THESIS

REGIONAL DATA REFINE LOCAL ABUNDACE MODELS: MODELING PLANT SPECIES ABUNDANCE DISTRIBUTIONS ON THE CENTRAL PLAINS

Submitted by

Nicholas E. Young

Department of Forest, Rangeland, and Watershed Stewardship

In partial fulfillment of the requirements

For the degree of Master of Science

Colorado State University

Fort Collins, Colorado

Fall 2010

Master's Committee:

Department Head: Frederick Smith

Advisor: Thomas J. Stohlgren

Eugene F. Kelly
James J. Graham
Paul H. Evangelista

ABSTRACT

REGIONAL DATA REFINE LOCAL ABUNDACE MODELS: MODELING PLANT SPECIES ABUNDANCE DISTRIBUTIONS ON THE CENTRAL PLAINS

Species distribution models are frequently used to predict species occurrences in novel conditions, yet few studies have examined the effects of extrapolating locally collected data to regional scale landscapes. Using boosted regression trees, I examined the issues of spatial scale and errors associated with extrapolating species distribution models developed using locally collected abundance data to regional extents for a native and alien plant species across a portion of the central plains in Colorado. Topographic, remotely sensed, land cover and soil taxonomic predictor variables were used to develop the models. Predicted means and ranges were compared among models and predictions were compared to observed values between local and regional extent models. All models had significant predictive ability ($p < 0.001$). My results suggested: (1) extrapolating local models to regional extents may restrict predictions; (2) modeling species abundance may prove more useful than models of species presence; (3) multiple sources of predictors may improve model results at different extents; and (4) regional data can help refine and improve local model predictions. Regional sampling designed in concert with large sampling frameworks such as the National Ecological Observatory Network, Inc (NEON) may improve our ability to monitor changes in local species abundance.

ACKNOWLEGEMENTS

The research for this study was funded by The United States Geological Survey (USGS), the National Science Foundation (NSF) and the National Ecological Observatory Network, Inc. (NEON). I would like to thank the Natural Resource Ecology Laboratory (NREL) at Colorado State University for facility use and expertise. In addition, I thank the herbarium and staff for access to their collection and for their expertise. For field assistance, I would like to acknowledge Paul Evangelista, Greg Newman and Heather Lindsey. I greatly appreciate the reviews provided by Greg Newman and Kirstin Holfelder and statistical and modeling assistance from Sunil Kumar.

I would also like to thank my committee: Dr. Thomas Stohlgren, Dr. Eugene Kelly, Dr. Jim Graham and Dr. Paul Evangelista. Tom Stohlgren's passion for science and vision for future ecology has been motivating. I'm grateful for Jim Graham's guidance and advice with project management and his devotion to project quality and, most importantly, to the people within a project. Paul Evangelista's mentorship and dedication to ecology and service has been inspirational during this process. I have taken so much from their mentorship, and to each I am grateful. I also would like to thank Mariah Coler, Catherine Jarneviche, and Lee Casuto for their help and support.

Finally, I would like to thank my family; Mike Young, Janeen Easley, Debbie Young, Mackenzie Young and especially my brother, Brendan Young, for providing me with and strong foundation and their love during this journey. In addition, I would like to

thank Heather Lindsey for her patience and unyielding support through this process. When it was dark they were my light and I am indebted to them.

TABLE OF CONTENTS

Introduction

The application of species distribution models (SDMs) has increased in the past decade. Advancements in computer capabilities and powerful geographic information systems have facilitated the modeling of complex ecological interactions to predict species distributions (Guisan & Thuiller, 2005; Elith & Leathwick, 2009). Species distribution models are increasingly being used to extrapolate information in space and time, often beyond the extent of the data used to develop the model (Elith *et al.*, 2010). Extrapolating models developed using local data to regional extents can save valuable and limited resources (e.g. personnel, time, and money). While insights may be gained through extrapolation, recent studies have suggested this may not always be the best approach when modeling species distributions in novel environments (Pearson *et al.*, 2006). The majority of SDMs use presence-absence or presence-only data (e.g., (Kumar *et al.*, 2009), and there has been little investigation into extrapolation of the spatial distribution of species abundance. Abundance data require more resources to collect and are less common than presence-absence and presence-only data. At the same time, land managers need accurate predictions of species abundance to guide decisions. Predicted abundance allows managers to prioritize management actions that may not be possible with predicted presence alone. Furthermore, the effects, accuracy and predictive power of extrapolating abundance species distribution models to regional extents using only locally collected data are largely unknown.

Scaling ecological patterns and processes has always been a challenge for ecologists (Levin, 1992). Spatial Scale can be thought of in two ways; spatial extent and resolution (also referred to as grain; Wiens, 1989). Spatial scale both in terms of extent

and resolution has significant implications on the ability to identify patterns within and among scales. Although there is no single spatial extent for ecological studies, the most studies usually only identify one scale. This can have significant limitations on our ability to not only identify ecological patterns, but also to understand the processes driving those patterns (Scott *et al.*, 2002). For example, the drivers of change at a local scale are often influenced more by past disturbances of that area, while drivers at a continental scale are primarily climatic (Brown *et al.*, 2008). The importance of scale in ecological study is increasingly being recommended and being included into more study designs, such as those of the National Ecological Observatory Network, Inc. (NEON).

Ecologists have been quantifying species distributions since Grinell's (1917) observation of the relationships between a species and the environmental conditions where it is found. Hutchinson (1957) later expanded this concept, describing this relationship as a n-dimensional hypervolume of biotic and abiotic interactions where a species can survive and persist. This has often been referred to as the fundamental niche or the potential distribution in a geographical space. A species realized niche is the portion of the fundamental niche that is actually occupied by the species, and includes all the constraints on the species' distribution. While environmental variables are the most common and readily available predictors used to define the niche, other factors such as competition, dispersal barriers, and land use (Pulliam, 2000) can also be important contributors but are more difficult to include in SDMs. These factors are not only more challenging to quantify, but can also vary significantly with time and space. The niche can be thought of in two spaces; the environmental space and the geographic space. The environmental space is the environmental values associated where a species is found,

while the geographic space is the physical location of a species on a landscape. Of the many terms that have been used to describe the species-environment relationship, I will use ecological niche to refer to this relationship.

Species distribution models are numeric tools that relate species response data (either occurrence or abundance) with environmental characteristics at those locations (Elith & Leathwick, 2009). These models have been used to meet many management objectives including identifying previously unknown populations of endangered species (Evangelista *et al.*, 2008b), predicting vulnerable habitats to species invasions (Stohlgren *et al.*, 2002), estimating species richness (Graham & Hijmans, 2006), and many others (Elith & Leathwick, 2009). Species distribution models allow ecologists to combine current knowledge of species-environment relationships with advanced algorithms that explore and test multiple interactions to model the ecological niche of a species. Many SDMs have been developed and are commonly used, including Maxent (Phillips *et al.*, 2006), boosted regression trees (BRT; Friedman *et al*. 2000), multivariate adaptive regression splines (Friedman, 1991), and Random Forests (Breiman, 2001). Each algorithm offers strengths and weaknesses for modeling species distributions and multiple studies compare these methods (Araujo & New, 2007; Elith & Graham, 2009; Kumar *et al.*, 2009; Parisien & Moritz, 2009). Although these models are often compared using the same data set, environmental predictors, and spatial scale, it is important to consider that SDMs are designed to handle different types of data sets and perform best under specific circumstance. For example, Maxent is designed specifically for presence-only data, while BRTs require presence and absence data. Therefore, when absence data are available, BRTs may be the more appropriate method to use. The primarily

assumptions of SDMs are the species being modeled are at an equilibrium with the environment (Guisan & Zimmermann, 2000) and the environmental variation has been adequately sampled within the extent being modeled.

While SDMs were developed to model species within the environment from which the data were collected, these models are now being used to predict species distributions in novel conditions not representative of the data used to develop the model. This has been referred to as model projecting, generalizing, transferring, and extrapolating (Fielding & Haworth, 1995; Randin *et al.*, 2006), hereafter referred to as extrapolation. Species distribution model extrapolation has been used to predict the distribution of a species under climate change (Penman *et al.*, 2010), in hypothesized susceptible regions of invasion (Medley, 2010), and to predict a species distribution over large extents (Mateo-Tomas & Olea, 2010). In these applications of model extrapolation in time and space, the model is being applied to novel environmental conditions not captured in the original data. Efforts have been made to improve extrapolated model predictions in space and time using ensemble modeling (Araujo & New, 2007), scaling functions (Miller *et al.*, 2004), or improving model calibration (Phillips & Elith, 2010). Previous studies have shown extrapolating SDMs to regions not representing the complete range of environmental conditions can lead to highly liberal predictions of occurrence (Thuiller *et al.*, 2004). On the other hand, regional data are more difficult to collect and require more resources. If accurate regional predictions can be made using only locally collected data through model extrapolation land managers can save valuable resources.

