UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

EXPLORING ECOLOGICAL AND BIOGEOGRAPHIC QUESTIONS USING BIOLOGICAL DATABASES DERIVED FROM NATURAL HISTORY COLLECTIONS AND SURVEYS

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

Priscilla H. C. Crawford
Norman, Oklahoma
2009

EXPLORING ECOLOGICAL AND BIOGEOGRAPHIC QUESTIONS USING BIOLOGICAL DATABASES DERIVED FROM NATURAL HISTORY COLLECTIONS AND SURVEYS

A DISSERTATION APPROVED FOR THE
GRADUATE COLLEGE

BY

____
Dr. Bruce Hoagland, Chair

____
Dr. Wayne Elisens

____
Dr. J. Scott Greene

____
Dr. Jeffery Kelly

____
Dr. Caryn Vaughn

for my Mom

Patricia Ann Callahan

ACKNOWLEDGEMENTS

I thank my committee Bruce Hoagland, Wayne Elisens, Scott Greene, Jeff Kelly, and Caryn Vaughn for assisting with my dissertation research and writing. I am grateful to Michael Patten for helpful ecological discussions and for assisting with statistical problems. I am thankful for the opportunity to attend the "Species Distribution Modeling Methods for Conservation Biologists" workshop hosted by the American Museum of Natural History and lead by Richard Pearson and Steven Phillips. I thank the Oklahoma Biological Survey, Oklahoma Natural Heritage Inventory, and Oklahoma Natural Areas Registry for supporting this research. I appreciate all the support and encouragement I received from my family and friends, especially Phillip, Rosie, Mom, Jenny, John, Pat, Shari, Jim, Leschen, Heather, and Paulette.

TABLE of CONTENTS

LIST OF TABLES

Chapter 1

Chapter 2

Chapter 3

Chapter 4

LIST OF FIGURES

Chapter 1

Chapter 2

Chapter 3

Chapter 4

ABSTRACT

Biogeographic research has benefited from the digitizing of large databases derived from natural history collections and biological surveys. These resources made available via the Internet can be accessed by biogeographers around the world to address a multitude of ecological and geographic questions. Utilizing this data taps into hundreds of years of study and countless hours of research conducted by biologists across the globe. This dissertation could not have been completed without the availability of data collected by legions of researchers from museums, herbaria, and government agencies. By taking advantage of data collected by others, I was able to work at a geographic scale that would have been impossible had I gathered all my own data.

In chapter one, I use herbarium data to describe the temporal and spatial patterns of invasive and expansive species for the entire state of Oklahoma. Because of the inherent bias in collections of natural history specimens. I test techniques for eliminating temporal collecting bias: regression models and proportion curves. I found that patterns of species invasion and expansion in Oklahoma could be detected using these techniques which were developed for regions with longer collecting plant histories. The proportion curve analysis eliminated some biases inherent in herbarium data by reducing the effect of collecting effort. Both the regression model and proportion curve analyses illustrate the temporal invasion patterns of alien, invasive species. However, the native species did not show a clear expansion pattern. The information found in recently established herbaria may not be sensitive enough to detect the increase of abundance of native species.

Currently species distribution modelling is one of the most popular methods of utilizing large, georeferenced, biological databases. Chapter two is a brief review of the overabundant literature on species distribution modelling. Topics covered are the theoretical basis for distribution modelling, species and predictor data, modelling techniques, model evaluation, and uses for predictive maps created by modelling.

Using survey data collected for the U.S. Fish and Wildlife Service, I apply species distribution modelling techniques to predict suitable habitat for the endangered American burying beetle (*Nicrophorus americanus*). Using a suite of predictor variable thought to influence a burrowing insect, I built several models using a variety of modelling techniques. The Maxent modelling algorithm performed the best. However, being a generalist species, the suitable habitat for *N. americanus* was not well modelled. Model performance could be improved by incorporating information on the cause of *N. americanus*'s endangered status and its population shrinkage. To improve the models and consequently the recovery effort for the species, I need to take into account interactions including congener and vertebrate competition and a reduction in optimally sized prey. Creating an accurate spatial layer of this data will be a future challenge. My hope was to produce a map of potentially suitable habitat for *N. americanus* that would guide conservation efforts within the state of Oklahoma. Although the model was not highly accurate, the map of suitable habitat can help to inform conservation biologists of areas that have suitable habitat for the *N. americanus*.

In chapter four, I return to the invasive species theme by addressing the question of whether the introduced distribution of invasive species can be predicted from its native range. I modelled the potential distribution within the United States of three alien

invasive species native to Europe using the Maxent modelling technique. Using occurrence data from both the native (Europe) and introduced (US) ranges, I used reciprocal modelling to evaluate habitat discrepancies between the introduced and native ranges. This modelling approach can help to determine which environmental factors within the introduced range are different from the native range and which habitats within the native range are not represented in the introduced range. Further, reciprocal modelling can reveal potential problems with occurrence data and predictor variables in both native and introduced ranges, but it also has also been used to investigate ecological phenomena, such as niche shifts of invasive species in their introduced range. The native occurrences in Europe accurately predicted the distribution within Europe; and introduced occurrences in the US accurately predicted the US distribution. However, the reciprocal models did not perform well. The explanations for the dissociated ranges of each species in Europe and US can possibly be related to the hypotheses postulated for invasive species success. The characteristics that make a species invasive may be the cause of the species' environmental range to be different in the native and introduced regions. My aim was to see if we could use easily obtained data to model the potential areas of invasion within our state and use this information to assist conservation efforts such as early detection and rapid response. My model results indicate that the occupied niches are too inconsistent between the native and introduced ranges to make models useful at the scale we are interested in. Further modeling attempts will utilize more introduced occurrence data from areas within our region of the United States. This will entail a more concerted effort to locate available data in the areas where the species may be expanding.

INTRODUCTION

This dissertation began with an interest in utilizing for research the vast storehouse of data collected in the herbaria of Oklahoma, now digitized as the Oklahoma Vascular Plants Database (OVPD; Hoagland et al. 2009). I wanted to explore biogeographic questions at the state level by mining the data collected by botanists over the past 100+ years. My interest in invasive plants led me to chose my first dissertation problem - Can we use data in the OVPD to map the historic invasion of plants across Oklahoma? And can we apply the same techniques to species that are native, but exhibit invasive behavior in response to human disturbance? The results of my investigation into these questions make up the contents of Chapter 1.

Having explored the historic spread of invasives, I was interested to see if we could predict the future distribution of invasive species that have not yet become well established in Oklahoma. A recently developed and growing sub-field of biogeography - species distribution modelling - became an excellent tool to study the potential distribution of new invasive species. Species distribution modelling (SDM) is currently the trendy line of research and the literature is extensive and rapidly growing. Because of its relatively new status, there were few texts or articles that compile and review the literature when I began my research into SDM. I conducted a review of the literature for my own use to better understand the background and proper use and interpretation of the models produced by these techniques (Chapter 2). During the course of researching and writing the literature review, I thought it wise to introduce myself to SDM using a small data set that contained both presence and absence data. Survey data for the American burying beetle were available and a model of its habitat preference would be useful for conservation efforts within Oklahoma for this endangered species.

I was lucky enough to attend the “Species Distribution Modeling Methods for Conservation Biologists” workshop hosted by the American Museum of Natural History and lead by Richard Pearson and Steven Phillips who have authored many articles on the topic. At the workshop, I was inspired to take my invasive species modeling international and use the native range data to explore the potential range invasive species in a new area. My intention was to use the result to help locate areas in Oklahoma that had the potential habitat for particular invasive species, but my results illustrate a not uncommon problem - species do not necessarily occupy the same climatic niche in their native and introduced ranges.

CHAPTER 1

Can herbarium records be used to map alien species invasion and native species expansion over the past 100 years?

Priscilla H. C. Crawford [1, *]

& Bruce W. Hoagland [1, 2]

[1] Oklahoma Biological Survey, University of Oklahoma,
111 E. Chesapeake St., Norman, OK 73019, USA;
[2] Department of Geography, University of Oklahoma,
Norman, OK 73019, USA

* Correspondence: Priscilla H. C. Crawford, Oklahoma Biological Survey,
University of Oklahoma, 111 E. Chesapeake St., Norman, OK 73019, USA.
E-mail: prill@ou.edu.

formatted for submission to Journal of Biogeography

(published in Vol 36: pg 651-661; doi: 10.1111/j.1365-2699.2008.02043.x)

Abstract

Aim To determine if the temporal and spatial pattern of alien plant invasion and native plant expansion can be observed using 100 years of herbarium data from Oklahoma, USA and to eliminate herbarium collection biases in such analyses.

Location Oklahoma, USA.

Methods Using herbaria records from the Oklahoma Vascular Plants Database from 1903 to 2004, we reconstructed the spatial and temporal collection history of two alien, invasive taxa (*Lonicera japonica* and *Tamarix* spp.) and three native, expansive species (*Ambrosia psilostachya*, *Amphiachyris dracunculoides* and *Juniperus virginiana*). To compare the overall collecting trend, groups of native, non-expansive taxa were selected as counterparts. We recorded the year of the first collection in each township in Oklahoma for all taxa. The cumulative number of occupied townships was log-transformed, plotted against time, and modelled with linear regression. The slope of the linear regression represented collection trend over time for the non-expansive counterpart group. However, for the invasive and expansive species, the regression slope represented the collection effort ***plus*** invasion or expansion rate. We calculated the proportion of invasive and expansive species to non-expansive species by dividing the cumulative number of townships for each invasive or expansive species by the cumulative number of townships occupied by the counterpart group (proportion curve).

Results Maps of the collection records of invasive and expansive taxa illustrated no discernable spatial invasion or expansion pattern. The slopes of the linear regression for alien, invasive taxa were significantly steeper than those of their associated native, non-expansive counterparts, indicating an increase in abundance. *Juniperus virginiana, L.*

japonica and *Tamarix* spp. exhibited one or more periods during which they were collected at a disproportionately higher rate than their native, non-expansive counterparts.

Main conclusions Patterns of species invasion and expansion in Oklahoma were detected using techniques developed for regions with longer collecting plant histories. The proportion curve analysis eliminated some biases inherent in herbarium data by reducing the effect of collecting effort. Both the regression model and proportion curve analyses illustrate the temporal invasion patterns of alien, invasive species. The native species did not show a clear expansion pattern. The information found in recently established herbaria may not be sensitive enough to detect the increase of abundance of native species.

INTRODUCTION

Understanding the temporal and spatial dynamics of invasive and expansive species has become an important research topic for biogeographers, ecologists, weed scientists and conservation biologists. To understand the geographic history of alien plant invasions and native plant expansions many researchers are turning to the vast storehouses of information associated with herbarium specimens. Collections of alien plant species in herbaria around the world are being analysed to help ecologists recognize the spatio-temporal patterns of plant invasions (Stadler *et al*., 1998; Delisle *et al*., 2003; Woods *et al*., 2005; Wu *et al*., 2005; Barney, 2006; Chauvel *et al*., 2006; Fuentes *et al*., 2008). Herbaria are underutilized institutions that contain a large repository of historical and geographical information. Pyšek, using European herbarium specimens, developed a technique to quantify invasion rate (Pyšek, 1991; Pyšek & Prach, 1993; Mihulka & Pyšek, 2001; Pyšek *et al*., 2003). He used the term “invasion curve” to represent a regression model of the cumulative number of localities of an invasive plant plotted against the year of collection. The slope of the regression was considered a quantification of the invasion rate (Pyšek & Prach, 1993).

However, we must be cautious interpreting regression models calculated from herbarium data because of the non-random sampling bias inherent in plant collections. Few studies take into consideration the biased nature of natural history collections such as: unequal sampling effort over time, non-random geographic representation, poor location information, incorrect identification, and disproportionately represented taxa. Therefore, methods must be developed to remove such biases to reveal the true pattern of invasion. Temporal variation in plant collection effort is apparent when the number of herbarium specimens is plotted against year. For example, in the herbaria of the state of Oklahoma, USA, the number of specimens collected per year since 1883 varies from zero to 6365, with a mean of 1752 per year (Hoagland *et al*., 2006). The intensity of floristic inventory is therefore highly variable and should be taken into account when studying invasive species. The increase in the number of specimens of an invasive species may indicate an increase in abundance, or simply may mean an increase in the overall collecting effort that year or decade. Mihulka & Pysek (2001), using data from herbaria across Europe, corrected for collection rate among countries to account for the variation in plant collecting intensity. Delisle *et al*. (2003) also developed a method to account for the bias associated with temporal variation in plant collections in riparian areas of southern Québec, Canada. They selected widespread, native, non-invasive species to provide a picture of collecting trends in the region. In addition to comparing collection rates, they calculated the ratio of invasive and non-invasive plant records for each year, termed the "proportion curve" (Delisle *et al*., 2003). If the proportion of invasive species collections increases over a period of time, this suggests that the invasive species increased in range or abundance. This differs from Pyšek's invasion curve, which evaluates the overall invasion rate of a species since its first collection and does not take into account specific time periods during which invasion may have occurred rapidly.

Pyšek also recognized that herbarium data had limitations and believed that a "strong, long-term florisitic tradition" in the region was important to produce reliable analysis of plant invasion (Pyšek & Prach, 1993). Yet, Fuentes *et al*. (2008) in Chile, Woods *et al*. (2005) in Kansas, USA, Delisle *et al*. (2003) in Québec, Canada, and Stadler *et al*. (1998) in Kenya all produced analyses with data sets that were significantly more recent than the several hundred years of data available in Europe. In Kansas the earliest specimen was collected in 1869, while in Québec the earliest specimen was collected in 1820. For Chile, Fuentes *et al*. (2008) only analysed the herbarium specimens collected since 1900. In Kenya a few specimens were collected before 1940, but most were collected after 1960. Wu *et al*. (2005) were concerned with the adequacy of using herbarium data to map the distribution of alien, invasive species because of their short-term history in Korea. They studied *Crotolaria* species that had only been naturalized for 70 years. Not all European studies have the benefit of a long-term data set. Chauvel *et al*. (2006) examined the increase of North American *Ambrosia* species in France using only approximately 150 years of data.

We were interested in testing these methods on herbarium data found in the Oklahoma Vascular Plants Database (OVPD), the repository for the plant collection data of the state of Oklahoma. The OVPD represents slightly over 120 years of plant collecting, with the earliest specimen collected in 1883, though significant numbers of plant collections were not made until the 1910s (Hoagland *et al*., 2006). Not only are we interested in applying these methods to truly invasive species, but are also interested in detecting the patterns of increase of native, expansive species. Invasive taxa are alien species that have spread over a considerable area after introduction from another region by humans (Richardson *et al*., 2007). Expansive species are native plants that are moving into new areas and increasing in abundance because of human-induced changes to the landscape. Some of the expansive species are considered agricultural weeds, but some, especially in the Great Plains of North America, are woody species encroaching on grasslands.   In this paper, we address the following questions. (1) Will we be able to detect the spatial and temporal invasion pattern of alien plants in Oklahoma using the relatively recent collecting history represented in the OVPD? (2) Can we effectively eliminate regional and temporal biases using previously developed research methods? (3) Will these methods be suitable for illustrating expansion patterns of native, weedy plant species?

MATERIALS AND METHODS

We reconstructed the spatial and temporal collection history of: two alien, invasive taxa; three native, expansive species; and three native, non-expansive counterpart groups using records in the OVPD. We chose taxa that are both alien and native to see if we would be able to detect a spatio-temporal pattern of increase from herbarium records. Nomenclature follows the PLANTS Database (USDA, NRCS, 2006). We selected four species and one genus that are considered "weeds" in the Great Plains (Stubbendieck *et al*., 1994; Southern Weed Science Society, 1998; Coppedge *et al*., 2002; Friedman *et al*., 2005; USDA, NRCS, 2006): *Ambrosia psilostachya* DC. (Asteraceae), *Amphiachyris dracunculoides* (DC.) Nutt. (Asteraceae), *Juniperus virginiana* L. (Cupressaceae), *Lonicera japonica* Thunb. (Caprifoliaceae), and *Tamarix* L. (Tamaricaceae). *Ambrosia psilostachya* and *A. dracunculoides* are native to Oklahoma and are considered agricultural weeds (USDA, NRCS, 2006). *Juniperus virginiana* is a woody species native to Oklahoma that is known to increase in abundance in grasslands in the absence of fire (Coppedge *et al*., 2002; USDA, NRCS, 2006). *Lonicera japonica* and *Tamarix* are alien, invasive taxa that originated in Asia and Eurasia, respectively (USDA, NRCS, 2006). Species of *Tamarix* known to occur in Oklahoma are *T.*

parviflora, *T. ramosissima* and *T. chinensis* (Tyrl *et al*., 2006). We grouped all species of *Tamarix* for our analysis due to the difficulties in identification, current confusion in the taxonomy, and similar ecological functional roles.

