5-2016

Modeling Feral Hogs in Great Smoky Mountains National Park

Benjamin Anthony Levy
University of Tennessee - Knoxville, blevy@vols.utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Benjamin Anthony Levy entitled "Modeling Feral Hogs in Great Smoky Mountains National Park." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Mathematics.

Suzanne Lenhart, Major Professor

We have read this dissertation and recommend its acceptance:

Charles Collins, Agricola Odoi, Paul Armsworth

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Modeling Feral Hogs in Great Smoky Mountains National Park

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Benjamin Anthony Levy

May 2016

I dedicate this dissertation to my wife Rachel. Your love throughout the years made this possible. Thank you for supporting me through this journey.

Acknowledgements

I would first like to thank my advisors, Suzanne Lenhart and Charles Collins. I have learned a great deal from you both and I am grateful for the time and effort you dedicated to my success. Dr. Collins, thank you for helping me develop as a teacher, your important role in this dissertation, and always taking time out of your schedule to chat. Dr Lenhart, thank you for being a great role model, your superlative guidance, and for all the opportunities you made possible. You each provided a different type of guidance and support that was crucial to my development as both a researcher and educator.

I would like to thank my committee members, Paul Armsworth and Agricola Odoi. I deeply respect your opinions and I thank you for considering as well as supporting my work. Judy Day, thank you for introducing me to mathematical biology and for being so supportive. I would also like to thank Lou Gross whose help and advice was key to my success.

Thank you to the mathematics department at the University of Tennessee for supporting me for three years. Thank you Henry Simpson, Pam Armentrout, Ben Walker, and Angela Woofter for providing help and support throughout my graduate studies. I would also like to thank the National Institute for Mathematical and Biological Synthesis (NIMBioS) for supporting me through their research assistantship. The opportunities and experiences provided by NIMBioS truly shaped me as a researcher. Thank you Great Smoky Mountains National Park for its pivotal role in this research.

Finally, thank you to my family, Rachel, Callie, Thomasina, Steve, Zach, Carol, Gary, Sara, and Mike. This would not have been possible without your unwavering support.

The book of nature is written in the language of mathematics.

-Galileo Galilei

Abstract

Feral Hogs (*Sus scrofa*) are an invasive species that have occupied the Great Smoky Mountains National Park since the early 1900s. Recent studies have revitalized interest in the pest and have produced useful data. The Park has kept detailed records on mast abundance as well as every removal since 1980 including geographic location and disease sampling. Data obtained via Lidar includes both overstory as well as understory vegetation information. In this dissertation, three models were created and analyzed using the detailed data on vegetation, mast, and harvest history. The first model is discrete in time and space and was formulated to represent hog dynamics in the park. The second is a spatial model of the niche of the population that relates known presence locations to environmental predictors. The third model is a compartmental disease model for pseudorabies in the population. Together, these projects assess the importance of the existing control program, predict suitable locations for hog presence in the Park, and quantify possible transmission routes for Pseudorabies.

Table of Contents

Chapter 1

Introduction

Invasive species are among the world's most significant and urgent environmental concerns as they inflict ecological and economic damages that are both costly as well as detrimental to the environment (Olson et al., 2006). These concerns are exacerbated as an invasive species becomes established in an area (Epanchin-Niell and Hastings, 2010). European Wild Boar (*Sus scrofa*) were brought to the United States in the early 1900s by settlers as a source of food. Since their introduction, feral hogs have been expanding their range, increasing in density, disrupting natural ecosystems, and posing a significant disease threat to livestock and native animals (Witmer et al., 2003). See Engeman et al. (2003), Engeman et al. (2004), Engeman et al. (2007) and Olson et al. (2006) for more details on economic and ecological impacts of feral hogs in the United States.

In 1912, hunters near Hoopers Bald, North Carolina, imported European wild boar (*Sus scrofa*) to populate a hunting preserve and were left to breed and expand their population for a span of 8-10 years (Jones, 1957). During this time a number escaped and dispersed throughout the surrounding area (Stegeman, 1938). They bred with hogs of domestic ancestry and have since spread throughout the Great Smoky Mountains National Park (GSMNP). We refer to this hybrid population using the term "feral hogs".

Great Smoky Mountains National Park is a 2,080 km^2 plot of land that straddles the border between Tennessee and North Carolina, the vast majority of which is undeveloped

forest. GSMNP is surrounded by 3 National Forests, the eastern border of a Cherokee Indian reservation, and Fontana reservoir (Stiver and DeLozier, 2005). Elevation throughout the park ranges from as low as 270 m to as high as 2,024 m. The park is characterized by a high elevation ridgeline that runs diagonally through the center of the park and by its unique and flourishing habitats. Due to its undeveloped nature, elevation gradient and rich environmental viability, GSMNP is home to over 6,000 flora and 400 fauna (Stiver and DeLozier, 2005).

Feral hogs in the Park consume acorns, known as hard mast, which fall from oak trees at the end of the summer. They also scavenge for tubers, roots and other food that can be found underground (Scott and Pelton, 1975). We refer to these additional food sources as the base food source. Since the feral hogs depend heavily on oak mast, it plays a significant role in their life cycle affecting reproduction and movement patterns (Johnson et al., 1982; Singer et al., 1981). We capture these dynamics by updating the corresponding mast-dependent parameters each month in our model.

One main concern about the presence of the invasive species *Sus scrofa* in GSMNP is that they compete with native species and are destructive to the surrounding environment. Rooting activities are extremely disruptive to vegetative communities, alter nutrient cycles and may even alter forest succession patterns in the long term (Bratton, 1974; Howe and Bratton, 1976). Feral hogs are in direct competition for oak mast with black bears and are known to scavenge for and consume salamanders (Singer, 1981). There are several National Park policies that state that the control or eradication of non-native species is necessary when such species endanger the protection and interpretation of natural resources in the park. Since the Park is considered the salamander capitol of the world and the black bear is perhaps the most beloved and publicized species in the park, in addition to the damage done to the grounds, direct negative impacts on these species by feral hogs is of great concern.

Another impact of *Sus scrofa* is the potential for transmission of Porcine parvovirus, leptospirosis, toxoplasmosis and pseudorabies, as each were found during various serological

surveys of feral hogs in GSMNP (Cavendish et al., 2008; Sandfoss et al., 2012). More alarming, the prevalence of pseudorabies has increased dramatically in recent years (Cavendish et al., 2008; Sandfoss et al., 2012). The spread of these diseases have serious implications for a large number of domestic and wild animals throughout the region. For example, although studies have shown the impact of pseudorabies on feral hogs is minor, clinical signs in commercial swine is well documented (Miller et al., 2013). Further, park officials believe the transportation of feral hogs due to illegal hunting interests may be increasing the presence and spread of these various diseases.

Due to the aforementioned concerns, Great Smoky Mountains National Park implemented a feral hog control program beginning in 1959. Although the program's procedures have varied since its implementation, recent efforts to control the feral hog population have been both opportunistic as well as active. Opportunistic activities include dispatching feral hogs when encountered by park rangers as well as setting traps in suspected high activity areas that are convenient for park employees to access and maintain. The Park also hires seasonal employees to actively search for and harvest feral hogs. Most of the active hunting takes place between January and May, as this is a time during which bears are hibernating, foliage is at a minimum and the feral hog population is concentrated in the lower elevation regions. Throughout the remainder of the year harvesting is much more limited taking place only in the absence of more pressing park needs. Hunting laws in both Tennessee and North Carolina have changed a number of times over the last 30 years and we have no data to measure its impact outside the park.

The negative impacts that feral hogs have on natural resources, in addition to the fact that feral hogs are hosts for infectious diseases, has resulted in government and park officials having a vested interest in knowing the whereabouts, threat levels and optimal management strategies for controlling the feral hog population (Stiver, 2012a,b). These factors and others propelled the creation of a working group at the National Institute for Mathematical and Biological Synthesis (NIMBioS) entitled "Feral swine/pseudo-rabies in Great Smoky

Mountains National Park." This group provided data, background information and input relevant to the formulation of our model.

Its pristine conditions and biodiversity has prompted a great deal of scientific research and documentation related to the Park. Important to this project are detailed vegetation data and corresponding records of yearly acorn crop levels over the last 30 years. We also make use of the harvest data that has resulted from the control efforts conducted by GSMNP (National-Park-Service, 1980). As a result, the main goals of this work was to use modeling in coordination with available harvest and mast data to estimate the population, to assess the effect of harvesting on the population, to create a habitat suitability map to help guide control efforts, and to model the presence of Pseudorabies in the population.

The remainder of the document will take the following form: First, the formulation of a data-driven metapopulation model will be discussed. This chapter will include a description of the data used, an overview of the study area, a summary of the population dynamics, the assumptions made, and how all of components combine in a mathematical formulation. Discussion of parameters being used will follow and will include how we estimated their values using the harvest data. A discussion of the results and any conclusions that can be made will also be presented, which will include specific uses of the model and future work related to modeling feral hogs in the Park. This project is in collaboration with Suzanne Lenhart, Charles Collins, Marguerite Madden, Joseph Corn, Rene Salinas, and William Stiver. Second, a niche model will be presented for the population in the Park which relates the harvest data and ecological variables using statistical techniques. An environmental niche factor analysis is used to relate presence locations to environmental predictors in order to spatially model locations within GSMNP that are suitable for hog presence. This project produces results that will guide control efforts using scientific analysis. This project is in collaboration with Suzanne Lenhart, Charles Collins, and William Stiver. The final chapter will discuss building a compartmental disease model for pseudorabies into the metapopulation model that is discrete is both time and space. The goal of this project is to explore possible transmission routes, estimate transmission parameters, and model the spread of the disease

in the population. This project is in collaboration with Suzanne Lenhart, Charles Collins, and William Stiver.

Chapter 2

Metapopulation Model

2.1 Introduction

There has been previous work modeling feral hogs in different parts of the world, each of which differs depending on the specific goal of the research. Spatially explicit models using partial differential equations have been formulated to model feral hog populations in different geographic areas (Clayton et al., 1997; Gaines et al., 2005; Keeling et al., 1999). Though insightful, these initial models are centered around basic ecological concepts such as logistic growth and contain limited feral hog features. An age structured model (without spatial features) has also been considered in an attempt to determine the structure and characteristics of specific population dynamics (Focardi et al., 1996). Most recently, an individual-based model was constructed for feral hogs in GSMNP that resulted from the same NIMBioS working group and is based on the similar harvest data (Salinas et al., 2015). In the individual-based model, the annual total of the harvest data was only used to estimate constant harvest and mortality rates. Models constructed with differential equations, agent-based models and discrete formulations each have benefits and drawbacks that depend upon the available information and specific goals of a project.

With the harvest and mast data (National-Park-Service, 1980, 1981) given in monthly intervals, we were able to carefully estimate a range for monthly mast-dependent parameters

to best match the observations of feral hog behavior in the Park. We fit yearly harvest rates in each region over two season per year. With our goals of analyzing general population dynamics and measuring the importance of the control program, we formulate a discrete, data-driven metapopulation model to describe the feral hog population in the Park.

2.2 Model Formulation

2.2.1 Data and Regions

Increased interest and technological advances have made research relating to feral hogs in GSMNP more tractable in recent years. One key contribution from the working group was data expressing vegetation types and distributions obtained through remote sensing (Madden et al., 2004). The data were used to create a digital vegetation map and database of overstory and understory flora found throughout GSMNP. This information is used to divide the Park and its immediate surrounding area into 8 regions as determined by vegetation type and also to establish where oak trees and other food sources are located.

Other significant data for our model were provided by GSMNP officials. The Park provided harvest and oak mast data in GSMNP from 1980-2010 (National-Park-Service, 1981, 1980). The harvest data contains over 11,000 entries and includes quantity, age, month and geographical location of feral hogs harvested. Mast index data consists of visual estimations of white and red oak acorns that existed throughout the Park in each given year. Values range from 0.45 to 5.1 with a higher value indicating a more bountiful year. We use the harvest data to set an initial distribution of feral hogs throughout the Park, to estimate parameter values and to check the accuracy of the model. The driving force behind feral hog behavior is their primary diet of oak acorns (Scott and Pelton, 1975; Singer, 1981). Therefore, knowing the quantity and location of the key food supply is ideal for this model. The aforementioned data was paramount in the formulation of our data-driven metapopulation model that is discrete is both time and space.

The discrete time step is 1 month as we wish to model several discrete events that occur on a scale that is not less than a month including births, mast deployment and seasonal movement.

The current month is denoted by t. We take $t = 1$ to be January. To denote any periodic or other type of time based events, we use m for the month (1-12) and y for the year.

The model spatial domain is divided into 8 regions based on overstory vegetation types as they produce the food that drives population dynamics (See Figure 2.1) (Scott and Pelton, 1975). The overstory data are from (Madden et al., 2004) and is detailed in Table 2.2. There are six regions inside the park (regions $1 - 6$) with region 6 constituting an upper-elevation ridge line that runs diagonally through the center of the park. Two of the regions are outside the park, one on the north side (region 8) and one on the south side (region 7).

For each region r, we record the area in acres (A_r) and the length of boundary between connected regions r and s (given in $BL_{r,s}$). These are used in determining yearly food supplies and governing movement between regions, each of which be explained in detail in later sections.

There is a feral hog population in each region that varies over time and is not differentiated by sex nor by age. The initial conditions for the model are based on harvest data from 1988 (National-Park-Service, 1980). The initial population in each region is reconstructed by dividing the number of feral hogs harvested from each region in 1988 by rough estimates from the Park of yearly on- and off-season monthly harvest rate of 0.03 and 0.01.

The population in region r at time t is denoted by $P_{r,t}$. The specific population in a given region r at time t depends on all of the parameters and variables that comprise the model.

2.2.2 Order of Events

The order of events in a discrete model is very important as it impacts the dynamics of the system (Bodine et al., 2012). Given the Mast supply ($M_{r,t}$) and the Populations ($P_{r,t}$) from the previous month, the events proceed in the following order:

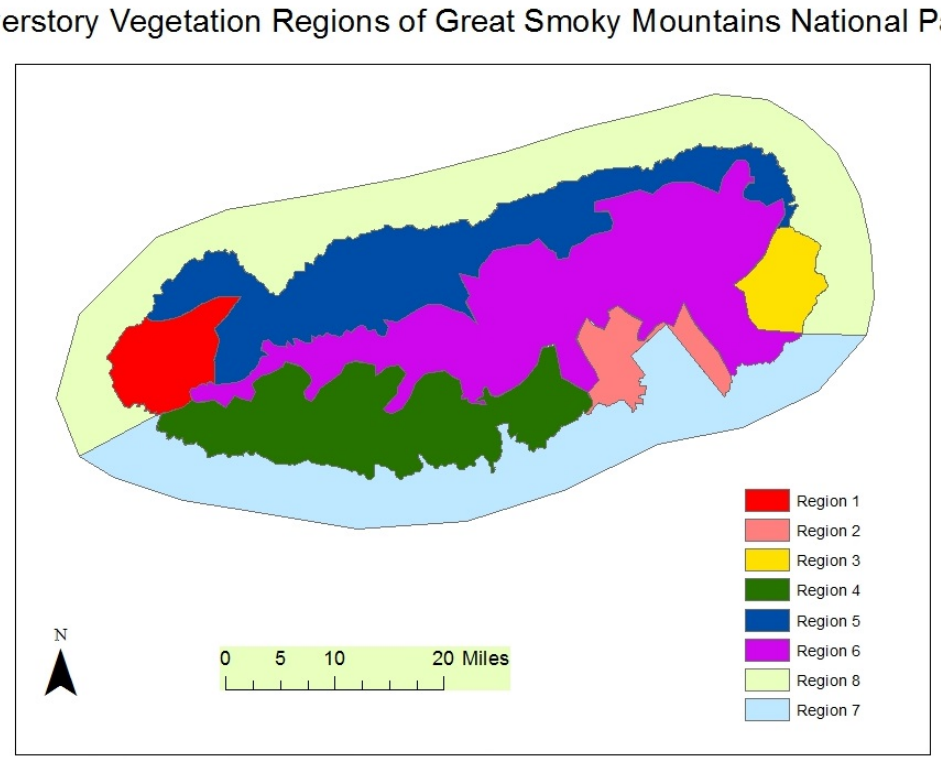


Figure 2.1: The Great Smoky Mountains National Park is broken into 8 regions based on overstory vegetation types.

1. Update the Mast for the month since many of the parameters that govern feral hog dynamics in GSMNP are driven by hard mast availability (Singer, 1981; Scott and Pelton, 1975).

2. Harvest at a rate determined by the month.

3. Compute the portion of the post-harvest population that survives. We do this before adding births because only surviving adults can reproduce.

4. If the month is January, we then compute the number of births based on the surviving population and mast supplies. We keep track of the births and apply a different survival rate before adding them into the general population.

5. Perform movement, using either general movement or seasonal movement, dependent upon the time of year.

Mast Dynamics

Each region has two food sources- the base food source and hard mast. The hard mast source is from white and red oak trees, which is believed to be preferred by feral hogs in GSMNP (Scott and Pelton, 1975). Hard mast becomes available in August, but varies in amount from year to year. The amount of hard mast in each region in each year is measured in kilocalories and based on a hard mast index for years 1981-2010 produced by (National-Park-Service, 1981). The mast index ranges from $0.45 - 5.1$ with lower values indicating poor mast years and higher values for ample mast years. See Table 2.1 for a list of the 25 overstory types found in GSMNP, their total acreage and corresponding kilocalorie per acre (Inman, 1997). Of course, the location of hard mast is an important driver of the population and the top four overstory types in each region by total area covered can be seen in Table 2.2. Hard mast levels in a given year and region are derived by pairing the distribution of the 25 overstory types in each region with the assumed kilocalorie per acre of each to create a baseline level of hard mast. This value is scaled further according to the recorded hard mast data from 1981 to 2010 to properly model the impact changes in food levels has on the population (National-Park-Service, 1981). With the aforementioned data in hand, the available kilocalories of hard mast that becomes available each year in August in each region is calculated.

The other food source is tubers, roots, small animals and other items feral hogs can scavenge off the ground, which we refer to as the base food source (Scott and Pelton, 1975). The amount of available kilocalories of the base food source is assumed to be available in each region at the constant rate of 1000 kilocalories per acre. We consider this food source as a constant amount proportional to region size because at any given time there are an unknown level of renewable food sources found on the forest floor in each region and because the base food source plays a minor role in a hogs diet. The reduced role it plays compared to hard mast is because it is less nutritious as well as less abundant and is thus less desirable. The role of the base food source in the life cycle of feral hogs is to ensure minimal sustenance

in times and regions where there is not sufficient hard mast. This is modeled by including a smaller amount of soft mast in the model which allows hard mast to more significantly influence parameter values that govern population dynamics.

Feral hog behavior and movement are believed to be driven by hard mast availability (Singer, 1981). Denote the time varying hard mast in region r at time t by $HM_{r,t}$. Hard mast is dropped from trees at the end of August and decreases over time due to feeding by feral hogs and competitors as well as natural decay. We will initialize the model in August using available mast, vegetation and harvest data for this month from (Madden et al., 2004; National-Park-Service, 1980, 1981). The hard mast index is a single number that indicates the level of oak mast in a given year. This single number is paired with the known acreage of oak trees in each region to produce $MI_{r,y}$, the amount of oak mast produced in region r in year y. After mast is dropped in August we assume each feral hog consumes at rate C_P, which takes the value of 5000 kilocalories per day (Inman, 1997). Hard mast also depletes as a result of natural decay and consumption by other animals at rate δ, which is assumed to be 8% per month to ensure most food is consumed each year. To be ecologically consistent, we ensure the hard mast value in each region does not become negative. All of the previous years hard mast is entirely depleted before the following August when the next hard mast drop occurs. Thus, the specific amount of hard mast available in region r at time t is dependent on time, the specific region, the hard mast index value for the given year and the number of feral hogs in the region and is given by

$$HM_{r,t+1} = \begin{cases} MI_{r,y} & m = 8, \\ ((1-\delta)HM_{r,t} - C_P P_{r,t})^+ & m \neq 8. \end{cases}$$

Denote the constant amount of the base food source in region r by SM_r. Since the base food source is assumed to exist at the constant rate of 1000 kilocalories per acre, $SM_r = 1000A_r$.

The total mast in each region r and time t is denoted by $M_{r,t}$ and is given by

$$M_{r,t} = HM_{r,t} + SM_r. \tag{2.1}$$

The specific amount of total mast available in each region influences parameter values at each time step as determined by the scale function.

Table 2.1: Overstory Vegetation Types. Listed above are the 25 overstory types that exist in Great Smoky Mountains National Park, their corresponding total acreage and kilocalorie per acre.

Overstory Type	**Abbreviation**	**Total Acreage**	**Kilocalorie per Acre**
Bare Ground	Bare	1,223	0
Cove Hardwood Forest	CHx	78,655	19,154
Dead	Dd	335	0
Human Influence	Hi	4,828	0
Rhododendrom	Rhd	8,054	0
Mixed Hardwood Forest	Hx	34,781	15,465
Montane Alluvial Forest	MAL	6,605	4,000
Montane Oak Forest- White Oak	MOa	2,414	8,000
Montane Oak Forets- Red Oak	MOr	18,554	8,000
Northern Hardwood (Birch)	NHx	77,184	4,066
Meisic Chestnut Oak Forest	OcH	9,163	8,000
Meisic Oak Forest (Red Oak)	OmH	103,221	8,337
Xeric Oak Forest (Red &White Oak)	OzH	79,886	8,127
Xeric Oak Forest (Pine Mix)	OzH-P	1,447	6,000
Yellow Pine Forest	P	27,283	7,117
Meisic Oak Forest Mix (Oak & Pine)	P-OmH	548	5,915
Xeric Oak Forest Mix (Oak & Pine)	P-OzH	22,678	5,915
Pasture	Pa	449	0
Rock	Rk	528	0
Spruce-Fir Mix	S-F	37,278	4,578
Shrub	Sb	2,363	0
Hemlock	T	15,762	6,000
Vines	V	1,116	0
Water	W	7,498	0
Wetland	Wtl	108	0
Total		**219,440**	**542,015**

Table 2.2: Dominant Overstory Type by Region. This table displays the top four overstory types by total acreage in each region, with percentage of total area covered in parenthesis (Madden et al., 2004).

Region	Total Acreage	Type 1 (% area)	Type 2 (% area)	Type 3 (% area)	Type 4 (% area)
1	13,091	OzH (32.2)	P-OzH (21.5)	OmH (16.3)	CHx (15)
2	18,546	OmH (21.9)	CHx (21.7)	OzH (11.9)	Sb (10.4)
3	18,546	OmH (21.9)	CHx (21.7)	OzH (11.9)	Sb (10.4)
4	44,441	OcH (38)	OzH (19.7)	CHx (13.9)	Hx (7.1)
5	55,412	OzH (24.4)	OmH (21.5)	CHx (20.1)	Hx (9.8)
6	74,988	NHx (39)	S-F (20.1)	CHx (10.2)	MOr (6.6)
7	22,220	OcH (38)	OzH (19.7)	CHx (13.9)	Hx (7.1)
8	27,706	OzH (24.4)	OmH (21.5)	CHx (20.1)	Hx (9.8)

The Scale Function

Many of the parameters that comprise the model vary in time and space and are based on mast availability. Such parameters include percentage of adults that survive ($Surv$), how likely it is that feral hogs will move to an adjacent region during general movement ($Move$) as well as the yearly birth rate (BR). Due to this fact, each of these parameters are determined at each time step via a scaling function that produces appropriate values for each parameter based on the mast availability.

Let $Param_{r,t}$ denote one of the above mentioned mast-dependent parameters in region r at time t.