Identifying and predicting the spatial pattern of species abundance has advanced through the increased use of geographic information systems and spatial models (Sagarin *et al.*, 2006). Before the advent of powerful computers and geographic information systems, previous studies predicting abundance primarily used regression methods (Evangelista *et al.*, 2004; Crall *et al.*, 2006). While these methods rarely produced a map, they were still capable of predicting species abundance given a set of environmental variables. These methods are still a foundation to many of the recently developed models (e.g. random forests, boosted regression trees). Managers are not only interested in the pattern and probability of presence, but also predicted abundance. The number of species distribution models using abundance data is small in comparison to those using presence-absence or presence-only data. This is largely due to the scarcity of abundance data. Modeling abundance data requires more robust statistical models than presence-absence data (Austin, 2002). Most SDMs are designed exclusively for presence-only or presence-absence data and are not compatible with abundance data. These models predict the probability of presence or probability of suitable habitat rather than a measure of the number of species. Still, abundance models have occasionally been used to predict densities and dominance of native and non-native species across the landscape (Strubbe *et al.*, 2010). Although abundance data are difficult to acquire, these data can be used with boosted regression trees and provide better predictions of species abundance on a landscape.

Alien species continue to be an economic burden for the organizations responsible for maintaining ecosystem integrity and processes (Mack *et al.*, 2000). For land managers, the task of surveying an entire area for alien species is unrealistic. Modeling

abundances of alien species can help managers spatially prioritize detection, control and prevention efforts. Furthermore, monitoring the distribution of alien and native species can identify locations vulnerable to risk (Stohlgren *et al.*, 2002). Species distribution models can also serve as a monitoring tool for detecting alien species invasions(Barnett *et al.*, 2007). For alien species, knowledge of predicted abundance in addition to predicted presence may be important to better guide management priorities.

The purpose of this study was to examine the issue of scale (in terms of spatial extent) and errors associated with extrapolating models developed using locally collected data to regional extents for a native and alien plant species on a portion of the Central Plains. My objectives were to: (1) investigate model performance and prediction errors associated with extrapolating local models to regional extents, and (2) evaluate the ability of boosted regression trees to predict abundance using percent cover of plant species from two extents on a portion of the Central Plains in Colorado. I used abundance data (i.e. percent foliar cover) and boosted regression trees to compare models using (1) local data extrapolated to regional extents with (2) models created using data at both the local and regional extents.

Methods

Study area

I examined two extents on the central plains of eastern Colorado (Figure 1). I chose these extents because they represent two of the four strategic designs NEON has identified in their continental-scale research platform for discovering and understanding the impacts of climate change, land-use change, and alien species on ecosystem processes

(NEON, 2010). There are 20 ecoclimatic domains established by NEON across the United States, and the local and regional extents used in this study represent the Core Wildland Site and the Airborne Observatory Platform, respectively, within the Central Plains Domain (Domain10).

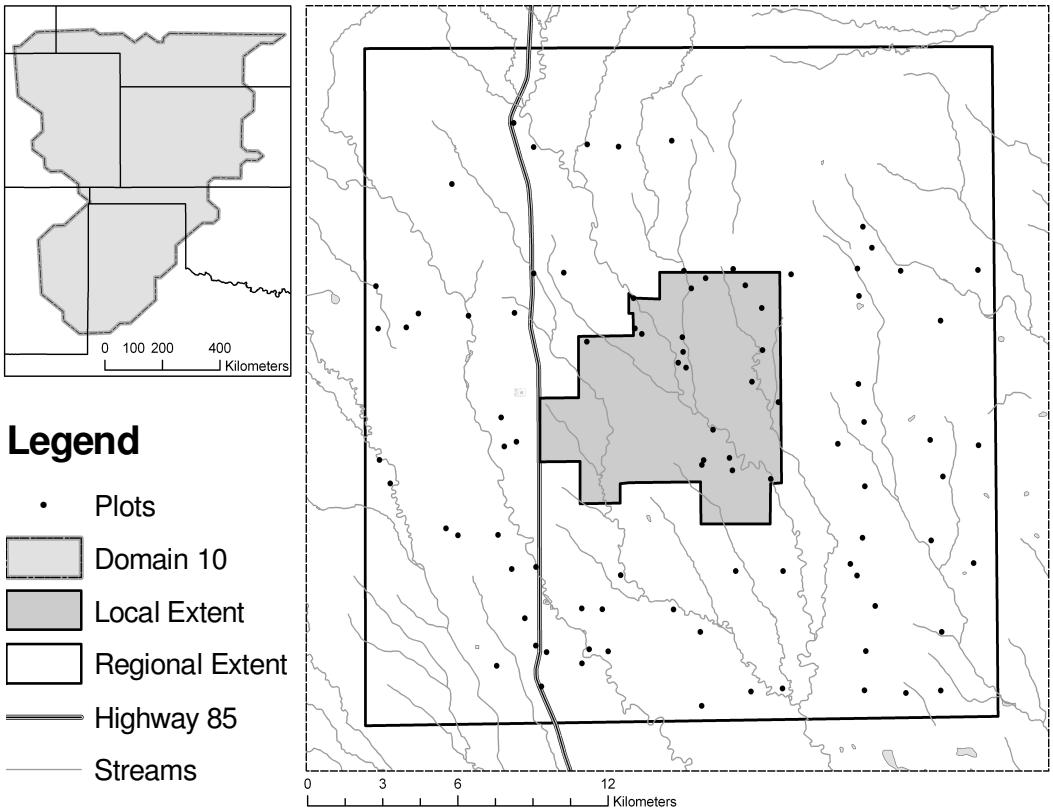

Figure 1. Study area showing sampled plots at the local and regional extents. The local and regional extents are within the larger central plains domain (National Ecological Observatory Network domain 10). Data for figure downloaded from Colorado Department of Transportation and NEON. Figure displayed in World Geodetic System 1984 projection Universal Trans Mercator zone 13 North datum.

The Core Wildland Site for Domain 10 is at the Central Plains Experimental Range, which is located in the Colorado Piedmont section of the Great Plains (40°49' N and 104°46' W). Covering 6,798 ha, the site represents the local extent of my study area and focal point of my research. This area is a semi-arid, C_4-dominated native shortgrass

steppe ecosystem. Most of the precipitation occurs during the growing season from April to September. Grazing by domestic cattle is the dominant land use in conjunction with research and monitoring projects including prescribed fire (Shortgrass Steppe Long Term Ecological Research; http://www.sgslter.colostate.edu).

The regional extent my study area covers an area of 40,000 ha, which represents the 20 x 20 km Airborne Observatory Platform defined by NEON for this domain. The regional area contains the local extent and is similar in terms of climate and ecological characteristics; however, land use is more diverse. Much of the regional extent is a mosaic of shortgrass steppe, agricultural land, rangeland, and human development. Highway 85 runs north-south down the middle of the regional extent and the large developed area surrounding the town of Nunn is located in the southern portion.

Species

I chose to model common native and alien species to the central plains. *Bouteloua gracilis* (blue grama), a warm season perennial bunchgrass, is native to the shortgrass steppe and considered a dominant species in the Colorado Piedmont. Although it has evolved on the shortgrass steppe, *B. gracilis* has been observed to recover poorly on disturbed sites (Marilyn & Hart, 1994). *Sisymbrium altissimum* (tall tumblemustard) is an alien annual or biannual species found in disturbed sites with other alien and native annuals (Allen & Knight, 1984). *Sisymbrium altissimum* can be found on many different soil types including sand (Patman & Hugh, 1961). Although both species are considered generalist in the shortgrass steppe ecosystem, *B. gracilis* is generally a dominant species, while *S. altissimum* is rarely dominant.

Field data

Local and regional abundance data were augmented from two separate studies. Data used for the local extent were collected in 2008 as a part of a NEON preliminary assessment for the Central Plains Experimental Range (Evangelista *et al.*, 2009a) and consisted of 20 sampled plots. Vegetation cover abundances were recorded by estimating the percent cover within a 168-m^2 circular, multi-scale vegetation plot modified from the National Forest Service Inventory and Analysis Program (Barnett *et al.*, 2007; Frayer & Furnival, 1999). Regional abundance data were collected using the Braun-Blanquet method (Braun-Blanquet, 1932) and totaled 72 sampled plots (see Appendix A for regional extent species list). This relatively quick method of sampling is suited for species-environment relationships (Wikum & Shanholtzer, 1978). These data were collected outside of the local extent with the exception of two locations which were sampled inside the local extent. These two samples were also added to the local extent dataset. From the augmented dataset, I selected the native and alien species with the largest number of occurrences. While the number of observed abundance plots at the regional extent for *B. gracilis* (n=62) and *S. altissimum* (n=38) were adequate, the number of observed abundance plots in the local extent were relatively small (*B. gracilis* n=20, *S. altissimum* n=7).

Environmental variables

I used soil, land cover, topographic and remotely sensed environmental data as my predictor variables (Appendix B). All predictor variables had a 30 m resolution.

Topographic variables consisted of elevation, slope, aspect, solar radiation, eastness, and northness. No climatic variables were used in the models because the spatial extent was too small for these predictors to be important drivers.