To compare the overall collecting trend, groups of non-expansive species native to Oklahoma were selected as counterparts for each invasive or expansive taxon. Species chosen for counterpart groups were selected based on the following criteria: represented in the OVPD with at least 200 specimens; distribution similar to the invasive or expansive taxa; similar life form or habit; readily identifiable; and not taxonomically confusing. We used a combination of several species to diminish possible collecting bias found in any particular species.

The following species in the Asteraceae were assigned to the non-invasive counterpart group for *A. dracunculoides* and *A. psilostachya*: *Engelmannia peristenia* (Raf.) Goodman & Lawson, *Gaillardia pulchella* Foug., *Liatris squarrosa* (L.) Michx., *Pyrrhopappus grandiflorus* (Nutt.) Nutt. and *Ratibida columnifera* (Nutt.) Woot. & Standl. An effort was made to choose species within the same family, approximately the same size, and found in similar habitats. The following common, woody species were chosen as native, non-expansive counterparts for both *J. virginiana* and *Tamarix* spp.: *Morus rubra* L. (Moraceae), *Prunus angustifolia* Marsh. (Rosaceae), *Rhus aromatica* Ait. (Anacardiaceae) and *Sapindus saponaria* L. (Sapindaceae). Similar to the invasive and expansive species to which they will be compared, these woody species are large shrubs or small trees and are widely distributed throughout the study area. We chose two congeneric species, *Lonicera flava* Sims and *Lonicera sempervirens* L., as native, non-expansive counterparts for *L. japonica*. These were chosen based on similar taxonomy (within the same genus), habit (vining perennials), habitat (woodland edges and fencerows), and distribution (eastern Oklahoma). By comparing the temporal and spatial collection rates of invasive and expansive taxa to non-expansive taxa, we attempt to understand the general collecting trend so that attention could be drawn to the invasion and expansion history. We hope to de-emphasize the general collecting trend of the native, non-expansive taxa from the collecting trend of invasive species to emphasize the increase in abundance over time of the invasive and expansive species.

All specimen records for invasive and expansive species and their non-expansive, native counterpart groups were selected from the OVPD, which includes all plant collections from the following major herbaria: OKL, OKLA, TULS, OCLA, CSU, and DUR (for institution names and locations, see Holmgren & Holmgren, 2006; Hoagland *et al*., 2006). At the time of this research, minor plant collections represented in the OVPD were from Oklahoma Panhandle State University at Goodwell and the University of Oklahoma Biological Station at Kingston. In general, herbarium specimens have the following associated data: species name, location of collection, collector, collection date and collector's collection number. However, there is no standard label format or data requirements and many specimens lack even basic data. The variable nature of information provided on herbarium specimen labels required the elimination of some specimens from our study. First, specimens lacking specific collection date were removed from analysis. Cultivated specimens were also removed from analysis. Specimens with unknown or imprecise location information were excluded from analysis. Specimens of the same species with identical collectors, collection dates, collection numbers and locations were considered duplicate records and treated as one collection.

Specimens in the resulting data set were georeferenced to township (93.3 km^2) and mapped using ArcGIS 9.1 (ESRI®, Redlands, CA, USA). Townships, established in Oklahoma during the Public Land Survey of 1871, are quadrangles approximately 6 miles (9.66 km) on each side and contain 36 equal sections (Hoagland, 2006). If not recorded, the township was determined by interpreting directions to collection location provided on the herbarium label. The date and location of the first collection in each township was identified and the total number of townships in which the invasive and non-invasive counterpart groups were found was calculated. For a better comparison of the uneven sample sizes of the invasive and expansive species with their counterpart groups, we log-transformed ($\log_{10}$) the cumulative number of occupied townships. Beginning with the first collection of the invasive or expansive taxa, the data were plotted against time, and linear regression models were calculated. The slope of the linear regression model was used to quantify the collection and invasion or expansion rate of the taxa in this study. The slope of the curve represented collection effort over time for the non-invasive counterpart group and collection effort ***plus*** invasion rate for invasive species. The steeper the slope of the curve, the faster the rate of collection or invasion (Pyšek & Prach, 1993). We then tested equality of the slopes of the regressions (Sokal & Rohlf, 1995). We also employed the method developed by Delisle *et al*. (2003) to compare the trend in general collecting of non-invasive species to the collection trend of invasive species because this method does not correct for the temporal variability of plant collections. We calculated the proportion of invasive to non-invasive plant collections by dividing the cumulative number of townships for each invasive species by the cumulative number of townships occupied by the non-invasive counterpart group (proportional curve). This proportion illustrated in graphical format, the proportional curve, allowed us to examine collection rate during short time periods.

RESULTS

Herbarium specimens

Following the removal of unusable and duplicate specimens, 3696 records remained for analysis (Table 1). Of those, township was recorded on the specimen label for 1103 records. 3114 were manually georeferenced. Although the first specimen used in this analysis was collected in 1903, relatively few specimens of taxa of interest were collected in Oklahoma before 1935.

Spatial and temporal distribution

The native, non-expansive counterpart groups of woody species and Asteraceae taxa were found throughout Oklahoma and were not concentrated in any geographic region (Fig. 1a,c). The native, expansive taxa, *A. dracunculoides*, *A. psilostachya* and *J. virginiana*, also were not limited to a particular region of the state (Fig. 2a,b,c). *Lonicera* collections, both native and alien, were generally restricted to the eastern half of Oklahoma (Fig. 1b, 2d). *Tamarix* was found across Oklahoma with the exception of the south-eastern corner (Fig. 2e).

The maps generated from specimen location information illustrated no discernable spatial invasion or expansion pattern by any of the invasive or expansive taxa; new localities in different regions of the state were collected simultaneously (Fig. 2). The earliest collections of *A. dracunculoides*, *A. psilostachya* and *J. virginiana* were scattered across Oklahoma in a pattern that did not suggest an expansion front or radial expansion pattern (Fig. 2a,b,c). The first four collections of *L. japonica* were made in north-central Oklahoma in the 1930s (Fig. 2d). However, subsequent collections were scattered throughout the eastern half of the state and did not follow a radial pattern of invasion. The first *Tamarix* collection was made in the centre of the state in 1910. There was no apparent radial or linear (such as along a river corridor) invasion of *Tamarix* based on initial analysis of the early collections points (Fig. 2e). The lack of evidence of an invasion front could indicate that the alien species were first introduced to the state in multiple locations.

Invasion and expansion rates

The linear regression models for the native, expansive species, *A. dracunculoides*, *A. psilostachya*, and *J. virginiana*, were not significantly steeper than the models of the associated non-invasive counterpart groups (P > 0.05; Fig. 3a,b,c). The regression models for both the alien, invasive taxa, *L. japonica* and *Tamarix*, had significantly steeper slopes than the associated non-invasive counterparts (P < 0.01; Fig. 3d,e). This indicates that the rate at which *L. japonica* and *Tamarix* have been collected over the last 100 years has increased in comparison to the collection rate of their associated non-invasive counterpart taxa. The comparisons of the regressions of *A. dracunculoides*, *A. psilostachya* and *J. virginiana* to their native counterparts indicate that the collection rates of these species are not significantly different from the overall collection rate.

The proportion curve analysis indicates a time period during which for some of the invasive and expansive taxa were collected disproportionately more compared to their native counterpart group (Fig. 4). *Juniperus virginiana* shows a likely increase in abundance during the 1930s, but, interestingly, appears to decline from that period to the present (Fig. 4c). *Lonicera japonica* has a dramatic spike after its initial collection in the 1930s and the proportion curve illustrates a steady increase in abundance relative to its native congeners since 1970 (Fig. 4d). Tamarix also increased in abundance in the 1930s and shows a slight increase during the 1960s (Fig. 4e). Neither *A. dracunculoides* nor *A. psilostachya* have proportion curves that illustrate remarkable expansion, with the exception of a small, short increase in the late 1930s by *A. dracunculoides* (Fig. 4a,b).

DISCUSSION

Regression models and proportion curves

Generally, after the initial introduction of an invasive species, the pattern of invasion begins with a lag period of few collections followed by a period of rapid, exponential expansion. Alien, invasive species recently studied in France (Chauvel, 2006), Kenya (Stadler *et al*., 1998), Quebec (Delisle *et al*., 2003), and across Europe (Pyšek & Prach, 1993) and North America (Barney, 2006) follow this temporal invasion pattern. Our data appear not to support a typical lag period because the short, flat portion of the curve at the beginning of the time period is also seen in the native, non-expansive taxa. This suggests that the pattern is an artefact of collection history. The absence of a true lag period may be the result of the OVPD not having records during this phase of the invasion. The alien species in our study were both introduced to North America before many specimens in the OVPD were collected. The lack of a lag phase may also be due to the generation time (time for the population to reproduce) of the alien species in our research. Pyšek & Prach (1993) found that the generation time of riparian species affected the rate of invasion. The shorter a species lifespan, the faster the invasion rate.

The alien species examined in our research are both long lived perennials, one a woody vine and the other a small tree/shrub. Both Pyšek & Prach (1993) and Delisle *et al*. (2005) were working with species in riparian areas, a habitat type that may see a faster rate of invasion. Water flow can be an important dispersal agent for both seed and vegetation fragments (Baker, 1974; Richardson *et al*., 2007).

We found, in spite of the short and variable plant collecting history in Oklahoma, that the regression models indicate an invasion trend in the alien taxa (*L. japonica* and *Tamarix*). Both regression models had steeper slopes than their non-invasive counterpart groups, signifying over the past 100 years that the cumulative number of townships occupied was increasing faster than the number of townships occupied by non-invasive species. Delisle *et al*. (2003) found that four of the six invasive species in their study exhibited steeper slopes than their native counterpart groups. The expansion trend was not clear for the native, expansive species that we studied. This may be due to the nature of native, expansive species. Native, expansive plants have presumably been present in the region since the arrival of Europeans in North America, but they increase in abundance over time, in response, mostly, to human disturbance. In Oklahoma, this may be the result of a variety of factors, such as fire suppression, regrowth in abandoned fields, or intensive grazing. By looking at native, expansive species, we are really looking at an increase in population abundance which differs greatly from alien plant invasion. Attempting to use herbarium data to understand population dynamics of native species will be extremely difficult, if not impossible, due to the irregular nature of plant collecting and herbarium data.

The proportion curves revealed temporal invasion and expansion patterns, but at a finer scale and therefore may better serve for analysis of data sets that cover a shorter time frame. *Juniperus virginiana*, *L. japonica* and *Tamarix* exhibited one or more periods during which they were collected at a disproportionately higher rate than their native, non-expansive counterparts (Fig. 4c,d,e). Because the proportion curve of *L. japonica* shows an increase compared to that of the native congeners over the past 30 years until the present, we may hypothesize that *L. japonica* continues to invade new locations (Fig. 4d). *Juniperus virginiana*'s proportion curve shows a significant increase in collections during the 1930s, but also has a steady decline for approximately the last 50 years. These results contradict other studies that clearly demonstrate that *J. virginiana* has expanded into grasslands in Oklahoma over the past 50 years (Coppedge *et al*., 2002). The differing results from the proportion curves of *J. virginiana* and *L. japonica* may be an indication of plant collector bias. The continued collection of *L. japonica* above the rate of its native congeners is evidence of continued expansion of *L. japonica* into ***new*** locations. Plant collectors are interested in collecting species new to an area or rare in a habitat. The decline in *J. virginiana* collections with respect to other native woody species may be counterintuitive evidence of its increase in abundance. Botanists generally have neglected to collect native species considered to be abundant weeds. One of the most ubiquitous species in North America, *Taraxacum officinale* (common dandelion), has only 202 records in the 210,000 records of the OVPD (Hoagland *et al*., 2006). However, Woods *et al*. (2005) found that early collections of alien species in Kansas were extensive and were consistent with the overall collecting pattern for the state. The possible lack of interest in collecting native "weedy" species makes analyses such as ours more complicated. While native, expansive species may be ignored, alien, invasive species may currently hold the interest of collectors who are trying to document their spread. The increase of *L. japonica* and *Tamarix* specimens in the past decade signify the recent trend to identify and control alien, invasive species and may not necessarily signify an increase in their real-world abundance.

Complications of herbarium data

The relatively short history of plant collecting in Oklahoma is problematic when one wants to understand long-term trends in biogeography of the region especially the invasion history of alien species. Pyšek & Prach (1993) believe that a long history of thorough plant collecting is necessary to produce reliable results. Initial collecting of the Oklahoma flora began late, when some alien species had already been introduced. Both *L. japonica* and *Tamarix* were introduced to North America in the early 1800s (Baum, 1967; USDA, ARS, 1970), well before the first herbaria were established in Oklahoma. However, this study demonstrates that the data from herbarium specimens in Oklahoma are sufficient to demonstrate periods of invasion by alien taxa. The history of plant collecting in Oklahoma may be too short for detailed analysis of spatial patterns and population increase of native, expansive species.

The nature of herbarium records, which involves opportunistic and non-systematic plant collecting, makes analysis difficult because this type of data gathering introduces several biases. Several historical events, beginning with the establishment of the state's universities, influenced the temporal plant collecting pattern of the records in the OVPD. The geographic pattern of plant collecting is determined by the preference of the plant collector, not based on a systematic grid of the state, or stratified random sampling of ecoregions. Taxonomic bias, overrepresentation of certain groups of taxa, can be found in many collections. All temporal, geographic, and taxonomic biases must be considered for one to be confident in the results obtained from herbarium data research. Through various methods we made an effort to reduce the power of these biases to control our results.

Maps of plant distributions made with records in the OVPD should give us a reasonably accurate picture of the current extent of a given species within Oklahoma. Wu *et al*. (2005) tested the adequacy of herbarium data to illustrate the distribution of alien taxa. By comparing herbarium data with extensive field surveys, they found that herbarium records gave an accurate picture of the distribution and frequency of several species introduced into Korea during the last 70 years. Plant distribution maps will be more accurate as the number of plant collections increases. Therefore, the longer the history of plant collecting in the region, the better documented the flora, and the more comprehensive the herbarium collections. The accumulation of specimens over 100 years should provide a good illustration of species distribution. Mapping the records from the earliest decades would be less likely to yield a reliable representation of species distribution because there simply are fewer specimens collected. Attempting to discern a pattern of invasion over time using the somewhat sparse data prior to 1930 is unlikely to represent the true invasion history of a plant; instead, we merely document the “invasion” of Oklahoma by botanists. Given the short history of the herbaria embodied in the OVPD, analysis of the change in species distribution over time can be misleading. In reality, we did not find a spatial invasion pattern in the maps generated in our analysis. Neither of the alien taxa illustrates the pattern of species introduction and subsequent exponential spread via a front or corridor. This could indicate that the alien taxa were introduced prior to most collections in the OVPD or were introduced at multiple sites at approximately the same point in time. Delisle *et al*. (2003) and Pyšek (1991) found invasive riparian species dispersing along river corridors, but our maps of *Tamarix* gave little indication that it was spreading up or down riparian zones. We believe that *Tamarix* is almost certainly spreading along rivers in Oklahoma (DiTomaso, 1998); however, our data are not sufficiently sensitive, either temporally or geographically, to map the pattern. Baker (1974) described the typical North American invasion pattern to be scattered populations expanding to fill in absences between populations. Both the invasive alien and native expansive taxa in our study appear to follow this pattern.

The geographic distribution of specimens collected in Oklahoma is not random, but instead follows a pattern correlated to population centres and botanically "interesting" areas. More species have been collected in counties with institutes of higher education than in neighboring counties, though one would expect the flora to be similarly diverse (Hoagland *et al*., 2006). Researchers in Kansas identified population centres as one of the problematic biases (Woods *et al*., 2005) and Iverson & Prasad (1998) actually took into account the number of botanists residing in a county when they modelled the diversity of the Illinois flora. Locations of canyons, mountains, unique rock outcrops, and other topographically outstanding elements have lured botanists to collect many specimens to document their distinctive flora. Counties with such features are overrepresented in the OVPD (Hoagland *et al*., 2006).