We will obtain the value of a given parameter in the current time step dependent upon the mast availability according to the following scale function (called F):

$$F(M, Param0, ParamMax, M_h) = \frac{Param0 \cdot M_h + ParamMax \cdot M}{M + M_h}, \tag{2.2}$$

where $M = M_{r,t}$ is the mast value in region r at time t, $Param0$ is the value of $Param_{r,t}$ when $M = 0$, and $ParamMax$ is the value of $Param_{r,t}$ as $M \to \infty$. The M_h is the half-saturation mast constant such that if $M = M_h$, then $F = \frac{Param0+ParamMax}{2}$. At the beginning of each month the specific value of mast-dependent parameters values is determined by the

scale function. See Figure 2.2 for a depiction of the form and asymptotic nature of the scale function.

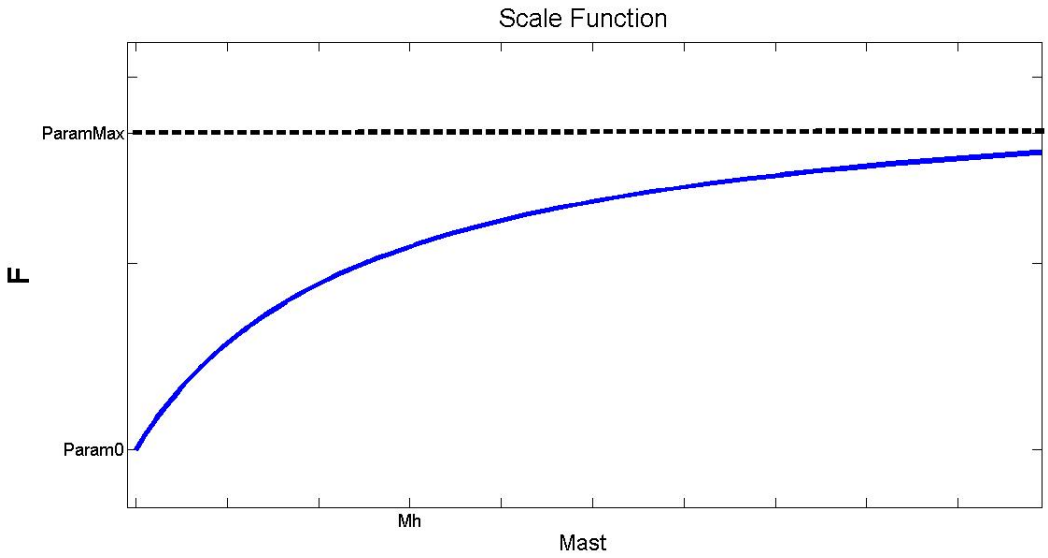


Figure 2.2: A number of the parameters that comprise the model are mast-dependent. The rational scale function ranges between a maximum and minimum parameter value as determined by the current mast level in each region.

2.2.3 Population Dynamics

The population dynamics are driven in each region by 3 factors, which occur in the order listed: survival, birth and movement. The remainder of the section is presented in the order in which the events take place in each region r at time t. Throughout the process we calculate how many feral hogs survive each step and update the population accordingly. After movement takes place, $P_{r,t+1}$, the population that will start out in each region in the next month, is calculated.

Survival

Although park rangers are constantly making some effort to harvest feral hogs, the most opportune time to hunt is when the bears are hibernating, the foliage is at a minimum and the feral hogs are at lower elevations. Due to this fact, the park hires additional employees to actively seek feral hogs from January through May. During the remaining months, harvesting is limited to convenient, and thus less frequent, harvesting. To reflect this recurring trend we have an relatively low off-season harvest rate between months $6-12$, a much higher on-season harvesting rate from months $1-5$. Since we have no data on harvesting outside of GSMNP we only model harvesting within its boundaries. Let $Hrate$ denote the seasonal harvest rate given by

$$Hrate = \begin{cases} rate_1, & m = 1, ..., 5 \ \& \ r = 1, ..., 6 \\ rate_2, & m = 6, ..., 12 \ \& \ r = 1, ..., 6 \\ 0, & r = 7, 8 \end{cases}.$$

To find parameter values and assess the general harvest trend, we initially assume that $Hrate$ is uniform in space and only varies in time according to the season. After estimating parameter values, we then vary harvesting in time and space in order to more accurately model the system and evaluate the effects the Park's efforts have the population.

The number of feral hogs harvested at time t in region $r = 1, ..., 6$ is denoted by

$$H_{r,t} = Hrate \cdot P_{r,t}. \tag{2.3}$$

The number of feral hogs that survive harvesting in each region continue on to the subsequent mast-dependent survival step.

The number of feral hogs that die due to natural causes is mast dependent and independent of piglet survival. We assume that there is an alternative food supply available (base food source) but that it is less desirable, less prevalent, and less nutritious than hard mast. Thus the feral hogs don't die when the primary supply is consumed, but they also don't survive as easily. These dynamics are captured by the previously mentioned scale function, paired with a constant, but significantly less, amount of the base food source. The adequacy

of the food supply is also related to the current population as more feral hogs need more available mast.

The percent of feral hogs that do not perish due to natural causes is given by $Surv_{r,t}$, the survival rate in region r at time t. $Surv_{r,t}$ is dependent on the available mast in the region in the given month and has a minimum value of $Surv0$ given no available mast and approaches a maximum value of $SurvMax$ as mast availability increases. Since the percent that survive is dependent on mast availability we first determine $Surv_{r,t}$ via the scale function,

$$Surv_{r,t} = F(M_{r,t}, Surv0, SurvMax, M_h), \tag{2.4}$$

and apply both the survival rate as well as the harvest rate to the existing population:

$$P_{r,t} \cdot (1 - Hrate) \cdot Surv_{r,t}. \tag{2.5}$$

Individuals that make it through harvesting and mast-dependent survival move on to the birth stage of the model.

Birth:

Female feral hogs go into estrus as soon as they can and usually give birth in January (Singer, 1981). The number of piglets produced per population in our model depends on the following factors:

1. Percent of the population that are mature females: We assume that 50% of the population is female, but that only 90% are mature. This means that 45% of the population are able to give birth, denoted by B_F.

2. Average litter size: Typical litter size is 3-8, so we will take it as 6, denoted by L_A (Singer, 1981).

3. Percent that are pregnant: Even if a feral hog reproduced in the previous year, they can still go into estrus. In each regions, we assume there are enough males to impregnate nearly all the available females but that the actual number of the mature females that

become pregnant is mast dependent (Stegeman, 1938; Stiver, 2014).

4. Percent of the pregnant females that give birth: Inevitably, a portion of the females that become pregnant will not successfully give birth. This portion depends on the mast level during the pregnancy. In a low mast years, the percent of pregnant females that will come to term is very low compared to high mast years where a much larger percentage of pregnant females will give birth (Singer, 1981).

5. Percent of the litter that survives the first month: This also depends on current mast level and is comparable to the survival for the population in general, but lower as there is a higher mortality for piglets in the first 6 months.

The last three factors are heavily mast dependent with inadequate mast causing production of far fewer births than when there is plentiful mast. Although the amount of mast available in each region in the months surrounding January determine these birth parameters, we use the amount of mast remaining in January in each region as an indication of the presence of mast before and after feral hogs give birth. Thus the percent of mature females that become pregnant is determined by the scale function and is dependent on the amount of mast available in each region in the month of January.

In our model we assume births only take place in January, denote m^* to be the month of January (Singer, 1981). When $m = m^*$, a percentage of the population in each region are pregnant females. Of those pregnancies, a portion actually come to term and produce piglets and then a fraction of the births survive their first month of life and are added into the general population. Denote the number of piglets produced by each individual in the in region r at time t by $BR_{r,t}$. The value of $BR_{r,t}$ is the product of the birth rate with the average litter size and number of mature females. Since the birth rate is highly dependent on mast availability, the exact level is determined by the

scale function:

$$BR_{r,t} = \begin{cases} B_F \cdot L_A \cdot F(M_{r,t}, BR0, BRMax, M_h) & m = 1 \\ 0 & m \neq 1. \end{cases}$$

The total number of births in each region will then be a product of $BR_{r,t}$ and the number of adult feral hogs that have lived through both survival stages. This value can immediately be added into the general population since the birth rate only includes piglets that survive. Let $SH_{r,t}$ denote the surviving feral hog population in region r at time t, which is composed of all individuals that have lived through the survival stage as well as any new births into the population given by

$$SH_{r,t} = P_{r,t} \cdot (1 - Hrate) \cdot Surv_{r,t} \cdot (1 + BR_{r,t}).$$

General and Seasonal Movement:

The reasons feral hogs move throughout GSMNP can be characterized in two ways: between areas of the park searching for food as well as making use of the topography of the park by moving up in elevation during the spring and down in elevation in the fall (Singer, 1981). The movement toward higher elevations takes place in the warmer months and is caused by the decline in mast availability paired with increasing temperatures in lower elevations. The movement towards lower elevations is a result of mast becoming available in these regions at the end of the summer. Independent of this seasonal movement, the food-based movement is hard mast dependent(Scott and Pelton, 1975; Singer, 1981).

General movement of the feral hogs refers to the movement between the various regions independent of the intrinsic characteristics of the region themselves or time of year, but instead as a result of food availability. Each month a percentage of the population will

move out of each region depending on the mast availability in each region. Lower food levels increase the percent of feral hogs moving out of that region where high levels of food results in limited movement. The specific level of general movement in each region is governed by the scale function. We assume that the local population density is low enough that it does not directly impact movement, but rather indirectly impacts general movement through mast availability. Feral hogs moving out of a region move to a neighboring region proportional to shared boundary length, which is a well studied and accepted method of modeling movement of a population (Holland et al., 2007). The proportions of shared boundary length is captured in this connectivity matrix:

$$BL = \begin{pmatrix} 0 & 0 & 0 & 0.106 & 0.316 & 0.20 & 0 & 0.378 \\ 0 & 0 & 0.189 & 0.055 & 0 & 0.20 & 0.555 & 0 \\ 0 & 0.366 & 0 & 0 & 0.006 & 0.20 & 0 & 0.428 \\ 0.044 & 0.031 & 0 & 0 & 0 & 0.20 & 0.725 & 0 \\ 0.093 & 0 & 0.001 & 0 & 0 & 0.20 & 0 & 0.706 \\ 0.017 & 0.157 & 0.073 & 0.326 & 0.427 & 0 & 0 & 0 \\ 0 & 0.316 & 0 & 0.684 & 0 & 0 & 0 & 0 \\ 0.141 & 0 & 0.107 & 0 & 0.751 & 0 & 0 & 0 \end{pmatrix} \tag{2.6}$$

where $BL_{i,j}$ is the percent of feral hogs moving out of region i that will move into region j. The matrix BL is a connectivity matrix derived from the proportion of shared boundary between regions.

For example, since $BL_{2,4} = 0.055$, this implies that region 2 shares a boundary with region 4 and that the length of their shared boundary constitutes 5.5% of the total boundary of region 2. As a result, when general movement takes place, 5.5% of feral hogs moving out of region 2 will move into region 4. All entries of $BL_{i,j}$ that are 0 imply that regions i and j do not share a boundary.

There are several clarifications that must be made regarding general movement as it relates to this matrix. First, recall that region 6 is the high-elevation region, which does not

contain any hard mast producing trees. Since general movement is entirely mast-driven, in practice, we limit the amount of general movement from the lower-elevation regions $1-5$ to region 6 by allowing only 20% of all feral hogs marked to move from each interior region to move up to region 6. To ensure rows of BL sum to 1, the adjusted decrease in shared boundary length between each interior region with region 6 was re-distributed to the other neighbors proportionally . Also reflected in BL is that all interior regions border region 6 while the two exterior regions do not.

Seasonal movement takes place when the population changes their location based on elevation from March through June and again in August (Singer, 1981).

There are two forms of seasonal movement: from the higher regions to the lower regions in the fall (August) and from the lower regions to the higher regions during the spring (March through June). Since the high to low movement culminates with the drop of the hard mast paired with decreasing temperatures, we assume all feral hogs will move down in elevation during the month of August (Singer, 1981). The low to high movement takes place from March through June as hard mast in the lower regions become depleted and temperatures increase (Singer, 1981). With this in mind, we have additional feral hogs move from each low region to the high region beginning in March at a rate which increases as we approach June. This ensures nearly all feral hogs reach the higher regions by mid-summer.

Although these movement patterns were derived from a previous study (Singer, 1981), a more recent telemetry study is currently being formulated by GSMNP that makes use of advanced collaring and tracking techniques (Stiver, 2012a). The data will be used to either confirm these movement dynamics or re-shape our current assumptions.

The number of feral hogs moving out of region r at time t is denoted by $MH_{r,t}$. In practice general movement and seasonal movement follow the same pattern of first deciding how many feral hogs will move out of each region, removing them from the surviving population and then re-distributing them to neighboring regions.

General movement is less of a driver than seasonal movement thus general movement occurs only in the absence of seasonal movement. In months when there is only general

movement the percent of feral hogs moving out of each region, $Move_{r,t}$, is dependent on the mast level in the region, $M_{r,t}$. Using the scale function we have

$$Move_{r,t} = F(M_{r,t}, Move0, MoveMax, M_h), \tag{2.7}$$

from which the number of surviving feral hogs moving out of each region can be computed:

$$MH_{r,t} = Move_{r,t} \cdot SH_{r,t}. \tag{2.8}$$

During general movement, feral hogs marked to move out of lower elevation regions $1, 2, 3, 4, 5, 7, 8$ and travel to adjacent regions according to the proportions given in BL. However, movement from the high elevation region 6 to lower-elevation regions is controlled differently. After spending the warmer months in region 6, at the end of August the entire upper elevation population simply move regions $1, ..., 5$ proportional by boundary length. Otherwise, during general movement months we employ a mechanism for region 6 that accounts for the low elevation region from which the feral hogs came. We achieve this by having feral hogs move down to one of regions $1, ..., 5$ weighted by the number of feral hogs that have moved out of each of those regions to the upper region over the past year. To achieve these movement patterns, we define $MP_{i,j}$ as the proportion of feral hogs moving out of region i and into region j as

$$MP_{i,j,t} = \begin{cases} BL_{i,j} & i \neq 6 \;\&\; m = 1, 2, 7, 9, 10, 11, 12 \\ MP^*_{j,t} & i = 6 \;\&\; m = 1, 2, 7, 9, 10, 11, 12 \end{cases}.$$

where

$$MP^*_{j,t} = \begin{cases} BL_{6,j} & 1 \leq j \leq 5 \;\&\; m = 8 \\ \frac{MP^*_{6,t-1} + 0.2 \cdot MH_{j,t-1}}{\sum_{j=1}^{5} MP^*_{j,t}} & 1 \leq j \leq 5 \;\&\; m = 1, 2, 7, 9, 10, 11, 12 \\ 0 & j \geq 7 \end{cases}.$$

The feral hogs marked to move from each region during general movement are then removed from the surviving population and then redistributed given by

$$P_{r,t+1} = P_{r,t} \cdot (1 - Hrate) \cdot Surv_{r,t} \cdot (1 + BR_{r,t}) - MH_{r,t} + \sum_{i=1}^{8} MP_{i,r,t} \cdot MH_{i,t} \quad \text{for } r = 1, ..., 8. \tag{2.9}$$

Seasonal movement takes place from March through June when depleted mast supplies and increasing temperatures cause feral hogs to move from lower elevation regions $1, ..., 5$ to higher elevation region 6. During this time we calculate an initial number of feral hogs that will move out of the lower regions by applying the scale function and then add an additional increase to the general movement amount in each subsequent month. This results in feral hogs migrating in increasing proportions to the high elevation region in the center of the park while allowing a small population to remain at a lower elevation. For the month of July, general movement then applies again as discussed previously. Since the understory is the only food source in the high elevation region, the scale function produces relatively high general movement rates in this month, which simulates the start of the migration back down to lower regions. Then, in August, hard mast falls from oak trees and draws remaining feral hogs back down to the lower the elevation regions. During this month we move all remaining feral hogs in region 6 to regions $1, ..., 5$ proportional to shared boundary length. Although the specific number of feral hogs moving to and from each region changes during this time from year to year, the proportions and distribution locations are deterministic and based on the specific month.

We first determine how many feral hogs will be moving out of each region:

$$MH_{r,t} = \begin{cases} Move_{r,t} \cdot SH_{r,t} + \frac{m-2}{5}(SH_{r,t} - Move_{r,t} \cdot SH_{r,t}) & r \neq 6 \,\&\, 3 \leq m \leq 6 \\ 0 & r = 6 \,\&\, 3 \leq m \leq 6 \\ Move_{r,t} \cdot SH_{r,t} & r \neq 6 \,\&\, m = 8 \\ SH_{r,t} & r = 6 \,\&\, m = 8. \end{cases}$$

Table 2.3: A list and description of variables found in the model.

Name	Description
t	Time (in months), start with t=1 in January and run for 20 years
m	Month in the given year
y	Number of years
N	Number of regions
r	Region number
$BL_{i,j}$	Proportional boundary length between regions i and j
$MI_{r,y}$	Hard mast produced in region r in August in year y (in kilocalories)
$SM_{r,t}$	Base food source in region r at time t in kilocalories
$M_{r,t}$	Total mast in region r at time t in kilocalories
$BR_{r,m=1}$	Births in regions r. Occur in January
$P_{r,t}$	Feral hog population in region r at time t
$H_{r,t}$	Feral hog population harvested in region r at time t
$SH_{r,t}$	Number of surviving feral hogs in region r at time t
$MH_{r,t}$	The number of feral hogs moving out of region r at time t
F	Scale function that determines mast-dependent parameters

We again define $MP_{i,j}$ as the proportion of feral hogs moving out of region i and into region j which will account for the migration of feral hogs to upper elevations during the warmer months by increasing the proportion of feral hogs moving from regions $1, ..., 5$ to region 6 from March through June:

$$MP_{i,j,t} = \begin{cases} BL_{i,j} & 1 \leq i, j \leq 8 \,\&\, m = 8 \\ MP_{i,j,t-1} + \frac{m-3}{5} & 1 \leq j \leq 5 \,\&\, j = 6 \,\&\, 3 \leq m \leq 6 \\ MP_{i,j,t-1} - \frac{m-3}{5} \cdot \frac{BL_{i,j}}{\sum_{r \neq 6} BL_{i,r}} & 1 \leq j \leq 5 \,\&\, j \neq 6 \,\&\, 3 \leq m \leq 6 \\ BL_{i,j} & i \geq 7 \,\&\, 3 \leq m \leq 6 \\ 0 & i = 6 \,\&\, j \neq 6 \,\&\, 3 \leq m \leq 6. \end{cases}$$

Then, feral hogs moving out of each region are removed from the surviving population and re-distributed during seasonal movement given by:

$$P_{r,t+1} = P_{r,t} \cdot (1 - Hrate) \cdot Surv_{r,t} \cdot (1 + BR_{r,t}) - MH_{r,t} + \sum_{i=1}^{8} MP_{i,r,t} \cdot MH_{i,t} \quad \text{for } r = 1, ..., 8.$$

Table 2.4: A list and description of parameters found in the model.

Name	Value	Description
A_r	13-55	Area of region r in thousands of acres
δ	0.08	Monthly food loss percentage due to decay and competitors
C_P	5000	Calories consumed per feral hog per day
M_h	$150,000$	Half saturation mast constant
B_F	0.45	Percent of population that are mature females
B_P	0.95	Percent of female population that are mature and can give birth
L_A	6	Average size of a litter
$Surv0$	0.88	Survival factor if there is no mast
$SurvMax$	0.97	Survival factor as mast approaches a maximum level
$BR0$	0.27	Percent of population that give birth and whose piglets survive the first month given no mast
$BRMax$	0.89	Percent of population that give birth and whose piglets survive the first month as mast approaches a maximum level
$Move0$	0.51	Percent of feral hogs moving with no available mast
$MoveMax$	0.16	Percent of feral hogs moving as mast approaches a maximum level
$rate_1$	0.35	On-Season harvest rate, from January through May
$rate_2$	0.15	Off-Season harvest rate, from June through December
$Surv_{r,t}$	$0.88 - 0.97$	Percent of feral hogs that survive in region r at time t. Mast dependent
$BR_{r,t}$	$0.27 - 0.89$	Percent of population that give birth and whose piglets survive the first month in region r at time t. Mast dependent
$Move_{r,t}$	$0.16 - 0.51$	Percent of feral hogs that move out of region r at time t

2.3 Parameter Estimation

Feral hog population dynamics vary across space. The dynamics of a specific feral hog population greatly depends on the local environment. Most research conducted on the feral hog population in GSMNP is outdated and thus many of the parameter values are unknown. Our metapopulation model contains the following eight unknown parameters as described in Table 2.4 $Surv0$, $SurvMax$, $BR0$, $BRMax$, $Move0$, $MoveMax$, $rate_1$ and $rate_2$.

It is important to note that all of the above parameters, except $rate_1$ and $rate_2$, are mast dependent and thus get updated each month using the scale function. With this in mind each "0" value will not ever be achieved due to a constant amount of available the base food source and each "Max" value shown above are approached asymptotically as a result of the rational form of the scale function.

We wish to find the parameter values that, when used in our model, produce harvest levels that best match the available harvest data. We use data from 1989 through 2000 since the harvesting strategies that took place during this time period were most consistent.
More specifically, the problem can be stated as

$$\underset{x}{\text{Minimize}} \quad J(x) = \frac{\sqrt{\sum_y \sum_r (H_{r,y} - H^*_{r,y})^2}}{\sqrt{\sum_y \sum_r (H^*_{r,y})^2}}$$

r represents all interior regions

y is the year ranging from 1989-2000

x represents all possible parameter values

$H^*_{r,y}$ is harvest data from region r in year y

$H_{r,y}$ is the computed harvest from region r in year y

In addition, we need to make sure that the parameters reflect conditions found in the Park. Specifically, minimum survival rate should be significantly less than maximum survival rate and on-season harvesting produces much higher yields than off-season rates. As a result, the above problem is also restricted by the following linear constraints order to ensure that the resulting values reflect these trends and the parameters are ordered correctly:

$$\begin{aligned} &\frac{11}{10}Surv0 \leq SurvMax \\ &BR0 \leq BRMax \\ &MoveMax \leq Move0 \\ &\frac{3}{2}rate_2 \leq rate_1 \end{aligned} \tag{2.10}$$

All parameters were constrained within the interval $[0, 1]$.

To solve the above optimization problem, we made use of the Global Optimization Toolbox from MATLAB$^{\text{TM}}$. Since our model contains a large number of complicated implicit functions, we employed a method that did not require input of any derivatives. Furthermore, given the overwhelming number of possible parameter combinations, we needed our local solver to work in concert with an algorithm that would test a large number of starting points. Thus, we chose to use the MultiStart Algorithm with fmincon as its local solver.

The MultiStart Algorithm was most appropriate for our problem as it allowed us to test a large number of evenly distributed starting points and stores all local solutions in a manageable way using the built-in manymins function. The MultiStart Algorithm generates uniformly distributed random starting points within the given bounds and passes them one-by-one to the local solver, fmincon, which attempts to find a local basin of attraction relative to each given start values. Any solution that is found is then stored increasing order of objective function output for later review using the manymins function.

The fmincon local solver was most appropriate for our problem as it accepts smooth, nonlinear objective functions, is a derivative-free solver and allows enforcement of the linear

inequality constraints and bounds given in (2.10). Instead of inputting a derivative, fmincon approximates the gradient numerically in order to move towards the basin of attraction given each starting point. In all of our trials, the exit flag produced by fmincon indicated that a convergent run occurred. Exit flags are integers that range from -3 to 5 with zero and negative values indicating poor or no solution and positive values indicating different strengths of quality solutions. More specifically, negative exit flags correspond to termination by the script itself, no feasible point found or a constraint violation and zero indicates the maximum number of iterations was exceeded. On the other hand, positive exit values correspond to various appropriate reasons for termination of the algorithm.

2.4 Results and Discussion

Recall that the initial population in the model is set using harvest data in each region from 1988 paired with a presumed yearly harvest rate. Due to the fact that our goal was to determine parameters based on how well they produced harvest numbers that matched our data, the initial population being used in the model greatly affected the computed harvest rates. In changing the initial population one also runs the risk of altering various parameter values. Due to this potential sensitivity to initial values, the previously stated optimization problem was run with the following differing initial populations: 454, 774, 1410 and 2924. These values were obtained the same way initial conditions were described previously except by assuming different uniform harvest rates.

