, B. B., M. E. Harmon, et al. (1978). Seasonal food habits of the European wild hog. Part II. Studies of the European wild boar in Great Smoky Mountains National Park. Report to Superintendent. Gatlinburg, TN, USA, Uplands Field Research Laboratory.

Aicher, S. (2011). Okefenokee National Wildlife Refuge. personal communication.

Allouche, O., A. Tsoar, et al. (2006). "Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS)." Journal of Applied Ecology **43**(6): 1223-1232.

Ambrosini, R., A. M. Bolzern, et al. (2002). "The distribution and colony size of barn swallows in relation to agricultural land use." Journal of Applied Ecology **39**(3): 524-534.

Anderson, J. R., E. E. Hardy, et al. (1976). A land use and land cover classification system for use with remote sensor data. Geological Survey Professional Paper 964, Washington DC, USA, United States Government Printing Office.

Anselin, L. (1995). "Local indicators of spatial association - LISA." Geographical Analysis **27**(2): 93-115.

Augustin, N. H., M. A. Mugglestone, et al. (1996). "An autologistic model for the spatial distribution of wildlife." Journal of Applied Ecology **33**(2): 339-347.

Bailey, T. and T. Gatrell (1996). Interactive Spatial Data Analysis. Harlow, UK, Prentice Hall.

Bailey, T. C. and A. C. Gatrell (1995). Interactive Spatial Data Analysis, Prentice Hall.

Behrens, C. E. (2010). Landsat and the Data Continuity Mission. CRS Report for Congress, Congressional Research Service.

Bernhardsen, T. (2002). Geographic Information Systems: An Introduction. New York, NY, USA, John Wiley & Sons.

Betts, M. G., A. W. Diamond, et al. (2006). "The importance of spatial autocorrelation, extent and resolution in predicting forest bird occurrence." Ecological Modelling **191**(2): 197-224.

Bissonette, J. A. and I. Storch (2003). Landscape Ecology and Resource Management: Linking Theory With Practice. Washington DC, USA, Island Press.

Blood, D. C., O. M. Radostits, et al. (1983). Veterinary Medicine. London, England, Bailhere Tindall.

Borcard, D., F. Gillet, et al. (2011). Numerical Ecology with R. New York, NY, USA, Springer.

Boyce, M. S. and L. L. McDonald (1999). "Relating populations to habitats using resource selection functions." Trends in Ecology & Evolution **14**(7): 268-272.

Brown, D. G., R. Aspinall, et al. (2006). "Landscape models and explanation in landscape ecology- A space for generative landscape science?" Professional Geographer **58**(4): 369-382.

Burel, F. and J. Baudry (2003). Landscape Ecology: Concepts, Methods, and Applications. Enfield, NH, USA, Science Publishers.

Burrough, P. A. and R. A. McDonnell (1998). Principles of Geographical Information Systems. Oxford, UK, Oxford University Press.

Carroll, M. L., C. M. DiMiceli, et al. (2004). 250m MODIS Normalized Difference Vegetation Index. College Park, Maryland, University of Maryland.

Cerezo, A., S. Perelman, et al. (2010). "Landscape-level impact of tropical forest loss and fragmentation on bird occurrence in eastern Guatemala." Ecological Modelling **221**(3): 512-526.

Chatterjee, S. and A. S. Hadi (2006). Regression Analysis by Example. Hoboken, NJ, USA, John Wiley & Sons.

Choquenot, D., J. McIlroy, et al. (1996). Managing Vertebrate Pests: Feral Pigs. Canberra, Bureau of Resource Sciences Australian Government Publishing Service.

Christensen, R. (1997). Log-Linear Models and Logistic Regression. New York, NY, USA, Springer.

Clark, W. A. V. and P. L. Hosking (1986). Statistical Methods for Geographers. New York, NY, USA, Wiley.

Clarke, K. C. (2000). Getting Started with Geographic Information Systems. New Jersey, NJ, USA, Prentice Hall.

Corn, J. L., J. C. Cumbee, et al. (2009). "Pathogen exposure in feral swine populations geographically associated with high densities of transitional swine premises and commercial swine production." Journal of Wildlife Diseases **45**(3): 713-721.