Other biases can be found in collections. Concentration on a particular group of plants will produce a taxonomic bias. Many systematists deposit their collection of a single genus or species in a herbarium. Being knowledgeable of the region's history can also be useful. For example, certain prairie species may be overrepresented if they are part of roadside plantings organized by the Department of Transportation. Small projects, such as these, maybe unknown and, alas, we cannot know all the nuances of bias in our data sets.

CONCLUSIONS

One could argue that too many uncontrolled variables in herbarium data sets cause inaccurate representations of the historical biogeography of taxa. Nonetheless, the techniques developed by other biogeographers to analyse patterns of species invasion and eliminate biases inherent in herbarium data have been successful, to a degree, in our research. We deliberately chose taxa that are known to have increased in abundance and to be invasive in Oklahoma. We found that the alien, invasive species demonstrate an invasion trend in both the regression model and proportion curve analyses. However, the native species that have been labelled "expansive" did not show a clear expansion pattern. The information found in herbaria, especially comparatively recently established herbaria, may not be sensitive enough to detect the increase of abundance of native species in response to human disturbance, for example. Yet, herbaria are important storehouses of phytogeographic data. Unfortunately they are threatened institutions; plant collecting in the U.S. is in decline (Prather *et al*., 2004), a trend confounded by a reduced interest in plant taxonomy (Wortley *et al*., 2002), and the elimination of herbaria at some universities in recent years. Herbaria represent many decades of plant collecting, thousands of miles travelled, and countless man-hours of identification. We hope research such as ours will encourage others to take advantage of information gathered by the scores of botanists before us and to design novel techniques and new avenues of research utilizing herbarium records.

ACKNOWLEDGEMENTS

The authors thank the herbarium curators who granted access to their collections so that they could be entered into the OVPD. We also thank all the undergraduate and graduate students who entered the more than 210,000 records into the OVPD. Dustin Woods and Katy Levings were responsible for georeferencing many of the records. Michael Patten provided important advice on statistical analysis. We are grateful to Petr Pyšek and an anonymous reviewer for useful comments that improved the manuscript. This work was partially funded by NSF grant #0237433.

LITERATURE CITED

Baker, H.G. (1974) Evolution of weeds. *Annual Review of Ecology and Systematics*, **5**, 1-24.

Barney, J.N. (2006) North American history of two invasive plant species: phytogeographic distribution, dispersal vectors, and multiple introductions. *Biological Invasions*, **8**, 703-717.

Baum, B.R. (1967) Introduced and naturalized tamarisks in the United States and Canada. *Baileya*, **15**, 19-25.

Chauvel, B., Dessaint, F., Cardinal-Legrand, C. & Bretagnolle, F. (2006) The historical spread of *Ambrosia artemisiifolia* L. in France from herbarium records. *Journal of Biogeography*, **33**, 665-673.

Coppedge, B.R., Engle, D.M., Fuhlendorf, S.D., Masters, R.E. & Gregory, M.S. (2002) Landscape cover type and pattern dynamics in fragmented southern Great Plains grasslands, USA. *Landscape Ecology*, **16**, 677-690.

Delisle, F., Lavoie, C., Martin, J. & Lachance, D. (2003) Reconstructing the spread of invasive plants: taking into consideration biases associated with herbarium specimens. *Journal of Biogeography*, **30**, 1033-1042.

DiTomaso, J.M. (1998) Impact, biology, and ecology of saltcedar (*Tamarix* spp.) in the southwestern United States. *Weed Technology*, **12**, 326-336.

Friedman, J.M., Merigliano, M.F., Freehling, M.D., Griffin, E.R., Auble, G.T., Shafroth, P.B., & Scott, M.L. (2005) Dominance of non-native riparian trees in western USA. *Biological Invasions*, **7**, 747-751.

Fuentes, N., Ugarte, E., Kühn, I., & Klotz S. (2008) Alien plants in Chile: inferring invasion periods from herbarium records. *Biological Invasions*, **10**, 649-657.

Hoagland, B.W. (2006) Township and range survey system. *Historical atlas of Oklahoma* (ed. by C.R. Goins and D. Goble), pp. 114-115. University of Oklahoma Press, Norman.

Hoagland, B.W., Buthod, A.K., Butler, I.H., Crawford, P.H.C., Elisens, W.J., & Tyrl, R.J. (2006) *Oklahoma Vascular Plants Database*. Oklahoma Biological Survey, University of Oklahoma, Norman, OK. Available at: http://www.biosurvey.ou.edu/atlasdesc.html (last assessed 1 May 2006).

Holmgren, P.K. & Holmgren, N.H. (2006) *Index Herbariorum*. New York Botanical Garden. Available at: http://sciweb.nybg.org/science2/IndexHerbariorum.asp (last accessed 1 June 2006).

Iverson, L.R. & Prasad, A. (1998) Estimating regional plant biodiversity with GIS modeling. *Diversity and Distributions*, **4**, 49-61.

Mihulka, S. & Pyšek, P. (2001) Invasion history of *Oenothera* congeners in Europe: a comparative study of spreading rates in the last 200 years. *Journal of Biogeography*, **28**, 597-609.

Prather, L.A., Alvarez-Fuentes, O., Mayfield, M.H. & Ferguson, C.J. (2004) The decline of plant collecting in the United States: a threat to the infrastructure of biodiversity studies. *Systematic Botany*, **29**, 15-28.

Pyšek, P. (1991) *Heracleum mantegassianum* in the Czech Republic: dynamics of spreading from the historical perspective. *Folia Geobotanica*, **26**, 439-454.

Pyšek, P. & Pracht, K. (1993) Plant invasions and the role of riparian habitats: a comparison of four species alien to central Europe. *Journal of Biogeography*, **20**, 413-420.

Pyšek, P., Sádlo, J., Mandák, B. & Jarošík, V. (2003) Czech alien flora and the historical pattern of its formation: what came first to Central Europe? *Oecologia*, **135**, 122-130.

Richardson, D.M., Holmes, P.M., Esler, K.J., Galatowitsch, S.M., Stromberg, J.C., Kirkman, S.P., Pyšek, P. & Hobbs, R.J. (2007) Riparian vegetation: degradation, alien plant invasions, and restoration prospects. *Diversity and Distributions*, **13**, 126-139.

Sokal, R.R. & Rohlf, F.J. (1995) *Biometry*, 3rd edn. W.H. Freeman & Co., New York.

Southern Weed Science Society (1998) *Weeds of the United States and Canada*. CD-ROM. Southern Weed Science Society, Champaign, IL.

Stadler, J., Mungai, G. & Brandl, R. (1998) Weed invasion in East Africa: insights from herbarium records. *African Journal of Ecology*, **36**, 15-22.

Stubbendiek, J., Frisoe, G.Y. & Bolick, M.R. (1994) *Weeds of Nebraska and the Great Plains*. Nebraska Department of Agriculture, Bureau of Plant Industry, Lincoln.

Tyrl, R.J., Barber, S.C., Buck, P., Elisens, W.J., Estes, J.R., Folley, P., Magrath, L.K., Murray, C. L., Smith, B. A., Taylor, C.E. & Thompson, R.A (2006) *Keys and descriptions for the vascular plants of Oklahoma*. Flora Oklahoma Inc., Noble.

USDA, ARS (1970) *Selected weeds of the United States*. Agricultural Handbook No. 366. U.S. Government Printing Office, Washington, D.C.

USDA, NRCS (2006) *The PLANTS Database*. National Plant Data Center, Baton Rouge, LA. Available at: http://plants.usda.gov (last accessed 15 May 2006).

Woods, T.M., Strakosh, S.C., Nepal, M.P., Chakrabarti, S., Simpson, N.B., Mayfield, M.H. & Ferguson, C.J. (2005) Introduced species in Kansas: floristic changes and patterns of collection based on an historical herbarium. *SIDA*, **21**, 1695-1725.

Wortley, A.H., Bennet, J.R. & Scotland, R.W. (2002) Taxonomy and phylogeny reconstruction: two distinct research agenda in systematics. *Edinburgh Journal of Botany*, **59**, 335-349.

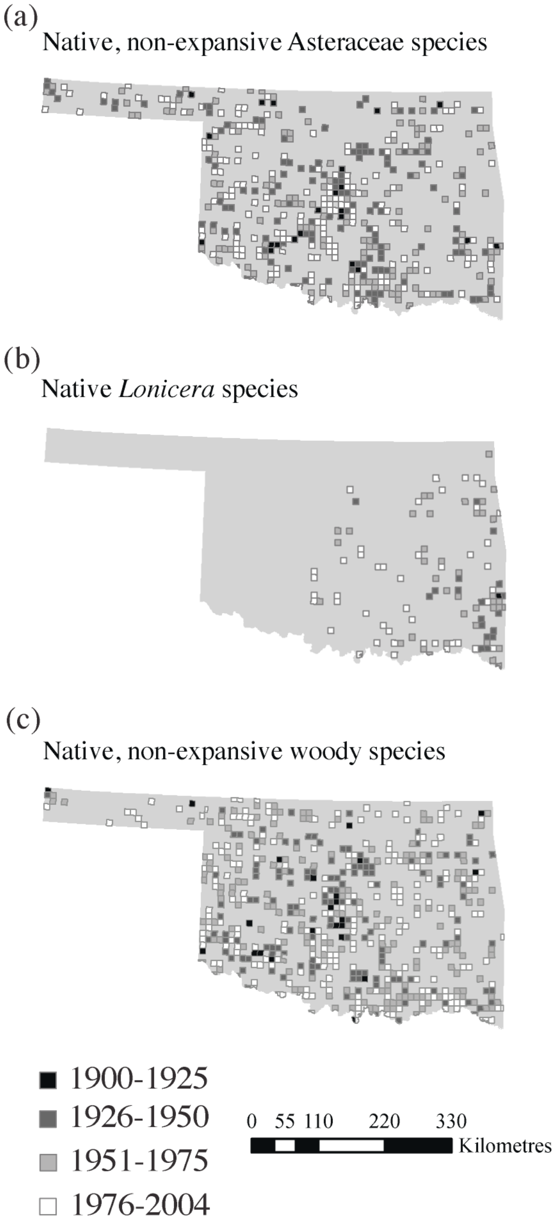
Wu, S., Rejmánek, M., Grotkopp, E. & DiTomaso, J.M. (2005) Herbarium records, actual distribution, and critical attributes of invasive plants: genus *Crotalaria* in Taiwan. *Taxon*, **54**, 133-138.

Table 1. The number of townships in Oklahoma, USA occupied by select alien, invasive taxa[1], native, expansive species[2], and native, non-expansive counterpart groups[3]. The total number of townships in Oklahoma is 2098. Specimens were recorded in the Oklahoma Vascular Plants Database (OVPD), the repository for the plant collecting data of the state of Oklahoma. * Specimens were removed from analysis if they could not be georeferenced, were missing collection year, were cultivated, or were duplicate collections.

	Total number of specimens in OVPD	Number of specimens used in analysis*	Number of townships in which taxa were found	Year of first collection
Ambrosia psilostachya[2]	240	201	140	1913
Amphiachyris dracunculoides [2]	277	236	168	1913
Juniperus virginiana [2]	603	466	236	1913
Lonicera japonica[1]	121	103	75	1936
Tamarix species[1]	398	297	178	1910
Native, non-expansive Asteraceae[3]	1002	859	463	1903
Native *Lonicera* species [3]	283	231	103	1913
Native, noninvasive woody[3]	1201	1003	555	1906

Figure 1. The spatial and temporal collection history of select native, non-expansive groups in Oklahoma, USA. Occupied townships (9.66 x 9.66 km) are shaded based on the time period during which the first collection of that taxon was made. Darker townships are the locations of the earliest plant records.

Figure 1



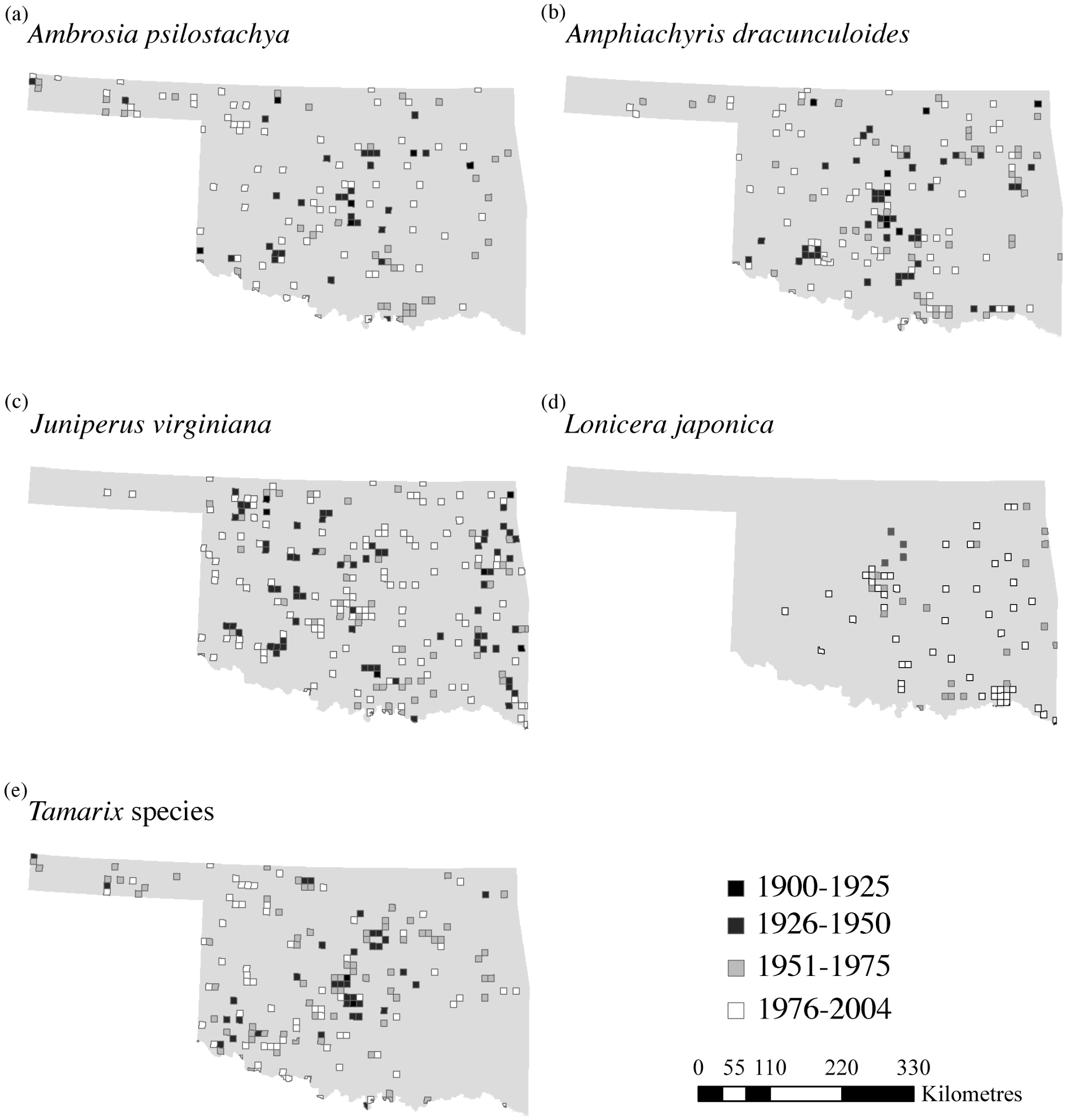
(a) Native, non-expansive Asteraceae species

(b) Native *Lonicera* species

(c) Native, non-expansive woody species

Figure 2. The spatial and temporal collection history of select alien, invasive and native, expansive taxa in Oklahoma, USA. Occupied townships (9.66 x 9.66 km) are shaded based on the time period during which the first collection of that taxon was made. Darker townships are the locations of the earliest plant records.