After the initial population was set in each scenario, the optimal harvest values were then included in the parameter estimation.

After a number of exploratory runs of the MultiStart Algorithm with fmincon as a local solver, it became obvious that there were a great number of viable solutions that both satisfy the constraints and produce comparable error outputs. These initial trials also imparted some intuition about the appropriate range for each parameter. This allowed us to repeat our process with more confined constraints to improve the speed and accuracy of the results.

Table 2.5: Results optimizing over all parameters for different initial population values. In the table, P_0 is the initial population value used in the optimization procedure. The average value and variance for the results of each initial population are also shown.

P_0		$Surv0$	$SurvMax$	$BR0$	$BRMax$	$Move0$	$MoveMax$	$rate_1$	$rate_2$
454	Mean	0.867	0.959	0.712	0.975	0.695	0.015	0.982	0.575
	Variance	0.00006	0.00001	0.03	0.002	0.002	0.0009	0.001	0.003
774	Mean	0.863	0.965	0.460	0.749	0.636	0.068	0.685	0.369
	Variance	0.0003	0.0001	0.01	0.02	0.008	0.001	0.003	0.005
1410	Mean	0.859	0.955	0.575	0.738	0.515	0.059	0.393	0.190
	Variance	0.00004	0.00001	0.005	0.002	0.005	0.004	0.0006	0.0006
2924	Mean	0.845	0.963	0.508	0.728	0.620	0.111	0.206	0.090
	Variance	0.0002	0.00003	0.007	0.004	0.006	0.001	0.0001	0.00007

With the above in mind, given each of the four initial populations, we used 500 starting values and tested both the less constrained as well as more constrained bounds. In all 500 trials a convergent local solution for all 8 parameters were found. Of these 500 values we only considered those within 20% of the lowest error output. As a result we were able to narrow down the values to more reasonable candidates from which an average was calculated, as shown in Table 2.5.

The second column displays the lowest output of the objective function $J(x)$ in the given trial. The third column of the table indicated how many of the 500 starting points produced an error within 20% of the lowest error output and thus were considered when calculating the mean. The displayed mean values are very consistent between trials with the only values that are not clustered across trials being the harvest values, which are directly related to the initial population in a given trial. In fact, each harvest solution settled very close to the presumed harvest rates that set the initial value in the first place. Also, notice that while values for $Surv0$ and $SurvMax$ needed to be manually forced to be different from each other using the inequalities shown in 2.10, the values for $BR0, BRMax, Move0$, and $MoveMax$ naturally settled on starkly different values. This fact further supports the notion that accurate and meaningful parameter values were estimated.

We chose the parameters from the third line in Table 2.5 because values were consistent across initial populations and Park officials believe that the current population of feral hogs in GSMNP is nearest to 1410 (Conversation with William Stiver, May 2012). Figure 2.3a illustrates the resulting computed harvest values when compared to the harvest data using these parameters. Keep in mind that harvest values were estimated using only the $1989-2000$ data and that any similarities between computed harvest and harvest data past the year 2000 are an indication that the model is capturing historical population dynamics. Furthermore, since we have assumed the same uniform harvest rate in each region that only varies by season, the computed harvest values only capture the general behavior of the data rather than closely approximating it, as shown in Figure 2.3b. For the purposes of estimating the non-harvest parameters we only wanted to mimic general historical population trends and thus we were satisfied with the resulting values. We used these results to estimate values for the survival, birth and movement parameters.

In reality, the Park's harvesting levels and locations vary week-to-week or even day-to-day. In fact, when using the estimated non-harvest parameters in a similar optimization scheme but instead varying the harvest values by region and by year, we are able to match the harvest data nearly exactly as illustrated in Figure 2.3c. These harvest values more closely match historical efforts, and thus from our interest in evaluating the importance of the control program, the rest of our analysis will be based on this set of parameters.

Although the Park has had a control program in place for over 50 years, little is known about the effectiveness of their efforts. Since a significant amount of time and money is spent on the control program throughout the year, there is question to whether there should be fewer resources spent harvesting feral hogs in GSMNP. Using the yearly harvest values for each region we derived, we are in a unique position to evaluate what would happen if harvest efforts were reduced for a period. The values were shown to match the data very closely (see Figure 2.3c). To evaluate the effects of having applied a different level of effort from 1994-2000, we tested what would happen if the Park had either reduced or increased

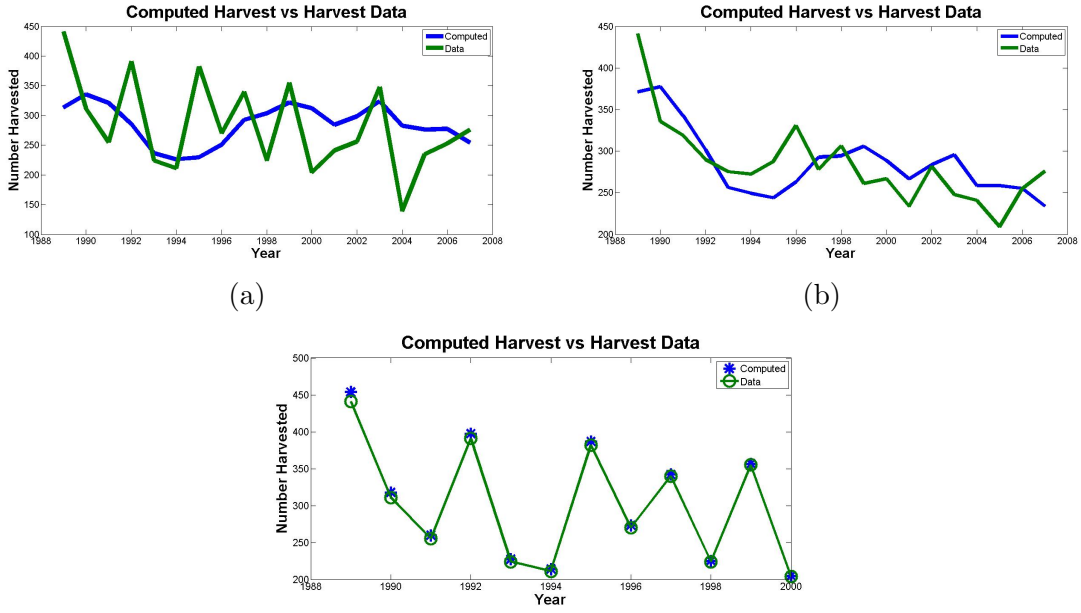
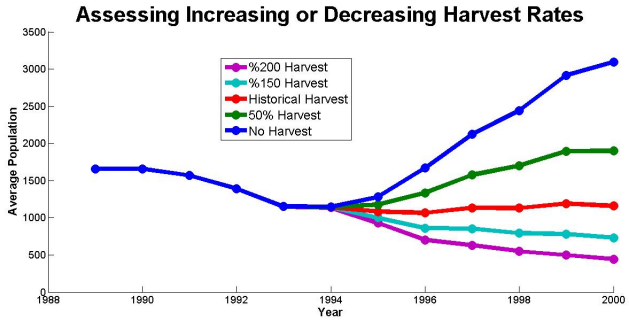


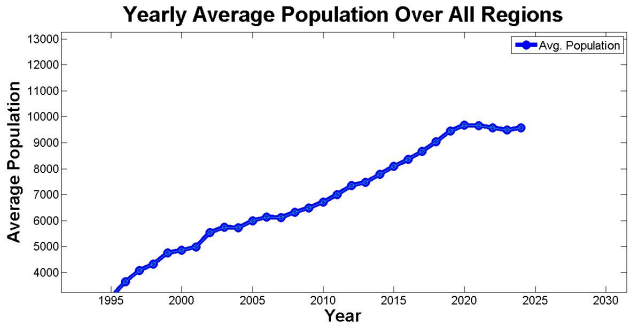
Figure 2.3: Each figure displays total computed harvest values produced by the model compared with the harvest data provided by GSMNP. Figure 2.3a depicts a model simulation using parameter values obtained from the trials that used 1410 as an initial population. Figure 2.3b displays the output from the same parameter values compared to a smoothed version of the harvest data where yearly values were derived from a three year average surrounding each data point. Figure 2.3c was produced from a model simulation using yearly harvest values for each region and illustrates how we are able to accurately quantify historical harvest efforts.

the harvest rates by 50% or 100%. The results are shown in Figure 2.4a and illustrate that historical Park efforts have been successful in limiting the size of the population. As Figure 2.4a shows, eliminating harvesting altogether for a six year period would result in an increase in the population by much as 260%. On the other hand, doubling harvest rates from 1994-2000 could have reduced the historical population by more than 60%. Supposing that the control program never existed and that instead the population was left to grow, the model levels off at a total population of nearly 10,000 individuals in and around GSMNP. This apparent carrying capacity is depicted in Figure 2.4b and was derived by a model simulation without harvest starting in 1988.

As previously mentioned, an individual-based model (IBM) was recently constructed for feral hogs in GSMNP. Our project and (Salinas et al., 2015) estimate a similar annual



(a)



(b)

Figure 2.4: Both figures began with, and use parameter values derived from, an initial population of 1410. Figure 2.4a shows the effects varying the estimated harvesting rates from 1994-2000. In just 6 years, decreasing the harvest rate can result in an increase in population by as much as 260%. The model shows that doubling the harvesting rate over 6 years could reduce the population by over 60%. Figure 2.4b illustrates that without harvesting the population could increase to nearly 10,000 feral hogs.

harvest rate, predict a similar population level and both emphasize the importance of the control program. Although based on similar data, the two models also differ in a number of ways. For instance, the individual-based model implements a density-dependent form of movement while the metapopulation model mimics believed movement patterns through mast-dependent and seasonal forms of movement. The distinctions in movement characteristics could account for the differences in estimated population growth and carrying capacity for feral hogs in GSMNP that the two models predict in the absence of harvesting. A telemetry study is currently being conducted by GSMNP where feral hogs are being tracked

via GPS collars. Information from this project can be used to refine movement mechanisms in both models to best reflect the system.

The structure of the metapopulation model also allowed us to carefully estimate parameters against available data, providing an accurate global perspective of the system. Although both models allow acorn availability to influence mast-dependent responses in birthrates and movement rates, the metapopulation model is well suited to handle population level responses with limited empirical data and thus may better model survival. While the IBM reveals how individual behavior influences the population as a whole, the metapopulation model captures large scale dynamics and trends.

2.5 Conclusions and Future Work

We formulated a metapopulation model for feral hogs in GSMNP. The parameters in this discrete model (in space and time) were carefully estimated using data involving harvest (of feral hogs), hard mast, understory and geographic size of the regions. Seasonal movement and appropriate demographic processes were included.

The structure of our model was shown to accurately estimate the amount of feral hogs harvested in the Park over the last 35 years. We can conclude that if the Park would reduce harvest levels in the future, the feral hog population can increase dramatically with an expected result of further habitat damage and negative impact on other species. Park official recognize that these results emphasize the need to continue the control program. Seeing how their efforts has reduced the population also provides support for future funding.

The strong dependence of estimating harvest rates on an uncertain initial population is a limitation of our work. We used expert opinion from Park officials to guide our choices about the initial population levels and other features of the model. Also, though we only consider births in January, a second birth pulse in the summer has been documented in feral hogs in the event of extreme food conditions (Johnson et al., 1982). This behavior was not deemed important to general dynamics and was thus not included in this model.

With the model in place we are in position to make use of the available data to learn more about feral hog behavior in the GSMNP. More specifically, we intend to consider a cost effective management strategy and analyze potential disease threats that feral hogs pose to the area. These endeavors will be aided further by new data resulting from ongoing related research and help guide control efforts.

One challenge in modeling feral hogs is that each population behaves differently depending on the local environment. Limited accessibility to most areas of the Park has deterred locating, tracking, eradication efforts, and the general study of the feral hogs in this region. However, three new grants were awarded related to feral hogs in GSMNP. Two of the grants provided funds to continue with a control and disease monitoring program in GSMNP that has provided the most current data for feral hogs locations(Stiver, 2011, 2012b). The third grant will fund a study in which the location and movement patterns of feral hogs in GSMNP and Big South Fork National River and Recreation Area. Feral hogs will be tracked using radio collars, which will provide more detailed information related to the movement of feral hogs throughout the Park (Stiver, 2012a). This unprecedented data about the movement and home range of feral hogs in GSMNP can be used to include more detailed movement structure in future models.

Since the implementation of the feral hog control program, over 13,000 harvest entries have been logged that can provide insight into the types of habitat in GSMNP that feral hogs prefer and can be used to determine potential locations of the invasive pest in the Park. Since the data we have is presence data only, it can be used to create a habitat suitability map via an Environmental Niche Factor Analysis (ENFA) or Maximum Entropy theory. Both approaches require only presence data and aim to find the relationship between known feral hog locations and environmental factors that drive the population.

This research related to feral hog dynamics in GSMNP provides the framework necessary to conduct additional analyses. To improve the efficiency of their efforts, we intend to use the model to consider an optimal harvest strategy. This may include specific regions, months and strategies that will maximize harvest yields. A habitat suitability map will contribute

to this process as it illustrates areas in the Park that have a high probability of containing feral hogs.

Another meaningful project related to feral hogs in GSMNP is to consider the threats and implications of feral hogs as a vessel for pseudorabies (Cavendish et al., 2008). Pseudorabies poses significant threats for animals such as canines and commercial livestock (Cavendish et al., 2008). Since feral hogs are currently the only reservoirs of pseudorabies, they are the only source of future outbreaks (Miller et al., 2013). A disease analysis will involve a compartmental model with spatial and temporal elements while considering the possibility of pseudorabies suddenly appearing in far reaches of the Park as a result of illegal feral hog release by hunters.

Chapter 3

Spatial Niche Model

3.1 Background and Theory

3.1.1 Problems with Multiple Correlations and Predictors

A common area of interest in mathematics is to determine the relationship between two of more predictor variables and a single criterion variable. Many related methods involve linear combinations of the predictor variable in order to explain the criterion variable. Such methods include simple and multiple linear regression, general linear models, least squares estimation and maximum likelihood estimation to name a few. Many of the aforementioned theory has applications in distribution modeling under the correct assumptions.

To illustrate the issues that correlations between predictor variables present to this problem, consider the following hypothetical scenario: Suppose we wish to predict the probability $\hat{p}$ that a focal species exists in cell j based on number of environmental variables Z_i for $i = 1, ..., k-1$. We can view this situation as attempting to relate $k-1$ predictors (environmental variables) to explain the criterion variable (probability of focal species existence) in k-dimensional space. That is, after learning about the interrelationships between the k elements in the vector space where a species is found, $k-1$ predictor scores can be computed that can be used to obtain the probability that the focal species exists in a new

cell j in the study area as a linear combination of the predictor scores and predictor values on cell j, i.e.,

$$\hat{p_j} = b_1 z_{1,j} + b_2 z_{2,j} + ... + b_{k-1} z_{k-1,j}$$

where b_j are the regression weights and $z_{i,j}$ is the value of environmental factor i in cell j.

Of course, this calculation is subject to some error, e, which can be used to adjust the weights, b_i. Since e is the difference between the actual and computed probability that the species exists in cell j, it can be quantified as

$$e_j = p_j - \hat{p}_j.$$

Naturally, for the purpose of regression and ultimately prediction, we wish to minimize this error. Formally, we can pose this problem as

$$\underset{b}{\text{Min}} f(e(b))$$

$$\text{where } f(e(b)) = \frac{\sum_{j=1}^{N} e_j^2}{N} = \frac{\sum_{j=1}^{N} (p_j - \hat{p}_j)^2}{N},$$

for $b = (b_1, ..., b_{k-1})$ and where N is a normalization factor and the square is introduced to account for negative values that might occur from the difference. Substituting the formula for $\hat{p}_j$ yields

$$f(e(b)) = \frac{\sum_{j=1}^{N} [p_j - (b_1 z_{1,j} + b_2 z_{2,j} + ... + b_{k-1} z_{k-1,j})]^2}{N}.$$

Differential calculus can be used to solve this minimization problem (Cooley and Lohnes, 1962). Let $r_{i,j}$ represent the value explaining the inter correlation between environmental variables (predictor variables) i and j, and $r_{i,k}$ be the correlation of environmental variable (predictor variable) i with the criterion variable (probability of focal species existence)

(Cooley and Lohnes, 1962). After taking partial derivatives and setting equal to zero, the problem can be represented in matrix form by

$$\mathbf{R_{1,1}b = R_{1,2}}$$

where

$$\mathbf{R_{1,1}} = \begin{bmatrix} 1 & r_{1,2} & r_{1,3} & \cdots & r_{1,k-1} \\ r_{2,1} & 1 & r_{2,3} & \cdots & r_{2,k-1} \\ r_{3,1} & r_{3,2} & 1 & \cdots & r_{3,k-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{k-1,1} & r_{k-1,2} & r_{k-1,3} & \cdots & r_{k-1,k} \end{bmatrix},$$

$$\mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_{k-1} \end{bmatrix}$$

and

$$\mathbf{R_{1,2}} = \begin{bmatrix} r1,k \\ r_{2,k} \\ r_{3,k} \\ \vdots \\ r_{k-1,k} \end{bmatrix}.$$

It becomes clear that in order to solve for the regression weights in the matrix $\mathbf{b}$, one needs to invert the matrix $\mathbf{R_{1,1}}$ so that

$$\mathbf{b = {R_{1,1}}^{-1}R_{1,2}}.$$

It is shown in (Cooley and Lohnes, 1962) that the squared multiple correlation coefficient R^2 is given by

$$R^2 = \mathbf{b}'\mathbf{R}_{\mathbf{1,2}} = \sum_{j=1}^{p-1} b_j r_{jp}.$$

To illustrate this complication consider a two predictor system where

$$R^2 = b_1^2 + b_2^2 + 2b_1 b_2 r_{12}.$$

This is extremely problematic because interpreting the regression weights now relies upon the interaction between the predictors and if the predictors are correlated, understanding the relationship is impossible (Cooley and Lohnes, 1962). This and a number of other problems illustrates a general need for uncorrelated variables (Cooley and Lohnes, 1962).

However, using uncorrelated variables is not always an option when conducting a regression. For example, it is generally accepted that environmental variables tend to be correlated. Thus, if we wish to predict the presence of a species based on the conditions in a study area, we need a way to transform the environmental variables into uncorrelated predictors.

3.1.2 Principal Component Analysis

We just saw how correlations between predictor variables is a confounding issue when conducting a linear regression. This fact can also be extended to many non-linear regression models as well as nearly all of the existing multivariate prediction models involve interpreting the regression weights, which is clouded by the correlations between predictors. This issue is made worse when considering environmental variables to predict a species distribution as most environmental variables are inherently correlated in many ways (Hirzel et al., 2002). Thus, if we want uncorrelated predictors, we need to construct them via a transformation that also preserves the information the data conveys. We will start by first describing the

concept of Principal Component Analysis (PCA) for a general case (and general dispersion matrix) before discussing the application to Environmental Niche Factor Analysis.

Conceptually, PCA aims to resolve correlation issues in m predictors by applying an orthogonal transformation to convert a set of potentially correlated observations into a set of linearly uncorrelated variables known as principal components. As an added bonus, we can achieve a reduction in dimensions as m predictors can typically be represented by $k < m$ principal components: those that are most important in explaining the data (Legendre and Legendre, 2012). Let $\mathbf{Z}$ be an mxn matrix of standardized data in which correlations between predictors exist. Standardized means that the data in $\mathbf{Z}$ has been transformed to have mean zero. This can be achieved via

$$z_{i,j} = \frac{x_{i,j} - \bar{x_j}}{\sigma x_j},$$

where $x_{i,j}$ is the value of data j in cell i, $\bar{x_j}$ is the mean of predictor j and σ_{x_j} is its standard deviation.

Since the correlations that exist within the data cause problems, the general concept is to perform a linear transformation on the data via matrix $\mathbf{P}$ to re-express the information given in $\mathbf{Z}$ in an uncorrelated matrix $\mathbf{Y}$:

$$\mathbf{PZ} = \mathbf{Y}.$$

The matrix $\mathbf{P}$ is what is of interest to us as its rows are a basis for the columns of $\mathbf{Z}$ that will determine the important dynamics of the predictors and filter out redundancies. We call the rows of $\mathbf{P}$ the principal components of $\mathbf{Z}$. Further clarification of the goals of the transformation must be made in order to find a unique $\mathbf{P}$.

While we wish to transform the information given by $\mathbf{Z}$ into an uncorrelated matrix $\mathbf{Y}$, it needs to be done in a way that best expresses the data. The concept of determining how exactly to "best express" the data was pioneered by Karl Pearson in 1902. Since his time others, including Truman Lee Kelley, Harold Hotelling and T.W. Anderson, have pondered

this question and agreed upon a general solution to maximize the variance in the data (Shlens, 2014). The reason for this decision stems from the at least two potential problems with the poor data in $\mathbf{Z}$: Either the data has noise, or it contains redundant information.

A common method to measure noise in data is the signal-to-noise ratio (SNR) which is achieved by dividing the variance of the signal (or global distribution), σ_s^2, by the variance of the noise (or predictor distribution), σ_n^2:

$$SNR = \frac{\sigma_s^2}{\sigma_n^2}.$$

Since we have already assumed that the data set characterizes the dynamics of the system, σ_n^2 should be less than σ_s^2. Thus, the more SNR exceeds 1, the more precise (less noisy) the data is. To reduce noise and best express the data in $\mathbf{Z}$, each principal component in $\mathbf{P}$ should aim to maximize the variance in $\mathbf{Y}$.

The second possible problem with the data in $\mathbf{Z}$ is redundancy. That is, one row of $\mathbf{Z}$ is strongly correlated with a different row and thus two rows provide essentially the same information. To combat this issue and ensure the transformed matrix is uncorrelated, we require that the columns of $\mathbf{P}$ be orthogonal to each other:

$$\mathbf{p_i p_j = 0 \quad for \;\; i \neq j}.$$

A further requirement on $\mathbf{P}$ is that its columns must be normalized. Doing so does not change the information gained from $\mathbf{P}$ and will allow us to implement relevant linear algebra techniques that require orthonormal matrices later on.

Now that we have more specific information of our goals we can state the problem we wish to solve as "find an orthonormal $\mathbf{P}$ such that $\mathbf{PZ} = \mathbf{Y}$, the variance in $\mathbf{Y}$ is maximized and the rows of $\mathbf{Y}$ are uncorrelated".

The question then becomes how to maximize variance between a set of predictors. The variance of vector $\mathbf{a} = (a_1, a_2, ..., a_n)$ with average μ is given by

$$\sigma_a^2 = \frac{1}{n}\sum_{i=1}^{n}(a_i - \mu)^2,$$

Since normalized vectors have mean zero, their variance is simply

$$\sigma_a^2 = \frac{1}{n}\sum_{i=1}^{n} a_i^2.$$

Extending this concept to multiple vectors introduces the concept of covariance. Given vectors $\mathbf{a} = (a_1, a_2, ..., a_n)$ and $\mathbf{b} = (b_1, b_2, ..., b_n)$, the covariance of $\mathbf{a}$ and $\mathbf{b}$ is a measurement of how the two variables change together and is the best representation of variance between two vectors. Covariance between vectors $\mathbf{a}$ and $\mathbf{b}$, σ_{ab}^2, is calculated as

$$\sigma_{ab}^2 = \frac{1}{n-1}\sum_{i=1}^{n} a_i b_i = \frac{1}{n-1}\mathbf{a}\mathbf{b}^T. \tag{3.1}$$

Note that if $\sigma_{ab}^2 = 0$ then $\mathbf{a}$ and $\mathbf{b}$ are completely uncorrelated, whereas $\sigma_{ab}^2 = \sigma_a^2$ or $\sigma_{ab}^2 = \sigma_b^2$ implies $\mathbf{a} = \mathbf{b}$.