Soil data were downloaded from Soil Data Mart provided by USDA NRCS (SoilDataMart@nrcs.usda.gov). The data were originally classified by map unit series. I classified the map series to soil great groups (Appendix C). Soil great groups are a classification of soil taxonomy that reflect assemblages of the horizons and the most significant properties of the whole soil (Soil Taxonomy, 1999). In cases where a map unit had multiple series, the series first listed was used for classification. For example, I used the soil great group Thedalund for the map unit Thedalund-Keota loams series. Certain map units did not have associated series (e.g. water, playas, badlands). These categories were left as their original classification.

I downloaded LANDFIRE existing Vegetation Type land cover data from the LANDFIRE website (http://www.landfire.gov/products_national.php). The LANDFIRE dataset was developed using a complied field database for reference plots along with biophysical gradients and Landsat imagery (Rollins, 2009). LANDFIRE uses land cover classifications defined by NatureServes's ecological systems classifications which are ecological units at mid-scale resolution (NatureServe 2009). The LANDFIRE values were grouped to represent nine land cover types (Appendix D). Open water (11), developed (21, 22, 23, 24), barren (31, 2007), agriculture (81, 82), shrubland (2072, 2081, 2086, 2107), grassland/forbland (2094, 2127, 2181, 2182, 2183), mixedgrass prairie (2132), shortgrass prairie (2149) and riparian (2159, 2162). I used these grouped

land cover types to represent classifications appropriate for the scales I was modeling and to allow for more intuitive interpretation of model results.

Six topographic predictor variables were used in the models. Using a U.S. Geological Survey 30 m digital elevation model (DEM), I calculated solar radiation in ArcGIS 9.3 (The Environmental System Research Institute, USA). I used the time period for solar radiation calculations from June 15, 2010 to June 29, 2010 which was when the regional extent sampling occurred. I chose to use a sky resolution of 1000 instead of the default 200 because of the relatively small extent of the digital elevation model used and short time period. Slope, aspect, northness and eastness were also derived from the DEM and calculated using ArcGIS 9.3.

In addition to land cover and topographic variables, remotely sensed Landsat 7 ETM+ satellite scene data were downloaded for July 7, 2000 from USGS Earth Resources Observation Center (EROS, http://glovis.usgs.gov/). The scenes were the most recent cloud free images obtained when the operational scene line corrector was functioning for the season the field data were collected. The scenes and derived vegetation indices were processed using ERDAS Imagine 2010 (ERDAS Atlanta, GA, USA) and ArcGIS 9.3 software. I generated three vegetation indices: Normalized Difference Vegetation Index (NDVI), Ratio Vegetation Index (RVI) and Soil-Adjusted Vegetation Index (SAVI). These indices are used for vegetation and land cover feature estimations. Tasselled cap transformations were also conducted for the Landsat 7 scenes using ERDAS Imagine 2010. These transformations provide measurements of soil brightness (tasselled cap, band 1), vegetation greenness (tasselled cap, band 2) and soil/vegetation wetness (tasselled cap, band 3). Tasselled cap bands and vegetation

indices have been shown to be effective predictors of plant occurrences when used with SDMs (Evangelista *et al.*, 2009b).

Analysis

For my spatial analysis, I used BRTs to model *B. gracilis* and *S. altissimum* at local and regional extents. Modeling species abundances using BRTs is a relatively new method in ecology. In addition to being able to model abundance data, I chose BRTs because they have been shown to perform well with small sample sizes compared to other SDMs (Wisz *et al.*, 2008). Boosted regression trees attempt to minimize the loss function by generalizing many simple classification and regression trees. Tree based models, such as BRTs, accomplish this by applying rules to the predictors that partition the data into rectangles with the most homogeneous response (Elith *et al.*, 2008). For each tree, the data are split into two groups based on a single predictor variable and a rule. The boosting part of BRTs can be thought of as an ensemble model of many tree models that allow for a more robust estimate of the response. Boosting is a form of resampling that, unlike other methods such as bagging or subsampling, applies a weighted probability of a response to be resampled based on previous classifications (Franklin, 2009). Therefore, BRTs decrease overfitting the data by averaging the predictions of many trees created using subsets of the data (Franklin, 2009). Boosted regression trees are also able to incorporate categorical predictors. The relative importance of the predictor variables can also be generated from the model. This is calculated based on the number of times a predictor variable was used as a splitting node and weighted based on the improvement to the model based on each split (Friedman & Meulman, 2003).

I used the generalized boosted models (gbm) package in R (R Development Core Team, 2010) to run BRT models (Friedman *et al.*, 2000). There are a few settings that can be adjusted when running BRTs. A low learning rate decreases the model over-learning, but requires more iterations (De'ath, 2007), I chose to use a learning rate at 0.001 and performed 5000 iterations. Optimizing both the learning rate in conjunction with the number of trees is similar to model regularization. Regularization prevents models from over-fitting training data. Interaction depth or tree complexity is the number of nodes in each tree created. By adding more nodes to the tree, more variable interactions are added. With smaller datasets, larger tree complexity provides no advantage (De'ath, 2007). I set the tree complexity to 3 and performed 5000 iterations (see appendix E for example R code and model settings).

Preliminary models for each species and each extent were constructed using all 15 predictor variables to identify those with the greatest predictive contributions and reduce the overall number of variables used for my analyses. From these results, I kept only those predictor variables that contributed over 5% to the model and removed the others. From those, I performed a Pearson's cross-correlation test using SYSTAT (version 12; SYSTAT Software, Port Richmond, California, USA) to remove highly correlated variables (Pearson correlation coefficient >0.8 or <-0.8). The variables remaining were used to develop final models.

With larger datasets, the model can be developed using a training dataset and tested against a separate test dataset. I had a small dataset, especially for the local extent study area. Had I split my data into a training and test dataset, I would have degraded the estimate of predictive error (Franklin, 2009). Therefore, I used cross-validation to test the

model. Cross-validation withholds a certain proportion of the data at each stage of model development, but uses all data in forming the final model.

I evaluated the difference of the mean, minimum and maximum between model predicted abundance and observed abundance to compare models developed with local data and models developed using regional data. The predicted abundance values were extracted from the local and regional models using Hawth's Tools point intersect function (Beyer, 2004) at the locations were abundance values were observed. The summary statistics were calculated using SYSTAT.

Results

Predictor variables

The local model for *B. gracilis* had five predictor variables and the final regional model had 10 predictor variables that were used in the final models (Table 1). *Sisymbrium altissimum* final models for local and regional data had five and four variables, respectively (Table 2). Soil and topographic environmental variables had a relative influence to the final models that was greater than vegetation indices and remotely sensed variables. For all models except for the *B. gracilis* regional model, the top three predictors had a total relative influence over 80% to the model. Soil great group was a key contributor (>20% relative influence) for each model with the exception of the *S. altissimum* regional model where soil great group was not included in the final model. Soil great group had a relative influence of over 40%, indicating soil is a key predictor for these species at the local and regional scales. Solar radiation was also an important

predictor, with a relative influence of over 65% to the *S. altissimum* regional model, and over 30% for the *B. gracilis* regional model.

Table 1. *B. gracilis* local and regional model environmental predictor relative influence

Local		Regional	
Predictor	Relative influence	Predictor	Relative influence
Soil Great Group	42	Soil Great Group	22
Solar Radiation	32	Eastness	12
Wetness	12	Aspect	11
Eastness	9	Slope	10
Northness	5	Ratio Vegetation Index	10
		LANDFIRE veg. class	8
		Wetness	7
		Solar Radiation	7
		Soil brightness	6
		Elevation	6

Table 2. *S. altissimum* local and regional model environmental predictor relative influence

Local		Regional	
Predictor	Relative influence	Predictor	Relative influence
Soil Great Group	43	Solar Radiation	65
Soil brightness	23	LANDFIRE veg. class	19
Eastness	17	Wetness	8
Soil-Adjusted Vegetation Index	9	Enhanced Vegetation Index	8
Ratio Vegetation Index	8		

Model performance

Local and regional models showed significant predictive ability for both *B. gracilis* and *S. altissimum* ($p < 0.001$). The *B. gracilis* regional model had the highest explained variance (adjusted $R^2 = 0.62$) while the *S. altissimum* regional model had the

lowest (adjusted R^2= 0.34). The local models both explained more than 50% of the variance (*B. gracilis*= 0.53, *S. altissimum* = 0.59).

Predictive map descriptions and model estimates

In general, the predicted abundance for the local model of *B. gracilis* was highest in the southern portion of the regional extent and extended northward in bands following suitable soil types (Figure 2a). The lowest abundance predictions were found in the northern portion of the regional extent where the land cover is more barren and includes a portion of the Pawnee Buttes National Grassland. The regional model of *B. gracilis* differs from the local model in that the highest predicted abundance values are found in a swath running from the east to the northwest (Figure 2b). In addition, where the local model predicted high abundance in the southern portion, the regional model predicted low to moderate abundance. When looking at the local extent, the local model predicted more uniform abundance with higher abundance in the east (Figure 2c). The regional model for the local extent showed much more variation in abundance predictions than the local model, but showed similar areas of high abundance (Figure 2d).