Figure 2



(a) *Ambrosia psilostachya*

(b) *Amphiachyris dracunculoides*

(c) *Juniperus virginiana*

(d) *Lonicera japonica*

(e) *Tamarix* species

Figure 3. Invasion and expansion curves generated for select invasive and expansive taxa compared to the general collection trend of the native, non-expansive counterpart group. The slope of the linear regression represented collection trend over time for the non-expansive counterpart group. However, for the alien, invasive taxa and native, expansive species the regression slope represented the collection effort ***plus*** invasion or expansion rate. All linear regressions were statistically significant ($P < 0.001$). Regression pairs with * indicate slopes that differ significantly from each other ($P < 0.01$).

Figure 3

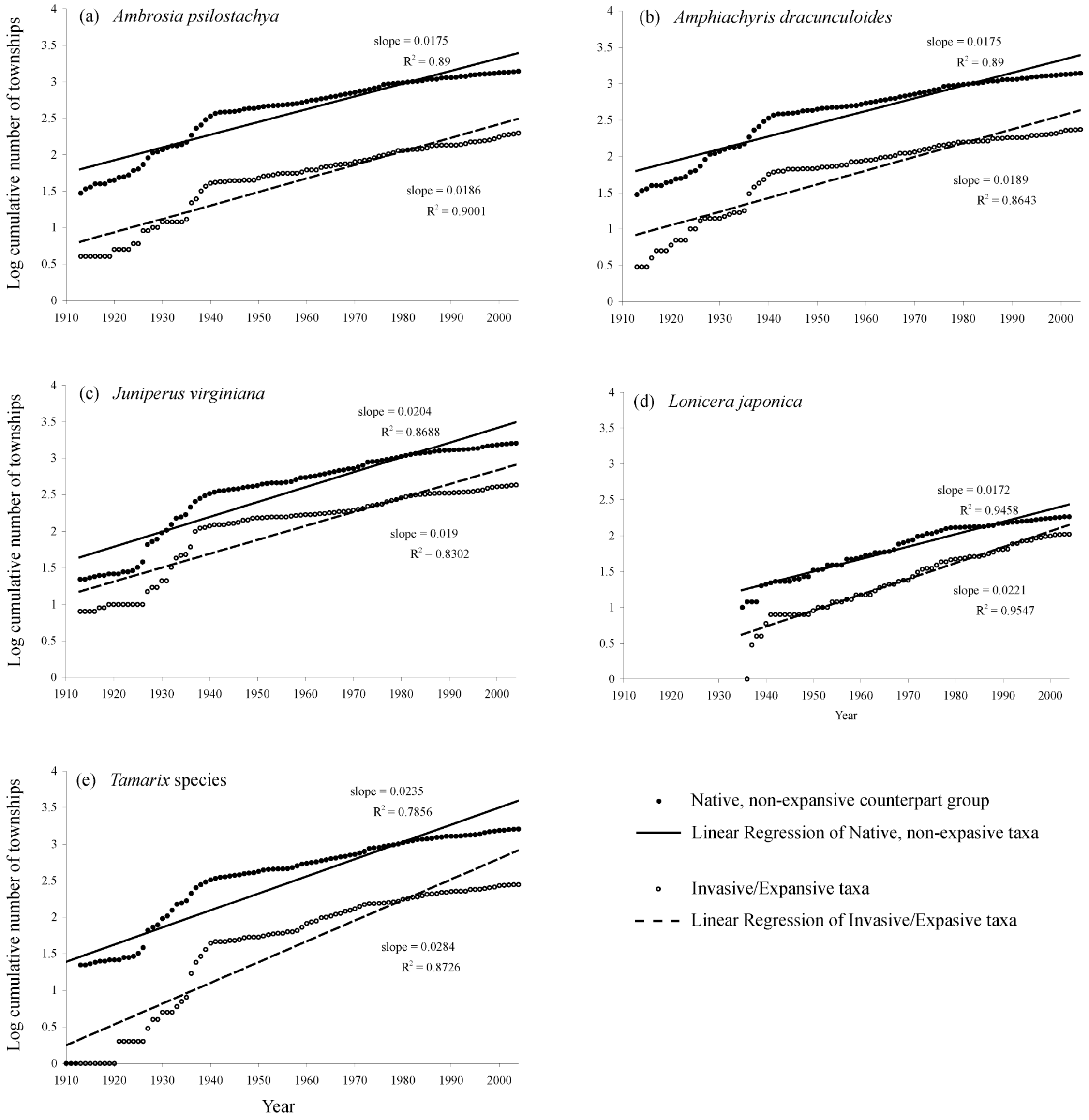
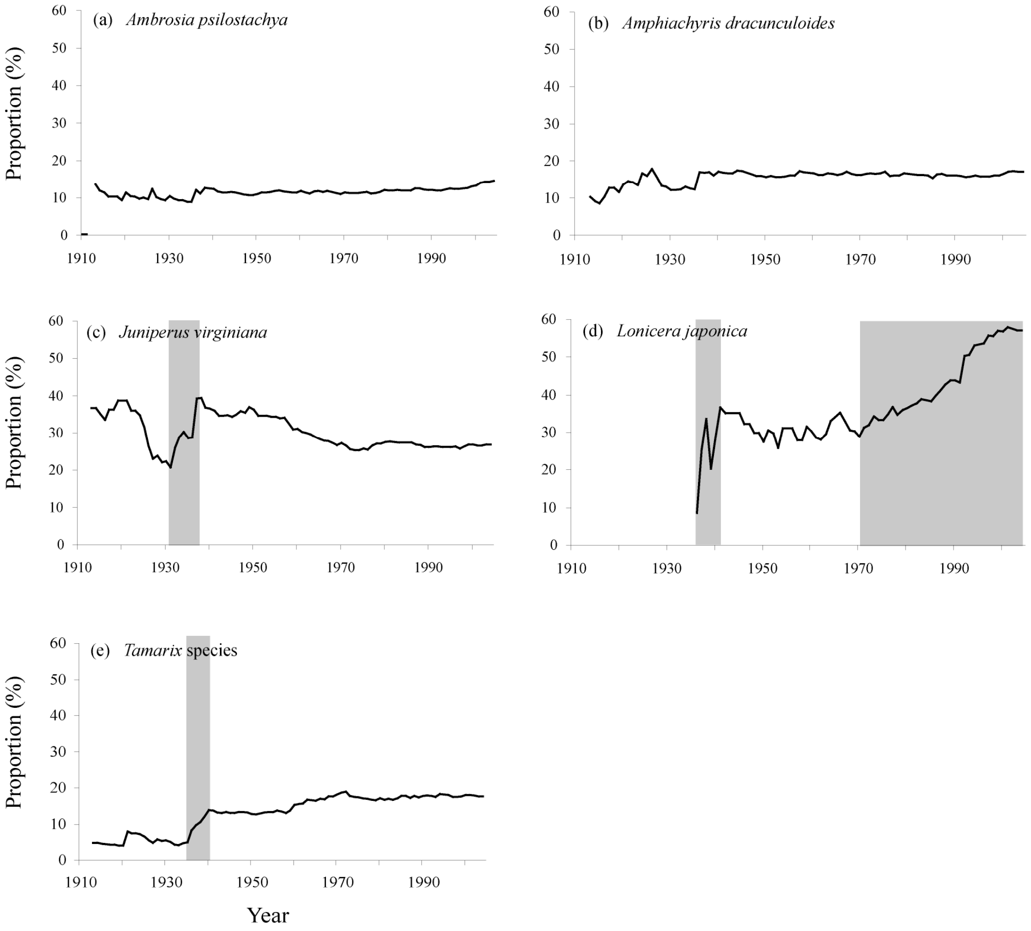

(a) *Ambrosia psilostachya*
slope = 0.0175
R^2 = 0.89
slope = 0.0186
R^2 = 0.9001
(b) *Amphiachyris dracunculoides*
slope = 0.0175
R^2 = 0.89
slope = 0.0189
R^2 = 0.8643
(c) *Juniperus virginiana*
slope = 0.0204
R^2 = 0.8688
slope = 0.019
R^2 = 0.8302
(d) *Lonicera japonica*
slope = 0.0172
R^2 = 0.9458
slope = 0.0221
R^2 = 0.9547
(e) *Tamarix* species
slope = 0.0235
R^2 = 0.7856
slope = 0.0284
R^2 = 0.8726
Log cumulative number of townships
0
0.5
1
1.5
2
2.5
3
3.5
4
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
Year
● Native, non-expansive counterpart group
— Linear Regression of Native, non-expasive taxa
○ Invasive/Expansive taxa
– – Linear Regression of Invasive/Expasive taxa

Figure 4. Proportion curves were calculated by dividing the cumulative number of townships in Oklahoma for each alien, invasive or native, expansive species by the cumulative number of townships occupied by the native, non-expansive counterpart group. Periods of increase, indicated by the shading, occur when the invasive or expansive taxa was collected more often than would be expected from the general collecting trend.

Figure 4


(a) *Ambrosia psilostachya*
(b) *Amphiachyris dracunculoides*
(c) *Juniperus virginiana*
(d) *Lonicera japonica*
(e) *Tamarix* species
Proportion (%)
0
10
20
30
40
50
60
1910
1930
1950
1970
1990
Year

CHAPTER 2

The use of species distribution models to answer ecological and biogeographic questions: a review of the literature

INTRODUCTION

Species distribution models (SDM) have become important tools for ecologists, biogeographers, conservation biologists, and restoration ecologists. While much of the SDM literature focuses on testing existing techniques (for examples see: (Elith and others 2006; Fielding and Bell 1997; Kadmon and others 2003; Meynard and Quinn 2007; Muñoz and Felicísimo 2004; Pearce and Boyce 2006; Segurado and Araújo 2004; Stockwell and Peterson 2002)), other researchers are using these tools for hypothesis generation or adding them to the suite of tools for conservation decision-makers. Like other multivariate statistical analyses, SDMs attempt to reduce the number of potential variables in a data set to determine those that best explain a species' distribution. Therefore, SDMs help to understand and possibly quantify the ecological requirements of a species (Box and others 1993; Costa and others 2007; Danks and Klein 2002; De'ath 2002; De'ath and Fabricius 2000; Laurent and others 2004; Murphy and Lovett-Doust 2007; Norris and others 2006). It has been argued that SDMs, in fact, model the niche of the species (this will be discussed further in the next section). However, a species' distribution is not simply a result of the physical environment matching the ecological requirements of a species. Evolutionary and historical factors also determine a species distribution and SDMs may illuminate the geographic or historical features that limit a species' modern distribution (Anderson 2003; Camarero and others 2005; Van Mannen and others 2002). If SDM results, based on ecological data, suggest a much wider distribution, what might be causing the limited distributions? Further hypothesis testing may lead to an understanding of the dispersal barriers or historical situation that created the current, seemingly limited, distribution.

In this literature review, I cover the essential topics associated with SDMs. First, I discuss the niche concept and the variety of theoretical interpretations of model output. I then consider issues associated with data, both species occurrence data and environmental data, that are generally used to build SDMs. Of course, there are a multitude of modelling techniques, a few of which I briefly describe and compare. Model comparison can be performed using a variety of methods that I summarize. Finally, I discuss the current challenges facing modelers and outline some potential improvements to this field of inquiry.

Species distribution modelling has proven useful for locating populations of rare, endangered, or even undiscovered species (Pearson and others 2007; Peppler-Lisbach and Schräder 2004). Although not widely published in the scientific literature, many biologists associated with state agencies are using SDMs to find populations of rare or endangered taxa and plant communities (Fertig and others 1998). For example, Natural Heritage Programs, which maintain spatial data of the occurrence of rare and endangered species at state and regional scales, have begun to apply SDMs for locating populations of rare species. The Wyoming Natural Diversity Database has successfully used SDMs to locate several new populations of pygmy rabbit in areas where experienced biologists did not expect to find the species or seriously consider as suitable locations (Beauvais and others 2004). The Oregon Natural Heritage Information Center biologists found nine new locations of grassy balds, a rare plant community, with information from only 35 original locations (Buechling and Tobalske 2007) allowing them to make more informed conservation recommendations. In relatively little known areas, researchers are using SDMs of related taxa to find rare and even undiscovered species.   Surrogate taxa are modelled with the expectation that similar species will have similar ecological requirements (Römermann and others 2007). New reptile species have been found in Madagascar using this modelling approach (Raxworthy and others 2003).

Locating new populations of rare species is just one conservation application of SDMs. Models have been used to help identify sites of high potential biotic diversity (Cowling and Samways 1994; Ferrier and Guisan 2006; Iverson and Prasad 1998a; Lira-Noriega and others 2007; Ortega-Huerta and Peterson 2004; ter Steege and others 2003). These model results can help to identify sites for land conservation or nature reserve systems (Danks and Klein 2002; Ortega-Huerta and Peterson 2004; Rodríguez and others 2007; Tole 2006). Making important conservation decisions based on species distribution models must be done with caution. Size of data set, bias in the data, and gaps in data coverage will affect the quality of the results (Hopkins 2007; Loiselle and others 2007; Stockwell and Peterson 2002; ter Steege and others 2003; Vaughan and Ormerod 2003). Model choice and even how the model is evaluated will determine the type and degree of error significantly affect results (Loiselle and others 2003; Pearson and others 2006).

Conservation biologists also are using SDMs to identify specific locations that are best suited for species reintroduction or translocation (Carey and Brown 1994). In chapter 3 of this dissertation, I use SDMs to create a map of habitat suitability for the American burying beetle. I expect these results will not only contribute to a better understanding of the ecological requirements and species distribution, but also be used by the U.S. Fish and Wildlife Service to determine the best locations for beetle translocation from road or pipeline construction sites.

Alien species invasion can also be explored with SDMs. When species are introduced into a new region by humans they are transported over evolutionary and biogeographic time and space. Having overcome dispersal barriers, these species attempt to carve out an ecological niche in a new region. Many alien species are currently in the process of invading a new region and have not reached their full potential (Peterson and others 2003; Peterson and Vieglais 2001; Welk and others 2002). By projecting the ecological requirements from the home range on the newly invaded region, the models can predict the potential extent of invasion in the new region (Anderson and others 2006; Collingham and others 2000; Hulme 2003; Peterson 2003; Peterson and Nakazawa 2007; Peterson and others 2003; Peterson and Vieglais 2001; Robertson and others 2001; Welk and others 2002; Zhu and others 2007). In chapter 4 of this dissertation, I attempt to model invasive species distribution using native and introduced range data. SDMs may also be able to predict what species are likely to become invasive before they have ever been introduced to a new region (Nyári and others 2006). Because of the potential economic and ecological impact of alien species invasion, many researchers are exploring the use of SDMs to help in the fight against invasive species (Dark 2004; Lippitt and others 2008).

Finally, the hottest topic in an already fiery field is using SDMs to predict future suitable habitat in the face of global climate change. Researchers build models with species current distributions under current climate conditions, then alter climate data to reflect various climate change scenarios and project the resulting hypothesized distributions (Araújo and Luoto 2007; Araújo and Pearson 2005; Araújo and others 2005a; Araújo and Rahbek 2006; Carey and Brown 1994; Iverson and Prasad 1998b;

Oberhauser and Peterson 2003; Papes 2007; Pearson and Dawson 2003; Pearson and others 2006; Thuiller and others 2005a; Thuiller and others 2005b). There are many complicating factors that affect the results of these models. Like all models, they can be significantly affected by the model algorithm, model assumptions, parameterization of the model, and the geographic range of the data, but the added uncertainty of climate models confounds the errors in the prediction (Araújo and others 2005b; Davis and others 1998). In addition, because these models are predicting future distributions based on potential climate change, model evaluation becomes problematic. In their review of distribution models based on future global warming, Botkin and others (2007) found that few of the models were evaluated and none were able to validate the model with independent data. Validation with a truly independent data set may be impossible (Araújo and others 2005a). However, work continues to improve the model output and model forecasts (Araújo and Luoto 2007; Araújo and New 2007).

NICHE CONCEPTS

In the SDM literature one can find varying opinions on the terminology and the most appropriate definitions of model outcomes, but at their theoretical base, SDMs rely on the niche concept. In fact, modelers cannot even agree on what to call these models: species distribution models, potential habitat models, climate envelope models, or ecological niche models, for example. However, there is no clear, uniform definition for niche in the discipline of ecology. Researchers continue to argue over theoretical semantics in using “niche” to explain the output of correlative, descriptive models. Most authors in the SDM literature use, or at least imply, the basic niche definition put forth by

Hutchinson (1957) where the niche is a multidimensional space in which the many axes represent gradients of variables that limit an organism's or population's fitness (as cited in (Olding-Smee and others 2003)). The niche concept attempts to explain species abundance patterns along multiple environmental gradients. Hutchinson distinguished between the fundamental and realized niche. The fundamental niche represents the theoretical space occupied by a species in which the combination of all relevant environmental variables allow the species to survive and reproduce. However, the general interpretation of the fundamental niche does not include biotic interactions, in particular interspecific competition. Therefore, it was necessary to define the realized niche as a portion of the fundamental niche where the species is competitively dominant and can successfully reproduce.