Considering 3.1, it is clear that the covariance matrix, $\mathbf{S_Z}$, that describes the interactions between all predictors can be found through

$$\mathbf{S_Z} = \frac{1}{n-1}\mathbf{Z}\mathbf{Z}^\mathbf{T},$$

since components $\mathbf{S_{Z_{i,j}}} = \mathbf{z_i}\mathbf{z_j}^T$ (Legendre and Legendre, 2012). The diagonal terms, $\mathbf{S_{Z}}_{i,i}$, are the variance of the specific predictor and off-diagonal terms $\mathbf{S_{Z}}_{i,j}$ describe the covariance between different predictors. Thus, since we wish to maximize variance, in addition to each principal component $\mathbf{p_i}$ being orthogonal to another, we require that each $\mathbf{p_i}$ maximize the variance in $\mathbf{Y}$. Furthermore, since one of our goals is to express the data given in $\mathbf{Z}$ as uncorrelated data in $\mathbf{Y}$, we need the covariance matrix for $\mathbf{Y}$ to be diagonal.

Generally speaking, a matrix is diagonalizable if it is similar to a diagonal matrix. That is, matrix $\mathbf{A}$ is diagonalizable if there exists non-singular matrix $\mathbf{B}$ such that matrix $\mathbf{B^{-1}AB}$ is diagonal. In finite-dimensional space F, $\mathbf{A}$ is diagonalizable if and only if there exists an ordered basis of F that consists of the eigenvectors of $\mathbf{A}$. Also applicable here is the fact that a matrix is symmetric if and only if it can be diagonalized by a matrix of its orthonormal eigenvectors. That is, given symmetric matrix $\mathbf{K}$,

$$\mathbf{D} = \mathbf{E^T K E},$$

where $\mathbf{D}$ is a diagonal matrix containing eigenvalues of $\mathbf{K}$ and the columns of $\mathbf{E}$ are the orthonormal eigenvectors of $\mathbf{K}$ (Legendre and Legendre, 2012). Since $\mathbf{ZZ^T}$ is clearly symmetric,

$$\mathbf{D} = \mathbf{E^T}(\mathbf{ZZ^T})\mathbf{E}$$

for diagonal eigenvalue matrix $\mathbf{D}$. Since $\mathbf{E}$ is orthonormal, $\mathbf{EE^T} = \mathbf{E^T E} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. Thus, we can derive the relationship

$$\mathbf{ZZ^T} = \mathbf{EDE^T}. \tag{3.2}$$

Notice that substituting $\mathbf{Y} = \mathbf{PZ}$ into $\mathbf{S_Y} = \frac{1}{n-1}\mathbf{YY^T}$ yields

$$\begin{aligned}
\mathbf{S_Y} &= \frac{1}{n-1}\mathbf{YY^T} && (3.3)\\
&= \frac{1}{n-1}(\mathbf{PZ})(\mathbf{PZ})^\mathbf{T} && (3.4)\\
&= \frac{1}{n-1}\mathbf{PZZ^TP^T} && (3.5)\\
&= \frac{1}{n-1}\mathbf{P}(\mathbf{ZZ^T})\mathbf{P^T}, && (3.6)
\end{aligned}$$

For this reason we will make the choice that the rows of matrix $\mathbf{P}$ should consist of the eigenvectors of matrix $\mathbf{ZZ^T}$ where eigenvalues λ_i and associated eigenvector $\mathbf{p_i}$ are arranged

in descending order. Thus

$$\mathbf{P} \equiv \mathbf{E^T} \tag{3.7}$$

where matrix $\mathbf{E}$ contains eigenvectors of $\mathbf{ZZ^T}$ as columns. Note that the eigenvalues are all positive due to the fact that all dispersion matrices are positive definite. Additionally, by construction $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n \geq 0$. Substituting 3.7 and 3.2 into 3.6 results in

$$\begin{aligned}
\mathbf{S_Y} &= \frac{1}{n-1}\mathbf{E^T}(\mathbf{ZZ^T})\mathbf{E} \\
&= \frac{1}{n-1}\mathbf{E^T}(\mathbf{EDE^T})\mathbf{E} \\
&= \frac{1}{n-1}\mathbf{IDI} \\
&= \frac{1}{n-1}\mathbf{D}.
\end{aligned}$$

As one can see, this particular choice of $\mathbf{P}$ diagonalizes $\mathbf{S_Y}$ like we wanted. The principal components of the original predictor matrix $\mathbf{Z}$ are eigenvectors of $\mathbf{ZZ^T}$ and $\mathbf{S_{Y}}_{ii}$ indicates the variance of $\mathbf{Z}$ along principal component $\mathbf{p_i}$. Furthermore, since each $\mathbf{p_i}$ aimed to maximize variance along each axis, the principal components and corresponding variance are arranged in decreasing order.

It should be clear at this point that the specific choice of $\mathbf{P}$ linearly transforms correlated data $\mathbf{Z}$ into uncorrelated data in $\mathbf{Y}$ (Legendre and Legendre, 2012). However, we still need to show that the resulting matrix $\mathbf{Y}$ best expresses the transformed data. That is, we wish to show that the variance in $\mathbf{Y}$ has been maximized while preserving its uncorrelated nature. To prove that eigenvectors $\mathbf{p_i} = \mathbf{e_i^T}$ and corresponding eigenvalues λ_i indeed maximize the variance of each row of $\mathbf{Y}$, consider the following. Let $\mathbf{z_i}$ represent *column* i in matrix $\mathbf{Z}$ and $\mathbf{p_i}$ be *row* i in matrix $\mathbf{P}$. Recall that, as a result of the proposed linear transformation on the data in $\mathbf{Z}$,

$$\mathbf{y_{1,i}} = \mathbf{p_1 z_i}.$$

Since we wish to maximize the variance in $\mathbf{Y}$, we want the first principal component, $\mathbf{p_1}$, to maximize the variance in the first row of $\mathbf{Y}$:

$$\text{Maximize } \frac{1}{n}\sum_{i=1}^{n}\mathbf{y_{1,i}^2}$$

where

$$\mathbf{p_1 p_1^T} = 1,$$

due to the fact that we have restricted the columns of $\mathbf{P}$ to be the orthonormal eigenvectors of $\mathbf{ZZ^T}$. However, notice that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\mathbf{y_{1,i}^2} &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{p_1 z_i})^{\mathbf{2}} \\ &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{p_1 z_i})(\mathbf{p_1 z_i})^{\mathbf{T}} \\ &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{p_1 z_i})(\mathbf{z_i^T p_1^T}) \\ &= \mathbf{p_1 S_Z p_1^T}.\end{aligned}$$

To maximize $\mathbf{p_1 S_Z p_1^T}$ subject to $\mathbf{p_1 p_1^T} = \mathbf{1}$, the method of Lagrange multipliers can be used (Legendre and Legendre, 2012). This results in the following scalar equation

$$\phi_1 = \mathbf{p_1 S_Z p_1^T} - \lambda_1(\mathbf{p_1 p_1^T} - 1),$$

where λ_1 is a Lagrange multiplier. We wish the variance in the first row of $\mathbf{Y}$, which is expressed by ϕ, to be maximized by the first principal component $\mathbf{p_i}$ and thus we consider the partial derivative

$$\frac{\partial \phi_1}{\partial \mathbf{p_1}} = 2\mathbf{S_Z p_1^T} - 2\lambda_1 \mathbf{p_1^T}.$$

Setting this expression equal to 0 and simplifying yields

$$(\mathbf{S_Z} - \lambda_1 \mathbf{I})\mathbf{p_1^T} = \mathbf{0},$$

or

$$\mathbf{S_Z p_1^T} = \lambda_1 \mathbf{p_1^T},$$

which is clearly the eigenstructure relating leading Lagrange multiplier and eigenvalue λ_1 with corresponding eigenvector $\mathbf{p_1}$ (Cooley and Lohnes, 1962). Multiplying on the left by $\mathbf{p_1}$ results in

$$\mathbf{p_1 S_Z p_1^T} = \lambda_1 \mathbf{p_1 p_1^T} = \lambda_1,$$

We wanted to maximize $\frac{1}{n}\sum_{i=1}^{n} \mathbf{y_{1,i}^2} = \mathbf{p_1 S_Z p_1^T}$, which has clearly been achieved by construction, this value is equal to the leading eigenvalue of $\mathbf{ZZ^T} = \mathbf{S_Z}$.

Subsequent principal components are found in a similar manner, only with more restrictions required. For the second principal component we wish to maximize the variance in the second row of $\mathbf{Y}$ while assuring that the transformed data is uncorrelated. Mathematically, this translates into

$$\text{Maximize } \frac{1}{n}\sum_{i=1}^{n} \mathbf{y_{2,i}^2} = \mathbf{p_2 S_Z p_2^T}$$

subject to

$$\mathbf{p_2 p_2^T} = \mathbf{1},$$

where

$$\mathbf{y_{2,i}} = \mathbf{p_2 z_i}.$$

The requirement that the rows of $\mathbf{Y}$ remain uncorrelated means

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{y_{2,i} y_{1,i}} = 0.$$

Notice that

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\mathbf{y_{2,i}y_{1,i}} &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{p_2 z_i})(\mathbf{p_1 z_i}) \\
&= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{p_2 z_i})(\mathbf{z_i^T p_1^T}) \\
&= \mathbf{p_2}\frac{1}{n}\sum_{i=1}^{n}(\mathbf{z_i z_i^T})\mathbf{p_1^T} \\
&= \mathbf{p_2}\frac{1}{n}\mathbf{Z Z^T p_1^T} \\
&= \mathbf{p_2 S_Z p_1^T} \\
&= \mathbf{p_2 p_1^T}\lambda_\mathbf{1}.
\end{aligned}$$

This implies that

$$\mathbf{p_2 p_1^T} = \mathbf{0}, \tag{3.8}$$

emphasizing our assumption of orthogonal eigenvectors.

We can again use a Lagrange multiplier approach and derive the scalar equation

$$\phi_2 = \mathbf{p_2 S_Z p_2^T} - \lambda_2(\mathbf{p_2 p_2^T} - 1) - m_1\mathbf{p_2 S_Z p_1^T} \tag{3.9}$$

where λ_2 and m_1 are Lagrange multipliers (Legendre and Legendre, 2012). Differentiating 3.9 with respect to $\mathbf{p_2}$, multiplying on the left by $\mathbf{p_1}$, setting equal to zero, and considering 3.8 results in

$$\mathbf{p1}\frac{\partial\phi}{\partial\mathbf{p_2}} = \mathbf{2p_1 S_Z p_2^T} - \mathbf{2}\lambda_\mathbf{2 p_1 p_2^T} - \mathbf{2m_1 p_1 S_Z p_1^T} = \mathbf{0}. \tag{3.10}$$

Equation 3.10 implies that

$$-2m_1\mathbf{p_1 S_Z p_1^T} = 0$$

and thus

$$m_1 = 0.$$

As a result, eigenvector $\mathbf{p_2}$ and associated second largest eigenvalue λ_2 maximize the variance in the second row of $\mathbf{Y}$ (Cooley and Lohnes, 1962).

This process is continued for each subsequent eigenvector and principal component $\mathbf{p_i}$. By the end of the process, the rows of $\mathbf{P}$ contain the eigenvectors associated with the ordered eigenvectors λ_i. This structure is the subject of the next section where we discuss how the number of predictors can be reduced, adding value to the process of principal component analysis.

3.1.3 Dimension Reduction

The structure that results from a principal component analysis is extremely useful in a number of ways. First, the eigenvectors p_i of the dispersion matrix $\mathbf{S_Z} = \mathbf{ZZ^T}$ are the basis for the linear transformation $\mathbf{PZ} = \mathbf{Y}$ (Legendre and Legendre, 2012). Furthermore, $\mathbf{p_1}$ describes the direction of maximum variance in the predictors and each subsequent $\mathbf{p_i}$ for $i = 2...n$ gives the direction of the next largest variance orthogonal to each of the previous directions. The construction of this transformation is what results in uncorrelated predictors. Second, the descending positive eigenvalues λ_i that correspond to eigenvectors $\mathbf{p_i}$ tell the variance that exist in the direction of each principal axis. This is immensely important as their values decrease rapidly, which indicates that the transformed information from $\mathbf{Z}$ can be summarized in far fewer dimensions via the axes described by the eigenvectors of its dispersion matrix (Legendre and Legendre, 2012).

Since it is generally agreed upon that maximizing variance best represents the transformed data and that the percent of variance being represented by principal axes $\mathbf{p_i}$ for $i = 1...n$ is given by

$$\frac{\lambda_i}{\sum_{j=1}^{n} \lambda_i},$$

where the magnitude of each eigenvalue gives us a clue of how many principal components should be used. There are several methods to determine how many axis are significant including the Kaiser-Guttman criterion (Kaiser, 1991) and the broken stick method (Frontier,

1976). The Kaiser-Guttman criterion indicates that only principal components with corresponding eigenvalues that are larger than the mean of all eigenvalues should be used. The broken stick method compares the values of decreasing eigenvalues to the decreasing values in the broken stick model to make a choice of which principal components are significant and should be used (Frontier, 1976).

In either case, usually only the first several principal components are deemed meaningful. This has profound implications for analysis as not only does PCA produce uncorrelated data, but it allows one to reduce the dimensions being used in subsequent analysis, which can greatly reduce computation time and lead to easier interpretation of results. Environmental Niche Factor Analysis modifies this theory and methodology for predictors that indicate the how likely an area is to support a given focal species.

3.2 Environmental Niche Factor Analysis

3.2.1 Conceptual Information

Distribution modeling of species is an exciting topic with numerous applications. From determining the precise areas a species inhabits to quantifying suitable conditions in an area that could support a specific species, researchers have been making use of distribution theory to explore what environmental factors are important for a given species' presence. However, each method is based on different assumptions and requires various inputs. For example, general linear models and additive models all require both presence and absence data. That is, one needs data relating to the conditions that result in a habitat being suitable for a species as well as those conditions that are unsuitable. While the applications of such models have been wide spread, absence data is not always obtainable and, even if it is, it often cannot be trusted. For this reason approaches such as Environmental Niche Factor Analysis (ENFA) and Maximum Entropy were developed to allow analysis using presence data only. Here we focus on ENFA, which is based on the concept that a species inhabits a certain niche, or areas where environmental variables interact to form a suitable habitat.

In order for a species to maintain a viable population, they need to inhabit an area where the interaction between environmental factors allows for sustained existence. Furthermore, when considering presence data of a species, their locations will be in areas with appropriate living conditions. Such areas are found within a species niche, which is explored through topic of niche theory (Hutchinson, 1957). Niche theory relies on the belief that species have unimodal distributions along environmental variables in that a given species has greater abundance along a single interval within important environmental factors. Practically, this corresponds to the average value of a given environmental variable in which a focal species can be found representing the optimum value for that species with respect to the given predictor. Comparing this average to the average of a global test space across number of environmental predictors provides the framework for quantifying the types of conditions that result in a suitable habitat for the focal species.

The theory is based on the concept that the species is not randomly distributed in a given area, but instead exist in locations based on the conditions that exist in the local environment. What makes this methodology particularly useful and robust is the need for presence data only. Given a set of presence points that have been collected as a result of a uniform process, information related to what constitutes the species niche can be extracted from the ecological conditions that exist in the specific geographical presence locations. Using the information conveyed by presence points only, ecological conditions in other locations in the study area can be examined to determine whether or not their interaction constitutes a habitable location (Hirzel et al., 2002). As a result, applications of ENFA are bountiful and can be applied to many situations including a hunted species, invasive species, or other populations where presence data is obtainable but absence data is unavailable or unreliable.

Each individual factor will hereafter be referred to as a ecogeographical variable (EGV) and a species will often require a specific combination of these variables in order for a particular area to be deemed a suitable habitat. For example, a species may require certain temperature, food, climate, or other important factors. The factors that are relevant for a given population depends on the particular species, the study area being considered, and the

environment found within the study area. One way explore how a species niche is the result of ecogeographical variables interacting in an environment to produce a preferred habitat, is to examine the distribution of ecogeographical variables in the areas where a species is located compared to the global distribution. The concept of marginality and specialization are introduced in order to help quantify a species niche.

Marginality is a quantification of how the average environmental conditions on the cells where a species is known to inhabit compare to that of the average on the global distribution. Since the average value of an EGV on known presence locations can be considered optimal for a species, marginality aims to describe the distance of the species optimum from the ecological condition in the study area. Specifically, the marginality (M) is stated as the absolute value of the difference between the global mean of an ecogeographical variable (m_G) and the mean in cells where species are present (m_S) (Hirzel et al., 2002). This value is then normalized by dividing by 1.96 standard deviations of the variance of the global distribution (σ_G):

$$M = \frac{\mid m_g - m_s \mid}{1.96\sigma_G}. \tag{3.11}$$

Due to the fact that a randomly chosen value from the global distribution would be expected to lie σ_G away from the mean, dividing by 1.96 σ_G ensures that marginality will most always be between zero and one, with a value closer to one indicating the species has specific environmental preferences relative to the global set. Of course, any species niche will depend on the interaction between a number of ecogeographical variables, and thus a multivariate version of equation 3.11 will be given in a later section.

Specialization is a measurement of how the standard deviation of the cells with species presence compares to the standard deviation of the global ecogeographical variable. Specialization (S) is quantified by dividing the standard deviation of the global distribution (σ_G) with that of the focal species (σ_S) (Hirzel et al., 2002):

$$S = \frac{\sigma_G}{\sigma_S}.$$

A randomly chosen set of cells would be expected to have the standard deviation of the global distribution and thus would have a specialization value close to 1. Similarly, a specialization value grater than 1 would mean that $\sigma_S < \sigma_G$, indicating the focal species is sensitive to changes in their preferences with respect to the given ecogeographical variable. Note again that this formula would only apply when considering a single ecogeographical and that the multidimensional version will be provided later.

The way we will describe the niche of a species in an area is the subset of the area where the species has a reasonable probability to occur, which can be quantified using the concepts of marginality and specialization. However, ecological variables are not necessarily independent. At least two will almost certainly be correlated with each other, and it is difficult to relate two correlated variables to quantify a species niche. This issue is especially worrisome as a species' niche depends on the interaction between a significant number of such variables. To circumvent this issue, a principle component analysis is applied from which a reduced number of predictors can be selected as explained previously.

Many problems in multivariate analysis are aided by methods of describing relevant information in the original observations with a reduced test space (Cooley and Lohnes, 1962). Principle components is one such method and, since it does not require a criterion variable, it is applicable for ENFA's purpose. As described in the previous section, in PCA , axes are generally chosen in to maximize the variance of the distribution. A key distinction in the principal component analysis performed by ENFA is that the first axis is chosen to account for all marginality of the species and the subsequent axes are chosen to maximize specialization. Data used in ENFA are obtained using Geographical Information Systems.

Geographical Information Systems (GIS) is a powerful tool that allows one to explore the connection between a species and its habitat. Environmental data paired with presence points allows one to explore the dynamic between the cells that contain a presence point and those that do not. In short, GIS data is the tool from which one can quantify environmental conditions in an area. This data can then be used explicitly in Environmental Niche Factor Analysis in order to determine the suitable habitat for a species in a given area.

Raster maps are the preferred format in this type of study and can be thought of as a partitioning of the study area in to smaller square pieces called cells, each of which contains a value relevant to the specific map. A raster map can be quantitative, such as elevation, average temperature or frequency of vegetation type, or qualitative such as road maps, soil type or land use. Practically, a raster map can be thought of as an $m \times n$ matrix of values where each i, j entry corresponds to a geographic location. Note that a fine scale could result in a more detailed analysis, but also means more entries in each raster map and higher computation time. One must choose the most relevant scale based on goals of the project, available resources, and computation power (Hirzel et al., 2002). In ENFA, raster maps for relevant ecogeographical variables that cover the whole study area are paired with a raster map of presence data indicating which cells have been known to support the species. After a particular focal species and study area have been determined, one must first obtain the relevant data.

Which ecogeographical variables one chooses are determined by the focal species and study area in question. For example, the distribution of the primary food source(s) of the focal species is usually related to their niche. Specifically, if considering the existence of panda bears one should have information related to the locations of bamboo (not acorns), whereas a study of feral hogs in the Smoky Mountains would instead require data on the distribution of oak trees (not bamboo). The size and composition of the study area can play a factor as well. For example, while ecological factors need always influence species distributions, considering areas with human inhabitants requires the need to measure human influence. Human influence can be quantified by locations and frequencies of roads, impervious surfaces, buildings, land use and other similar metrics.

Prior to conducting an analysis, knowledge of the focal species and what key factors determine their distribution helps choose relevant EGVs. However, since we will apply a principal component analysis that will extract most of the relevant information from the data in far fewer factors by transforming the data to uncorrelated axes, ENFA is rather robust with respect to redundant information in the data and correlations between predictors

(Hirzel et al., 2006). Thus many EGVs should be used initially and only need to be removed if extreme correlations exist. Of course, again due to the principal component analysis, removing an overly correlated EGV does not result in any loss of information as the same relationships will instead be derived from the correlated EGVs that remain. This is contrary to other methods, such as Maxent, in which removing EGVs to reduce correlations between predictors is often required and results in a loss of information (Hirzel et al., 2002).

Assume that we have partitioned the study area represented by an $m \times n$ raster matrix totaling N square cells. The species presence map should be obtained by unbiased sampling and a lack of evidence in a location is simply entered as a 0. Presence can be represented with a 1 or an integer value in cells that warrant weighting. A relatively small number of presence points are required to quantify a species niche and increasing the number of points does not generally provide much additional information (Hirzel et al., 2002).

Assume we have V ecogeographical variables that are relevant to the particular focal species and study area. The EGVs need to be arranged in raster maps that are overlay-able to the species presence/absence map. This means that each EGV has the same number of cells, which are the same size and the extent covers the exact same geographical space. Quantitative data, such as average rainfall or slope are ready to be used in their current form to understand the key characteristics of the locations with known species presence. Qualitative data, such as vegetation type and many human disturbances, must first be quantified. This is due to the fact that marginality is dependent on the mean of each EGV and if the current value of an EGV is not a meaningful numerical value, then the average is also meaningless and the marginality factor will be invalid. Given a raster map containing the location and characteristic of a qualitative variable there are a number of ways to translate the information into a quantitative format including measuring the frequency of occurrence centered around each cell within a certain radius or to calculate the nearest distance from the each cell to the each qualitative feature.

As previously mentioned niche theory relies on the fact that species distributions are unimodal. This comes into play in ENFA as the mean of each EGV is thought to represent

the optimal value and used as such. Though the method is robust to deviations from multinormality, it is theoretically needed in each EGV (Glass and Hopkins, 1970). A common box-cox transformation is performed on all EGVs in order to render them more normally distributed (Sokal et al., 1969).

Once all EGVs are quantitative and transformed via box-cox or a similar transformation, they can then be normalized further to prepare them for the principal component analysis. As mentioned previously, in order to ensure the multinormal data has mean zero we must subtract off the mean of each EGV from the value of each cell and divide by the standard deviation

$$z_{i,j} = \frac{x_{i,j} - \bar{x_j}}{\sigma_{x_j}},$$

where $x_{i,j}$ is the value of variable x_j in cell i, $\bar{x_j}$ is the mean of this variable over all cells and σ_{x_j} is its standard deviation. What results is an $N \times V$ matrix $\mathbf{Z}$ of multinormal data with mean zero and standard deviation of 1.

With the normalized matrix environmental variables $\mathbf{Z}$ in hand, one can compute the matrix

$$\mathbf{R_G} = \frac{\mathbf{1}}{\mathbf{N}}\mathbf{Z^T Z}.$$

Due to its standardization, the $V \times V$ matrix $\mathbf{R_G}$ is a covariance matrix as well as a correlation matrix which describes the interactions between predictors (Hirzel et al., 2002). Letting $\mathbf{S}$ be a subset of matrix $\mathbf{Z}$ where focal species presence is known results in an $N_S \times V$ matrix. Using $\mathbf{S}$, the $V \times V$ species covariance matrix can be obtained through

$$\mathbf{R_S} = \frac{1}{N_S - 1}\mathbf{S^T S}.$$

Note that $\mathbf{R_S}$ is not a correlation matrix as standardization was performed on the global set using global values and not on the species subset using their respective values.