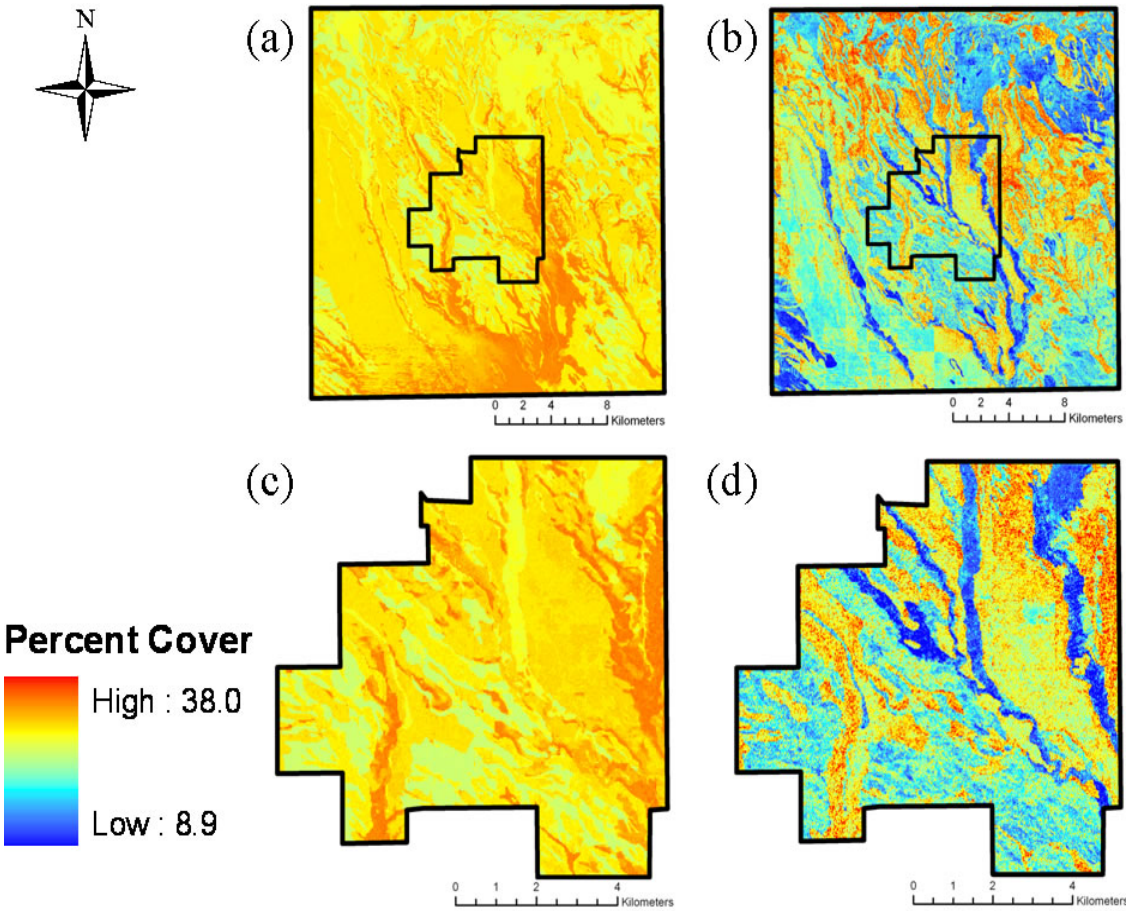

Figure 2. *B. gracilis* (a) model developed using local data extrapolated to regional extent. (b) Model developed using local data and regional data at regional extent. (c) Local extent modeled using local data and (d) local extent modeled using local and regional data. Figure displayed using World Geodetic System 1984 projection and Universal Transverse Mercator zone 13 north datum.

The mean abundance prediction for the local (mean=25.7% cover, S.E.± 1.3) and regional models (mean=22.4% cover, S.E.± 4.0) were similar, but the range of abundance for the local model (6.0% cover) was narrower than the regional model (29.2% cover). .The regional model predicted a maximum abundance of 38% cover while the local model predicted a much lower prediction (29%). Furthermore, the minimum predicted abundance for the regional model was 9%, while the local model minimum prediction was 23%.

The *S. altissimum* predicted map of the local model shows higher abundance in patches concentrated in the central portion of the regional extent (Figure 3a). Lower abundance was located in the northeast and southeast corners. The regional predicted map shows high abundance in the northwest that extends southeastward and into the regional extent (Figure 3b). These higher abundance locales were not predicted by the local model. Furthermore, higher abundance was predicted for the southwest where the land use is primarily agricultural and where the highest abundance estimates were observed. At the local extent, the local model predicted low abundance with little variation in estimates (Figure 3c). In contrast, the regional model predicted relatively higher abundance with more variation (Figure 3d).

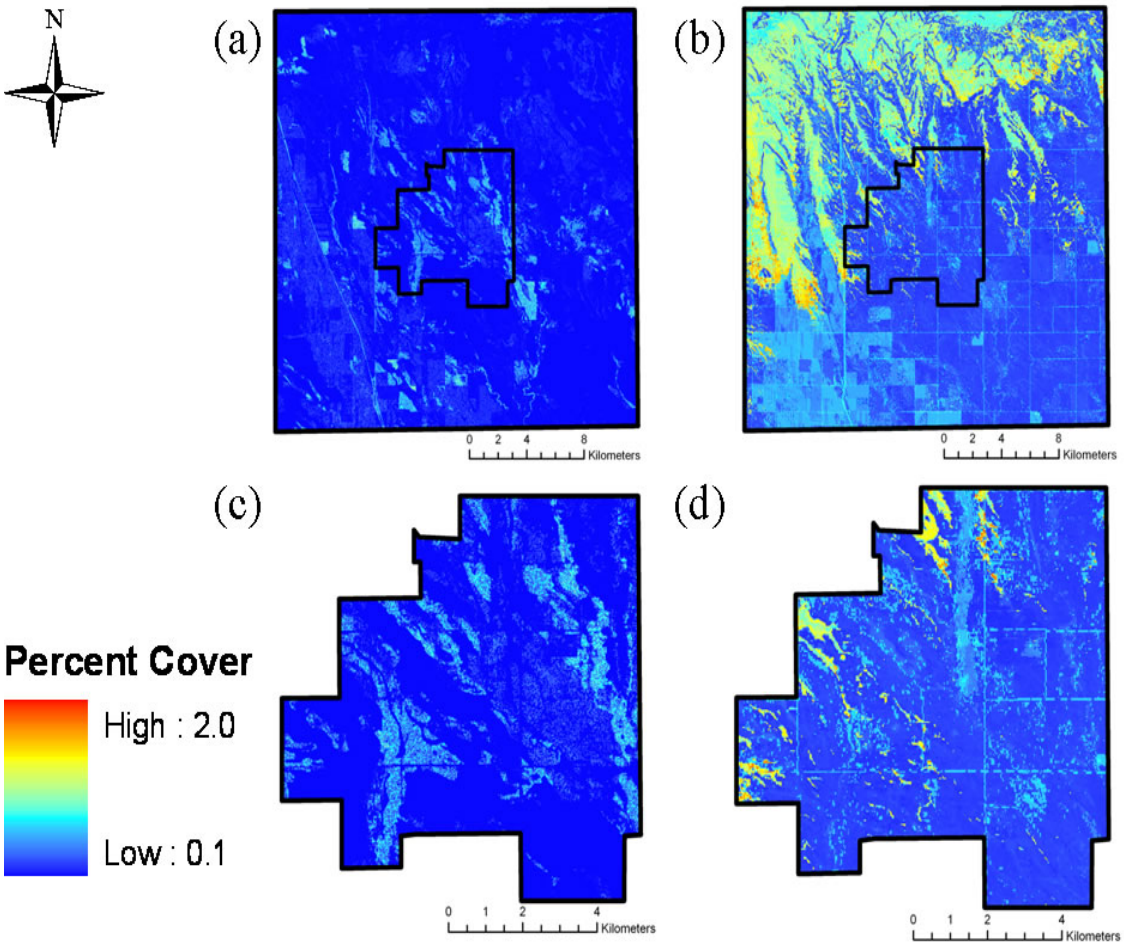

Figure 3. *S. altissimum* (a) model developed using local data extrapolated to regional extent. (b) Model developed using local data and regional data at regional extent. (c) Local extent modeled using local data and (d) local extent modeled using local and regional data. Figure displayed using World Geodetic System 1984 projection and Universal Transverse Mercator zone 13 north datum.

Sisymbrium altissimum regional and local models have similar predicted abundance ranges (local =1.0% cover, regional=1.2% cover) but the predicted mean of the local model (0.5% cover, S.E.± 0.2) was much less than the regional model (1.1% cover, S.E.± 0.2) Similar to *B. gracilis*, the maximum predicted abundance for the local *S. altissimum* model (1.1% cover) was much higher than the regional model (2.0% cover), but the minimum predicted abundance was lower for the local model (0.1%

cover). Furthermore, when comparing the area of the local extent (area=68.1 km^2) that was different between the local and regional models for both species, I found the total area either below or above the local model for *B. gracilis* was 36.9 km^2 and the area within both models predicted range was 31.3 km^2. *Sisymbrium altissimum* models differed from *B. gracilis* in that the local model predicted abundance values lower than the minimum abundance of the regional model, but only predicated a maximum abundance half that of the regional model. For the local model, the area predicted below the minimum value of the regional model was 52.8 km^2 while the area predicted by the regional model above the local model minimum was 11.4 km^2. The total area in the same range for both models for *S. altissimum* was 3.9 km^2.