Pulliam (2000) put forth several theoretical niche or distribution relationships. He proposed the following four possible scenarios:

1. Grinellian niche, or Hutchinson's fundamental niche — species will occur where the environmental variables are suitable.
2. Hutchinson's realized niche — a subset of the fundamental niche where it is limited by interspecific competition or other biotic interactions.
3. Source-sink dynamics — species may be found in locations that will not support reproduction, based on metapopulation theory (the study of the interactions of populations separated in geographic space).
4. Dispersal limitation — also related to metapopulation dynamics, suggests that species are absent from suitable habitat because of limitations in organism dispersal and the

time needed to establish a successful reproductive population in fitting habitat patches.

The niche or distribution of an individual species may be described with any of these theoretical frameworks. Knowledge of the environmental and physiological limitations of a species will improve the outcome of a distribution model because model choice and model parameters will have direct ecological meaning.

The source–sink scenario is one concept that is not often incorporated in distribution modelling. In general, it is not known if a record of presence in the data set is from a source or sink population. When recording species presence, especially a rare species, it is very difficult to know if you are collecting data from a self sustaining population. It is likely that data sets acquired from opportunistically collected records (records not collected as part of a methodical research study) contain observations of individuals from sink populations. Therefore, most models are built on data that do not represent the true niche, fundamental or realized, because data come from locations that may not allow for successful reproduction. Logically, the models constructed on these data should not be called “niche” models (Araújo and Guisan 2006; Kearney 2006).

Soberón and Peterson (2005) argue that the data set entered into the model is the spatial representation of the fundamental niche because the observations are correlated to abiotic variables and, therefore, they argue that these modelling techniques should be called niche models rather than distribution models. They contend that the distribution of a species is a “complex expression of its ecology and evolutionary history.” They assert that the modelling algorithms produce an estimate of the fundamental niche, which is more imprecise than a species distribution. The true distribution of a species, they reason, includes in its concept the limitations of the species due to dispersal/migration and the evolutionary capacity of a population to adapt to new environments. Both Kearney (2006) and Guisan and Thuiller (2005) strongly disagree with Soberón and Peterson (2005) and argue that correlative distribution models do not represent the fundamental niche of an organism. They hold that Soberón and Peterson (2005) are not taking into account that the observational data are already constrained by biotic factors. They assert that only mechanistic models based on direct measurements of physiology or behavior can produce the fundamental niche, any use of observational data are, in effect, reflecting the realized niche.

Araújo and Guisan (2006) want to dismiss any use of the formal definition of niche with respect to distribution modelling. They suggest that ignoring biotic interactions when defining the fundamental niche is incorrect. They believe that even Hutchinson recognized that positive biotic interactions influenced the fundamental niche. They quote Hutchinson's (1957) concluding remarks to support this argument:

> … all variables, both physical and biological, being considered, the fundamental niche of any species will completely define its ecological properties.
> …Interaction of any of the considered species [defining the realized niche] is regarded as competitive…All species other than those under consideration are regarded as part of the coordinate system.

They interpret Hutchinson's statements to mean that biotic interactions other than competition, such as pollination or parasitic relationships, should be included in the multidimensional space that defines the fundamental niche. They support Leibold's (1995) updated niche definition which combines Hutchinson's realized niche concept, but also adds the impact of organisms on their environment.

Not only are organisms affected by the environment, but the organisms themselves also can modify their environment. This concept is not routinely considered in distribution modelling. Elton (1939), MacArthur (1967), and Leibold (1995) support the resource-consumer or trophic level niche ideas that place an emphasis on the organism's role or function within the environment. Laland and others (1999) extend the concept further by describing how the evolutionary process can be affected by "niche construction." Niche construction occurs when organisms reshape both the abiotic and biotic relationships that determine their niche. This modification causes feedbacks that alters the pressures of natural selection and consequently the dynamics of the evolutionary process. While theoretically compelling, the niche construction hypothesis is too complex to be integrated into today's distribution models. Also, the scale at which an organism changes the environment is usually very fine, which may excuse this concept from applications in very large scale (continental or regional) distribution modelling.

Both Kearney (2006) and Araújo and Guisan (2006) endorse the term "habitat" to describe output of the SDMs. Kearney maintains that the term niche implies that we understand and take into account the behavioral, morphological, and physiological properties of a species. He believes it is more appropriate to characterize the output of correlative models, which do not imply cause and effect, as potential habitat maps thus emphasizing the descriptive nature of these modelling techniques and discouraging possible misuse. He also advocates reserving the term "niche" for situations that truly describe the direct effect of the environment on fitness or potential for reproduction of a population, such as in the mechanistic models. Araújo and Guisan (2006) also recommend using the phrase "potential geographic distribution" with modelling techniques that have a definite spatial aspect and model results and output is projected in a map format.

A sophisticated argument on the ecological theory behind SDM may not necessarily be important, however. What is important is trying to decide what occurrence data really represent. That is where biological expertise comes into action. The algorithm is used to find areas that are most similar to the occurrence data. The environmental data used to predict the distribution will also help to determine if the model represents the realized or fundamental niche or simply suitable climatic conditions. However, the use of niche seems to imply that the observational data represent occurrence data from individuals within their realized niche where they can successfully reproduce, which may not always be the case (i.e. sink populations or individuals caught during dispersal). The arguments for using the terms “habitat” or “distribution” modelling have won over this author. For the purposes of this dissertation and my subsequent research depending on the context, I will continue to use the phrase “distribution modelling” or suitable “habitat.”

MODEL TRAINING DATA

Bias and completeness

Species observations, or training data, are the most important component of SDM. Without a high quality data set of sufficient size and scope, you can expect great error in model output. The greatest difficulties with SDMs result from assumptions associated with the model techniques and the actual characteristics of the observational data set. Parametric procedures require a random sample of unbiased, independent observations.

However, these features are rarely found in data sets used in distribution modelling, because almost none of the data have been collected for the purpose of spatial modelling. Thus the methods used to analyze the data must take into consideration the unfavorable characteristics of the data, such as geographic bias or uneven sampling effort. Bias is represented in data sets in many ways depending upon the specific methods of data gathering (Barry and Elith 2006).

Several groups of researchers have tested biased data in models to explore the effect each type of bias has on model output. Kadmon and colleagues (2004) were concerned that roadside bias in observational data sets of woody plants would affect the results of models relating species to climate variables. They found, for Israel at least, that roadside observations and the road network did not have a climatic correlation. However, Canadian breeding bird survey data, which are based on road transects, were significantly biased with most points occurring in the south, thus over-representing warmer climates (Phillips and others 2009). Some roadside observations are very likely to have some environmental bias beyond climatic related variables, such as disproportionately representing disturbed or fencerow habitat. Models built with predictor variables other than climatic are likely to be affected by roadside bias. Loiselle and colleagues (2007) were concerned that underlying climate bias in herbarium data may be influencing model predictions. In the Neotropics, they found an increase in the number of plant collections within specific ranges of several climate gradients. Fortunately, they found that this bias did not greatly impact the model output. Instead, they found the greatest factor in model performance to be the number of observations in the training data. Hortal and others (2007) found that large databases for well sampled areas still have gaps and biases that affect model performance. Their work concerned the diversity of seed-plants in Tenerif, Canary Islands. They recommend assessing the completeness of a database with respect to the environmental variables used in the model building.

The principal source of bias in the training data is that observations are not spread evenly across the environment gradients on which the predictions are based, but it has been shown that stratifying the samples can improve model performance (Barry and Elith 2006; Vaughan and Ormerod 2003). If a data set is biased to one end of an environmental gradient, then this may lead to spurious relationships between prediction and response variables. To counter this affect, Araújo and Guisan (2006) suggest that subsampling observations may improve the quality of the data set. They suggest reducing identified bias by removing selected observations in the over-represented environmental space. However, this can result in a reduction of data points, which are highly valuable in model building. They also recommend additional stratified sampling based on the areas that are not well represented in the observation data set (Araújo and Guisan 2006).

Sample size, the number of observations used to train the model, appears to be the most important factor, after data accuracy, in model performance. All modelling techniques benefit from additional training data and suffer when training data are limited (Hernandez and others 2006; Loiselle and others 2007; Stockwell and Peterson 2002). The size of the data set necessary to accurately model a species distribution will be relative to the complexity of the species-environment relationship (Barry and Elith 2006). Researchers have had the greatest success modelling species with specialized or specific ecological requirements (Brotons and others 2004). Hernandez and colleagues (2006) were able to produce useful models with as few as 10 positive observations of a wide variety of animal species with specialized ecological requirements. Success with such a small data set could be a result of the researchers understanding the species' ecological requirements and a relatively simple relationship between the species and the environment. In generalist species, where the species can tolerate a wide range of environmental gradients in a variety of combinations, models are more mathematically complex and may not perform well (Brotons and others 2004). The smaller the training data set, the fewer the number of predictor variables that can used in the model building (Burnham and Anderson 2002).

Small data sets are typical of rare or poorly known species and in areas that have not been intensively surveyed. One purpose of modelling these species is to identify areas of potential habitat to focus further research. Surrogate taxa can be used to model the potential distribution of species that have very few recorded observations (Rushton and others 2004). Also, a survey of landowners for the presence of a conspicuous species may stand in for traditional occurrence data in areas that have had few inventories by biologists (Vaughan and Ormerod 2003).

Pseudo-absences

One of the short comings of both natural history collections and bird survey data is the lack of reliable absence data. Absence data are necessary for many of the older modelling techniques. Because many observation data sets are lacking absence records, pseudo-absences are generated and used in model building. Pseudo-absences, also known as background data, often are placed at random onto the study area (Stockwell and Peters 1999). However, an alternative to simply using randomly generated pseudo-absences is to limit their placement to areas where they are unlikely to be found. Chefaoui and Lobo (2008), working with a threatened, endemic moth, used presence-only modelling techniques to identify unsuitable habitats in which to focus the generation of pseudo-absences. They found that Generalized Linear Models performed better with expertly selected pseudo-absences than with the randomly chosen pseudo-absences. Lütolf and colleagues (2006) tested several approaches to generate pseudo-absence data for three butterfly species. They tried placing pseudo-absences in areas that had no observations of the model species or no records for species with similar habitat preferences generated the best models. Both techniques relied on preliminary model building from which to decide pseudo-absence locations. This may confound the subsequent models because both models take advantage of the same training data set. Also, using non-random pseudo-absences may over-fit the model to the training data, which will increase accuracy with the training data, but reduce transferability to independent data (Chefaoui and Lobo 2008). Another strategy they tested was to assume that areas with high numbers of butterfly records have been relatively thoroughly searched for butterfly species. Therefore, if there was no record for the butterfly species modelled in highly surveyed areas, then it is likely absent. Surprisingly, this hypothesis was not supported. Models made with pseudo-absence data based on this hypothesis performed poorly, in fact more poorly than models built on randomly generated pseudo-absence data (Lütolf and others 2006).

Old data are routinely removed from spatial analysis because it is assumed that they may represent environmental conditions that are no longer present and/or reflect past distribution patterns prior to the influence of anthropogenic land cover and land use and climate change (Raxworthy and others 2003). Also, older data tend not be accurate in location description (Rowe 2005). Yet, results from Lütolf and colleagues (2006) indicate that older data may improve model predictions. They found that when 100 year-old data were removed, a model's ability to predict present day occurrences significantly decreased.

Spatial Autocorrelation

Spatial autocorrelation exists when the value of a data point is more or less similar to the values of nearby data points than would be expected from a random distribution (Legendre 1993). Spatial autocorrelation is an assessment of the relationship of a variable to its spatial location. Spatial autocorrelation can be positive or negative; positive when points with similar values appear together spatially and negative when values are dissimilar (Legendre 1993). Generally, we encounter positive spatial autocorrelation in ecological data. Most ecological data have a spatial structure and the distribution of a species is neither uniform nor random; and the same can be said for environmental data (Henebry and Merchant 2002). The spatial patterns most often seen are patches or gradients. These patterns are often generated by multiple environmental and ecological factors.

Data that exhibit spatial autocorrelation should not be used in classical statistical tests because the data points are not independent observations (Beale and others 2007;

Legendre 1993; Legendre and others 2002; Lennon 2000). Most statistical tests are based on the assumption that data points, or observations, are independent (Gotelli and Ellison 2004). However, we can no longer make that assumption when the values of neighboring data points are interrelated. Lennon (2000) found that when spatial autocorrelation was not corrected the variables with high spatial autocorrelation were more likely to be "significant" in classical statistical tests. Spatial autocorrelation, a form of pseudoreplication, can lead to an overestimation of sample size and an inflation of statistical significance of correlations. For example, to improve model performance by increasing the sample size of a 10 point data set a researcher may collect 10 new points, each adjacent to one of the original 10. Although there are now 20 data points, they are not spatially independent. By using n = 20, the degrees of freedom will be overestimated, inaccurate p values will be calculated, and the standard errors of the correlation coefficients will be underestimated. This results in an increase in type I error, rejecting the null hypothesis, and assigning a false positive (Gotelli and Ellison 2004; Legendre and others 2002; Liebhold and Sharov 1998).

Spatial autocorrelation can be quantified by calculating Moran's *I*, which is based on the residuals of a regression analysis (Gotelli and Ellison 2004). The "*I* coefficient" compares the expected value and variance of spatially defined points and determines the number of pairs that have a spatial relationship. The values for Moran's *I* range from -1 to 1; values close to 1 indicate positive spatial autocorrelation and negative values a negative spatial autocorrelation. A value not statistically different from 0 means there is no spatial autocorrelation (Liebhold and Gurevitch 2002; Liebhold and Sharov 1998).

Once the degree of spatial autocorrelation has been determined, one must decide how to correct it. One simple, but imperfect, technique is to remove data to increase the separation distance of clustered points (Guisan and others 2006). For this method, points are assigned a buffer based on the species' biology and autecology that often represents an individual's home range. Buffers that overlap could be considered observations of the same individual. Data points are then removed until there are no overlapping buffers. This usually reduces spatial autocorrelation but also discards potentially valuable information. Accurate ecological information can be costly to obtain and discarding it could be considered wasteful.

The use of spatial autoregressive models can help eliminate the spatial autocorrelation effect within the data (Carl and Kühn 2007; Dark 2004; Lichstein and others 2002; Segurado and Araújo 2004). The use of spatially explicit models is more advantageous than throwing away data and will generate fewer errors in spatially autocorrelated data sets than classical statistical techniques such as regression models. Normally distributed data perform well in autoregressive models, but much of the occurrence data are presence-absence and not abundance, therefore a binary distribution. Carl and Kühn (2007) were able to remove spatial autocorrelation affects found in binary (presence-absence) data by using the generalized estimating equation model, a lesser-known method. Classification tree analyses appear to perform better with spatially autocorrelated data as well. Segurado and others (2006) and Cablk and others (2002) tested the effect of spatial autocorrelation in distribution models, and found in spite of autocorrelation in the original data, classification trees accurately modelled correlative relationships between species richness and several environmental variables. Legendre and others (2002) employed Dutilleul's modified t-test, which corrects for variance of the test statistic and degrees of freedom in response to spatial autocorrelation, and found it to effectively correct for spatial autocorrelation.

Hawkins and others (2007) suggest that spatial autocorrelation may not be a factor in analyses of very large scale data. They found that statistical analysis was not affected by the spatial autocorrelation of gridded data on a continent scale. Using ordinary least squares regression (OLS), they tested the assumption that spatial autocorrelation would significantly affect the OLS coefficients of data taken from 110 x 110 km cells across several continents. Moran's *I* indicated spatial autocorrelation at relatively short-distances (approximately 750-1500 km) given the geographic distances sampled (4500-9000 km). However, the OLS coefficients appeared to be unaffected by spatial autocorrelation (Hawkins and others 2007).