Recall that we wish to perform a principal component analysis to reduce the number of dimensions and uncorrelate the data found in $\mathbf{Z}$ and that PCA is essentially a linear

transformation on the data of the form

$$\mathbf{PZ} = \mathbf{Y}.$$

However, in the general descriptions of PCA, axes are chosen in the direction that maximizes variance. Here, we require the first axis to maximize marginality and the subsequent axes to maximize specialization. The data in $\mathbf{Z}$ by construction already represents of the marginality of the species on a specific variable. Thus, the vector that describes all marginality, $\mathbf{m}$, is given by

$$\mathbf{m_j} = \frac{1}{N_S} \sum_{i=1}^{N_S} z_{i,j} \ .$$

In practice, one calculates this vector first and then sets it aside for later use. Specifically, since we want our predictors to be uncorrelated, we need the remaining axes to be orthogonal to $\mathbf{m}$. This orthogonality requirement will result in the PCA producing $p-1$ non-zero factors for the p predictors and $\mathbf{m}$ will become the pth value by being substituted into the first column of the factor matrix. Before doing this, however, the other factors need to be computed which each maximize the specialization of the species on axes orthogonal to $\mathbf{m}$ (Hirzel et al., 2002).

Letting $\mathbf{u}$ be a normalized vector of the EGV space, the variance of the global distribution on this vector is given by

$$\mathbf{u^T R_G u}.$$

Similarly, the variance of this vector on the species distribution is

$$\mathbf{u^T R_S u}.$$

Thus, referring back to the definition of specialization, maximizing specialization clearly means maximizing $\mathbf{\Theta}(\mathbf{u})$ where

$$\mathbf{\Theta}(\mathbf{u}) = \frac{\mathbf{u^T R_G u}}{\mathbf{u^T R_S u}}. \tag{3.12}$$

The need for uncorrelated principal components results in the additional constraint $\mathbf{m^T u} = 0$. As a result, the previous problem can be stated as

Find a $\mathbf{u}$ such that such that

$$\mathbf{m^T u} = 0$$

$$\mathbf{u^T R_S u} = 1$$

$\mathbf{u^T R_G u}$ is as large as possible, which is needed to maximize specialization.

A change of variables is applied to this system with $\mathbf{v} = \mathbf{R_S^{.5}u}$, $\mathbf{W} = \mathbf{R_S^{-.5}R_G R_S^{-.5}}$ and $\mathbf{y} = \frac{\mathbf{z}}{(\mathbf{z^T z})^{.5}}$ where $\mathbf{z} = \mathbf{R_S^{-.5}m}$. The exponents with .5 are needed to obtain the square root of the matrix. Doing so results in the translated problem

Find a $\mathbf{v}$ such that such that

$$\mathbf{v^T y} = 0$$

$$\mathbf{v^T v} = 1$$

$\mathbf{v^T W v}$ is as large as possible.

The above, and each subsequent, problem is solved by successive eigenvectors of the matrix $\mathbf{H} = (\mathbf{I} - \mathbf{yy^T})\mathbf{W}(\mathbf{I} = \mathbf{yy^T})$, the proof of which is similar to that in the principal component section and can be found in (Hirzel et al., 2002). After solving for the eigenstructure of $\mathbf{H}$, the eigenvectors can be transformed back into vectors $\mathbf{u_i}$ for $i = 1...V$ that solve the original problem. The matrix $\mathbf{U}$ is then constructed, whose columns contain solutions $\mathbf{u_i}$. Recall that, since each $\mathbf{u_i}$ must be orthogonal to the marginality factor $\mathbf{m}$, $\mathbf{u_p}$ is null and thus removed while $\mathbf{m}$ is substituted into the first column of $\mathbf{U}$.

Recall that in a general PCA, eigenvectors of the dispersion matrix maximize variance and are thus the principal components and the corresponding eigenvalues describe how each direction contributes to the variance. Here, by design, the first factor accounts for 100% of the marginality of the species. To represent how much of the specialization of the species is

accounted for by each factor, we obtain each factor $\mathbf{u_i}$'s contribution to specialization. This is achieved by substituting each factor $\mathbf{u_i}$ into equation 3.12 and comparing output. From this, values λ_i for $i = 1...p$ are computed which represent the amount of specialization that each factor accounts for, including the marginality factor. In many cases, the marginality factor can account for more than the specialization factors combined (Hirzel et al., 2002).

3.2.2 Interpreting the Results

After completing all the necessary transformations and calculations with respect to the data one has all the tools to make meaningful statement regarding the focal species niche.

Matrix $\mathbf{U}$, also referred to as the factor matrix or score matrix, contains vital information about marginality and specialization of the species and how they are related to each predictors. Since the first column $\mathbf{u_1}$ accounts for 100% of the marginality of the species, the interpretation of this factor is easy. Values near 0 indicate the species is found in average conditions with respect to the global EGV where positive values means the species can be found in *higher* than average conditions and negative values show that the species is found in locations with *lower* than average values. The eigenvalue associated with this factor, λ_1, describes its contribution to explaining the specialization of the species. An overall marginality condition for the entire study area can be computed as

$$M = \frac{\sqrt{\sum_{i=1}^{V} m_i^2}}{1.96},$$

which is only useful to compare marginality between different species.

The remaining factors $\mathbf{u_i}$ for $i = 2, ..., p$ maximize specialization in orthogonal directions. Since specialization of the species is spread across the remaining $p-1$ factors, interpretation is not as straightforward as the first. The sign for these factors is arbitrary and thus values the farther from 0 simply indicate that the species is more restricted on the values it considers suitable with respect to the given variable. The eigenvalues associated with each factor

describe its given contribution to explaining the specialization of the species. Similar to marginality, a global specialization factor can be computed for the entire study area in order to compare between species. Global specialization is characterized by

$$S = \frac{\sqrt{\sum_{i=1}^{V} \lambda_i}}{V}.$$

Similar to the general PCA case, eigenvalues usually decrease rapidly allowing for significant reduction in dimensions. In fact, 100% of the marginality and nearly all of the specialization of a species are typically expressed by only the first several factors. Also similar to a PCA, how many factors to keep can be determined by one of several criterion. The creators of ENFA recommend using the Broken Stick Method (Wilson, 1991).

3.2.3 Creating a Habitat Suitability Map

One could think of a number of ways to use the aforementioned derived information to determine which cells in the entire study area are suitable to support the species. The authors of (Hirzel et al., 2002) considered a number of methods and have included four in their BIOMAPPER software (Hirzel and Perrin, 2002). We will discuss one of the four as it is the most robust and is what we used in our analysis.

We use the method referred to as the distance geometric-mean algorithm to determine a habitat suitability value for each cell in the study area. This method aims to measure the distance between each given cell and each observation points in the factor space (Hirzel et al., 2006). Specifically, for any point P in the transformed factor space, the geometric mean of the distances (H_G) to all observations O_i is calculated as

$$H_G(P) = \sqrt[N_s]{\prod_{i=1}^{N_s} \delta(\mathbf{P}, \mathbf{O_i})},$$

where δ represents the distance in the factor space (Hirzel et al., 2006). Since areas with dense presence points are thought to best represent the species niche, these locations reinforce the environmental conditions that exist in those locations and result in higher habitat suitability values. Furthermore, unlike the medians algorithm, the geometric mean algorithm makes no assumption about the shape of the species distribution and results in a more smooth map (Hirzel et al., 2006). Though this method is more computationally intensive and thus takes longer compared to other options, its results provide a good generalization of the species niche and is thus preferred. We also achieved far superior validation results when using the geometric mean algorithm.

3.2.4 Validating a Habitat Suitability Map

Most evaluation measures previously used were based on presence/absence (or presence/pseudoabsence) data and strongly depend on a subjective habitat suitability cutoff threshold. Since presence data used in ENFA is typically accompanied by nonexistent, or untrustworthy, absence data the authors of (Hirzel et al., 2002) developed a new approach that has gone through several stages of evolution. We will focus here on the evaluation techniques that currently exist in BIOMAPPER 3.0. These techniques include calculating an absolute validation index (AVI) and corresponding contrast validation index (CVI) in addition to the preferred method of the continuous Boyce index (Boyce et al., 2002).

The AVI and CVI metrics were originally derived for BIOMAPPER 1.0. The absolute validation index is described as the proportion of presence evaluation points falling above some fixed threshold (e.g. 0.5) and varies from 0 to 1. The contrast validation index is the difference between the AVI and a chance model that predicts presence everywhere. Similar to other validation approaches, the AVI and CVI values suffer from the fact that an arbitrary threshold must be chosen. Due to this fact the authors of (Hirzel et al., 2006) implement the a continuous version of the Boyce index.

Instead of enforcing a cutoff threshold, the Boyce index partitions the habitat suitability range into b classes. For each class i, the predicted frequency of evaluation points, P_i, and the expected frequency of evaluation points, E_i are calculated. Let p_i be the number of presence points predicted by the model to fall in habitat suitability class i and let a_i be the number of cells belonging to habitat suitability class i. Then

$$P_i = \frac{p_i}{\sum_{j=1}^{b} p_j}$$

and

$$E_i = \frac{a_i}{\sum_{j=1}^{b} a_j}.$$

Then, for each class i, a predicted-to-expected (P/E) ratio F_i can be calculated as

$$F_i = \frac{P_i}{E_i}.$$

If a habitat suitability model is functioning properly, it will correctly categorize areas with low, moderate and high suitability. Since low suitability areas should contain few presence points, one should expect $F_i < 1$ when i represents a low suitability class. Similarly, adequate suitability classes should have F_i values monotonically increasing past 1 as i increases. This monotonic increase is quantified by the Spearman rank correlation coefficient between F_i and class i and is referred to as the "Boyce Index", B_b (Boyce et al., 2002). Boyce index values range from -1 to 1 with positive values resulting from predictions that are consistent with the presence distribution, a value of 0 means the model does not differ from a chance model and negative values indicate an incorrect model is being used. The main shortcoming of this method as described is that it is sensitive to the number of classes b and to their boundaries. To combat this, a continuous version of the Boyce index was formulated. Instead of using fixed classes, a moving window approach is used where computation of the F curve is carried out on a small class size, say $[0, W]$ for $W = 0.1$, before the window is gradually increased

to 1 by a fixed, small step size. At each window size, the P/E curve is plotted against the average suitability value of the class, $W/2$. Thus, instead of a discrete curve subject to class size and boundary issues, a smooth and continuous curve is created from which a continuous Boyce index can be computed. This process is expanded further by performing a variance analysis via a cross-validation method.

A k-fold cross-validation method is a resampling approach that allows assessment of the robustness of the above measure (Hirzel et al., 2006). The approach begins by dividing the data set into k independent partitions. Then $k-1$ of them are used to calibrate the model and the final partition is used to evaluate the model. This process is repeated k times, using each partition sequentially to evaluate the model. What results is k estimations of the evaluator which allows one to determine the tendency of the model and calculate its variance (Hirzel et al., 2006). Since a variance calculation was one of the things lacking from ENFA, this new approach was a welcome addition to the BIOMAPPER software.

3.3 Habitat Suitability of Feral Hogs in Great Smoky Mountains National Park

3.3.1 Introduction

We now create a habitat suitability map for Feral Hogs in Great Smoky Mountains National Park using the previously described methodology. Two recent papers were written about wild hogs in GSMNP using removal data provided by the Park and vegetation data obtained via LIDAR (National-Park-Service, 1980; Madden et al., 2004). The first paper contains an agent based model in which individual hog behavior is modeled in order to examine population dynamics (Salinas et al., 2015). The second paper has a discrete metapopulation model that uses parameters that were carefully estimated using available data (Levy et al., 2015). Historical effort levels of the control program were carefully estimated by comparing output of the metapopulation model with the removal data. Model simulations indicate that

the control program is important in limiting the population (Levy et al., 2015). One aim of this paper is to scientifically inform control efforts by illustrating locations where hogs can be effectively and efficiently removed from the Park.

The control data only conveys information related to the presence (not absence) of hogs in the Park. For this reason the harvset data lends itself to presence-only methods such as Environmental Niche Factor Analysis (ENFA) (Hirzel et al., 2002). The theory aims to quantify a species niche by relating presence points to ecogeographic variables. We can then scientifically and mathematically assess potential wild hog locations with the goal of limiting their population within Great Smoky Mountains National Park (GSMNP). Using this approach we are able to derive a detailed map of suitable locations for hog presence, as well as a map showing locations where one would be most likely to encounter hogs. Thus, results produced from this method have applications in illustrating wild hog habitat in the Park while also suggesting potentially bountiful places to hunt. This information can be used to increase our understanding of the population, evaluate historical hunt sites, and provide insight into other potentially fruitful locations. This methodology could also prove to be useful for other species found in GSMNP such as flying squirrels and Indiana bats. Since both of these species are threatened, understanding and protecting the geographic locations of their habitat could be invaluable to their recovery.

A description of Environmental Niche Factor Analysis will first be presented. Then our methods and the data used in the analysis will be discussed. A Niche Factor Analysis will the be applied to the data in order to assess wild hog preferences in GSMNP. Using the resulting information, two map products are produced and validated using a continuous Boyce index (Boyce et al., 2002; Hirzel et al., 2006). Finally, we present conclusions from our map products and analysis.

3.3.2 Methods and Data

An Environmental Niche Factor Analysis uses the concepts of marginality and specialization to determine habitat preferences and model them spatially (Hirzel et al., 2002). Marginality

can be measured as the difference between the average conditions where the species is found and the average conditions of the study area. Specialization is the ratio of the variance of the conditions where the species can be found and the variance of the conditions of the study area. Together, marginality determines the types of conditions the species prefers and specialization measures how sensitive the species is to deviations from their preferences.

In order to measure these multiple variable dependent values, the powerful theory of principal component analysis is used in an innovative way to produce uncorrelated environmental predictors and conduct a factor analysis (Hirzel et al., 2002). Ecogeographic preferences can then be determined and used to create two maps relevant to hog presence in the Park. Results of these maps are validated using a predicted-to-expected ratio and a continuous Boyce index, which will be explained later.

Since each data point represents a location where a hog was removed, the conditions are suitable for both hog presence as well as a successful hunt. For this reason even though the presence-only data is likely biased by the behavior of hunters, a distribution map derived from successful hunting locations using ENFA conveys important information for the control program.

The study location is GSMNP, which is almost entirely undeveloped which limits the number of human-related ecogeographical variables to roads, trails and buildings. All of the data was initially processed into raster maps using ESRI ArcMap 10.1. All maps were re-projected to NAD 1983 UTM Zone 17N using a GSMNP boundary file to set the extent of each map and to clip the data to the study area. Given the size of the study area, available data and goals of the project, each raster map was constructed with a cell size of 30m by 30m. This resulted in 2899 columns and 1308 rows for a total of 3,791,892 cells. Some data preparation and all analysis were then completed in BIOMAPPER 4.0, a software developed for ENFA (Hirzel and Perrin, 2002).

All human related data was provided by GSMNP (National-Park-Service, 2011). Roads and buildings are fairly uncommon within the Park and these variables were quantified in each cell with a closest Euclidean distance analysis performed in ArcMap 10.1. Trails, however,

are more bountiful with certain areas of the Park containing larger densities of trails than others. Due to this fact, trails in the Park were quantified by the area covered by trails within a 2 km diameter centered on each cell using the frequency tool in BIOMAPPER (Hirzel and Perrin, 2002). Using a user supplied radius, the frequency tool measures the fraction of total area covered by the feature in a circle surrounding each cell. A 2 km radius was chosen after reviewing results using a number of different values.

Geographic data includes slope and elevation. Elevation was obtained from the National Elevation Dataset (NED) (Gesch et al., 2002). The digital elevation map was then used to derive the slope within the Park.

Not only is GSMNP almost entirely undeveloped, it is surrounded by national forests increasing the sheer magnitude of wilderness. As a result, it makes sense that hog preferences in the Park would be driven primarily by environmental factors related to vegetation, climate, food and water. A GIS vector map of locations of streams from GSMNP was quantified using the frequency tool in BIOMAPPER described previously (National-Park-Service, 2011).

Vegetative data includes understory vegetation and food preferences, as well as general growing conditions expressed by the normalized difference vegetation index. Detailed data related to the types of vegetation growth found throughout the Park was obtained via lidar and provided by (Madden et al., 2004). This data includes overstory and understory values, both of which were used in the model. Since hogs root for part of their food source, ground vegetation is an important part of a hog habitat. Understory vegetation was extracted from this source into three major categories that collectively cover 85% of the park. These categories include Rhododendron, Kalmia and Herbaceous and Deciduous understory. A frequency map was derived from each understory category in ArcMap. Another indication of ground vegetation is given by the normalized difference vegetation index (NDVI), which was developed by Rouse et al. (1974) and is commonly used to assess the growing condition of green vegetation. The NDVI is numeric data and was obtained from the Global Land Cover Facility (Carroll et al., 2003).

As previously discussed, food sources are believed to play a large role in the behavior and life cycle of hogs (Scott and Pelton, 1975; Singer et al., 1981). As such, overstory vegetation information from (Kirkpatrick and Pekins, 2002) was used to create a food preference map based on the amount of kilocalories produced per acre in each cell during the fall months when food is most bountiful. Categorical values range from 1 to 5 the summary of which can be seen in Table 3.1. As each of these maps is categorical, they were quantified using the same frequency method.

Table 3.1: Categorizing food sources based on the average amount of kilocalories produced per acre in the Fall. There is no category 4 as there are no locations in GSMNP produce between $9,000 - 11,000$ kilocalories per acre.

Category	Kilocalories per Acre	Dominant Species
1	$< 5,000$	Spruce Fir
2	$5,000 - 7,000$	Oak, Pine and Northern Hardwoods
3	$7,000 - 9,000$	Oak and Pine
5	$> 11,000$	Cove Hardwoods

To consider the influence the climate found in GSMNP has on the hog habitat, 19 variables derived from temperature and rainfall data over the last 50 years were obtained from (Hijmans et al., 2005). Since our study area is relatively small, temperature and rainfall values are similar across the Park which results in many of the maps being highly correlated. This results in most of the maps producing similar data in the factor analysis, requiring most maps to be discarded. Due to this fact, we only used average precipitation and average temperatures in the model. Even though most of the 19 variables could not be used, the two that were kept were the largest contributors to marginality and specialization. Furthermore, the high correlation between temperature and/or precipitation with the other variables mean they convey essentially the same information in the factor analysis.

Hog location data was provided by Great Smoky Mountains National Park and consists of the age, sex and geographic location of nearly all hogs removed in the Park since 1980 (National-Park-Service, 1980). The data can be categorized by hogs that were trapped or hunted. Traps draw hogs into certain locations and thus may not accurately represent the

species niche. However, hunted hog locations result from park employees hiking through the backcountry throughout the park in search of hogs and thus better represent the species niche. With this in mind we chose to use locations of hunted hogs as our presence points, which results in 1,553 unique entries. See Figure 3.1 for a map of the locations. Note that even if this data is biased due to hunter preferences and behavior, it is suitable for this analysis since we wish to determine hunting locations that best limit the population. Also, hunting takes place at all locations where a trap is placed. Thus, although we are not considering trap locations explicitly, our results can be used to determine appropriate locations to place traps as well.

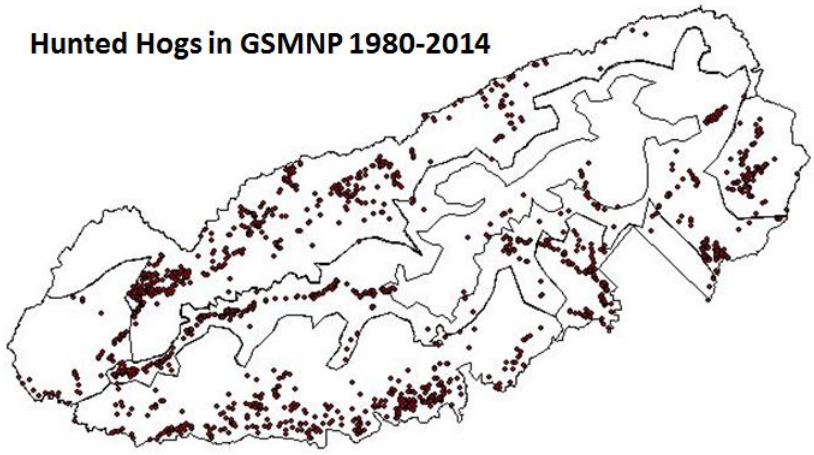


Figure 3.1: Locations of hunted hogs from 1980-2014 used as presence points in the analysis equals 1,553 in total.

Although all the data was obtained in ESRI format, BIOMAPPER requires Idrisi format. Converting into Idrisi format was carried out using Global Mapper and final preparation as well as analysis was conducted in BIOMAPPER 4.0. A summary of all EGVs used in our analysis can be found in Table 3.2. The mean and standard deviation values are useful when considering the results of the factor analysis in Table 3.3. In order to ensure the data took a Gaussian shape, the Box-Cox transformation was applied to all EGVs as recommended by (Hirzel et al., 2002). This assists in ensuring the data take a Gaussian shape, though the method is robust to deviations from normality (Hirzel et al., 2002).

Table 3.2: List of the 16 ecogeographical variables used in the analysis. Precipitation values are measured in mm, distances in m and frequency values in percent covered.

EGV	Minimum	Maximum	Mean	S.D.
Elevation	270	2,026	1,005	460
Slope	0	61	15	283
Distance to Roads	0	13,114	3,210	2,748
Distance to Buildings	0	24,306	7,191	4,863
Frequency of Trails	0	99	17	18
Frequency of Rhododendron	0	98	30	23
Frequency of Herbaceous Veg. & Deciduous Shrubs Understory	0	100	46	24
Frequency of Kalmia Understory	0	100	48	33
Calorie Level #1 Freq.	0	100	28	33
Calorie Level #2 Freq.	0	91	22	19
Calorie Level #3 Freq.	0	92	29	20
Calorie Level #5 Freq.	0	96	21	17
NDVI	20	250	162	23
Frequency of Streams	0	19	7	2
Average Temperature	6	14	11	1.5
Average Precipitation	100	148	129	9

3.3.3 ENFA Results

After performing the niche factor analysis relating our EGVs to presence locations, we wanted to retain enough factors to explain at least 80% of the total information in the data. The values of the EGVs on each of the 8 factors is shown below in Table 3.3. The first factor explains 100% of the marginality and factors 1-8 account for 70% of specialization for a combined 85% of total information explained by the model. Thus, eight factors were used to compute the habitat suitability map. For the marginality factor values, positive values indicate the data points are found in locations that contain higher than average values with respect to the given variable, values near zero indicate a preference for average conditions and negative values indicate presence in locations with lower than average values. Table 3.3 is ordered by decreasing absolute marginality values to illustrate the strongest preferences for feral hogs in GSMNP. Only the magnitude of the specialization factor values is important, not their sign. Larger magnitude indicate restricted ecological tolerance compared with the overall range of conditions in the study area, and magnitude closer to zero indicates that the population deviates from their preference.

The largest contributors to marginality are slope, frequency of rhododendron and frequency of herbaceous vegetation and deciduous shrubs. As Table 3.3 indicates, hogs prefer areas with slope and frequency of rhododendron values far less than average (-0.510 and -0.399). Furthermore, they prefer locations with a higher than average frequency of herbaceous vegetation and deciduous shrubs (0.365). Factors 2-8 each accounted for between 9-12% of the specialization in the model. Since one factor did not dominate, in order to interpret the information the average magnitude of each EGV was calculated weighted by the amount of specialization explained by each factor. The weighted average indicated that wild hogs in GSMNP are most sensitive to changes in elevation and average temperature. The weighted average of all other specialization values were fairly low relative to temperature and elevation, indicating that hogs are not highly specialized animals. These findings are consistent with the widespread and resilient nature of the species and agree with past research (Scott and Pelton, 1975; Singer et al., 1981).