Model predictions compared to observed values

Regional models predicted species abundance closer to observed values than local model predictions. When compared to the observed values, predicted abundance from models developed using local data were off by a factor of 9% (S.E.± 2.2, n=92) for *B. gracilis* and 0.5% (S.E.± 0.4, n=92) for *S. altissimum*. Conversely, the regional models were off by a factor of 5% (S.E.± 2.1, n=92) for *B. gracilis* and 0.1% (S.E.± 0.4, n=92) for *S. altissimum*. For both species, the regional models predicted abundance values closer to the observed values.

Discussion

Using regional data can refine local models

Incorporating regional data improved model predictions at local and regional extents. For both species, the models developed using regional data to model the local geographical extent showed a larger range of predicted abundances (Figure 2d and Figure 3d). This was especially true for *B. gracilis* where more than half of the area in the regional model was either below or above the range of the local model. The improved predictions may be attributed to the additional landscape elements included by increasing the extent (Wiens, 1989). Furthermore, although the *S. altissimum* local model predicted a lower minimum abundance value than the regional model, the regional model predicted a maximum value almost twice as much as the local model's maximum value. For both species, a larger range of predicted abundance can provide more detail and easier interpretations for location with extreme predictions. In addition, higher abundances were predicted in the northwest corner of the local extent for *S. altissimum,* possibly showing a leading edge of invasion into the area. This pattern was not detected in the local model, and provides important information when monitoring alien species. For example, large seed sources might exist just outside the area of interest being modeled and without including data outside the area of interest increases the possible risk of an invasion would go unnoticed. Modeling the potential distribution of invaders in the local area is essential to alien species risk characterization (Stohlgren & Schnase, 2006), and only possible by sampling outside the local area. My results are similar to those of Menke et al. (2009) who looked at extrapolation of an Argentine ant in southern California and found if predictions are to be made to larger unsampled regions, additional sampling is needed to

capture the environmental variation in those regions. My results also suggest the importance of collecting data outside the local area to not only capture the environmental variation but also species response variation. These additional samples may improve model predictions and reveal patterns missed by local models.

Extrapolation may restrict predictions

Using BRTs to extrapolate models using local data to regional extents may constrict the range of predicted values. My results show when extrapolating local models to regional extents, predictions in the regional extent will not exceed the range of predictions within the local extent. For both *S. altissimum* and *B. gracilis*, models developed using local extent data did not predict abundance values below the minimum or above the maximum predicted within the local extent (Figures 2 and 3). Randin *et al.* (2006) found similarly restricted predictions when extrapolating to a completely separate region. Similarly, additional studies of extrapolation have shown other SDMs may over predict or under predict when extrapolated to novel conditions (Peterson *et al.*, 2007). Thuiller *et al.* (2004) found limiting the environmental conditions used to train the model may cause unpredictable effects on the tails of the response curves leading to poor extrapolations. How a model will predict when extrapolated to novel environments appears to depend on the specific model being used (Pearson *et al.*, 2006). Boosted regression trees fit response curves for each predictor and, for environmental values outside the sample variation, the response curves remain constant (Elith & Graham, 2009). This explains why the abundance range for models developed only using local data was the same for both the local extent and the regional extent.

The term clamping has been used to refer to restricting the model to only those areas of the landscape within the range of values from which the data used to train the model were sampled (Phillips, 2008). Evaluating clamping is a recent addition to model interpretation, but is becoming more common practice (Anderson & Raza, 2010; Fouquet *et al.*, 2010). Even with the caution surrounding model extrapolation, extrapolating models will likely continue to prompt new studies to explore ways to improve model extrapolations (see Elith & Leathwick, 2009 for a summary of these studies). For example, Miller *et al.* (2004) recommend using simple mechanistic relationships that are well understood when extrapolating beyond narrow ranges. Elith *et al*. (2010) suggested smoothing the initial models to improve fitting a model to the species rather than the specific data set when the model will be used for extrapolation.

Modeling abundances provides additional information

My results show BRTs can be a useful tool to model plant species abundance. I generated accurate spatial distribution models for plant species abundance with both local and regional extents. All models performed well even with small sample sizes. Interestingly, the *S. altissimum* regional model preformed the poorest of all four models. This may be due to the relatively low abundance of *S. altissimum* (often only observed at 1% cover) or because *S. altissimum* is a generalist species, which are often more difficult to predict (Evangelista *et al.*, 2008a). While *S. altissimum* was generally observed at low abundance values, a few locations were observed to have over 20% cover. These high abundances are not common, but are important to recognize for alien species management and my results suggest that BRTs may not predict abundances that are

unusually high. Boosted regression trees have shown to perform well for other abundance distribution modeling (Pittman *et al.*, 2009), and when compared to other methods (Elith & Graham, 2009).The ability to predict abundance rather than just probability of presences may provide more than just where a species may occur, but also information on the quality of habitat (Pearce & Ferrier, 2001). In terms of alien species, this information may help managers identify possible susceptible life stages to control and prevent invasion (Brown *et al.*, 2008). When managing and monitoring alien species, abundance predictions can help prioritize control and prevention efforts in addition to early detection. Many SDMs are limited to presence-absence or presence-only data. This is most likely due to the costs associated with obtaining abundance data compared to presence-absences or presence-only data. This has prompted comparison studies that investigated possible correlations between probability of presence and abundance. Unfortunately, these studies found little correlation, and if so, only between high probability of occurrence and high abundance (Vanderwal *et al*., 2009; but see Pearce & Ferrier, 2001). Boosted regression trees can be a useful tool to model plant species abundance on the central plains, even with small sample sizes.

I have shown BRTs can provide accurate and informative abundance models of native and alien plant species on the central plains. Often species distribution models are developed using presence-absence or presence-only data and rarely use abundance values. NEON is focusing on collecting abundance response values and will most likely use distribution models that can provide abundance predictions. More importantly, abundance data will become more available with the development of large databases collecting and disseminating ecological data (Graham *et al.*, 2007). While BRTs are still

largely unused in ecology (De'ath, 2007), the recent increase of BRTs in the literature is promising.

Integrating multiple sources of predictors

I chose to use a variety of different predictors from different sources. While some studies have focused on using only a single source of predictor variables (Lahoz-Monfort *et al.*, 2010), I integrated multiple sources of predictors. Selecting predictor variables relevant to the extent of the study area is important for accurate predictions (Wiens, 1989). I used soil data, land cover data, remotely sensed data, and topographic data as environmental predictors for both species. These predictors included both continuous and categorical data. While it may have been convenient to only use predictors from one source, the risk of not including an important predictor is increased. Boosted regression trees allow the use of both categorical and continuous predictor variables. This allowed me to include soil and land cover data not supported by other SDM methods (e.g. generalized linear models and generalized additive models) and, in my case, soil great group proved to be a significant contributor to my models. Furthermore, by reclassifying the land cover and soil data, I was able to generate layers that provided more interpretation and ecological relevance for the extents being modeled.

Caveats

All models have assumptions and SDMs assumptions are more likely to be violated to some degree when extrapolating a model in time or space (Wiens *et al.*, 2009). This is especially true for alien species because they are not at equilibrium with the

environment. While the use of BRTs to predict abundance appears promising, uncertainty is inherent to all models and results should be carefully interpreted (Elith & Leathwick, 2009). More specifically, decision trees are sensitive to the response data and predictors being modeled (Berk, 2008). Modifying either of these can result in very different models that have similar measured predictive abilities (Scull *et al.*, 2005). I did not test the ability of other SDMs that can model abundance and results from these methods are likely to be different. Small data sets may be more prone to varying results because of the importance of the addition or removal of a single or a few data points. Likewise, I may have overlooked an important predictor. For example, I did not include any dispersal or competition predictors which could impact abundance predictions (Austin, 2002). My results are for two generalist plant species and different patterns may be observed for species with rarer occurrences or higher abundance. The extents I used in my study were small and different predictors may be more important if the same study was preformed with larger extents (Wiens, 1989). An iterative approach to surveys and modeling may gain a more comprehensive understanding of modeling abundances and possible errors stemming from model extrapolation.

Conclusion

Extrapolating local models to regional extents is likely to predict abundance further from observed values when compared to models that included regional data. When possible, additional samples should be collected in the regional extent to improve predictions. In addition, local models may be improved by including data outside the local extent (at the regional extent). These additional data can provide insights into

populations that may be just outside the local extent and would otherwise go unnoticed. This information can be important for regional and local conservation planning in prioritizing management efforts. Future work may investigate the number of additional regional samples required and their optimal location to provide the best predictions. Boosted regression trees can be a useful tool for modeling and predicting species abundance, especially when using multiple sources for predictor variables. As species distribution models become more robust and flexible, integrating multiple sources of predictor variables to include key predictors for the species being modeled will become increasingly more important. More research into the use and extrapolation of species distribution models to predict species abundance is needed to fully understand the errors and benefits of this application.