Corn, J. L., D. E. Stallknecht, et al. (2004). "Persistence of pseudorabies virus in feral swine populations." Journal of Wildlife Diseases **40**(2): 307-310.

De Frutos, A., P. P. Olea, et al. (2007). "Analyzing and modelling spatial distribution of summering lesser kestrel: The role of spatial autocorrelation." Ecological Modelling **200**(1-2): 33-44.

Decker, E. (1978). Exotics. Big game of North America: Ecology and Management. J. L. Schmidt and D. L. Gilbert. Harrisburg, PA, USA, Stackpole Books**:** 249-256.

Ecol, P. (1992). Investigation of feral pig populations and control measures: Cape Tribulation Section of the wet tropics World Heritage Area. Cairns, Queensland, Australia, Consultancy report to Wet Tropics Management Authority.

Engeman, R. M., B. Constantin, et al. (2001). "Monitoring changes in feral swine abundance and spatial distribution." Environmental Conservation **28**(3): 235-240.

Engeman, R. M., H. T. Smith, et al. (2003). "Prevalence and economic value of feral swine damage to native habitat in three Florida state parks." Environmental Conservation **30**(4): 319-324.

Engeman, R. M., A. Stevens, et al. (2007). "Feral swine management for conservation of an imperiled wetland habitat: Florida's vanishing seepage slopes." Biological Conservation **134**(3): 440-446.

ESRI (2009). ArcGIS Desktop Help 9.3.1. Redlands, CA, USA, ESRI.

Fenner, F., P. A. Bachmann, et al. (1987). Herpesviridae. Veterinary virology. F. Fenner, P. A. Bachmann, E. P. J. Gibbset al. San Diego, CA, USA, Academic Press, Inc**:** 353-356.

Fielding, A. H. and 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.

Fletcher, W. O., T. E. Creekmore, et al. (1990). "A field trial to determine the feasibility of delivering oral vaccines to wild swine." Journal of Wildlife Diseases **26**(4): 502-510.

Foresman, T. W. (1998). The History of Geographic Information Systems: Perspectives from the Pioneers. Upper Saddle River, NJ, USA, Prentice Hall.

Forman, R. T. T. (1995). Land Mosaics: The Ecology of Landscapes and Regions. Cambridge, UK, Cambridge University Press.

Forman, R. T. T. and M. Godron (1986). Landscape ecology. New York, NY, USA, Wiley.

Frederick, J. M. (1998). "Overview of wild pig damage in California." Proceedings of the Vertebrate Pest Conference **18**: 82-86.

Fuller, R. M., G. B. Groom, et al. (1994). "The land-cover map of Great Britain: An automated classification of Landsat thematic mapper data." Photogrammetric Engineering and Remote Sensing **60**(5): 553-562.

Gaines, K. F., D. E. Porter, et al. (2005). "A spatially explicit model of the wild hog for ecological risk assessment activities at the department of energy's Savannah river site." Human and Ecological Risk Assessment **11**(3): 567-589.

Georgia State Climate Office. (1998). "Climatology Of The Georgia Coastal Plain." from http://climate.engr.uga.edu/pubs/coastal.pdf.

Gesch, D., Oimoen, M., Greenlee, S., Nelson, C., Steuck, M., and Tyler, D. (2002). The National Elevation Dataset: Photogrammetric Engineering and Remote Sensing. **68:** 5-11.

Gesch, D. B. (2007). The National Elevation Dataset. Digital Elevation Model Technologies and Applications: The DEM Users Manual. D. Maune. Bethesda, Maryland, American Society for Photogrammetry and Remote Sensing**:** 99-118.

Ghilani, C. D. and P. R. Wolf (2002). Elementary Surveying: An Introduction to Geomatics. Englewood Cliffs, NJ, USA, Prentice Hall.

Gibson, L. A., B. A. Wilson, et al. (2004). "Modelling habitat suitability of the swamp antechinus (Antechinus minimus maritimus) in the coastal heathlands of southern Victoria, Australia." Biological Conservation **117**(2): 143-150.