Table 3.3: Coefficients of the variables generated by the principal component analysis in ENFA arranged in decreasing order by their value on the marginality factor.

EGV	**Marg.**	**Spec. 1**	**Spec. 2**	**Spec. 3**	**Spec. 4**	**Spec. 5**	**Spec. 6**	**Spec. 7**
Slope	-0.510	-0.019	0.064	0.012	0.016	0.059	-0.247	-0.128
Frequency of Rhododendron Understory	-0.399	0.061	-0.175	-0.015	0.303	-0.216	0.078	0.175
Frequency of Herbaceous Veg. & Deciduous Shrubs Understory	0.365	-0.035	-0.082	-0.010	0.076	0.002	0.107	-0.092
Distance to Roads	-0.260	0.010	0.044	0.014	-0.054	0.021	0.184	0.093
Frequency of Trails	0.239	-0.002	0.010	-0.048	-0.054	-0.040	-0.221	-0.015
Elevation	-0.238	-0.641	-0.486	-0.588	0.472	-0.704	-0.311	-0.734
Average Precipitation	0.236	0.153	0.215	0.013	0.066	-0.237	-0.361	-0.041
Average Temperature	0.232	-0.727	-0.680	-0.757	0.138	-0.600	-0.226	-0.542
NDVI	-0.181	-0.017	-0.030	-0.001	0.251	0.063	-0.064	-0.127
Calorie Level #3 Freq.	0.169	-0.066	-0.030	0.202	0.434	-0.077	-0.323	0.096
Frequency of Streams	0.158	0.010	-0.012	0.041	-0.002	-0.027	0.474	-0.233
Calorie Level #5 Freq.	-0.143	-0.011	-0.169	0.125	-0.356	-0.021	-0.030	-0.094
Calorie Level #1 Freq.	0.141	0.123	-0.332	0.079	0.035	0.080	-0.225	-0.018
Frequency of Kalmia Understory	-0.116	-0.057	-0.181	-0.095	-0.227	0.086	-0.074	-0.005
Calorie Level #2 Freq.	0.097	-0.092	-0.147	0.066	0.446	0.104	0.356	0.112
Distance to Buildings	-0.083	-0.006	-0.128	0.000	-0.147	-0.004	0.209	-0.025

3.3.4 Map Creation and Validation

Two map products were produced from the niche factor analysis. The first is a habitat suitability map that ranks each cell in the study area from unsuitable to most suitable (Figure 3.3). This map was derived based on hog preferences evident from the factor analysis and categorizes the areas in GSMNP in terms of conditions needed to support wild hogs. The second map product was constructed using validation results and labels each cell in the Park with the likelihood of encountering a hog at the given location (Figure 3.6). This is accomplished by emphasizing the regions within the suitable classes that contain the most predicted presence points by the model. Both maps convey information related to hog presence in the Park, but each with a different purpose. While the habitat suitability map ranks locations in terms of environmental conditions, the likelihood map may be more appropriate for finding hogs to remove for the control program.

The first map product was derived using the results of the factor analysis paired with the geometric mean algorithm in BIOMAPPER (see Figure 3.3) (Hirzel and Perrin, 2002). In order to estimate a habitat suitability value at each cell in the study area the geometric mean algorithm measures the cumulative distance, in the factor space, of the environmental conditions at each cell from the conditions at all presence locations. The performance of the model was evaluated using a k-fold cross-validation procedure that produces k estimates of the number of predicted points (P) and expected points(E) for increasing habitat suitability levels. The value of E is proportional to the area of the map covered by the given habitat suitability range. The value of P is derived from the number of presence points predicted by the model to fall within the given habitat suitability range. The ratio of these values can be represented as a continuous plot for increasing habitat suitability levels and is known as the P/E curve. We took $k = 10$ partitions along with a random seed following the method described in (Hirzel et al., 2006). The method produces 10 estimates of the P/E curve, which allows us to determine the accuracy and consistency of the model. A model is deemed accurate if the P/E ratio is less than 1 for low habitat suitability values and increases monotonically past 1 as the habitat suitability range is expanded. The Boyce index

ranges from -1 to 1 and measures this monotonic trend (Boyce et al., 2002; Hirzel et al., 2006). Positive values indicate the model predictions are consistent with the distribution of presence points with values near 1 indicating the most accurate models (Hirzel et al., 2006). Values near zero indicate the model's predictions are no better than chance. Negative Boyce index values indicates the model predicts too many presence points fall within poor suitability areas. As a measurement of the consistency of the model, a variance can be calculated using all 10 trials and appended to the Boyce index value. See Figure 3.2 to view the general trend of the P/E curve surrounded by the variance derived from all 10 trials. The curve is generally monotonic with any decrease reasonably within the displayed variance. A Boyce index value of 0.936 ± 0.049 measures this trend and indicates that the habitat suitability map is both accurate and consistent.

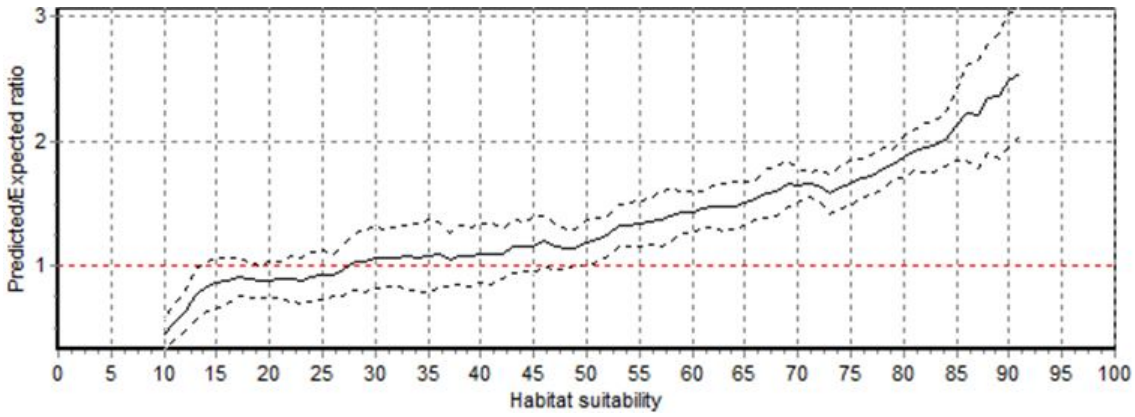


Figure 3.2: Ratio of predicted presence points to expected presence points for increasing habitat suitability values.

The P/E curve was then used to re-classify the suitability map into the 5 classes seen in Figure 3.3. Since the confidence interval surrounding the curve was fairly uniform, distinctions between classes were solely based on the shape and values of the curve. How each class was partitioned is shown in Table 3.3. By comparing presence locations to model output, Figure 3.7 allows one to visually verify the trend of the P/E curve and Boyce index values.

Table 3.3: Explanation of how each class in Figure 3.3 was partitioned using the information in Figure 3.2.

Classification	Habitat Suitability Range	Reasoning
Unsuitable	$0 - 10$	No presence points
Mildly Suitable	$10 - 25$	$P < E$ in this range
Moderately Suitable	$25 - 50$	P slightly larger than E in this range
Suitable	$50 - 70$	P significantly larger than E in this range
Most Suitable	> 70	Steepest part of the P/E curve

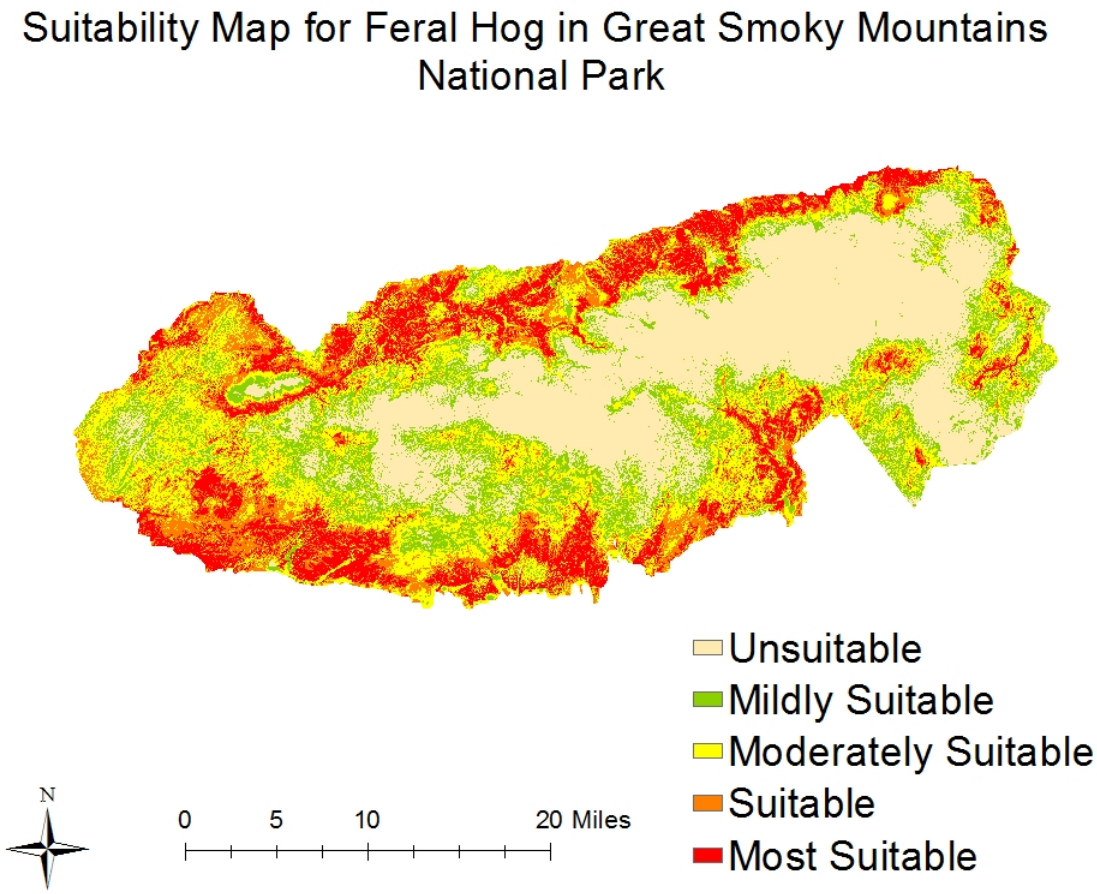


Figure 3.3: The 5 classes were partitioned using trends and standard deviation of the P/E validation curve shown in Figure 3.2.

In addition to a reaffirming continuous Boyce index value indicating an accurate and precise model, the suitability map is also consistent with past research and qualitative information received by Park rangers (Scott and Pelton, 1975; Singer, 1981; Singer et al., 1981; Stiver, 2014). The unsuitable class is predominately located in areas of the Park that exhibit high elevation. High elevation is related to lower average temperatures as well as the types of vegetation and food sources that hogs do not prefer (Scott and Pelton, 1975; Singer et al., 1981). Although hogs are believed to range into the higher elevations during the summer months, it is out of necessity, not preference. In contrast, the top three suitable classes are at low elevations where the slope is relatively flat and where oak trees, their favorite food source, are dominant. These locations are also where herbaceous vegetation and deciduous shrubs understory are most prevalent. Due to their lower elevation, these areas have a higher average temperature and precipitation as well, for which hogs evidently have a preference based on their values in the marginality factor.

One can see the implications of ENFA results when examining the output from our model. For example, the highest contributions to marginality were shown to be preferences for slope and two understory categories. When carefully examining the habitat suitability map it is clear that slope is significant in determining suitable locations (see Figure 3.5a compared to Figure 3.5b). Further distinctions between appropriately flat mild/moderate suitability locations and appropriately flat suitable/high suitability locations are being made in part due to understory values (see Figure 3.5a compared to Figure 3.5b and Figure 3.5c). Though some results can be deciphered using the naked eye, some cannot. For example, it is not obvious how the red areas in the likelihood map are determined and is likely a result of the model picking up on important and complex relationships that exist between the population and the environment. For this reason, in order to learn more about the wild hog population, these locations in the habitat likelihood map should be investigated in person in order to further validate the results of the model.

Another useful result from this analysis is the creation of a map depicting where one is most likely to encounter hogs based on the model findings. Again using the P/E curve,

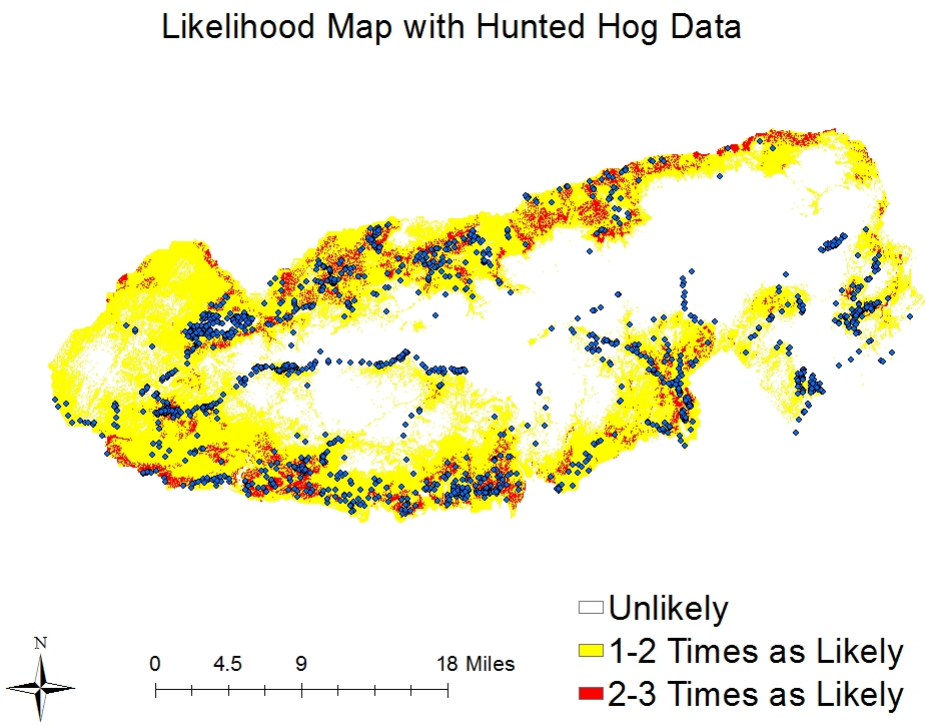
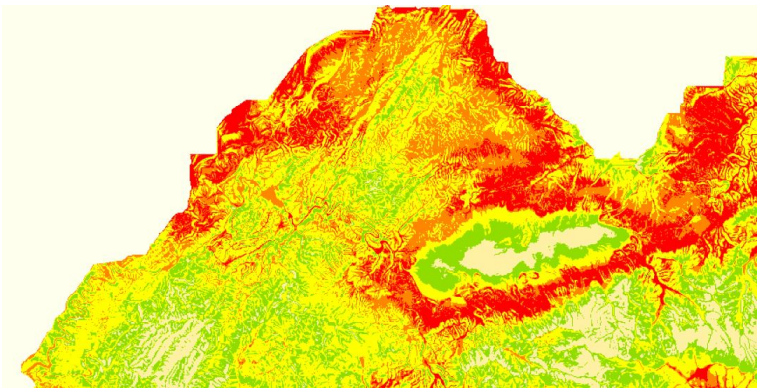


Figure 3.4: The likelihood of encountering a hog map superimposed with hunted hog data.

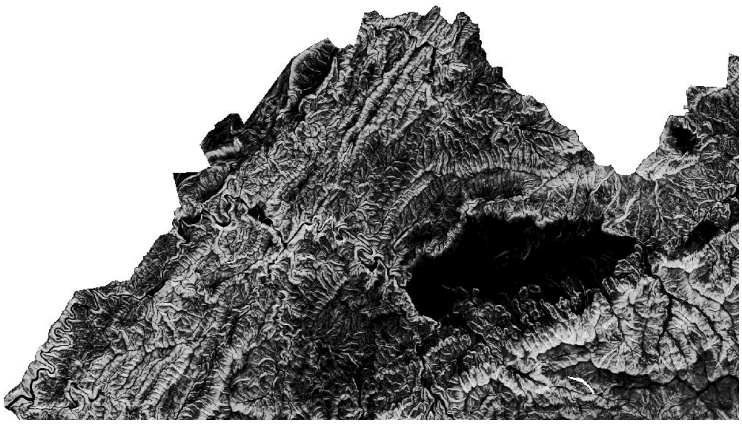
we partition the results into locations where one is less likely to encounter a hog compared with random chance ($P/E \leq 1$), cells where one is 1-2 times as likely to encounter a hog compared with random chance ($1 < P/E < 2$) and locations where one is 2-3 times as likely to encounter a hog ($2 \leq P/E \leq 3$). This map in Figure 3.6 illustrates subsections within the suitable areas in Figure 3.3 that one is most likely to encounter a wild hog. These areas could potentially have the highest hog densities in the Park, which would shed light on hog preferences while also being invaluable to park officials in terms of informing management strategies.

3.3.5 Conclusions

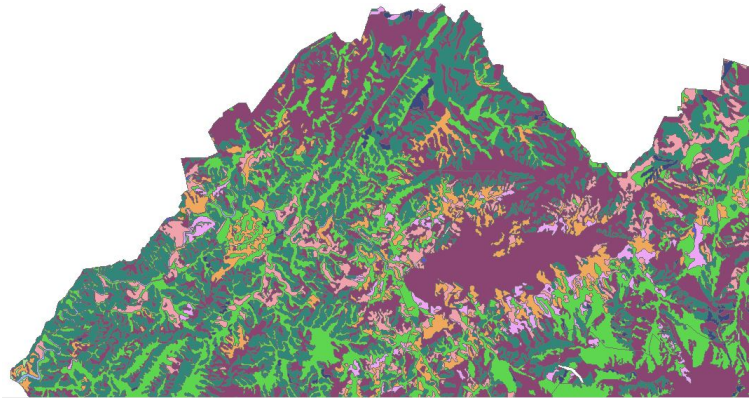
The presence of wild hogs (*Sus scrofa*) has been a concern to officials of Great Smoky Mountain National Park since the late 1940s. Their destructive nature and status as disease



(a) Habitat Suitability



(b) Slope



(c) Understory

Figure 3.5: A zoomed view of the northwest corner of Great Smoky Mountains National Park for three different maps. The oval region is a high-elevation cove known as Cades Cove.

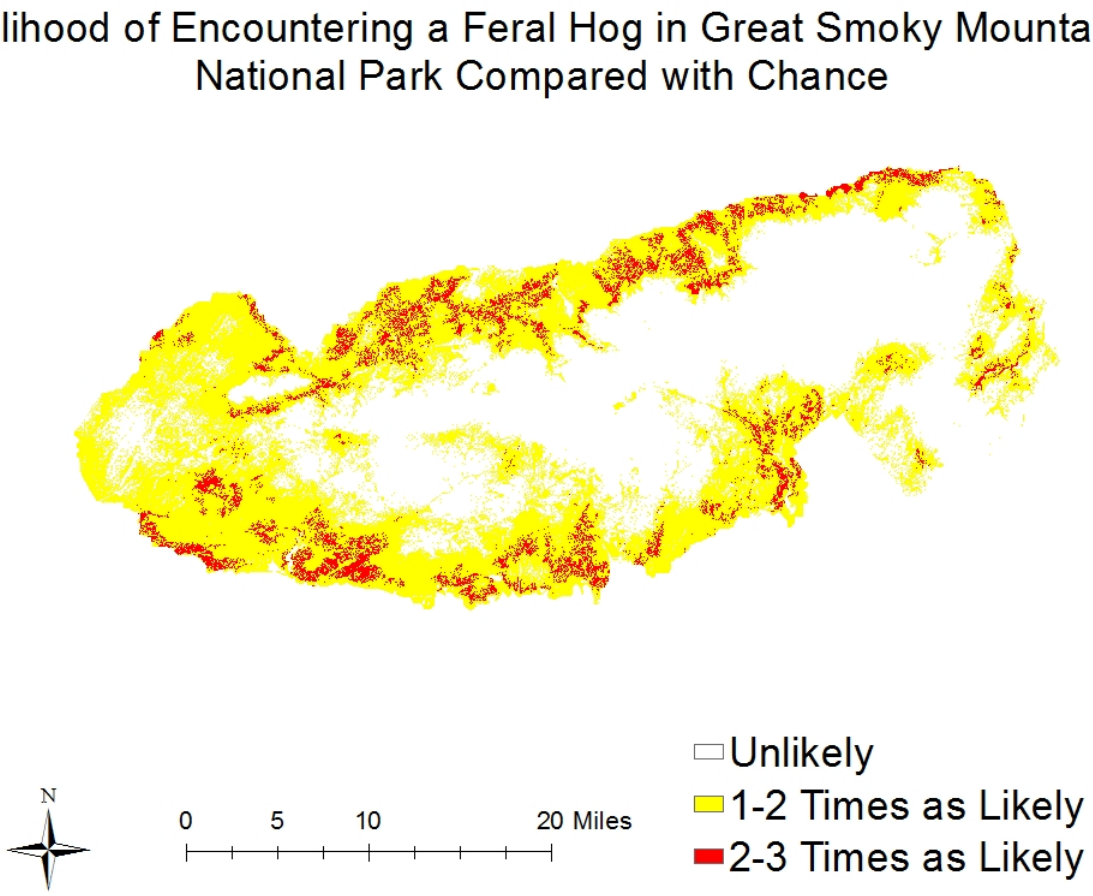


Figure 3.6: Map showing the likelihood of encountering a hog throughout Great Smoky Mountains National Park partitioned using the P/E curve shown in Figure 3.2.

carriers prompted the creation of a control program over 50 years ago with records that have aided this research. However, since little is known about the population, locations of focused control efforts have been based solely on historical success and recent reports of hog presence. We have used control data from hunted hogs over the past 34 years and relevant ecogeographical variables to create two maps to assist the understanding of wild hog preferences and help determine good locations to hunt. We are able to make use of the two map products when assessing each objective.

Both maps increase our understanding of the preferences and whereabouts of this invasive exotic species. Since successful hunt locations convey adequate conditions for wild hog presence, we are able to create a general habitat suitability map for the population in the Park. The model examines the presence location across all EGVs and illustrates other locations throughout the park whose environmental conditions are also deemed to be suitable for hog presence, shown in Figure 3.3. This can be explored further by looking at Figure 3.6, as this second map displays areas throughout GSMNP that may harbor a high density of hogs. These apparent hot spots of hog activity may convey unknown information related to the behaviors and preferences of this population.

The map products can also directly relate to the control program as they allow us to evaluate past hunting locations while also highlighting other potentially fruitful areas to explore for hog presence. As you can see in Figure 3.7, while there are a number of additional habitats for hunters to explore, the vast majority of removals lie within the cells deemed most suitable by the habitat suitability model. The second map we created (Figure 3.6) is also highly relevant to hunting efforts. Notice that all cells are a subset of the locations from the general habitat suitability map with the red areas illustrating possibly the best places to hunt as the model has deemed them most likely to contain hogs. Similar to the previous map, the red areas contain many of the historic removals with plenty of additional untapped locations being clearly displayed as well (see Figure 3.4). Thus, historic hunting sites can be appreciated while also examining alternative locations for future hunts in the orange/red areas throughout the park that have yet to result in a hog kill.