APPENDICES

APPENDIX A

Common name and scientific name of the species observed at the regional extent. 72 total plots sampled for foliar cover using Braun-Blanquet method.

Common name	Scientific name
blue grama	*Bouteloua gracilis*
western wheatgrass	*Pascopyrum smithii*
fourwing saltbush	*Atriplex canescens*
carex stenophylla spp.	*Eleocharis* spp.
plains prickypear	*Opuntia polyacantha*
scarlet globemallow	*Sphaeralcea coccinea*
prickly russian thistle	*Salsola. pestifer*
kochia	*Kochia scoparia*
cheatgrass	*Bromus tectorum*
tall tumblemustard	*Sisymbrium altissimum*
yellow sweet clover	*Melilotus officinalis*
needle and thread	*Hesperostipa comata*
smooth brome	*Bromus inermis*
sand dropseed	*Sporobolus cryptandrus*
meadow barley	*Hordeum brachyantherum*
alfalfa	*Medicago sativa*
fremont cottonwood	*Populus fremontii*
ponderosa pine	*Pinus ponderosa*
common barley	*Hordeum vulgare*
mountain rush	*Juncus arcticus*
meadow grass	*Poa annua*
yellow salsify	*Tragopogon dubius*
bush dry	*Atriplex canescens*
wolly plantain	*Plantago patagonica*
buffalo grass	*Buchloe dactyloides*
purple three awn	*Aristida purpurea*
purple locoweed	*Oxytropis lambertii*
sixweeks fescue	*Vulpia octoflora*
stiff greenthread	*Thelesperma filifolium*
yucca	*Yucca filamentosa*
scurfy pea	*Psoralea tenuiflora*
prairie sagewort	*Artemisia frigida*

APPENDIX B

List of all environmental predictors used in BRT models and their source.

Environmental Predictor	Source
Aspect	Calculated from Elevation
Eastness	Calculated from Elevation
Elevation	U.S. Geological Survey (http://eros.usgs.gov)
Enhanced Vegetation Index	Calculated from Landsat bands
Greeness	Calculated from Landsat bands
LANDFIRE veg. class	http://www.landfire.gov/products_national.php
Normalized Difference Vegetation Index	Calculated from Landsat bands
Northness	Calculated from Elevation
Ratio Vegetation Index	Calculated from Landsat bands
Slope	Calculated from Elevation
Soil brightness	Calculated from Landsat bands
Soil Great Group	Soil Data Mart (SoilDataMart@nrcs.usda.gov)
Soil-Adjusted Vegetation Index	Calculated from Landsat bands
Solar Radiation	Calculated from Elevation
Wetness	Calculated from Landsat bands

APPENDIX C

LANDFIRE original values and labels with the classified values and labels. LANDFIRE data can be downloaded from http://www.landfire.gov/products_national.php). The original classifications are defined by NatureServes's ecological systems.

Original Values	Original label	Classified value	Classified values	Classified Label
11	Open Water	1	1	Open Water
21	Developed-Open Space	2	2	Developed
22	Developed-Low Intensity	2	3	Barren
23	Developed-Medium Intensity	2	4	Agriculture
24	Developed-High Intensity	2	5	Shrubland
31	Barren	3	6	Grassland/forbland
81	Agriculture-Pasture/Hay	4	7	Mixedgrass Prairie
82	Agriculture-Cultivated Crops and Irrigated Agriculture	4	8	Shortgrass Prairie
2007	Western Great Plains Sparsely Vegetated Systems	3	9	Riparian
2072	Wyoming Basins Low Sagebrush Shrubland	5		
2081	Inter-Mountain Basins Mixed Salt Desert Scrub	5		
2086	Rocky Mountain Lower Montane-Foothill Shrubland	5		
2094	Western Great Plains Sandhill Steppe	6		
2107	Rocky Mountain Gambel Oak-Mixed Montane Shrubland	5		
2127	Inter-Mountain Basins Semi-Desert Shrub-Steppe	6		
2132	Central Mixedgrass Prairie	7		
2149	Western Great Plains Shortgrass Prairie	8		
2159	Rocky Mountain Montane Riparian Systems	9		
2162	Western Great Plains Floodplain Systems	9		
2181	Introduced Upland Vegetation - Annual Grassland	6		
2182	Introduced Upland Vegetation - Perennial Grassland and Forbland	6		
2183	Introduced Upland Vegetation - Annual and Biennial Forbland	6		

APPENDIX D

Original soil data values and map unit names downloaded from Soil Data Mart provided by USDA NRCS (SoilDataMart@nrcs.usda.gov) with associated soil great group taxonomy classified using Soil Taxonomy, 1999.

Map unit name	Soil great group	Soil map unit value	Great group value
Altvan fine sandy loam, 0 to 6 percent slopes	Argiustolls	1	1
Badland	Badland	11	2
Bankard loamy fine sand, 0 to 3 percent slopes	Torrifluvents	12	10
Bresser sandy loam, 3 to 9 percent slopes	Argiustolls	16	1
Bushman fine sandy loam, 0 to 3 percent slopes	Haplustolls	17	6
Bushman fine sandy loam, 3 to 9 percent slopes	Haplustolls	18	6
Cascajo gravelly sandy loam, 5 to 20 percent slopes	Haplocalcids	20	5
Dacono clay loam, 0 to 6 percent slopes	Argiustolls	23	1
Epping silt loam, 0 to 9 percent slopes	Torriorthents	27	11
Haverson loam, 0 to 3 percent slopes	Ustifluvents	29	11
Kim-Mitchell complex, 0 to 6 percent slopes	Torriorthents	31	11
Kim-Mitchell complex, 6 to 9 percent slopes	Torriorthents	32	11
Manter sandy loam, 0 to 6 percent slopes	Argiustolls	34	1
Manter sandy loam, 3 to 9 percent slopes	Argiustolls	35	1
Manzanola clay loam, 0 to 3 percent slopes	Haplargids	36	4
Midway clay loam, 0 to 9 percent slopes	Torriorthents	37	11
Nucla loam, 3 to 9 percent slopes	Haplustolls	39	6
Ascalon fine sandy loam, 0 to 6 percent slopes	Argiustolls	4	1

Nunn loam, 0 to 6 percent slopes	Argiustolls	40	1
Nunn clay loam, 0 to 6 percent slopes	Argiustolls	41	1
Olney fine sandy loam, 0 to 6 percent slopes	Haplargids	44	4
Olney fine sandy loam, 6 to 9 percent slopes	Haplargids	45	4
Otero sandy loam, 0 to 3 percent slopes	Ustorthents	46	11
Otero sandy loam, 3 to 9 percent slopes	Ustorthents	47	11
Paoli fine sandy loam, 0 to 6 percent slopes	Haplustolls	49	6
Ascalon fine sandy loam, 6 to 9 percent slopes	Argiustolls	5	1
Paoli fine sandy loam, 6 to 9 percent slopes	Haplustolls	50	6
Peetz gravelly sandy loam, 5 to 20 percent slopes	Calciustolls	51	3
Peetz-Altvan complex, 0 to 20 percent slopes	Calciustolls	52	3
Peetz-Rock outcrop complex, 9 to 40 percent slopes	Calciustolls	53	3
Platner loam, 0 to 3 percent slopes	Paleustolls	54	8
Renohill fine sandy loam, 0 to 6 percent slopes	Haplargids	55	4
Renohill fine sandy loam, 6 to 9 percent slopes	Haplargids	56	4
Renohill-Shingle complex, 3 to 9 percent slopes	Haplargids	57	4
Rosebud fine sandy loam, 0 to 6 percent slopes	Argiustolls	58	1
Rosebud fine sandy loam, 6 to 9 percent slopes	Argiustolls	59	1
Shingle clay loam, 0 to 9 percent slopes	Torriorthents	60	11
Stoneham fine sandy loam, 0 to 6 percent	Haplustalfs	61	6

slopes			
Stoneham fine sandy loam, 6 to 9 percent slopes	Haplustalfs	62	6
Tassel loamy fine sand, 5 to 20 percent slopes	Torriorthents	63	11
Terry sandy loam, 0 to 3 percent slopes	Haplargids	64	4
Terry sandy loam, 3 to 9 percent slopes	Haplargids	65	4
Thedalund-Keota loams, 0 to 3 percent slopes	Torriorthents	66	11
Thedalund-Keota loams, 3 to 9 percent slopes	Torriorthents	67	11
Ascalon-Bushman-Curabith complex, 0 to 3 percent slopes	Argiustolls	7	1
Vona loamy sand, 0 to 3 percent slopes	Haplustalfs	71	6
Vona loamy sand, 3 to 9 percent slopes	Haplustalfs	72	6
Vona sandy loam, 0 to 3 percent slopes	Haplustalfs	73	6
Vona sandy loam, 3 to 9 percent slopes	Haplustalfs	74	6
Wages fine sandy loam, 0 to 6 percent slopes	Argiustolls	75	1
Wages fine sandy loam, 6 to 9 percent slopes	Argiustolls	76	1
Weld loam, 0 to 6 percent slopes	Argiustolls	77	1
Water	Water	85	12
Playas	Playas	86	9
Avar fine sandy loam	Natrargids	9	7