Gibson, L. A., B. A. Wilson, et al. (2004). "Spatial prediction of rufous bristlebird habitat in a coastal heathland: a GIS-based approach." Journal of Applied Ecology **41**(2): 213-223.

Global Land Cover Facility. (2010). "Landsat Imagery." from http://www.glcf.umd.edu/data/landsat/.

Goward, S. N. and D. L. Williams (1997). "Landsat and Earth Systems Science: Development of terrestrial monitoring." Photogrammetric Engineering and Remote Sensing **63**(7): 887-900.

Griffith, D. A. (2003). Spatial Autocorrelation and Spatial Filtering. Heidelberg, Germany, Springer.

Guisan, A. and N. E. Zimmermann (2000). "Predictive habitat distribution models in ecology." Ecological Modelling **135**(2-3): 147-186.

Guo, Q., S. Y. Chen, et al. (2010). "Covariance intersection based image fusion technique with application to pansharpening in remote sensing." Information Sciences **180**(18): 3434-3443.

Hinely, A. J. (2006). GIS-based Habitat Modeling Related to Bearded Capuchin Monkey Tool Use. Department of Geography. Athens, GA, University of Georgia. **Master:** 90.

Hodler, T. W. and H. A. Schretter (1986). The Atlas of Georgia. Athens, GA, USA, The Institute of Community and Area Development - The University of Georgia.

Hosmer, D. W. and S. Lemeshow (2000). Applied Logistic Regression. New York, Wiley.

Huete, A., K. Didan, et al. (2002). "Overview of the radiometric and biophysical performance of the MODIS vegetation indices." Remote Sensing of Environment **83**(1-2): 195-213.

Jiang, G. S., J. Z. Ma, et al. (2009). "Multiple spatial-scale resource selection function models in relation to human disturbance for moose in northeastern China." Ecological Research **24**(2): 423-440.

Keitt, T. H., O. N. Bjornstad, et al. (2002). "Accounting for spatial pattern when modeling organism-environment interactions." Ecography **25**(5): 616-625.

Kleinbaum, D. G., M. Klein, et al. (2010). Logistic Regression: A Self-learning Text. New York, NY, USA, Springer.

Knapp, R. A., K. R. Matthews, et al. (2003). "Developing probabilistic models to predict amphibian site occupancy in a patchy landscape." Ecological Applications **13**(4): 1069-1082.

Lang, M. and M. Mengak (2007). Managing wildlife damage: Feral swine. Wildlife Management, Warnell School of Forestry and Natural Resources, University of Georgia. **WMS-07-14**.

Legendre, P. (1993). "Spatial Autocorrelation: Trouble or New Paradigm?" Ecology **74**(6): 1659-1673.

Li, Y. C. (2008). "Land cover dynamic changes in northern China: 1989-2003." Journal of Geographical Sciences **18**(1): 85-94.

Lillesand, T. M., R. Kiefer, W., et al. (2007). Remote Sensing and Image Interpretation. Hoboken, NJ, USA, John Wiley & Sons, Inc.

Linzey, D. W. (2008). A Natural History Guide: Great Smoky Mountains National Park. Knoxville, The University of Tennessee Press.

Liu, J. and W. W. Taylor (2002). Integrating Landscape Ecology into Natural Resource Management. Cambridge, UK, Cambridge University Press.

Long, J. S., J. Freese, et al. (2006). Regression Models for Categorical Dependent Variables Using Stata. College Station, TX, USA, Stata Press.

Longley, P. A., M. Goodchild, et al. (2005). Geographical Information Systems and Science. Chichester, England, John Wiley & Sons.

Lopez-Lopez, P., C. Garcia-Ripolles, et al. (2006). "Modelling breeding habitat preferences of Bonelli's eagle (Hieraaetus fasciatus) in relation to topography, disturbance, climate and land use at different spatial scales." Journal of Ornithology **147**(1): 97-106.

Loyn, R. H., E. G. McNabb, et al. (2001). "Modelling landscape distributions of large forest owls as applied to managing forests in north-east Victoria, Australia." Biological Conservation **97**(3): 361-376.