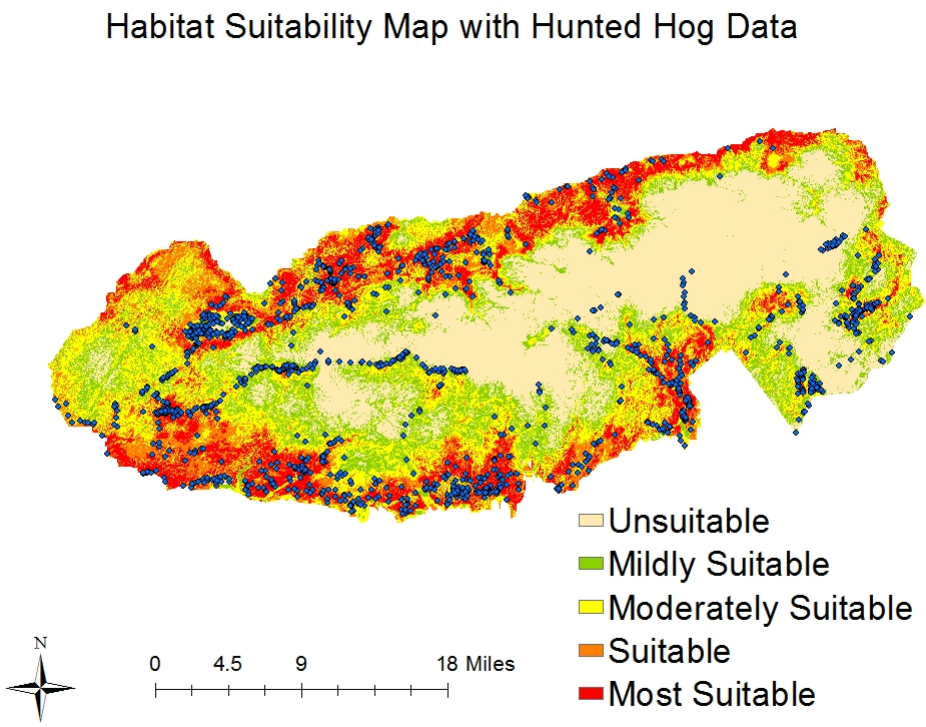


Figure 3.7: The habitat suitability map superimposed with hunted hog data.

One might argue that the results presented here are skewed because of bias that may exist in the data points as a result of hunter preferences. However, even if this were true, the bias introduced by the hunters will still produce results relevant to the control program. That is, whatever bias that exists is a result of the preferences intrinsic to hunters and each location lies in the intersection of locations that are both accessible to hunters as well as suitable for hog presence. For example, though the model indicates that presence locations occur in cells with below average frequency of rhododendron, it is not clear whether this is due to hog preferences or hunter behavior. Nevertheless what *is apparent* is how kills are more likely to occur in locations that lack rhododendron, which is a useful piece of information regardless of its driving cause. In this light, the hunting data may even better lend itself to evaluating hunting choices compared with unbiased presence data.

We can view both maps and make recommendations for future control activities. For example, Figure 3.4 shows that the total area where the likelihood of encountering a hog is $1-3$ times as likely compared with random chance dwarfs the area where control efforts have historical taken place illustrating additional potential hunting sites. It is also clear that a number of areas near the edge of the Park generally contain more suitable conditions and should to be exploited. Furthermore, since many of the suitable locations are on the edge of the Park, it begs the question of the hog-related activities taking place directly adjacent to these locations but on the outside of the Park. Areas of particular interest include Nantahala National Forest, Cherokee Reservation and the Cosby/Gatlinburg area. Although few historic removals have taken place in the northwest corner and far eastern border of the Park, Figure 3.6 suggests a high likelihood of encountering a hog in these areas indicating that they may be effective and efficient places to hunt. Finally, although many kills occur in open fields in areas such as Cades Cove and Cataloochee, the model indicates the wooded areas surrounding these fields are the suitable habitat for wild hogs (see Figure 3.5). For this reason, efforts should continue in the woods surrounding these fields.

It is easy to understand why some of locations that the model predicts to be highly suitable for hog presence have yet to produce kills, while other locations need to be studied further. For example, it is easy to understand that although the region in the northwest corner of GSMNP has good conditions for hog presence, not a single control kill has been produced in this area of the Park as it very remote with few trails and only a single dirt road. On the other hand, locations such as the far eastern border that are in fact accessible but are yet to produce hog kills are particularly interesting and may be worth exploring for signs of wild hogs. Such inaccessible or unexploited locations where the model indicates quality conditions for wild hog presence are concerning as they may serve as reservoirs for the population that work against the control efforts of the Park. The only way to determine this for sure is to carefully examine such locations for hog presence.

Chapter 4

Modeling Pseudorabies in the Population

4.1 introduction

Hogs are a vessel for disease with porcine parvovirus, leptospirosis, toxoplasmosis and pseudorabies each found in various serological surveys of hogs in GSMNP (Cavendish et al., 2008; Sandfoss et al., 2012). Beginning in 2001, GSMNP partnered with the North Carolina Department of Agriculture and Consumer Services (NCDACS) as well as the Tennessee Department of Agriculture (TDA) and the U.S. Department of Agriculture, Animal and Plant Health Inspection Service, Veterinary Services (APHIS) to begin monitoring wild hogs in the Park for disease. Specifically, 42.5% of harvested hogs from 2001-2013 were tested for pseudorabies and brucellosis, as they pose the greatest threat to humans and the domestic swine industry (Cavendish et al., 2008). Though no hogs have tested positive for brucellosis, an increasing number have tested positive for pseudorabies over the years.

From 2001-2004, all blood samples taken from harvested hogs in GSMNP tested negative for pseudorabies (PRV). However, hogs began testing positive for PRV starting in 2005 with the prevalence increasing steadily reaching as high as 56.9% (see Figure 4.1 and Table 4.1) (Cavendish et al., 2008; National-Park-Service, 1980).

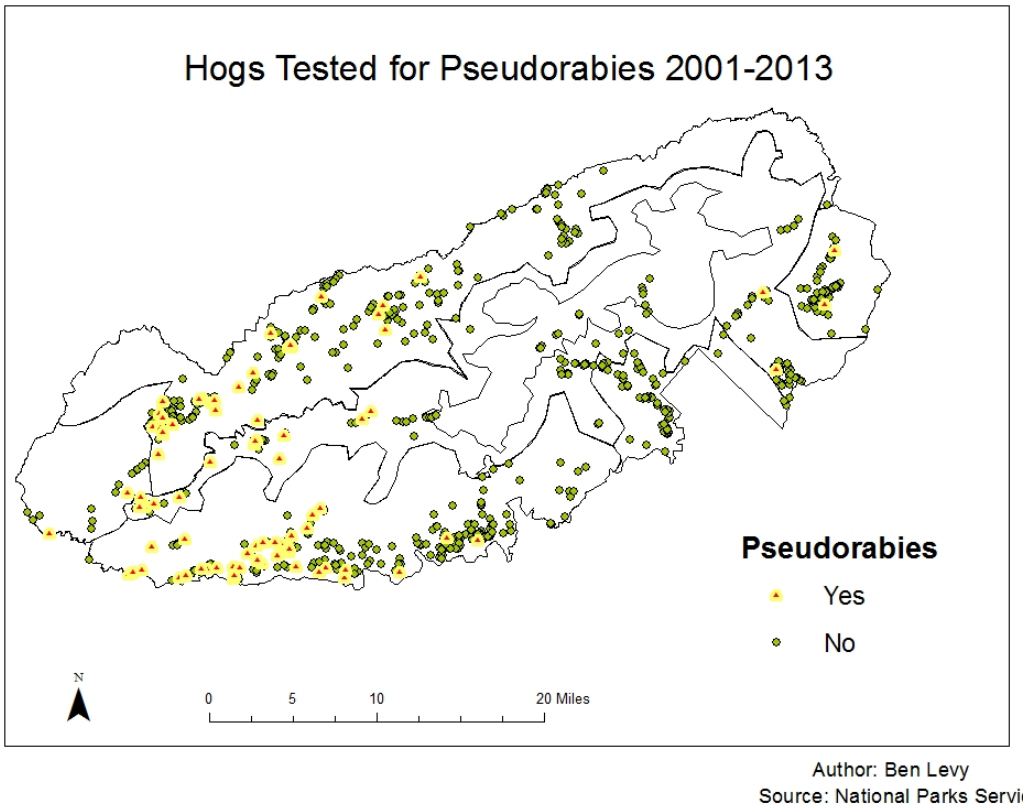


Figure 4.1: Depiction of the 42.5% of hog harvests tested for pseudorabies from 2001-2013. Although the disease is concentrated in the western half of the Park, a small pocket of disease is present in the far eastern regions.

It is the aim of this study to model pseudorabies in the population in order to better understand transmission routes and important dynamics of the disease. This is achieved by building a compartmental disease model into the existing metapopulation framework from Chapter 2.

4.2 Disease Dynamics

Pseudorabies, or Aujeszky's Disease, was first identified in 1902. This herpes viral infection is highly contagious and causes respiratory illness in adult and high mortality rate for piglets

Table 4.1: Number of blood samples taken, amount that positive tests for pseudorabies and resulting prevalence from 2005-2013. Though a relatively small number of samples were taken each year, an general trend of increasing prevalence can be seen in the data (National-Park-Service, 1980).

Year	**2005**	**2006**	**2007**	**2009**	**2009**	**2010**	**2011**	**2012**	**2013**
Blood Samples	150	208	64	106	155	105	90	58	69
Positive for Pseudorabies	2	4	10	9	9	4	19	33	20
Prevalence	0.013	0.019	0.156	0.085	0.058	0.038	0.211	0.569	0.290

in domestic swine (Müller et al., 2011). The disease is usually spread by nose-to-nose contact and through venereal transmission. Transmission via dead carcasses can also contribute to new infections. Transmission of antibodies to piglets during nursing and close contact with mothers is also possible. Since this transmission route has not been well studied, we will use the terminology "close contact with mothers" to represent all possible mother-to-piglet transmission routes, including vertical transmission. After an initial shedding period of about 7 days, adults recover but carry the disease for life in a latent form and can shed the virions periodically as a result of stress (Müller et al., 2011). Thus all hosts who contract PRV will test positive for life. Although the disease is not a significant threat to feral swine, it is can be contracted by other animals such as bear and coyotes, and is especially deadly for canines (Cavendish et al., 2008). Furthermore, decreases in birthrate due to the disease pose a threat to the domestic swine industry. Since the U.S. swine industry become PRV-free in 2004, there is vested interest in limiting the spread of the disease (Müller et al., 2011; Smith, 2012).

To model the transmission of PRV in feral swine, we first turn to the work done by Gary Smith (Smith, 2012). Dr. Smith participated in the working group "Feral swine/pseudorabies in Great Smoky Mountains National Park" at the National Institute for Mathematical and Biological Synthesis (NIMBioS) and his work analyzed what modes of transmission could account for the reported seroprevalence of pseudorabies found throughout the United States. His findings show that simply nose-to-nose transmission (direct transmission) can alone account for the reported seroprevalence, but sexual transmission alone cannot (Smith, 2012). We initially simply tried a different direct transmission rate for each region, but

results did not closely match the data with this approach. We then considered an increased transmission during mating season, transmission from mothers to piglets, and the possibility that carriers would become reinfected. By considering one additional transmission route at a time, we tested to see if a single mode of transmission could account for the prevalence of pseuodrabies seen in the data. However, each transmission route alone was not able to approximate disease dynamics present in the data. Only after including all transmission routes together were we able to mimic the correct dynamics. As such, we will apply a model for pseudorabies within the framework of our existing metapopulation model with nose-to-nose transmission, increased transmission during mating season, transmission from dead carcasses, reinfection of recovered individuals carrying the virus, and transmission from mothers to piglets.

4.3 Modeling Pseudorabies

Since the disease has an infectious period of one week, we first adapt the metapopulation model described previously from a one month time step to a one week time step. This was achieved by estimating all parameters in a similar manner as before as described in Chapter 2. Resulting estimated weekly parameter values can be see in Table 4.2, including average weekly on-season and off-season harvest rates. Following the methodology in Chapter 2, we then fix non-harvest parameters and the same optimization procedure was carried out to estimate harvest rates that varied by region and year. Results allowed the new model with a time step of one week to almost identically mimic the model with a time step of one month described in Chapter 2.

We consider three classes within each population in the 8 regions: Susceptibles (S), Infected (I), and Carriers (C). Susceptible individuals have never contracted the disease and are vulnerable to infection. Infected individuals are those who are symptomatic and are able to transmit the disease. Carriers have recovered from the disease and are no longer infectious but, since they still carry the virus, could experience symptoms again and transition to

infected at rate ϕ. To clarify, in our model carriers do not transmit the disease, but piglet born to carriers and infecteds can become carriers. Since each region has a different total area and will contain different numbers of individuals, we will apply a unique transmission rate to each region, denoted by β_r. We also keep track of all dead carcasses from infected individuals that result from harvesting as well as natural death as they can also transmit the disease at rate β_D. We only include deaths from infected individuals as they are more likely to transmit the virus. We also include an increased transmission rate ($\gamma\beta_r$) during mating season in September under the assumption of increased contact between individuals during this time. All parameters and variables that are found in this pseudorabies model that were not previously described in Chapter 2 can be see in Table 4.3.

Table 4.2: A list and description of estimated weekly parameters found in the pseudorabies model.

Name	Value	Description
$Surv0$	0.96	Survival factor if there is no mast
$SurvMax$	0.99	Survival factor as mast approaches a maximum level
$BR0$	0.10	Percent of population that give birth and whose piglets survive the first month given no mast
$BRMax$	0.13	Percent of population that give birth and whose piglets survive the first month as mast approaches a maximum level
$Move0$	0.14	Percent of feral hogs moving with no available mast
$MoveMax$	0.01	Percent of feral hogs moving as mast approaches a maximum level
$rate_1$	0.09	On-Season harvest rate, from January through May
$rate_2$	0.06	Off-Season harvest rate, from June through December

Similar to Chapter 2, we will have the following order of events with the addition of a disease transmission event taking place between births and movement, as well as the characteristic that the population is partitioned into three classes and that we must keep track of dead carcasses from the infected that can transmit the virus.

The model described in Chapter 2 therefore takes the following altered form:

1. Update the mast for the month since many of the parameters that govern feral hog dynamics in GSMNP are driven by hard mast availability (Singer, 1981; Scott and Pelton, 1975). We also consider a constant amount of soft mast per acre in each region. Let $HM_{r,t}$ and $SM_{r,t}$ represent all hard mast and soft mast that exists in region r at time t:

$$HM_{r,t+1} = \begin{cases} MI_{r,y} & m = 8, \\ ((1-\delta)HM_{r,t} - C_P P_{r,t})^+ & m \neq 8. \end{cases}$$
$$M_{r,t} = HM_{r,t} + SM_r. \tag{4.1}$$

2. Harvest at a rate determined by the specific region and time:

$$H_{r,t} = Hrate_{r,t}(S_{r,t} + I_{r,t} + C_{r,t})$$

3. Compute the portion of the post-harvest population that survives. We do this before adding births because only surviving adults can reproduce:

$$Surv_{r,t} = F(M_{r,t}, Surv0, SurvMax, M_h)$$
$$S_{r,t} \cdot (1 - Hrate_{r,t}) \cdot Surv_{r,t}$$
$$I_{r,t} \cdot (1 - Hrate_{r,t}) \cdot Surv_{r,t}$$
$$C_{r,t} \cdot (1 - Hrate_{r,t}) \cdot Surv_{r,t}$$

4. Update dead carcasses (D) that are capable of spreading the virus. Since Park employees leave harvested hogs to decay where they were destroyed, this will include both harvested hogs as well as those that die during the survival step. We assume the

carcasses decay at a rate of .5 at each time step:

$$D_{r,t} = .5D_{r,t-1} + I_{r,t} \cdot (1 - Hrate_{r,t}) \cdot (1 - Surv_{r,t})$$

5. If the month is January, we then compute the number of births based on the surviving population and mast supplies. A percent of piglets (α) will contract a latent form of the infection from the C and I compartments via close contact with mothers and become members of the carrier class. We assume piglets are equally as likely to contract the disease from either class. We note the number of surviving susceptibles ($\tilde{S}$), surviving infected ($\tilde{I}$), and surviving carriers ($\tilde{C}$):

$$BR_{r,t} = \begin{cases} B_F \cdot L_A \cdot F(M_{r,t}, BR0, BRMax, M_h) & m = 1 \\ 0 & m \neq 1. \end{cases}$$

$$\begin{aligned} \tilde{S}_{r,t} &= S_{r,t} \cdot (1 - Hrate_{r,t}) \cdot Surv_{r,t} \\ &\quad + (S_{r,t} + I_{r,t} + C_{r,t}) \cdot (1 - \alpha) \cdot (1 - Hrate_{r,t}) \cdot Surv_{r,t} \cdot BR_{r,t} \\ \tilde{I}_{r,t} &= I_{r,t} \cdot (1 - Hrate_{r,t}) \cdot Surv_{r,t} \\ \tilde{C}_{r,t} &= C_{r.t}(1 - Hrate_{r,t}) \cdot Surv_{r,t} + (I_{r,t} + C_{r,t}) \cdot \alpha \cdot (1 - Hrate_{r,t}) \cdot Surv_{r,t} \cdot BR_{r,t} \end{aligned}$$

6. Disease transmission for the given time step is based on the surviving populations for each class ($\tilde{S}, \tilde{I}$, and $\tilde{C}$). Note that since we reduced the model to a weekly time step, new additions to the carrier class (C) are simply those who survived from the infected class ($\tilde{I}$):

$$\beta_r = \begin{cases} \beta_r & m \neq 9. \\ \gamma\beta_r & m = 9 \end{cases}$$

$$
\begin{aligned}
S_{r,t} &= \tilde{S}_{r,t} e^{-\beta_r \tilde{I}_{r,t} - \beta_{D_r} D_{r,t}} \\
I_{r,t} &= \tilde{S}_{r,t}(1 - e^{-\beta_r \tilde{I}_{r,t} - \beta_{D_r} D_{r,t}}) + \tilde{C}_{r,t}\phi \\
C_{r,t} &= \tilde{C}_{r,t}(1 - \phi) + \tilde{I}_{r,t}
\end{aligned}
$$

where $\tilde{S}, \tilde{I}$, and $\tilde{C}$ are the number of individuals in each class that have lived through the survival events, β_r is the transmission rate in region r, β_{D_r} is the transmission rate in region r due to dead carcasses, and ϕ is the percent of carriers that become reinfected.

7. Perform movement, using either general movement or seasonal movement, dependent upon the time of year. This is carried out exactly as described in Chapter 2 by applying movement rates equally to each class. Note that we obtain $S_{r,t+1}, I_{r,t+1}$, and $C_{r,t+1}$ during this event.

Since pseudorabies was not detected in GSMNP until February of 2005, we will initialize the model in 2004, introduce 5 infected individuals into region 4 in February 2005, and run the model until 2013, which is the most recent year of which we have disease data. We choose 5 infected individuals since this number most closely reproduced observed prevalence values obtained in 2005.

4.4 Parameter Estimation

For a complete list of parameters and variables not found in the metapopulation model see Table 4.3. Park officials began sampling harvested hogs for disease in 2001 and the first case of pseudorabies was detected in region 4 in February 2005 (Cavendish et al., 2008). Although as many as 208 hogs were tested in a given year, the vast majority of tests were conducted on hogs found in regions 4 and 5 with very few hogs tested from other regions (see Table

4.4). For this reason, when estimating parameters we will use data from only regions 4 and 5 to estimate β_4 and β_5.

Scaling transmission rates is necessary when using a parameter estimated from one geographic location for use in a different geographic location (Hu et al., 2013). Since the regions in our model have different sizes, the contact rate of hogs varies between each discrete area. We will therefore scale β_4 by relative area of each region to determine a transmission rate for other regions in the model. This method allows us to use a different transmission rate for each region relative to the area of the given region, and also reduces the number of parameters to be estimated. Specifically,

$$\beta_r = \beta_4 \frac{A_4}{A_r},$$

for $r\epsilon\{1, 2, 3, 6, 7, 8\}$ where A_r represents the area of region r. The transmission rate for infected carcasses (β_D) will be implemented in a similar manner where we will estimate a value for region 4 and scale it for use in other regions.

Table 4.3: A list and description of new parameters and variables found in the pseudorabies model.

Name	Description
β_r	Transmission rate from infected class in region r
$\beta_{D,r}$	Transmission rate from dead infected carcasses in region r
α	Percent piglets that become carriers as a result of close contact with mothers
γ	Percent increase in transmission rate during mating season
ϕ	Percent of carrier class that becomes reinfected at each time step
$D_{r,t}$	Number of dead carcasses that carry the pseudorabies virus in region r at time t

Our pseudorabies model contains the following six unknown parameters as described in Table 4.3: β_4, β_5, β_D, α, γ, and ϕ. We wish to find the parameter values that, when used in our model, produce harvest prevalence levels that best match the available harvest prevalence data. Since carriers test positive for the disease, they are considered infected for the purposes of a prevalence calculation. We use data from 2005 through 2013 and we will only attempt

Table 4.4: Blood samples, detected cases and resulting prevalence for pseudorabies in each region of GSMNP.

Disease Samples from Region 1

	2005	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**
Total Blood Samples	0	3	1	1	6	1	7	5	1
Detected Cases	0	0	0	0	0	0	1	2	1
Prevalence	0	0	0	0	0	0	0.14	0.4	1

Disease Samples from Region 2

	2005	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**
Total Blood Samples	0	0	0	0	0	0	0	33	10
Detected Cases	0	0	0	0	0	0	0	1	0
Prevalence	0	0	0	0	0	0	0	.03	0

Disease Samples from Region 3

	2005	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**
Total Blood Samples	0	0	0	0	0	25	24	0	4
Detected Cases	0	0	0	0	0	1	1	0	0
Prevalence	0	0	0	0	0	.03	.04	0	0

Disease Samples from Region 4

	2005	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**
Total Blood Samples	19	49	71	39	26	36	53	63	35
Detected Cases	2	4	6	2	1	3	11	15	12
Prevalence	0.11	0.082	0.085	0.052	0.038	0.083	0.21	0.24	0.34

Disease Samples from Region 5

	2005	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**
Total Blood Samples	0	48	53	34	78	19	31	50	19
Detected Cases	0	1	2	2	8	1	4	16	5
Prevalence	0	0.021	0.038	0.059	0.10	0.053	0.13	0.32	0.26

Disease Samples from Region 6

	2005	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**
Total Blood Samples	10	13	11	35	26	16	18	22	21
Detected Cases	0	1	2	7	4	2	0	1	2
Prevalence	0	0.077	0.18	0.2	0.15	0.13	0	0.045	0.095

to match data from regions 4 and 5 as the majority of tests were conducted on hogs found in these regions. Even within regions 4 and 5, a consistent number of disease tests were not conducted year-to-year. We will therefore weight prevalence values by the number of disease samples in each region. In doing so our parameter estimates will better approximate the most reliable data. Let s_r represent the vector containing the number of hogs sampled for disease in region r in each year from 2005-2013, let HI_r represent the vector containing the prevalence of hogs harvested in the model from region r in each year from 2005-2013, and let HI_r^* represent the vector containing the prevalence of infected hogs from region r in the harvest data from 2005-2013. Each of these vectors is 1×9 as we are using yearly data from 2005 through 2013. The optimization problem for parameter estimation can be stated as

$$\underset{x}{\text{Minimize}} \quad J(x) = ||s_4 \cdot HI_4 - s_4 \cdot HI_4^*||_2 + ||s_5 \cdot HI_5 - s_5 \cdot HI_5^*||_2,$$

where x represents all values of β_4, β_5, β_D, α, γ, and ϕ in our own chosen range.

Additionally, we require that the parameters reflect conditions found in the Park. Similar to other transmission rates, β_5 should have an appropriate scale relative β_4, β_D should be less than both β_4 and β_5, and all parameters should fall within a reasonable range. As a result, the above problem is also restricted by the following linear constraints:

$$\begin{aligned} 1.25\beta_4 &= \beta_5 \\ \beta_D &\leq \beta_4 \\ \alpha &\leq 1 \\ \phi &\leq 1 \\ \gamma &\geq 1. \end{aligned} \tag{4.2}$$

To solve the above optimization problem we use the Global Optimization Toolbox from MATLAB$^{\text{TM}}$. Specifically, we use the fmincon local solver in coordination with the MultiStart Algorithm. The fmincon solver is a derivative-free solver, which accepts smooth, nonlinear

objective functions, and allows enforcement of the linear constraints and bounds given in (4.2). Since fmincon will only find local minimums, the MultiStart Algorithm allowed us to test a large number of evenly distributed starting points and store all local solutions in a manageable way using the built-in manymins option. The MultiStart Algorithm generates uniformly distributed random starting points within the given bounds and passes them one-by-one to the local solver, fmincon, which attempts to find a local basin of attraction relative to each given start values. Any solution that is found is then stored in increasing order of objective function output for later review using the manymins function.