We expect environmental data to have spatial autocorrelation. Although this poses potential statistical difficulties, we can also use it as an opportunity to identify the significance and understand the basis of the spatial patterns of the data.

PREDICTOR DATA

Environmental predictor variables fall into two major groups: indirect and direct. Direct variables are elements of the environment that directly affect the distribution of a species. Direct variables often have a physiological influence on the species (Austin 2002). For plants, direct variables would include: soil nutrients, solar radiation, precipitation, and days under 0°C. For animals, some examples of direct variables are: nesting sites, host plants, water temperature, and vegetation height. Indirect variables do not have a physiological affect, instead they are correlated to an environmental factor that directly impacts the species. For instance, altitude, longitude, or mean annual temperature do not directly limit species distribution, but instead it is their correlation with night time temperature, precipitation gradient, or evapotranspiration (respectively) that is the direct cause (Austin 2002; Korner 2007; Vaughan and Ormerod 2003). When direct variables are not easily measured, indirect variables are used as surrogates and integrated in the model.

Although predictor variables must contain some amount of error, few researchers acknowledge error and attempt to correct for it (Barry and Elith 2006). Error, or inaccuracy, can be a product of the nature of the data layer. For example, ecotones between the vegetation types are rarely classified. Ecotones blur the lines between vegetation types and create fuzzy boundaries. However, much vegetation classification data were originally digitized into distinct categorical polygons. The blurred line representing the transition from forest to grassland is not easily represented in the GIS.

Transferability of the model will be compromised if error in environmental data influenced the original model building, because error in environmental variables may have a greater effect when applied to a new area. A new area to which the model is applied may not have the same degree of error and the model will perform poorly in the new situation (Barry and Elith 2006). Rowe (2005) found that the quality of the georeferencing of historic specimens can significantly affect the attribution of the environmental data. The accuracy of recently georeferenced records is quite good due to the widespread use of GPS units. However, many natural history collection records must be assigned coordinates based on textual descriptions of the location found on the specimen label. Rowe (2005) calculated an accuracy buffer for all occurrence points based on the specificity of the location description. She plotted the points and buffers on a digital elevation model (DEM) to determine the potential inaccuracy that is transferred during elevation attribution. She found over 50% of the mammal collections in Utah could have elevation errors of over 400 m due to the lack of precision in georeferenced specimens.

The quality and accuracy of the DEM itself is an important factor in species distribution models (Barry and Elith 2006). While DEMs, at large scales, are quite accurate, they may be inaccurate at local scales. The errors within a DEM can be attributed to several causes. In particular, interpolation of digitized contour lines from topographic maps introduces error into the DEM which compounds errors inherent in the original data source and the digitizing process itself (Barry and Elith 2006). Because many environmental variables — such as slope, aspect, and elevation — are derived from DEM, it is important to understand how error in the DEM will propagate error in the derived variables. Van Niel and colleagues (2004) wanted to determine to what extent error in a DEM propagates error in secondary and tertiary derived variables. Logically, one would assume that error would increase with the level of derivation, but Van Niel and colleagues (2004) did not find this to be the case. Secondary variables, slope and aspect, had lower levels of accuracy compared to the tertiary variable solar radiation. Therefore, less derived does not necessarily mean less error.

Some of the original species distribution modelling techniques were solely based on climatic variables as predictors of distribution. Since then, researchers have moved beyond using simple environmental layers, such as climate, vegetation, and topography as predictor variables recognizing that interactions occur between environmental variables, though some of the modelling techniques neglect to incorporate them. Adding interaction coefficients into the model or creating environmental layers quantifying the interaction will make model interpretability and validation more complicated. Interaction terms in the algorithm greatly increases the number of parameters in the model (Guisan and Thuiller 2005). In addition, biotic interactions help to constrain species distributions and more researchers are trying to include the distribution or abundance of host, predator, or competitor species (Araújo and Luoto 2007; Davis and others 1998). Predictor variables that represent human influence also are being used in models, for example population density, airport density, and distance to roads (Kadmon and others 2004; Lippitt and others 2008). As a result, the use of remotely sensed data in species distribution models is increasing (Prates-Clark and others 2008). Satellite images are easier to acquire and are at high enough resolutions for fine ecological analysis. Many different environmental variables can be generated from satellite imagery — such as vegetation type and density, land cover and use, evergreen tree cover, or surface geology. Remotely sensed data are becoming easier to use given the computational power of current desktop computers and the availability of high resolution images (Lillesand and others 2004).

MODEL TYPES

Introduction

To relate known species occurrence data to the environment, many modelling methods have been developed and are currently in use. All modelling methods, to a great extent, possess similar characteristics fundamental to species distribution modelling. These characteristics are as follows:

1. Region under evaluation is represented in a GIS using raster format layers (grid cells).
2. Response variable is a data set made up of points of species observations — the values may be simply presence, presence–absence, or abundance.
3. Predictor variables are usually environmental layers in the GIS that, ideally, have an effect on the distribution of the response variable.
4. A function, which maybe simple or complex, is calculated to relate the response and predictor variables. This function will then classify each raster cell of the study region as suitable or unsuitable for the species.

The greatest variation among the modelling techniques is the type of function that determines the response–predictor relationship (Austin 2002). In this section I briefly review the popular modelling techniques found in the current literature and explore some of the advantages, disadvantages, and other noteworthy aspects of these methods.

Envelope Models

Envelope techniques have traditionally focused on the relationship between species distribution and climatic variables only; and consequently, are often referred to as climate envelope models (Kadmon and others 2003). BIOCLIM is one of the available software packages for envelope models. For its foundation in environmental space, the envelope model draws a rectangle resembling an envelope — hence the name — that bounds the range of climate variables suitable for the species (Nix 1986) as reported in (Carpenter and others 1993). Figure 2a is a simple example using two predictor variables, but envelope models are multidimensional, similar to the niche concept. One of the weaknesses of the envelope model is the rectangular shape which may include unsuitable habitat in the “corners” of the envelope. To address this problem, more complex shapes have been used to better characterize the species–climate relationship. Walker and Cox (1991) developed a variation of the climate envelope using irregular polygon envelopes, also known as convex hull (Fig. 2b). Convex hull methods, such as Habitat, eliminate the extra environmental space within the rectangles that is unlikely to have accurate presence–absence discrimination (Carpenter and others 1993). Both rectangular and convex hull envelopes define potential environments as “core” or “marginal.” Robertson and colleagues (2004) refer to these as crisp envelopes because the predictions are classified into three values—core, marginal, or absent. In turn, they developed a new modelling technique called fuzzy envelopment modelling that uses fuzzy logic. Robertson and colleagues (2004) have refined the crisp envelope by changing how the model copes with uncertainty and classification — the fuzzy model defines a continuous classification. The use of fuzzy logic in species distribution models is still in its early stages, but poses to be an ecologically realistic approach after further evaluation by researchers.

Domain

The Domain procedure uses a point-to-point similarity metric to assign a classification value to each grid cell based on its proximity in environmental space to the most similar species presence location (Carpenter and others 1993). Environmental similarity between the grid cell and the known presence site is calculated by summing the standard distance, in environmental space, between two points for each environmental variable. Standardized distance is then calculated by dividing the standard distance by the range of the environmental variables and equalizing the contribution from each environmental variable. The equal weight given to all the predictor variables may be considered a disadvantage of the Domain procedure. The output for Domain is the maximum similarity values between each grid cell and the known presence observations. Output in the form of positive values indicate presence; negative values are a prediction of absence. The output is a measure of the classification confidence — not a prediction of the probability of occurrence (Carpenter and others 1993).

Ecological Niche Factor Analysis

Ecological Niche Factor Analysis (ENFA) is a relatively new approach explicitly created to model species niches from presence–only data. Hirzel and colleagues (2002) emphasize that the model attempts to be ecologically realistic by assuming a unimodal relationship between the species and the environmental variables. In this factor analysis, the first factor (called the marginality factor) attempts to maximize the distance, in ecological space, between the species optimum for an environmental variable and the mean value of that variable for the entire geographic study area. The other factors maximize the specialization of the species along the environmental gradient by analyzing the ecological variance. The eigenvectors and eigenvalues are then used to map habitat suitability (Hirzel and others 2002). Few studies have been published comparing ENFA to more widely used algorithms. So far, evidence indicates that ENFA may be a promising technique to use with presence–only data sets (Sattler and others 2007; Tole

2006). ENFA has also been used to help choose unsuitable habitat in which to generate pseudo–absence points for other modelling algorithms (Chefaoui and Lobo 2007; 2008).

Ordination

Canonical Correspondence Analysis (CCA) is one of the more common ordination techniques used in distribution modelling. CCA is an indirect gradient analysis technique that relates environmental gradients to the distribution or abundance of a species (ter Braak 1986). Like ENFA, CCA assumes a unimodal relationship between the species and the environmental variable. CCA is a combination of correspondence analysis and multiple regression; using the reciprocal averaging algorithm of correspondence analysis combined with a multiple regression which is performed at each averaging cycle. The axes of the CCA are two dimensional combinations of the environmental and occurrence data. CCA is a "constrained" technique because the resulting ordination is constrained by the environmental variables (ter Braak 1986). The assumptions of a CCA, however, are difficult to satisfy with data typically available for distribution modelling. CCA requires: normally distributed data with symmetrical tails on the bell curve; species having equal amplitude in response to the environmental variable; and species optima evenly spaced along the environmental gradient (ter Braak 1986). Further research on CCA capabilities acknowledge that function performance may not be significantly affected if the assumptions are violated (Palmer 1993). The advantages of CCA are that it can use abundance data in addition to presence–absence data. CCA can also be used for multiple species at a time, but uses the same environmental variables for all species.

Regression

This suite of modelling techniques is the most widely used for species distribution prediction (Guisan and Zimmermann 2000). Regression has been thoroughly studied and produces models that are easily interpreted. Generalized linear models (GLMs) and generalized additive models (GAMs) are applied extensively in SDM because of their statistical power and their potential to realistically model species–environment relationships (Austin 2002; Yee and Mitchell 1991). GLMs are parametric techniques that assume a linear relationship, which may not always be ecologically appropriate. However, at finer scale a linear relationship may be the best representation of the relationship (Fig. 1). GAMs are considered more ecologically realistic because they use non-parametric functions that are more capable of modelling complex response–predictor relationships. GAMs may create models that fit the training data better than GLMs, but this appears to come at a cost. When validated with independent evaluation data, GAMs do not perform as well because of over-fitting, which limits the transferability of the model to different areas or time periods (Randin and others 2006). Unfortunately because of the complexity of the algorithm used to determine the shape of the species–environment relationship, GAMs require a large training data set to produce an accurate model (Yee and Mitchell 1991).

For all regression techniques, occurrence data should be independent and therefore not exhibit spatial autocorrelation. Stratified sampling across environmental gradients will improve regression models. This can be done by either removing data points, which may eliminate valuable data, or by additional field sampling, which may be costly and impractical. The use of spatial autoregressive models can help to eliminate the spatial autocorrelation effect within the data (Carl and Kühn 2007; Collingham and others 2000; Dark 2004; Lichstein and others 2002; Maggini and others 2006; Segurado and Araújo 2004).

A relatively new strategy for improving regression models is the use of information-theoretic approaches to select the best model based on the number of predictive variables and predictive accuracy (Johnson and Omland 2004). As the number of predictive variables in a model increases, the ability of the model to fit the training data increases. Maximizing accuracy or fit of the model, without considering model complexity, will favor a model that utilizes all possible parameters. With a large collection of predictor variables, it is possible to over-fit the model. The model becomes extremely good at predicting the training data, but poorly predicts data outside the original range. This, of course, reduces the potential usefulness of the model. To combat over-fitting and increasing complexity of models, model selection methods, such as Akaike's information criterion (AIC), have become increasingly popular in species distribution model research (Gibson and others 2004; Johnson and Omland 2004; Rushton and others 2004). AIC not only takes into consideration the model fit, but also imposes a penalty based on the number of predictor variables within the model. AIC is used to identify the most parsimonious set of models given the number of predictor variables and the ability of the model to correctly predict presence and absence (Burnham and Anderson 2002).

Another common method of reducing the number of predictive variables is to run a multivariate analysis on the correlation matrix to determine which variables are most important to the species distribution (Manel and others 2001). This also can help to explore the potential relationships between environmental data and observational data before model building. Analysis of the environmental data prior to model building is necessary to determine multi-collinearity among the variables. Most modelling procedures assume the predictor data sets are independent. Removing predictors that are highly correlated will improve the model performance. Thuiller and colleagues (2003) found that AIC allowed for additional predictor redundancy even after variables were selected with a PCA.

Classification and Regression Trees

Classification and regression tree (CART) methods can create predictive maps by either determining classes or average values for each grid cell of the study area. The algorithm divides the training data into two sub-sets, iteratively, based on the environmental variable that best reduces the variance in the response variable. A tree is constructed by further divisions causing dichotomous branching for each split of the data. This continues with all new sub-sets until all occurrences have been classified. The branches of the tree can lead to presence or absence based on the environmental variable used to sort the data (De'ath and Fabricius 2000). The CART method allows for species to be present in two different habitat types because CART can identify multiple combinations of environmental variables that may be suitable for presence — multiple branches of the tree may lead to presence (Norris and others 2006).

Random Forest is a form of CART that increases the power of the classification tree by generating multiple models from repeatedly sub-sampled training data sets (bootstrapping). The multiple models grow a “forest” of trees of which each tree is "grown" from a randomized subset of environmental variables. Each species data point is classified by all trees in the "forest." The classification backed by the greatest number of trees becomes the value for the data point (Breiman 2001). Although increasing the number of trees does not appear to increase over-fitting in Random Forests (Prasad and others 2006), it does complicate model interpretability (De'ath 2002)

Yet another advanced CART method is Boosted Regression Trees (BRT). The BRT models incorporate the regression tree algorithm of CART with a boosting algorithm that combines and summarizes a collection of many — 100s to 1000s — trees. In contrast, conventional regression finds a single tree or model that is the best. Boosting works on the premise that "it is easier to find many rough rules of thumb that it is to find a single highly accurate prediction rule" (Schapire 2002). The boosting procedure builds many "mediocre" models then combines them to produce an average. The addition of the boosting algorithm also enables the BRT models to better represent smooth species response curves by averaging many — 100s to 1000s — trees (Elith and others 2008). The models also are able to represent non-linear relationships and interactions between predictor variables (Elith and others 2008).

Although BRT modelling could be considered a "black-box" method, as many other machine learning methods have been labeled, it appears in initial modelling research that BRT results are making ecological sense (Elith and others 2006). One significant drawback of BRT is current implementation requires absence records in the training data set. Although modelers have had good results by using random background data or pseudo-absences (Elith and others 2006) (Elith and others 2008). Because of its complexity, BRT models can easily over-fit the training data. Elith and colleagues

(2008) have developed and published a tutorial and guidelines to facilitate the proper implementation and parameterization of BRT (see online supplement for (Elith and others 2008)). However, additional studies using BRT for SDM are necessary to fully understand their parameterization for a variety of species in many different regions.

GARP

Genetic algorithm for rule-set prediction (GARP) is a machine learning algorithm that takes an artificial intelligence approach to species distribution modelling. GARP develops rules for the distribution based on an iterative process of selection, evaluation, testing, and incorporation or rejection. GARP can improve the algorithms based on its calculations. This process is handled solely by the software without additional user input. Selection occurs when GARP chooses and implements one of several modelling algorithms to the training data. That algorithm, or rule, evolves to maximize accuracy of the model predictions. This evolutionary process is said to be analogous to DNA evolution — point mutations, deletions, crossing over — and accordingly the term “genetic” reflects this method. The accuracy procedure occurs up to 1000 times or until newly evolved rules do not improve the accuracy. The resulting rule-set is the model of the potential distribution and is mapped as predicted presence or absence (Stockwell and Peters 1999). GARP should, theoretically, perform better than individual modelling algorithms because it applies and selects the most accurate models (Peterson and Nakazawa 2007; Stockwell 2007). GARP, however, performed poorly in comparison to many other modelling techniques (Elith and others 2006). Additional research indicates that GARP may be most useful in situations where presence-only data are available in a very small data set (Pearson and others 2007; Stockwell and Peterson 2002).