Madden, M., H. Zhao, et al. (2009). Context-aware analysis, geovisualization and virtural globes for managing Earth resources. International Archives of Photogrammetry and Remote Sensing. J.Gong and Q. Zhou. **38 Part 4/W10**.

Maggini, R., A. Guisan, et al. (2002). "A stratified approach for modeling the distribution of a threatened ant species in the Swiss National Park." Biodiversity and Conservation **11**(12): 2117-2141.

Mapston, M. A. (2004). Feral hogs in Texas. Bulletin 6149, Texas Cooperative Extension. College Station, Texas, USA.

McKnight, T. (1964). Feral Livestock in Anglo-America. Berkley, CA, USA, University of California Press.

Menard, S. (1995). Applied Logistic Regression Analysis. Thousand Oaks, CA, USA, Sage Publications.

Mengeling, W. L. and E. C. Pirtle (1989). Susceptibility of feral swine to acute and latent infections with pseudorabies virus. Proceedings of the Feral Pig Symposium, Orlando, FL, USA, Livestock Conservation Institute.

Microsoft (2007). Microsoft Excel. Redmond, WA, USA, Computer Software.

Mitchell, M. S., R. A. Lancia, et al. (2001). "Using landscape-level data to predict the distribution of birds on a managed forest: Effects of scale." Ecological Applications **11**(6): 1692-1708.

Moran, P. A. P. (1950). "Notes on continuous stochastic phenomena." Biometrika **37**: 17-23.

Morrison, M. L., B. G. Marcot, et al. (1998). Wildlife-habitat Relationships: Concepts and Applications. Madison, WI, USA, University of Wisconsin Press.

Mountrakis, G. and L. Luo (2010). "Enhancing and replacing spectral information with intermediate structural inputs: A case study on impervious surface detection." Remote Sensing of Environment **115**(5): 1162-1170.

Mungall, E. C. (2001). Exotics. Ecology and management of large mammals in North America. S. Demarais and P. R. Krausman. Upper Saddle River, NJ, USA, Prentice Hall.

NASA. (2011). "Technical Information of Landsat Program." from http://landsat.gsfc.nasa.gov/about/technical.html.

NASS (1991). Sheep and goat predator loss. Washington, D. C., USA, USDA, Agricultural Statistics Board.

Natural Resources Spatial Analysis Lab. (2011). "Frequently Asked Questions." from http://narsal.uga.edu/projects/glut/faq.

Natural Resources Spatial Analysis Lab. (2011). "Landcover Descriptions." from http://narsal.uga.edu/projects/glut/classdescriptions.

O'Neill, R. V., C. T. Hunsaker, et al. (1996). "Scale problems in reporting landscape pattern at the regional scale." Landscape Ecology **11**(3): 169-180.

Oliveira, F. S. C., M. Kampel, et al. (2008). "Multitemporal assessment of the geomorphologic evolution of the Restinga of Marambaia, Rio de Janeiro, Brazil." International Journal of Remote Sensing **29**(19): 5585-5594.

Olsen, S. C. (2010). "Brucellosis in the United States: Role and significance of wildlife reservoirs." Vaccine **28**: F73-F76.

Osborne, P. E., J. C. Alonso, et al. (2001). "Modelling landscape-scale habitat use using GIS and remote sensing: a case study with great bustards." Journal of Applied Ecology **38**(2): 458-471.

Parker, D. C., S. M. Manson, et al. (2003). "Multi-agent systems for the simulation of land-use and land-cover change: A review." Annals of the Association of American Geographers **93**(2): 314-337.

Pearce, J. and S. Ferrier (2000). "Evaluating the predictive performance of habitat models developed using logistic regression." Ecological Modelling **133**(3): 225-245.

Pensaert, M. B. and J. P. Kluge (1989). Pseudorabies virus (Aujeszky's disease). Virus Infections of Porcines. M. B. Pensaert. New York, NY, USA, Elsevier Science & Technology**:** 39-64.