APPENDIX E

Example R code using boosted regression trees to model and predict abundance. This code uses the gbm package to fit the BRT.

```
#############################################################################
#Title: Boosted Regression Tree: Bouteloua gracilis Regional extent model
#10 variables
#Date: 07/29/2010
###########################################
#This code takes a CSV file with location points with cover of species
#and also the predictor variable values for each sample point to create a boosted
#regression
#tree model and prediction surface. Also calculates R2.
###########################################
#clear memory
rm(list=ls())

#Select a file and get path
#file.choose()

#Set working directory
#setwd("C:\\neyoung\\Projects\\R\\CPER_NEON\\")

#Data contains the response of all the species collected and the predictor vairables
data.all=read.csv("C:\\neyoung\\Projects\\CPER_NEON\\Data\\FinalData\\BRT_csv\\All
_bogr_10.csv")

#make generic data label so that we can move from dataset to dataset (CPER->AOP-
>All)
data<-data.all

#Data check command to see if data appear correct
head(data)
names(data)

#Look at the names of the predictors. These columns are the predictors (total = 15)
names(data[15:22])

################Organize data###############
#Set Categorical data
data$soil_gg<-factor(data$soil_gg, levels=c(1:14))
data$landfir<-factor(data$landfir, levels=c(1:9))

#set species to model
Y=data$bogr
```

```
#for loop to create vector of presence and absence for species
#This can be used to calculate AUC or generate a probability of presence model
#rather than predicted abundance
num.responses=length(Y)
pres.abs= numeric(num.responses)
for(i in 1:num.responses){
        if (Y[i]==0) pres.abs[i]=0
        else pres.abs[i]=1
}
#Count the number of occurances
num_occurances=sum(pres.abs)
print(num_occurances)

#Combine Response with Predictors to go into gbm
data.gbm<-cbind(Y,data[15:22])
##################FIT BRT MODEL###################
#call to load the gbm package
library(gbm)

#set the seed to get repeatable results
set.seed(1)

#set formula for BRT. Additional variables can be added here
formula<-Y~
        tc3 +
        tc1 +
        solarra +
        soil_gg +
        slope_d +
        rvi +
        landfir +
        elev +
        eastnes +
        aspect

#Fit the brt model
fit<- gbm(
        formula = formula,
        distribution = "gaussian",      # bernoulli, adaboost, gaussian,
                                                # poisson, coxph, and quantile available
        data = data.gbm,                        # dataset including response
        shrinkage = 0.001,              # shrinkage or learning rate
        n.tree = 5000,                  # number of trees
        interaction.depth = 3,          # 1: additive model, 2: two-way interactions, etc.
        #train.fraction = 1,            # fraction of data for training
        bag.fraction = 0.5,             # subsampling fraction, 0.5 is probably best
```

```
        cv.folds = 5,                   # do 5-fold cross-validation
        #keep.data = TRUE,              # keep a copy of the dataset with the object
        )
summary(fit)

###############REVIEW MODEL###############

#Set up plotting window to view results
par(mfrow=c(1,3))

#Set best iteration
best.iter<-fit[[1]]$best
print(best.iter)

# plot variable influence
summary(fit,n.trees=best.iter) # based on the estimated best number of trees

fit.summary<-summary(fit, n.trees=best.iter)
#Write the realtive influence file out
#write.table(fit_summary,"BRT_varimp_AOP_bogr.txt")

####################PREDICTIONS####################

#Make predictions from the boosted trees for the training data
pred.train <- predict.gbm(fit, data.gbm, best.iter)

#transform predicted values between 0-1
pred.train_transformed<-(pred.train-min(pred.train))*1/(max(pred.train)- min(pred.train))
Pred.test_transformed<-(pred.train-min(pred.train))*1/(max(pred.train)- min(pred.train))
summary(pred.train_transformed)

###############Spatial Predictions################

myfun<-function(x)round(unlist(predict.gbm(fit, x, best.iter)),4)

#Predictor and ASC names (the ASCIIs and predictor names must match)
fnames<-names(data[15:22])

#Path to a folder containing the ASCII predictor files
Fpath ->
"C:/neyoung/Projects/CPER_NEON/EnvironmentalPredictors/20x20/WGS84_UTM_13
N/ASCII/"

#Open Source R code for processing ascii by line
source("C:/neyoung/Projects/R/CPER_NEON/process_asc_by_line_v3.R")
```

```
#Spatial prediction function
#proc.asc.byline(fnames,fpath,myfun,n=150,outfile="all_bogr_BRT.asc")

###############ROC/AUC Calculations################
#Load the presenceAbsence package to calculate AUC
library(PresenceAbsence)

#Set up a data frame for the format used by the PresenceAbsence function
(ID,PresenceAbsence,Prediction)
#This is also used to calculate roc (see next section)
data.AUC=data.frame(ID=c(1:nrow(data)),response=pres.abs,pred.train_transformed)
#set up new plotting window for AUC plot
par(mfrow=c(1,1))

#Plot AUC/ROC plot
auc.roc.plot(data.AUC)

#Calculate auc using dataframe
auc.cal<-auc(data.AUC)
print(auc.cal)
###################Calculate R2##################
#Make predictions for the data
pred.train <- predict.gbm(fit, data.gbm, best.iter)
pred.observ <-cbind(Y,pred.train)

# calculate correlation between obsevred vs. predicted values from BRT
correlation<-cor(Y,pred.train)

# Run linear regression between obsevred and predicted to calculate R2
reg.predct.observe<-lm(Y~pred.train)
summary(reg.predct.observe)

#########################################################################
```

REFERENCES

Allen, E.B. & Knight, D.H. (1984) The effects of introduced annuals on secondary succession in sagebrush-grassland, wyoming. *Southwestern Naturalist*, **29**, 407-421

Anderson, R.P. & Raza, A. (2010) The effect of the extent of the study region on gis models of species geographic distributions and estimates of niche evolution: Preliminary tests with montane rodents (genus nephelomys) in venezuela. *Journal of Biogeography*, **37**, 1378-1393

Araujo, M.B. & New, M. (2007) Ensemble forecasting of species distributions. *Trends in Ecology & Evolution*, **22**, 42-47

Austin, M.P. (2002) Spatial prediction of species distribution: An interface between ecological theory and statistical modelling. *Ecological Modelling*, **157**, 101-118

Barnett, D., Stohlgren, T., Jarnevich, C., Chong, G., Ericson, J., Davern, T. & Simonson, S. (2007) The art and science of weed mapping. *Environmental Monitoring and Assessment*, **132**, 235-252

Berk, R.A. (2008) *Statistical learning from a regression perspective*. Springer Science+Buisness Media, New York, New York.

Beyer, H.L. (2004) Hawth's analysis tools for arcgis. URL http://www.spatialecology.com/htools

Braun-Blanquet, J. (1932) *Plant sociology. The study of plant communities*. McGraw-Hill book company, Inc, New York and London.

Breiman, L. (2001) Random forests. In. University of California Berkeley, CA

Brown, K.A., Spector, S. & Wu, W. (2008) Multi-scale analysis of species introductions: Combining landscape and demographic models to improve management decisions about non-native species. *Journal of Applied Ecology*, **45**, 1639-1648

Crall, A.W., Newman, G.J., Stohlgren, T.J., Jarnevich, C.S., Evangelista, P. & Guenther, D. (2006) Evaluating dominance as a component of non-native species invasions. *Diversity and Distributions*, **12**, 195-204

De'ath, G. (2007) Boosted trees for ecological modeling and prediction. *Ecology*, **88**, 243-251

Elith, J. & Graham, C.H. (2009) Do they? How do they? Why do they differ? On finding reasons for differing performances of species distribution models. *Ecography*, **32**, 66-77

Elith, J., Kearney, M. & Phillips, S. (2010) The art of modelling range-shifting species. *Methods in Ecology and Evolution*, no-no

Elith, J. & Leathwick, J.R. (2009) Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology Evolution and Systematics*, **40**, 677-697

Elith, J., Leathwick, J.R. & Hastie, T. (2008) A working guide to boosted regression trees. *Journal of Animal Ecology*, **77**, 802-813

Evangelista, P., Barnett, D., Stohlgren, T.J., Stapp, P., Jarnevich, C., Kumar, S. & Rauth, S. (2009a) Field and costs assessment for the fundamental sentinel unit (fsu) at the central plains experimental range, colorado. In. National Ecological Observatory Network, Inc.