Because of the random procedures built into GARP, the output will be different every time it is run despite the identical input and parameterization (Anderson and others 2003). This is an important concern for ecologists using and evaluating the model — results are not easily replicated or interpreted. The ecological relationship between species presence and the environmental variable is hidden within the software.

Maximum Entropy

Maximum Entropy (Maxent), like other machine learning techniques, improves the modelling algorithm automatically through a series of trainings with the data set. The creators of the Maxent technique used in species distribution modelling state that it is able to predict a species' distribution based on "incomplete information" — species observation data that do not necessarily cover the entire suitable range of environmental variables. Maxent estimates the distribution with maximum entropy (the most uniform or spread out distribution) of the known presence points given the constraints put on the distribution with respect to the point's relationship to the environmental layers. This relationship is quantified by using the empirical average of the environmental variable at all presence records (Phillips and others 2006). The implementation of Maxent for species distribution modelling was specifically designed for use with presence-only data. In comparison with other presence-only methods, it performs significantly better. Maxent also performs well when compared to presence–absence procedures that utilize both real and pseudo-absence data (Elith and others 2006; Hernandez and others 2006; Pearson and others 2007).

Maxent has several features that improve the models predictive performance and interpretability. Maxent can take into account the interaction between environmental variables. Maxent output is the probability of distribution, which is mathematically defined. Maxent also has a built in procedure to counteract over-fitting of the model; it employs a relaxation that allows the estimated distribution to go beyond the empirical average within the error bounds. This smoothing procedure, called regularization, can potentially correct for small sample size (Phillips and others 2006). Yet, recent research indicates mixed results of models built from small data sets (see both (Pearson and others 2007; Peterson and others 2007)).

Because it is a new technique, Maxent has not been thoroughly tested for potential weaknesses. The effect of spatial autocorrelation within a data set has not been tested by independent researchers. Also because of its recent application to species distribution modelling, there are fewer known rules that help to guide the use, and reduce the misuse or misinterpretation, of this technique. Although Maximum entropy modelling is new to ecology, this technique has been used for many different applications and research into its uses, problems, and advantages is active and growing (Phillips and others 2006).

MODEL COMPARISON

Several recent papers have systematically compared the performance of multiple modelling techniques. The most comprehensive comparison to date was carried out by Elith and colleagues (2006) who tested 16 different modelling techniques using many species in several geographic regions. One of their main objectives was to demonstrate the utility of presence-only data in species distribution modelling. Therefore, they did not use true absence data, but they did generate pseudo-absence data for techniques that required it. Generalized dissimilarity modelling (GDM), Maxent, boosted regression trees (BRT), and multivariate adaptive regression splines for community data (MARS-COMM) performed the best on average for all regions and species. Elith and colleagues (2006) suggest that future models will perform better through the use of some of these newer, more advanced techniques — such as BRT, MARS-COMM, and GDM.   Elith and colleagues (2006) believe that the best performing models share some key characteristics: ability to model complex species–predictor relationships and, by using smoothing or regularization techniques, do not over-fit the data. Techniques that responded poorly to the data were some of the older and more established methods: BIOCLIM (one type of envelope model), multivariate adaptive regression splines for individual species data (MARS-INT), Domain, and the desktop application of GARP (DK-GARP). However, almost all tests resulted in models that predicted species occurrence better than random. Elith and others (2006) comparison of model performance also illustrates the variability of modelling success across regions. Some regions, particularly Canada and the Australian wet tropics have more difficulty producing reliable model results. This reduction in model performance is most likely related to the quality of available data — both species and environmental — going into the model for these regions. For example, in Canada, the species occurrence data are biased toward the southern portion of the country, leaving a large geographic gap in the training data.

Where as Elith and others (2006) provide a thorough analysis of modelling methods, they do not consider finer details of the modelling process such as variable selection and modelling choice using information-theoretic approaches. Additional research on these topics needs to be done to test how models may be enhanced by refining their use.

Other comparison studies have not tested as many techniques, but their results have helped us to understand the circumstances that cause good, or bad, model results. Size of the occurrence data set has a significant effect on the model results with some modelling techniques producing useful models with small sample sizes. Maynard and Quinn (2007), using artificially generated data, found that GARP performed very well with extremely small sample sizes. Hernandez and colleagues (2006) also found GARP, in addition to Maxent, to perform reasonably well with small occurrence data sets. Prevalence, the ratio of presence to absence points, in species occurrence data will reduce the effective sample size. Maynard and Quinn (2007) found that a prevalence of 5% in a 2000 point data set was equivalent to having a sample size of 200 with a 50% prevalence.

The scale at which a model is built will also be an important factor in model outcome. Although scale was not directly addressed by Elith and others (2006), Thuiller and colleagues (2003) did evaluate model performance at different scales. Of the three model types tested (GLM, GAM, and CART), they found some models performed better at larger scales. They suggest models that can handle complex relationships will be better able to model at a variety of scales. Models that rely on a particular way to describe relationships, linear for example, may not be good for large scale analyses because it is less likely for species responses to be linear across the entire gradient of environmental variables. Linear models may be very useful at finer scales because the response is more likely to be linear over a shorter distance along the gradient (Fig. 1) (Segurado and Araújo 2004; Thuiller and others 2003).

In addition to scale, model accuracy will be affected by the range of a species' ecological requirements and tolerances. A specialist species with a narrow geographic range and specific ecological requirements are easier to model — relationships between specialist species and environment can be simply expressed mathematically. The distribution of generalist species with a high tolerance of a wide range of ecological situations across a large geographic extent will be much more difficult to predict. The model's capability to represent these broad relationships is limited (Elith and others 2006; Hernandez and others 2006; Segurado and Araújo 2004). Because SDM attempts to characterize and quantify the species relationship to the environment, the more specific, and simple, the relationship is the better.

Published comparisons of models show that there is no one technique that is superior for all circumstances, but certain modelling algorithms and software packages perform better in general (Hernandez and others 2006; Meynard and Quinn 2007; Muñoz and Felicísimo 2004; Segurado and Araújo 2004; Thuiller and others 2003; Vayssieres and others 2000). When deciding on a modelling technique you must take into account the available training data. Questions that should be asked are: How big is the data set? Does it cover the entire range of the species? Does it include absence data? Is there significant spatial autocorrelation? Does the environmental data include categorical values? Answers to these questions will help to determine technique type. Answering additional questions may lead to good model choices, such as: What is the purpose of the model; how will the results be used? If the intent is to identify land as endangered species habitat, then choose a model that minimizes false presence. Is the objective to understand the ecological relationship between predictor and response variables? Then choose a mathematical model that is interpretable. Even practical considerations have merit. How easily is the model implemented? Can existing data be used and are computer resources available? Table 2 outlines some of the important distinguishing features of each model type. These model properties will help to determine the most appropriate method for a given situation.

The models discussed and employed in this literature review and dissertation are correlative in nature. In other words, all SDMs incorporate algorithms that correlate species point occurrence data with a variety of environmental data. The algorithms attempt to find areas that are environmentally similar to those areas where the species is known to be present or absent. However, another branch of distribution modelling is interested in understanding the underlying mechanisms that determine species distributions. The mechanistic approach directly measures the individual’s response to abiotic variables, and, thus, determining the direct cause of a species geographic limitation. Kearney (2006) highly recommends more research be done to apply spatially referenced data to mechanistic models of niches.

MODEL EVALUATION

Model evaluation is the testing process that helps determine the validity of the model predictions. Testing must be conducted to defend the applicability of a model to the given data and to the true distribution. In general, models are evaluated based on the percentage of prediction errors, which are either false presence or false absence. The results of model evaluation are cross-tabulated in a confusion matrix (also known as an error matrix or contingency table) that compares the predicted and actual presence-absence points, which can be reported as either counts or percentages (Tab. 1). False presence errors are type I or commission errors; false absences are type II or omission errors. The confusion matrix also tallies true presence and true absence (Gotelli and Ellison 2004). Conventional statistical tests on contingency tables are inappropriate for evaluating model performance. Tests such as chi-square would result in highly significant values for situations that were either very accurate (high values of TP and TA) or very inaccurate (high values of FP and FN) (Gotelli and Ellison 2004).

Instead, from these four simple counts many accuracy measures can be derived. The most common of these are prediction success, sensitivity, specificity, and Cohen's kappa. Prediction success is the simple calculation of the percentage of points for which presence or absence is accurately predicted. Sensitivity (TP/(TP+FP)) is the likelihood that a predicted presence point should actually be absent. Specificity (TA/(TA+FA)) is the likelihood that a predicted absence point should really be classified as present. Cohen's kappa is a one of the few measures that uses all the data within a confusion matrix, taking into account commission and omission errors as well as sensitivity and specificity, to produce an index value. The index ranges from -1 to 1 — with high values meaning the predictions match the observation data, 0 indicating random agreement, and low values meaning the predictions are opposite of the observations (Elith and others 2006; Fielding and Bell 1997; Manel and others 2001).

Due to the nature of the available information, most models are built with binary observation data — simple presence-absence records for a given location. The environmental data used to build models are generally not binary, but are categorical with several possibilities or a continuous range of values. Consequently the model output is a continuous range of possibilities of presence. Each pixel or grid cell contains a value representing percentage of presence likelihood or percent suitability. Traditional model evaluation techniques cannot use the continuous model output, instead, the data must be converted to binary format (presence-absence) and a threshold percentage must be chosen. A threshold value of 0.5 is often chosen because it is the point at which the percentage of false presence and false absences are equal. However, when the data set does not have an equal number of absence and presence points, the threshold is biased towards the more common point (Manel and others 2001) (Jiménez-Valverde and Lobo 2006). When the number of absence points is equal to the number of presence points, it is said that the data set has a prevalence of 0.5. Prevalence is higher when the presence to absence point ratio is higher.

Liu and others (2005) conducted a comparison of twelve threshold selecting approaches using data sets with seven levels of prevalence. They found that most threshold determining procedures worked well in data sets with a prevalence of 0.5 and that model output is always biased toward the larger of the two groups, presence or absence. This especially poses a problem with modelling techniques that rely on presence-only data sets or randomly generated pseudo-absence points, which usually outnumber the original present point data 100 fold. The choice of the threshold must be adjusted based on the most common point in the data set (Collingham and others 2000; Jiménez-Valverde and Lobo 2006; Liu and others 2005; Manel and others 2001).

Research by Manel and others (2001) and Liu and others (2005) also illustrates that some threshold dependent model evaluation procedures are affected by data prevalence. Predictive success, sensitivity, and specificity are all significantly affected by prevalence in the data set. However, Manel and his colleagues found that Cohen's Kappa was only "marginally affected by prevalence" and recommend it as a simple calculation for model evaluation. Another advantage of Cohen's kappa is that it is always calculable despite the occurrence of zeros in the confusion matrix (Manel and others 2001).

Receiver-operating characteristic (ROC) plots have been widely used in recent years as a threshold independent evaluation technique for distribution models. Before their acceptance in the ecological modelling discipline, ROC plots have been used to discriminate radar signals, medical diagnostic test results, and weather predictions (Fielding and Bell 1997). ROC plots appear to be useful for species distribution modelling because they are not significantly affected by prevalence (Manel and others 2001) and their use eliminates the need to subjectively choose a threshold for model evaluation. The ROC curve plots sensitivity as a function of (1 - specificity) over the entire range of thresholds. A curve that maximizes sensitivity for low values of (1 - specificity) is characteristic of good model performance. This is illustrated by a curve that comes close to the upper left corner of the ROC plot (Zweig and Cambell 1993). The Area Underneath the Curve (AUC) is calculated and becomes a score of the model's accuracy for all possible thresholds. The score can range from 0.5 to 1 — 1 indicating perfect discrimination between present and absent points and 0.5 indicating the chance of being present or absent is 50% and therefore no discrimination between the two. An index has been developed for AUC values: a value of 0.5-0.7 is considered low accuracy; 0.7-0.9 is considered useful; and 0.9 and above is considered high accuracy (Swets 1988).

Despite its wide use, the validity of AUC as a measure of model accuracy has been questioned recently. Lobo and others (2007) recommend not using AUC for several reasons. First, they argue that AUC does not measure accuracy, but instead simply measures discrimination. If the predicted probabilities of species occurrence range from 0.4 to 0.6 in the region, the discrimination between suitable habitat and unsuitable habitat is low. The accuracy of the model may be very high, meaning the species occurrence probabilities may be accurate even though the discrimination between presence and absence is poor. Lobo and others (2007) also point out that it is not useful to have one score represent the entire range of thresholds because it is unlikely that the extremes of the threshold range contain useful information. The far edges of the threshold range correspond to very high type I or type II errors. The range of thresholds of interest are found in the middle where type I and type II errors are nearer to equal. Additionally, Lobo and others (2007) believe the main argument for using AUC — because it is threshold independent — is questionable. In the past, threshold choice has been considered subjective, but thresholds can be chosen using several tested methods (Liu and others 2005).

Models are evaluated both internally and externally. Internal evaluation is how well the model fits the training data. In the literature it is also known as resubstitution because it reuses the training data to verify the model. This estimation of model accuracy is, obviously, biased. Models tend to over-fit the training data because the model is built on the subtle variation of each point in the training data (Fielding 2002). While the model may fit the training data well, the additional variation found in the species occurrence in the real world may not be accurately predicted. Therefore, it is recommended that an independent data set be used to conduct an external evaluation — how well the model is able to fit a separate, independent set of evaluation data (Elith and others 2006; Loiselle and others 2007; Peterson 2005). Evaluation data ideally would be a truly independent data set, possibly obtained via different methods, during a different time period, or in a different region (Araújo and Guisan 2006; Manel and others 1999; Manel and others 2001). Unfortunately a genuinely independent data set is usually not available (however, see (Elith and others 2006; Fielding 2002)). Instead many modelers simply "hold out" a random selection of observations to be used in the external evaluation. A basic rule of thumb for the amount of evaluation data is 20-30% of the available observation points (Araújo and others 2005a; Pearson and others 2006; Thuiller 2003). However, Huberty and Olejnik (2006) developed a method for determining the percentage of data that should be held out for evaluation purposes. They propose that the amount of evaluation data should be based on the number of predictor or environmental variables used in the model. They recommend using:

$$[1 + (p-1)^{1/2}]^{-1}$$

where p is the number of predictors or environmental layers. As the number of environmental layers to build the model is increased the percentage of points used to build the model (the training data) should increase.

Instead of simply removing data for evaluation, more sophisticated data partitioning techniques have been developed to allow all available data to be used for model building.

Bootstrapping and jack-knifing procedures are common when models are built with small observational data sets (see review in (Fielding and Bell 1997). These procedures build the models repeatedly with random observations taken out for evaluation then replaced in the training data and models are built and evaluated again with a different selection of data (Fielding 2002). An average of the results is then reported. This may be the best compromise for small data sets representing rare species or relatively unknown regions, for which each data point is necessary for model building (Pearson and others 2007).

The modelling objective should help to determine the best method for model evaluation. When determining the appropriate threshold, the types of errors to minimize based on the goal or the modelling project must be ascertained (Lobo and others 2007; Loiselle and others 2003). For example, thresholds should be optimized to reduce type I error, false presence, attempting to locate populations for research purposes. However, reducing type II errors might enhance accuracy for inventories of an endangered species in a region of rapid human development. Loiselle and others (2003) analyzed error type and how it could affect conservation planning. She and her colleagues were exploring the usefulness of distribution models for identifying potential land for conservation reserves. They found that models that tended to minimize false positives, type I errors, were more likely to agree with expert ecologist opinions on good locations for land reserves. They conclude that the models, in general, may overestimate species habitat and possibly misdirect conservation effort.

FUTURE DIRECTIONS

The current state of most species distribution modelling focuses on basic implementation. Modelling algorithms correlate environmental data with species occurrence data. This is not a new concept — not in the least (Forbes 1844; Humboldt 1815). Because of the emergence of spatial technologies and advanced computing power SDMs can consider large areas, the whole globe in fact, and dozens of predictor variables. Models can also incorporate data from satellite images that allow the study of remote and little known regions. Advances have been made developing different techniques to manage spatial autocorrelation, presence-only data, and small training data sets.

Currently SDMs have problems that need to be addressed in the future to improve the reliability of their predictions. One of the current challenges is modelling species that are not at equilibrium with their environment, such as invading species being modelled in their new region. Native species may be still responding to past disturbances such as fire or even glacial retreat of the last ice age. Most modelling techniques assume equilibrium, but new techniques need to be developed to help account for this situation (Guisan and Thuiller 2005).