4.5 Results and Discussion

Results from the optimization procedure described above yielded the estimated parameters shown in Table 4.5. Resulting output from the parameters can be seen in Figure 4.2 and Figure 4.3. In Figure 4.2, we compare the weighted model prevalence and weighted prevalence data in regions 4 and 5. Notice in Figure 4.2 that the weighted output fits the data fairly well, which results in the non-weighted output matching the general prevalence trend. From Figure 4.2, the years that do not match well such as 2005 and 2009 in region 4, and 2010 and 2011 in region 5 are years where the fewest samples were taken from each region. The overall prevalence trend in GSMNP is also generally followed as seen in Figure 4.3. The years where the prevalence in the model does not match as well in the whole Park such as 2007 and 2012 are also years where the fewest samples were taken overall. In general, by matching prevalence values weighted by the number of observations we succeeded in closely approximating the prevalence values that were best supported by the data.

Table 4.5: Estimated values for new parameters and variables found in the pseudorabies model.

Name	Estimated Value
β_4	6.8×10^{-4}
β_5	8.52×10^{-4}
β_D	6.8×10^{-4}
α	0.8
γ	1.5
ϕ	0.1

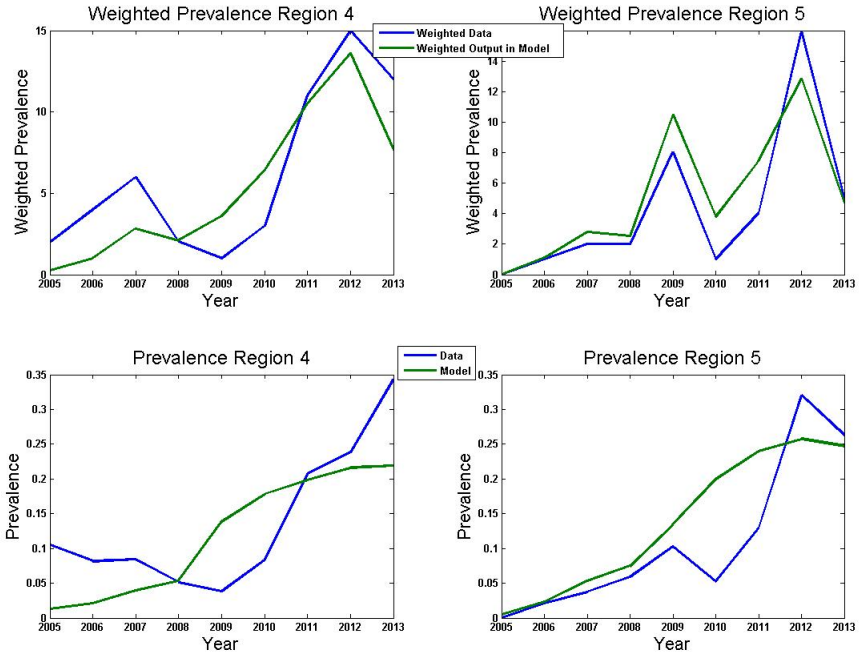


Figure 4.2: Model output using the parameter estimates shown in Table 4.5. The top two plots show the weighted prevalence in the model and weighted prevalence in the data weighted by the number of observations in each year. The bottom two plots show the unweighted prevalence in the model and prevalence in the data.

4.6 Conclusions

Since the model incorporates discrete regions, each with a different area and fluctuating population size, we needed to apply a different transmission rate for each region. We achieved this by estimating a transmission rate for regions 4 and 5 using available data for these

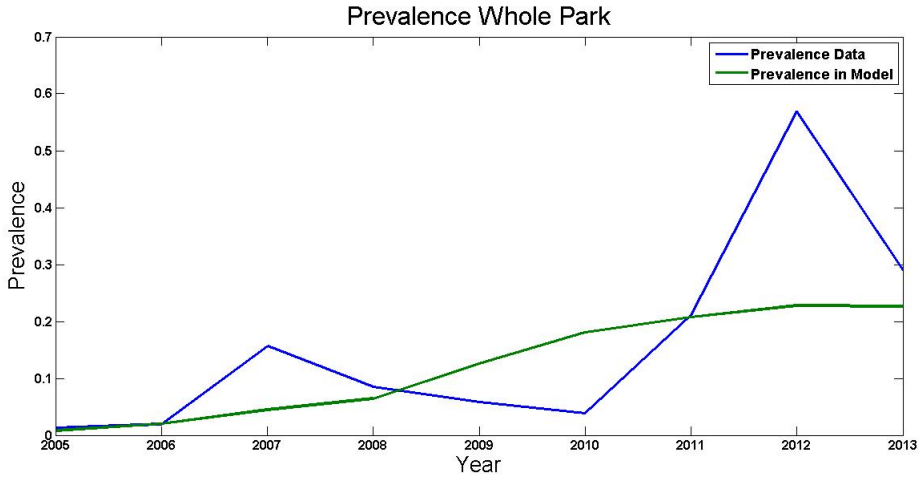


Figure 4.3: Model output using the parameter estimates shown in Table 4.5 comparing model prevalence in the entire Park compared to the prevalence data.

regions and then scaling the rate for region 4 for use in other regions by the relative area of other regions. All possible transmission routes were considered when fitting this model for pseudorabies in feral hogs in GSMNP. We initially tried simply a different direct transmission rate for each region, but were not able to match the data with this approach. We then considered increased transmission during mating season, piglets becoming carriers due to close contact with infected mothers, and the possibility that carriers would become reinfected. We added one of each additional parameter at a time, but also were not able to match disease dynamics present in the data. Only after including all additional parameters together were we able to mimic the correct dynamics. This is evidence that all the transmission routes used in the model are indeed present in the population within GSMNP.

Chapter 5

Conclusions

Feral hogs have occupied within Great Smoky Mountains National Park since the early 1900s. They are in direct competition with native flora and fauna, their rooting behavior causes significant ecological damage, and they are a reservoir for diseases such as pseudorabies. To limit these negative impacts, the Park has had a control program in place since 1959 where hogs are harvested via traps and active hunting. Since the population is relatively understudied, a working group at the National Institute of Mathematical and Biological Synthesis was formed to begin studying and addressing the negative impacts caused by feral hogs in the Park.

Key information was shared during the working group including data on vegetation, oak mast levels, and harvest records within GSMNP. These different forms of data were used to study the feral hog population in GSMNP by formulating a discrete metapopulation model, a spatial niche model, and a compartmental disease model for pseudorabies. The data was instrumental in each project as it was used to estimate key parameters, account for mast variability found in the system, and serve as predictor variables for potential presence locations. While much of the data was carefully obtained using scientific sampling techniques (Madden et al., 2004; National-Park-Service, 1981), the harvest data is a product of the control program and is therefore subject to sampling bias and error (National-Park-Service,

1980). Nevertheless, the harvest data conveys a great deal of information and proved to be useful in each project.

The purpose of the metapopulation model was to simulate mast-dependent population dynamics and analyze the effectiveness of the Park's control program. We achieved this by integrating mast index values with vegetation data to replicate seasonality in oak mast, by allowing mast-dependent parameters to vary in each region at each time step depending on food availability, and by estimating harvest efforts using the control data. Key results from this project include parameter estimates for the population rates in GSMNP and a clear indication that the control program has effectively limited the population. Specifically, model output indicates the population of feral hogs in GSMNP ranges between 1,000-2,000 individuals with the presence of the control program, but can approach a carrying capacity of nearly 10,000 individuals without the control program in place. Park officials have cited these findings when applying for future control program funding.

While many features of the metapopulation model are tailored to Great Smoky Mountains National Park, it can be adapted to model other feral hog populations. The dynamics between oak acorn availability and population growth can inform other populations models where seasonality impacts the life cycle of the animal. Additionally, the parameter estimation techniques can be used in numerous modeling scenarios.

Since the control program was found to be important in limiting feral hog presence, the purpose of the spatial niche model was to guide harvesting efforts by categorizing each cell within GSMNP with one of five options ranging from unsuitable for hog presence to most suitable for hog presence. To achieve this, hunting locations obtained via the control program were related to environmental predictors to quantify the niche conveyed by the presence data using a methodology known as an Environmental Niche Factor Analysis. While the control data provides a representation of the species' niche, bias exists in the data due to hunter preferences. However, since we wish to depict suitable hunting locations, the control data contain important and relevant information. Key results from this project quantify feral hog preferences with respect to each predictor variable and how sensitive the population is

to changes in its preferences. Specifically, presence locations exhibit the more significant preferences for below average values of slope and rhododendron, and above average levels of herbaceous vegetation and deciduous shrubs. Moreover, the presence points were most sensitive to changes in elevation and temperature. These and other preferences were used to derive a habitat suitability map for the species. This map can be used to validate past hunting locations and to exploit new regions in the Park that may contain bountiful hog presence.

The spatial niche model provides insight into the preferences of the population with respect to environmental variables. The technique presented can be applied to better understand any harvested population. Specific results can be used to analyze similar areas occupied by feral hogs, especially those where a control program is in place. The information can also be applied when considering new areas suitable for invasion. This is particularly useful for the population in and around GSMNP as feral hogs are not as wide spread in this area compared to other parts of the country (Singer, 1981).

To test the relevance of seasonality with respect to suitable locations, two distinct groups of data were analyzed: one where hogs are concentrated in the lower regions and one where they are concentrated in the upper regions. From August through February, oak acorns are available in lower elevations and the population remains in these regions. From March through June, the depletion of oak acorns paired with increasing temperatures cause the hogs to move up in elevation. Despite this general trend, each analysis produced very similar results and habitat suitability maps. This may be evidence that seasonality is not as important to the population as we originally thought. The telemetry study currently being conducted in the Park should be used to clarify the seasonal movement patterns of the population.

Pseudorabies is well studied in domestic pigs but understudied in wild pigs. Specifically, transmission routes for the disease in wild populations is not understood. Due to this fact, the pseudorabies model aimed to analyze potential transmission routes that exist in the feral hog population in GSMNP. To achieve this, the metapopulation model described in Chapter

2 was adapted to a one week time step to accommodate a compartmental disease model for pseudorabies. We include a different basic transmission rate in each region and consider increased transmission during mating season, transmission from mothers to piglets, and the possibility that carriers become reinfected and able to transmit the disease. Parameters for the model were estimated using the pseudorabies prevalence found in the harvest data. Each transmission possibility was considered separately as well as in combination with every other potential transmission route. Only after incorporating all potential transmission routes were we able to accurately mimic the disease dynamics present in the data. This provides evidence that an increased transmission during mating season, transmission from mothers to piglets, and carriers becoming reinfected all exist within GSMNP.

Both the structure as well as the results of the disease model contribute to the pseudorabies literature. Adjusting a transmission rate estimated for one location and using it in a different region is a useful concept in disease modeling (Hu et al., 2013). By scaling a transmission rate for region 4 for use in other regions, we add to this undeveloped concept. Support for the existence of numerous transmission routes for pseuodrabies in wild pig populations encourages empirical tests to clearly define transmission routes so we can better understand disease dynamics.

The results of the collective study are broadly applicable. First and foremost, they contribute to our knowledge of the understudied feral hog population in GSMNP. Our work encourages the existence of a control program, informs removal efforts, and warns about the potential risks of pseudorabies. Furthermore, we have illustrated how to combine different data sources and methods to address a broad question. In doing so, the techniques applied in this project can be used to inform similar research that incorporate numerous data sources.

Bibliography

Bodine, E. N., Gross, L. J., and Lenhart, S. (2012). Order of events matter: Comparing discrete models for optimal control of species augmentation. *Journal of Biological Dynamics*, 6(sup2):31–49. 8

Boyce, M. S., Vernier, P. R., Nielsen, S. E., and Schmiegelow, F. K. (2002). Evaluating resource selection functions. *Ecological Modelling*, 157(2):281–300. 59, 60, 62, 71

Bratton, S. P. (1974). The effect of the european wild boar (sus scrofa) on the high-elevation vernal flora in Great Smoky Mountains National Park. *Bulletin of the Torrey Botanical Club*, pages 198–206. 2

Carroll, M., Townshend, J., DiMiceli, C., and Sohlberg, R. (2003). Modis 250 meter normalized difference vegetation index, collection 4. University of Maryland, College Park, Maryland. 64

Cavendish, T., Stiver, W., and Delozier, E. K. (2008). Disease surveillance of wild hogs in Great Smoky Mountains National Park–a focus on pseudorabies. 3, 34, 80, 82, 87

Clayton, L., Keeling, M., and Milner-Gulland, E. (1997). Bringing home the bacon: a spatial model of wild pig hunting in Sulawesi, Indonesia. *Ecological Applications*, 7(2):642–652. 6

Cooley, W. W. and Lohnes, P. R. (1962). Multivariate procedures for the behavioral sciences. 36, 37, 38, 45, 47, 51

Engeman, R. M., Smith, H. T., Severson, R., Severson, M. A., Woolard, J., Shwiff, S. A., Constantin, B., and Griffin, D. (2004). Damage reduction estimates and benefit-cost ratios for feral swine control from the last remnant of a basin marsh system in Florida. *Environmental Conservation*, 31(03):207–211. 1

Engeman, R. M., Smith, H. T., Shwiff, S. A., Constantin, B., Woolard, J., Nelson, M., and Griffin, D. (2003). Prevalence and economic value of feral swine damage to native habitat in three Florida state parks. *Environmental Conservation*, 30(04):319–324. 1

Engeman, R. M., Woolard, J., Smith, H. T., Bourassa, J., Constantin, B. U., and Griffin, D. (2007). An extraordinary patch of feral hog damage in Florida before and after initiating hog removal. 1

Epanchin-Niell, R. S. and Hastings, A. (2010). Controlling established invaders: integrating economics and spread dynamics to determine optimal management. *Ecology Letters*, 13(4):528–541. 1

Focardi, S., Toso, S., and Pecchioli, E. (1996). The population modelling of fallow deer and wild boar in a mediterranean ecosystem. *Forest Ecology and Management*, 88(1):7–14. 6

Frontier, S. (1976). Étude de la décroissance des valeurs propres dans une analyse en composantes principales: Comparaison avec le moddle du bâton brisé. *Journal of Experimental Marine Biology and Ecology*, 25(1):67–75. 47, 48

Gaines, K. F., Porter, D. E., Punshon, T., and Lehr Brisbin Jr, I. (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. 6

Gesch, D., Oimoen, M., Greenlee, S., Nelson, C., Steuck, M., and Tyler, D. (2002). The national elevation dataset. *Photogrammetric engineering and remote sensing*, 68(1):5–32. 64

Glass, G. V. and Hopkins, K. D. (1970). *Statistical methods in education and psychology*. Prentice-Hall Englewood Cliffs, NJ. 54

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., and Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 25(15):1965–1978. 65

Hirzel, A. and Perrin, J. (2002). Biomapper 3.1. lausanne lab for conservation biology. http://www.unil.ch/biomapper. 58, 63, 64, 70

Hirzel, A. H., Hausser, J., Chessel, D., and Perrin, N. (2002). Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? *Ecology*, 83(7):2027–2036. 38, 49, 50, 52, 53, 54, 55, 56, 57, 58, 59, 62, 63, 66

Hirzel, A. H., Le Lay, G., Helfer, V., Randin, C., and Guisan, A. (2006). Evaluating the ability of habitat suitability models to predict species presences. *Ecological Modelling*, 199(2):142–152. 53, 58, 59, 61, 62, 70, 71

Holland, E. P., Aegerter, J. N., Dytham, C., and Smith, G. C. (2007). Landscape as a model: the importance of geometry. *PLoS Comput Biol*, 3(10):e200. 19

Howe, T. D. and Bratton, S. P. (1976). Winter rooting activity of the european wild boar in the Great Smoky Mountains National Park. *Castanea*, pages 256–264. 2

Hu, H., Nigmatulina, K., and Eckhoff, P. (2013). The scaling of contact rates with population density for the infectious disease models. *Mathematical Biosciences*, 244(2):125–134. 88, 97

Hutchinson, G. (1957). The multivariate niche. In *Cold Spr. Harb. Symp. Quant. Biol*, volume 22, pages 415–421. 49

Inman, R. M. (1997). Caloric production of black bear foods in Great Smoky Mountains National Park. 10, 11

Johnson, K. G., Duncan, R. W., and Pelton, M. R. (1982). Reproductive biology of european wild hogs in the Great Smoky Mountains National Park. In *Proceedings of the Annual Conference of the Southeastern Fish and Wildlife Agencies*, volume 36, pages 552–564. 2, 32

Jones, P. (1957). A historical study of the european wild boar in North Carolina. *Mast of Arts in Education Thesis, Appalachian State Teachers College, Boone, North Carolina*. 1

Kaiser, H. F. (1991). Coefficient alpha for a principal component and the kaiser-guttman rule. *Psychological Reports*, 68(3):855–858. 47

Keeling, M., Milner-Gulland, E., and Clayton, L. (1999). Spatial dynamics of two harvested wild pig populations. *Natural Resource Modeling*, 12(1):147–169. 6

Kirkpatrick, R. and Pekins, P. (2002). Nutritional value of acorns for wildlife. *Oak Forest Ecosystems. The Johns Hopkins University Press, Baltimore*, pages 173–181. 65

Legendre, P. and Legendre, L. F. (2012). *Numerical Ecology*, volume 24. Elsevier. 39, 41, 42, 43, 44, 46, 47

Levy, B., Collins, C., Lenhart, S., Madden, M., Corn, J., Salinas, R. A., and Stiver, W. (2015). A metapopulation model for feral hogs in Great Smoky Mountains National Park. *Natural Resource Modeling*. 61, 62

Madden, M., Welch, R., Jordan, T., Jackson, P., Seavey, R., and Seavey, J. (2004). Digital vegetation maps for the Great Smoky Mountains National Park. *The University of Georgia, Department of Geography, Athens, GA*. 7, 8, 11, 13, 61, 64, 94

Miller, R. S., Farnsworth, M. L., and Malmberg, J. L. (2013). Diseases at the livestock–wildlife interface: status, challenges, and opportunities in the United States. *Preventive Veterinary Medicine*, 110(2):119–132. 3, 34

Müller, T., Hahn, E., Tottewitz, F., Kramer, M., Klupp, B., Mettenleiter, T., and Freuling, C. (2011). Pseudorabies virus in wild swine: a global perspective. *Archives of Virology*, 156(10):1691–1705. 82

National-Park-Service (1980). Great Smoky Mountains National Park feral hog harvest data from 1980-2013. Accessed: 2011-08-30. 4, 6, 7, 8, 11, 61, 65, 80, 82, 94

National-Park-Service (1981). Great Smoky Mountains National Park hard mast index from 1981-2010. Accessed: 2011-08-30. 6, 7, 10, 11, 94

National-Park-Service (2011). Great Smoky Mountains National Park gis shapefile data. Accessed: 2014-10-15. 63, 64

Olson, L. J. et al. (2006). The economics of terrestrial invasive species: a review of the literature. *Agricultural and Resource Economics Review*, 35(1):178. 1

Salinas, R. A., Stiver, W. H., Corn, J. L., Lenhart, S., Collins, C., Madden, M., Vercauteren, K. C., Schmit, B. B., Kasari, E., ODOI, A., et al. (2015). An individual-based model for feral hogs in Great Smoky Mountains National Park. *Natural Resource Modeling*, 28(1):18–36. 6, 30, 61

Sandfoss, M. R., DePerno, C. S., Betsill, C. W., Palamar, M. B., Erickson, G., and Kennedy-Stoskopf, S. (2012). A serosurvey for brucella suis, classical swine fever virus, porcine circovirus type 2, and pseudorabies virus in feral swine (sus scrofa) of eastern North Carolina. *Journal of Wildlife Diseases*, 48(2):462–466. 3, 80

Scott, C. and Pelton, M. (1975). Seasonal food habits of the european wild hog in the Great Smoky Mountains National Park. *Proceedings of the Southeastern Association of Game and Fish Commissioners*, 29:585–593. 2, 7, 8, 9, 10, 18, 65, 68, 73, 85

Shlens, J. (2014). A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*. 40

Singer, F. J. (1981). Wild pig populations in the national parks. *Environmental Management*, 5(3):263–270. 2, 7, 9, 11, 16, 17, 18, 20, 73, 85, 96

Singer, F. J., Otto, D. K., Tipton, A. R., and Hable, C. P. (1981). Home ranges, movements, and habitat use of european wild boar in Tennessee. *The Journal of Wildlife Management*, pages 343–353. 2, 65, 68, 73

Smith, G. (2012). Preferential sexual transmission of pseudorabies virus in feral swine populations may not account for observed seroprevalence in the USA. *Preventive Veterinary Medicine*, 103(2):145–156. 82

Sokal, R. R., Rohlf, F. J., et al. (1969). *The principles and practice of statistics in biological research*. WH Freeman and Company San Francisco:. 54

Stegeman, L. C. (1938). The european wild boar in the Cherokee National Forest, Tennessee. *Journal of Mammalogy*, 19(3):279–290. 1, 17

Stiver, W. (2011). Continued intensive wild hog control and disease monitoring in the southwestern portion of Great Smoky Mountains National Park. *Tallassee Fund*. 33

Stiver, W. (2012a). Determining movements of wild hogs for disease modeling and control efforts in the Big South Fork National River and Recreation Area and Great Smoky Mountains National Park. *National Park Service*. 3, 20, 33

Stiver, W. (2012b). Proposal: Continue intensive wild hog control and disease monitoring in the southwestern portion of Great Smoky Mountains National Park. *Tallassee Fund*. 3, 33

Stiver, W. (2014). Conversation with William Stiver about feral hog dependence on hard mast availability as well as behavior and historical locations of hogs in Great Smoky Mountains National Park. *Personal Communication*. 17, 73

Stiver, W. and DeLozier, K. (2005). Great Smoky Mountains National Park wild hog control program. In *Official Park Report for Wild Pig Symposium*. 2

Wilson, J. B. (1991). Methods for fitting dominance/diversity curves. *Journal of Vegetation Science*, 2(1):35–46. 58

Witmer, G. W., Sanders, R. B., and Taft, A. C. (2003). Feral swine—are they a disease threat to livestock in the United States? *USDA National Wildlife Research Center-Staff Publications*, page 292. 1

Vita

Benjamin Levy was born in Litchfield, CT to the parents of Steven and Thomasina Levy. He has a younger brother, Zachary, who he is very close with. He attended Litchfield Montessori School, Rumsey Hall School, and Litchfield High School. Ben began dating his future wife, Rachel, in high school. He then attended Franklin and Marshall College where he studied Mathematics and Philosophy, all while cultivating his interest in Biology and the natural world. He graduated with a Bachelors of Arts degree from Franklin and Marshall in May 2008 with a major in Mathematics and minor in Philosophy. After graduating, Ben served on the faculty at Rumsey Hall School for two years where he taught 6th grade Earth Science and 8th grade Physical Science during the school year and Algebra I and II during the summer. He was accepted to a graduate program in Mathematics in 2010 at the University of Tennessee where he would study Mathematical Ecology. He held a teaching assistantship in the Mathematics Department for four years before obtaining a two year research assistantship through the National Institute for Mathematical and Biological Synthesis (NIMBioS). He earned a Masters of Science in December 2013 from the University of Tennessee before obtaining a Doctorate of Philosophy in May 2016. After graduating, Ben will become an Assistant Professor in the mathematics department at Fitchburg State University in Fitchburg, Massachusetts.