Pereira, J. M. C. and R. M. Itami (1991). "GIS-Based habitat modelling using logistic multiple regression: A study of the Mt. Graham red squirrel." Photogrammetric Engineering and Remote Sensing **57**(11): 1475-1486.

Pimentel, D., L. Lach, et al. (2002). Environmental and economic costs associated with non-indigenous species in the United States. Biological Invasions: Economic and

Environmental Costs of Alien Plant, Animal, and Microbe Species. Boca Raton, FL, USA, CRC Press: 285-303.

Plant, J. W., R. Marchant, et al. (1978). "Neonatal lamb losses due to feral pig predation." Australian Veterinary Journal **54**(9): 426-429.

Reese, G. C., K. R. Wilson, et al. (2005). "Factors affecting species distribution predictions: A simulation modeling experiment." Ecological Applications **15**(2): 554-564.

Rollins, D. (1993). Statewide attitude survey on feral hogs in Texas. Feral swine: A compendium for resource managers. C. W. Hanselka and J. F. Cadenhead. Kerrville, TX, USA, Texas Agricultural Extension Service.

Rosenberg, M. S. and C. D. Anderson (2011). PASSaGE 2 Help.

Rosenberg, M. S. and C. D. Anderson (2011). PASSaGE: Pattern Analysis, Spatial Statistics, and Geographic Exegesis. Version 2, Methods in Ecology and Evolution 2: in press.

Rouse, J. W., R. H. Haas, et al. (1974). Monitoring vegetation systems in the Great Plains with ERTS. Third Earth Resources Technology Satellite-1 Symposium, Greenbelt, MD.

Ryan, T. P. (1997). Modern Regression Methods. New York, NY, USA, John Wiley & Sons.

Santika, T. and M. F. Hutchinson (2009). "The effect of species response form on species distribution model prediction and inference." Ecological Modelling **220**(19): 2365-2379.

Seward, N., K. VerCauteren, et al. (2004). "Feral swine impacts on agriculture and the environment." Sheep and Goat Research Journal **19**: 34–40.

Shumway, R. H. and D. S. Stoffer (2006). Time Series Analysis and Its Applications With R Examples. New York, NY, USA, Springer

Simmons, S. (2010). Feral Hog Management Plan. Vian, Oklahoma, Sequoyah National Wildlife Refuge.

Singer, F. J., D. K. Otto, et al. (1981). "Home ranges, movements, and habitat use of European wild boar in Tennessee

" Journal of Wildlife Management **45**(2): 343-353.

Slonecker, E. T., D. Jennings, et al. (2001). "Remote sensing of impervious surface: a review." Remote Sensing Reviews **20**: 227–255.

Smolik, M. G., S. Dullinger, et al. (2010). "Integrating species distribution models and interacting particle systems to predict the spread of an invasive alien plant." Journal of Biogeography **37**(3): 411-422.

Sokal, R. R. and D. E. Wartenberg (1983). "A test of spatial autocorrelation analysis using an isolation-by-distance model." Genetics **105**: 219-237.

StataCorp (2005). Stata Statistical Software: Release 9. College Station, TX, USA, StataCorp LP.

Stevens, L. (1996). The Feral Hog in Oklahoma Samuel Roberts Noble Foundation.

Taylor, R. B., E. C. Hellgren, et al. (1998). "Reproduction of feral pigs in southern Texas." Journal of Mammalogy **79**(4): 1325-1331.

Texas Animal Health Commission (2008). Required Authorization for Movement of Feral Swine. T. S. o. State, Texas Secretary of State. **55.9**.

Thomas, C., T. Ranchin, et al. (2008). "Synthesis of multispectral images to high spatial resolution: A critical review of fusion methods based on remote sensing physics." Ieee Transactions on Geoscience and Remote Sensing **46**(5): 1301-1312.

Thornton, D. H., L. C. Branch, et al. (2011). "The influence of landscape, patch, and within-patch factors on species presence and abundance: a review of focal patch studies." Landscape Ecology **26**(1): 7-18.

Tobler, W. (1970). "A computer movie simulating urban growth in the Detroit region." Economic Geography **46**(2): 234-240.