Evangelista, P., Stohlgren, T., Morisette, J. & Kumar, S. (2009b) Mapping invasive tamarisk (tamarix): A comparison of single-scene and time-series analyses of remotely sensed data. *Remote Sensing*, **1**, 519-533

Evangelista, P., Stohlgren, T.J., Guenther, D. & Stewart, S. (2004) Vegetation response to fire and postburn seeding treatments in juniper woodlands of the grand staircase-escalante national monument, utah. *Western North American Naturalist*, **64**, 293-305

Evangelista, P.H., Kumar, S., Stohlgren, T.J., Jarnevich, C.S., Crall, A.W., Norman, J.B. & Barnett, D.T. (2008a) Modelling invasion for a habitat generalist and a specialist plant species. *Diversity and Distributions*, **14**, 808-817

Evangelista, P.H., Norman, J., Berhanu, L., Kumar, S. & Alley, N. (2008b) Predicting habitat suitability for the endemic mountain nyala (tragelaphus buxtoni) in ethiopia. *Wildlife Research*, **35**, 409-416

Fielding, A.H. & Haworth, P.F. (1995) Testing the generality of bird-habitat models. *Conservation Biology*, **9**, 1466-1481

Fouquet, A., Ficetola, G.F., Haigh, A. & Gemmell, N. (2010) Using ecological niche modelling to infer past, present and future environmental suitability for leiopelma hochstetteri, an endangered new zealand native frog. *Biological Conservation*, **143**, 1375-1384

Franklin, J. (2009) *Mapping species distributions*. Cambridge University Press.

Frayer, W.E. & Furnival, G.M. (1999) Forest survey sampling designs: A history. *Journal of Forestry*, **97**, 4-10

Friedman, J., Hastie, T. & Tibshirani, R. (2000) Additive logistic regression: A statistical view of boosting. *Annals of Statistics*, **28**, 337-374

Friedman, J.H. (1991) Multivariate adaptive regression splines. *Annals of Statistics*, **19**, 1-67

Friedman, J.H. & Meulman, J.J. (2003) Multiple additive regression trees with application in epidemiology. *Statistics in Medicine*, **22**, 1365-1381

Graham, C.H. & Hijmans, R.J. (2006) A comparison of methods for mapping species ranges and species richness. *Global Ecology and Biogeography*, **15**, 578-587

Graham, J., Newman, G., Jarnevich, C., Shory, R. & Stohlgren, T.J. (2007) A global organism detection and monitoring system for non-native species. *Ecological Informatics*, **2**, 177-183

Grinnell, J. (1917) Field tests of theories concerning distributional control. *American Naturalist*, **51**, 115-128

Guisan, A. & Thuiller, W. (2005) Predicting species distribution: Offering more than simple habitat models. *Ecology Letters*, **8**, 993-1009

Guisan, A. & Zimmermann, N.E. (2000) Predictive habitat distribution models in ecology. *Ecological Modelling*, **135**, 147-186

Hutchinson, G.E. (1957) Population studies - animal ecology and demography - concluding remarks. *Cold Spring Harbor Symposia on Quantitative Biology*, **22**, 415-427

Kumar, S., Spaulding, S.A., Stohlgren, T.J., Hermann, K.A., Schmidt, T.S. & Bahls, L.L. (2009) Potential habitat distribution for the freshwater diatom didymosphenia geminata in the continental us. In, pp. 415-420

Lahoz-Monfort, J.J., Guillera-Arroita, G., Milner-Gulland, E.J., Young, R.P. & Nicholson, E. (2010) Satellite imagery as a single source of predictor variables for habitat suitability modelling: How landsat can inform the conservation of a critically endangered lemur. *Journal of Applied Ecology*, **47**, 1094-1102

Levin, S.A. (1992) The problem of pattern and scale in ecology. *Ecology*, **73**, 1943-1967

Mack, R.N., Simberloff, D., Lonsdale, W.M., Evans, H., Clout, M. & Bazzaz, F.A. (2000) Biotic invasions: Causes, epidemiology, global consequences, and control. *Ecological Applications*, **10**, 689-710

Marilyn, S.J. & Hart, R.H. (1994) 61 years of secondary succession on rangelands of the wyoming high-plains. *Journal of Range Management*, **47**, 184-191

Mateo-Tomas, P. & Olea, P.P. (2010) Anticipating knowledge to inform species management: Predicting spatially explicit habitat suitability of a colonial vulture spreading its range. *Plos One*, **5**

Medley, K.A. (2010) Niche shifts during the global invasion of the asian tiger mosquito, aedes albopictus skuse (culicidae), revealed by reciprocal distribution models. *Global Ecology and Biogeography*, **19**, 122-133

Miller, J.R., Turner, M.G., Smithwick, E.a.H., Dent, C.L. & Stanley, E.H. (2004) Spatial extrapolation: The science of predicting ecological patterns and processes. *Bioscience*, **54**, 310-320

NEON (2010) National ecological observatory network inc. URL http://www.neoninc.org/

Parisien, M.A. & Moritz, M.A. (2009) Environmental controls on the distribution of wildfire at multiple spatial scales. *Ecological Monographs*, **79**, 127-154

Patman, J.P. & Hugh, I.H. (1961) Preliminary reports on the flora of wisconsin. No. 44. Cruciferae--mustard family. *Wisconsin Academy of Science, Arts and Letters*, **50**, 17-73

Pearce, J. & Ferrier, S. (2001) The practical value of modelling relative abundance of species for regional conservation planning: A case study. *Biological Conservation*, **98**, 33-43

Pearson, R.G., Thuiller, W., Araujo, M.B., Martinez-Meyer, E., Brotons, L., Mcclean, C., Miles, L., Segurado, P., Dawson, T.P. & Lees, D.C. (2006) Model-based uncertainty in species range prediction. *Journal of Biogeography*, **33**, 1704-1711

Penman, T.D., Pike, D.A., Webb, J.K. & Shine, R. (2010) Predicting the impact of climate change on australia's most endangered snake, hoplocephalus bungaroides. *Diversity and Distributions*, **16**, 109-118

Peterson, A.T., Papes, M. & Eaton, M. (2007) Transferability and model evaluation in ecological niche modeling: A comparison of garp and maxent. *Ecography*, **30**, 550-560

Phillips, S.J., Anderson, R.P. & Schapire, R.E. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231-259

Phillips, S.J. & Elith, J. (2010) Poc plots: Calibrating species distribution models with presence-only data. *Ecology*, **91**, 2476-2484

Pittman, S.J., Costa, B.M. & Battista, T.A. (2009) Using lidar bathymetry and boosted regression trees to predict the diversity and abundance of fish and corals. *Journal of Coastal Research*, **25**, 27-38

Pulliam, H.R. (2000) On the relationship between niche and distribution. *Ecology Letters*, **3**, 349-361

Randin, C.F., Dirnbock, T., Dullinger, S., Zimmermann, N.E., Zappa, M. & Guisan, A. (2006) Are niche-based species distribution models transferable in space? *Journal of Biogeography*, **33**, 1689-1703

Rollins, M.G. (2009) Landfire: A nationally consistent vegetation, wildland fire, and fuel assessment. *International Journal of Wildland Fire*, **18**, 235-249

Sagarin, R.D., Gaines, S.D. & Gaylord, B. (2006) Moving beyond assumptions to understand abundance distributions across the ranges of species. *Trends in Ecology & Evolution*, **21**, 524-530

Scott, M.J., Heglund, P.J. & Morrison, M.L. (2002) *Predicting species occurrences: Issues of accuracy and scale*. Island Press.

Scull, P., Franklin, J. & Chadwick, O.A. (2005) The application of classification tree analysis to soil type prediction in a desert landscape. *Ecological Modelling*, **181**, 1-15

Stohlgren, T.J., Chong, G.W., Schell, L.D., Rimar, K.A., Otsuki, Y., Lee, M., Kalkhan, M.A. & Villa, C.A. (2002) Assessing vulnerability to invasion by nonnative plant species at multiple spatial scales. *Environmental Management*, **29**, 566-577

Stohlgren, T.J. & Schnase, J.L. (2006) Risk analysis for biological hazards: What we need to know about invasive species. *Risk Analysis*, **26**, 163-173

Strubbe, D., Matthysen, E. & Graham, C.H. (2010) Assessing the potential impact of invasive ring-necked parakeets psittacula krameri on native nuthatches sitta europeae in belgium. *Journal of Applied Ecology*, **47**, 549-557

Team, R.D.C. (2010) R: A language and environment for statistical computing. In. R Foundation for Statistical Computing, Vienna, Austria

Thuiller, W., Brotons, L., Araujo, M.B. & Lavorel, S. (2004) Effects of restricting environmental range of data to project current and future species distributions. *Ecography*, **27**, 165-172

Wiens, J.A. (1989) Spatial scaling in ecology. *British Ecological Society*, **3**, 385-397

Wiens, J.A., Stralberg, D., Jongsomjit, D., Howell, C.A. & Snyder, M.A. (2009) Niches, models, and climate change: Assessing the assumptions and uncertainties.

Proceedings of the National Academy of Sciences of the United States of America, **106**, 19729-19736

Wikum, D.A. & Shanholtzer, G.F. (1978) Application of braun-blanquet cover-abundance scale for vegetation analysis in land-development studies. *Environmental Management*, **2**, 323-329

Wisz, M.S., Hijmans, R.J., Li, J., Peterson, A.T., Graham, C.H., Guisan, A. & Distribut, N.P.S. (2008) Effects of sample size on the performance of species distribution models. *Diversity and Distributions*, **14**, 763-773