Tolleson, D., D. Rollins, et al. (1993). Impact of feral hogs on ground-nesting gamebirds. Feral swine: A compendium for resource managers. C. W. Hanselka and J. F. Cadenhead. Kerrville, TX, USA, Texas Agricultural Extension Service**:** 76-83.

Towne, C. W. and E. N. Wentworth (1950). Pigs, from Cave to Corn Belt. OK, USA, Norman, University of Oklahoma Press.

Tozzini, F., A. Poli, et al. (1982). "Experimental infection of European wild swine (Sus scrofa L.) with pseudorabies virus." J Wildl Dis **18**(4): 425-428.

Turner, M. G., R. H. Gardner, et al. (2001). Landscape Ecology in Theory and Practice: Pattern and Process. New York, NY, USA, Springer.

Tyler Campbell, J. C. (2008). "Recent Range Expansion Of Invasive Feral Hogs." South Texas Wildlife **12**(2): 3 - 4.

U.S. Geological Survey. (2010). "Landsat Missions Timeline." from http://landsat.usgs.gov/about_mission_history.php.

UGA Department of Geology. (2011). "The Geology of Georgia." from http://www.gly.uga.edu/default.php?PK=0&&iPage=5.

US Department of the Interior Fish and Wildlife Service (2006). Okefenokee National Wildlife Refuge Comprehensive Conservation Plan. Atlanta, GA, USA.

USDA Forest Service and USGS (2002). Forest Cover Types. Reston, VA, National Atlas of the United States.

USGS. (2007). "The Road Indicator Project (TRIP) NORM ED Datasets." from http://rmgsc.cr.usgs.gov/trip/data/.

USGS. (2010). "The Road Indicator Project (TRIP) - Roadless Space in America." from http://rmgsc.cr.usgs.gov/trip/.

USGS Seamless Data Warehouse. (2010). "What is Elevation?", from http://seamless.usgs.gov/about_elevation.php.

Vanderleek, M. L., H. N. Becker, et al. (1993). "Prevalence of pseudorabies (Aujeszky's disease) virus antibodies in feral swine in Florida." Journal of Wildlife Diseases **29**(3): 403-409.

Wadsworth, R. and J. Treweek (1999). Geographical Information Systems for Ecology. New York, NY, USA, Springer.

Weisberg, S. (2005). Applied Linear Regression. Hoboken, NY, USA, John Wiley and Sons.

Williams, E. S. and I. K. Barker. (2001). Infectious Diseases of Wild Mammals. Ames, IA, USA, Iowa State University Press.

Wing, M. G. and P. Bettinger (2008). Geographic Information Systems: Applications In Natural Resource Management. Don Mills, Ontario, Canada, Oxford University Press.

Wiser, S. K., R. K. Peet, et al. (1998). "Prediction of rare-plant occurrence: A southern Appalachian example." Ecological Applications **8**(4): 909-920.

Wong, D. W. S. and J. Lee (2005). Statistical Analysis of Geographic Information with ArcView GIS and ArcGIS. Hoboken, New Jersey, John Wiley & Sons, Inc.

Wrigley, N. (1985). Categorical Data Analysis For Geographers And Environmental Scientists. New York, Longman.

Wu, J. and R. J. Hobbs (2007). Key Topics in Landscape Ecology. Cambridge, UK, Cambridge University Press.

Wu, J. G. and R. Hobbs (2002). "Key issues and research priorities in landscape ecology: An idiosyncratic synthesis." Landscape Ecology **17**(4): 355-365.

Wyckoff, A. C., S. E. Henke, et al. (2009). "Feral swine contact with domestic swine: A serologic survey and assessment of potential for disease transmission." Journal of Wildlife Diseases **45**(2): 422-429.

Zaniewski, A. E., A. Lehmann, et al. (2002). "Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns." Ecological Modelling **157**(2-3): 261-280.

Zhang, X. Y., M. A. Friedl, et al. (2003). "Monitoring vegetation phenology using MODIS." Remote Sensing of Environment **84**(3): 471-475.