In the recent literature several articles debate the most appropriate evaluation methods for SDMs. Researchers disagree about the validity of certain model validation procedures (Araújo and Guisan 2006; Austin 2007; Guisan and Thuiller 2005; Jimenéz-Valverde and others 2008; Lobo and others 2007). Basic model evaluation needs to be standardized so that models can be compared across species, regions, and time periods. Model evaluation can be improved through additional, yet simple, reporting. Vaughan and Ormerod (2005) found that sufficient model testing was “scarce and errors were seldom diagnosed.” They go on to suggest some straightforward practices that do not include novel calculations, but simply provide the reader with a better understanding of how the model was evaluated. They recommend that modelers report on the model’s overall performance, including its ability to be generalized and transferred. They also believe researchers should explain their evaluation parameters, such as threshold determination, to indicate the possible uses of the model. Finally, they advise researchers to identify the model’s weaknesses and communicate the possible causes (Vaughan and Ormerod 2005).

While some species distribution modelling software packages allow data to be dumped in and models to be built with little guidance from the biologist, it may be better for model choice to be directed by expert knowledge. Relevant predictor variables can be selected by biologists, who can then analyze them to determine which explain the most variation in the occurrence data. This is not new. What is relatively new is the use of the information-theoretic approach to model selection (Burnham and Anderson 2002). The information-theoretic approach will assist in choosing the most parsimonious model — the model that explains most of the variation weighted by the number of parameters used. Biologists and modelers will benefit from using this approach to choose elegant and ecologically significant models (Guisan and Thuiller 2005).

The future of species distribution modelling promises to reveal some exciting techniques to cope with the dilemmas of current modelling approaches. Austin (2002) argues that there needs to be a better connection between ecological theory and statistical modelling. Researchers are stepping back and looking at the theoretical principles at the root of species distributions and finding that a fundamental part of niche theory is missing — biotic interactions. The inclusion of interactions between species that, in part, govern species distribution will increase the ecological relevance of the model. The distribution of competitor, predator, and mutualist species can easily be added to a model, but the interaction coefficients may be more complicated. Nonetheless, knowledge of species life history is needed to produce good models. The inclusion of additional relevant predictor variables is a necessary challenge. Future modelers are obliged to consider migration and dispersal as important determinants of a species distribution. The literature is burgeoning with studies on the effect climate change will have on species distributions. It is becoming more evident that migration and dispersal characteristics of a species will become important factors as human caused habitat and climate change transforms species distributions. Species distribution models may also benefit by including theoretical concepts of population ecology. Metapopulation theory may improve the model's ecological relevance. Understanding and adding source-sink dynamics of the target species into models will lead to results that better represent the ecology of the organism (Austin 2002; Guisan and others 2006; Guisan and Thuiller 2005).

Communities and functional groups will be better modelled in the next several years. Already techniques have been designed to work with multiple species to build models of communities. The theoretical challenge for ecologists will be: How to reconcile individual responses to the environment with the desire to model entire communities as one? Modelling functional groups may, therefore, be less formidable because — depending upon how the group is defined — they may respond similarly to an environmental gradient.

LITERATURE CITED

Anderson RP. (2003). Real vs. artefactual absences in species distributions: tests for *Oryzomys albigularis* (Rodentia: Muridae) in Venezuela. Journal of Biogeography **30**,591-605.

Anderson RP, Lew D, Peterson AT. (2003). Evaluating predictive models of species' distributions: criteria for selecting optimal models. Ecological Modelling **162**,211-232.

Anderson RP, Peterson AT, Egbert SL. (2006). Vegetation-index models predict areas vulnerable to purple loosestrife (*Lythrum salicaria*) invasion in Kansas. The Southwestern Naturalist **51**,471-480.

Araújo MB, Guisan A. (2006). Five (or so) challenges for species distribution modelling. Journal of Biogeography **33**,1677-1688.

Araújo MB, Luoto M. (2007). The importance of biotic interactions for modelling species distributions under climate change. Global Ecology and Biogeography **16**,743-753.

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

Araújo MB, Pearson RG. (2005). Equilibrium of species' distributions with climate. Ecography **28**,693-695.

Araújo MB, Pearson RG, Thuiller W, Erhard M. (2005a). Validation of species-climate impact models under climate change. Global Change Biology **11**,1504-1513.

Araújo MB, Rahbek C. (2006). How does climate change affect biodiversity? Science **313**,1396-1397.

Araújo MB, Whittaker RJ, Ladle RJ, Erhard M. (2005b). Reducing uncertainty in projections of extinction risk from climate change. Global Ecology and Biogeography **14**,529-538.

Austin MP. (2002). Spatial prediction of species distribution: an interface between ecological theory and statistical modeling. Ecological Modelling **157**,101-118.

Austin MP. (2007). Species distribution models and ecological theory: A critical assessment and some possible new approaches. Ecological Modelling **200**,1-19.

Barry S, Elith J. (2006). Error and uncertainty in habitat models. Journal of Applied Ecology **43**,413-423.

Beale CM, Lennon JJ, Elston DA, Brewer MJ, Yearsley JM. (2007). Red herrings remain in geographical ecology: a reply to Hawkins et al. (2007). Ecography **30**,845-847.

Beauvais GP, Keinath D, Thurston R. (2004). Predictive range maps for 5 species of management concern in southwestern Wyoming. (Wyoming Natural Diversity Database, University of Wyoming, Laramie, WY). pp 1-11.

Box EO, Crumpacker DW, Hardin ED. (1993). A climatic model for location of plant species in Florida, USA. Journal of Biogeography **20**,629-644.

Brotons L, Thuiller W, Araujo MB, Hirzel AH. (2004). Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. Ecography **27**,437-448.

Buechling A, Tobalske C. (2007). Habitat modeling of rare plant species in Pacific Northwest forests. (Oregon Natural Heritage Information Center, Portland, OR). pp 35.

Burnham KP, Anderson DR. (2002). *Model selection and multimodel inference: a practical information-theoretic approach*. (Springer, New York, NY).

Camarero JJ, Gutiérrez E, Fortin MJ, Ribbens E. (2005). Spatial patterns of tree recruitment in a relict population of *Pinus uncinata*: Forest expansion through stratified diffusion. Journal of Biogeography **32**,1979-1992.

Carey PD, Brown NJ. (1994). The use of GIS to identify sites that will become suitable for a rare orchid, *Himantoglossum hircinum* L., in a future changed climate. Biodiversity Letters **2**,117-123.

Carl G, Kühn I. (2007). Analyzing spatial autocorrelation in species distributions using Gaussian and logit models. Ecological Modelling **207**,159-170.

Carpenter G, Gillison AN, Winter J. (1993). DOMAIN: a flexible modelling procedure for mapping potential distributions of plants and animals. Biodiversity and Conservation **2**,667-680.

Chefaoui RM, Lobo JM. (2007). Assessing the conservation status of an Iberian moth using pseudo-absences. Journal of Wildife Management **71**,2507-2516.

Chefaoui RM, Lobo JM. (2008). Assessing the effects of pseudo-absences on predictive distribution model performance. Ecological Modelling **210**,478-486.

Collingham YC, Wadsworth RA, Huntley B, Hulme PE. (2000). Predicting the spatial distribution of non-indigenous riparian weeds: issues of spatial scale and extent. Journal of Applied Ecology **37**,13-27.

Costa GC, Wolfe C, Shepard DB, Caldwell JP, Vitt LJ. (2007). Detecting the influence of climatic variables on species distributions: a test using GIS niche-based models along a steep longitudinal environmental gradient. Journal of Biogeography **35**,637-646.

Cowling RM, Samways MJ. (1994). Predicting global patterns of endemic plant species richness. Biodiversity Letters **2**,127-131.

Danks FS, Klein DR. (2002). Using GIS to predict potential wildlife habitat: A case study of muskoxen in northern Alaska. International Journal of Remote Sensing **23**,4611-4632.

Dark SJ. (2004). The biogeography of invasive alien plants in California: an application of GIS and spatial regression analysis. Diversity and Distributions **10**,1-9.

Davis AJ, Jenkinson LS, Lawton JH, Shorrocks B, Wood S. (1998). Making mistakes when predicting shifts in species range in response to global warming. Nature **391**,783-786.

De'ath G. (2002). Multivariate regression trees: A new technique for modeling species-environment relationships. Ecology **83**,1105-1117.

De'ath G, Fabricius KE. (2000). Classification and regression trees: A powerful yet simple technique for ecological data analysis. Ecology **81**,3178-3192.

Elith J, Graham CH, Anderson RP, Dudik M, Ferrier S, Guisan A, Hijmans RJ, Huettmann F, Leathwick JR, Lehmann A and others. (2006). Novel methods improve prediction of species' distributions from occurrence data. Ecography **29**,129-151.

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

Ferrier S, Guisan A. (2006). Spatial modelling of biodiversity at the community level. Journal of Applied Ecology **43**,393-404.

Fertig W, Reiner WA, Hartman RL. (1998). Gap analysis for plant species. GAP Analysis Program Bulletin **7**,24-25.

Fielding AH. (2002). What are the appropriate characteristics of an accuracy measure? In *Predicting species occurrences: issues of accuracy and scale*. Scott JM, Heglund PJ, Morrison ML, Haufler JB, Raphael MG, Wall WA, Samson FB, eds. (Island Press, Washington, DC). pp 271-280.

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

Forbes E. (1844). Report on the mollusca and radiata of the Aegean Sea and on their distribution, considered as bearing on geology. In *Reports on the British Association of Science for 1843*. pp 130-193.

Gibson LA, Wilson BA, Cahill DM, Hill J. (2004). Spatial prediction of rufous bristlebird habitat in a coastal heathland: a GIS-based approach. Journal of Applied Ecology **41**,213-223.

Gotelli NJ, Ellison AM. (2004). *A primer of ecological statistics*. (Sinauer Associates Inc., Sunderland, MA).

Guisan A, Overton JMC, Aspinall R, Hastie T, Lehmann A, Ferrier S, Austin M. (2006). Making better biogeographical predictions of species' distributions. Journal of Applied Ecology **43**,386-392.

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

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

Hawkins BA, Diniz-Filho JAF, Mauricio Bini L, De Marco P, Blackburn TM. (2007). Red herrings revisited: spatial autocorrelation and parameter estimation in geographical ecology. Ecography **0**,0-0.

Henebry GM, Merchant JW. (2002). Geospatial data in time: limits and prospects for predicting species occurrences. In *Predicting species occurrences: issues of accuracy and scale*. Scott JM, Heglund PJ, Morrison ML, Haufler JB, Raphael MG, Wall WA, Samson FB, eds. (Island Press, Washington, DC). pp 291-302.

Hernandez P, Graham CH, Master L, Albert DL. (2006). The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography **29**,773-785.

Hirzel AH, Hausser J, Chessel D, Perrin N. (2002). Ecological-niche factor analysis: how to compute habitat-suitability maps witout absence data? Ecology **83**,2027-2036.

Hopkins MJ. (2007). Modelling the known and unknown plant biodiversity of the Amazon Basin. Journal of Biogeography.

Hulme PE. (2003). Biological invasions: winning the science battles but losing the conservation war. Oryx **37**,178-193.

Humboldt AV. (1815). *Personal narrative of travels to the equinoctial regions of the New Continent during the years 1799-1804*. (M. Carey, Philadelphia, PA).

Iverson LR, Prasad AM. (1998a). Estimating regional plant diversity with GIS modelling. Diversity and Distributions **4**,49-61.

Iverson LR, Prasad AM. (1998b). Predicting abundance of 80 tree species following climate change in the eastern United States. Ecological Monographs **68**,465-485.

Jiménez-Valverde A, Lobo JM. (2006). The ghost of unbalanced species distribution data in geographical model predictions. Diversity and Distributions **12**,521-524.

Jimenéz-Valverde A, Lobo JM, Hortal J. (2008). Not as good as they seem: the importance of concepts in species distribution modelling. Diversity and Distributions **14**,885-890.

Johnson JB, Omland KS. (2004). Model selection in ecology and evolution. Trends in Ecology & Evolution **19**,101-108.

Kadmon R, Farber O, Danin A. (2003). A systematic analysis of factors affecting the performance of climatic envelope models. Ecological Applications **13**,853-867.

Kadmon R, Farber O, Danin A. (2004). Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. Ecological Applications **14**,401-413.

Kearney M. (2006). Habitat, requirements, and niche: what are we modelling? Oikos **115**,186-191.

Korner C. (2007). The use of 'altitude' in ecological research. Trends in Ecology & Evolution **22**,569-574.

Laurent JM, Cheddadi R, Bar-Hen A, François L, Ghislain M. (2004). Refining vegetation simulation models: From plant functional types to bioclimatic affinity groups of plants. Journal of Vegetation Science **15**,739-746.

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

Legendre P, Dale MRT, Fortin M-J, Gurevitch J, Hohn M, Myers D. (2002). The consequences of spatial structure for the design and analysis of ecological field surveys. Ecography **25**,601-615.

Lennon JJ. (2000). Red-shifts and red herrings in geographical ecology. Ecography **23**,101-113.

Lichstein JW, Simons TR, Shriner SA, Franzreb KE. (2002). Spatial autocorrleation and autoregressive models in ecology. Ecological Monographs **72**,445-463.

Liebhold AM, Gurevitch J. (2002). Integrating the statistical analysis of spatial data in ecology. Ecography **25**,553-557.

Liebhold AM, Sharov AA. (1998). Testing for correlation in the presence of spatial autocorrelation in insect count data. In *Population and community ecology for insect management and conservation*. Baumgartner J, Brandmayr P, Manly BFJ, eds. (Balkema, Rotterdam). pp 11-17.

Lillesand TM, Kiefer RW, Chipman JW. (2004). *Remote sensing and image interpretation*. (John Wiley & Sons, New York, NY).

Lippitt CD, Rogan J, Toledano J, Sangermano F, Eastman JR, Mastro V, Sawyer A. (2008). Incorporating anthropogenic variables into a species distribution model to map gypsy moth risk. Ecological Modelling **210**,339-350.

Lira-Noriega A, Soberon J, Navarro-Siguenza AG, Nakazawa Y, Peterson AT. (2007). Scale dependency of diversity components estimated from primary biodiversity data and distribution maps. Diversity and Distributions **13**,185-195.

Liu C, Berry PM, Dawson TP, Pearson RG. (2005). Selecting thresholds of occurrence in the prediction of species distributions. Ecography **28**,385-393.

Lobo JM, Jiménez-Valverde A, Real R. (2007). AUC: a misleading measure of the performance of predictive distribution models. Global Ecology and Biogeography **17**,145-151.

Loiselle BA, Brooks T, Smith KG, Williams PH, Howell CA, Graham CH, Goerck JM. (2003). Avoiding Pitfalls of Using Species Distribution Models in Conservation Planning. Conservation Biology **17**,1591-1600.

Loiselle BA, Jorgensen PM, Consiglio T, Jimenez I, Blake JG, Lohmann LG, Montiel OM. (2007). Predicting species distributions from herbarium collections: does climate bias in collection sampling influence model outcomes? Journal of Biogeography **35**,105-116.

Lütolf M, Kienast F, Guisan A. (2006). The ghost of past species occurrence: improving species distribution models for presence-only data. Journal of Applied Ecology **43**.

Maggini R, Lehmann A, Zimmermann NE, Guisan A. (2006). Improving generalized regression analysis for the spatial prediction of forest communities. Journal of Biogeography **33**,1729-1749.

Manel S, Dias JM, Buckton ST, Ormerod SJ. (1999). Alternative methods for predicting species distribution: an illustration with Himalayan river birds. Journal of Applied Ecology **36**,734-747.

Manel S, Williams HC, Ormerod SJ. (2001). Evaluating presence-absence models in ecology: the need to account for prevalence. Journal of Applied Ecology **38**.

Meynard CN, Quinn JF. (2007). Predicting species distributions: a critical comparison of the most common statistical models using artificial species. Journal of Biogeography **34**,1455-1469.

Muñoz J, Felicísimo ÁM. (2004). Comparison of statistical methods commonly used in predictive modelling. Journal of Vegetation Science **15**,285-292.

Murphy HT, Lovett-Doust J. (2007). Accounting for regional niche variation in habitat suitability models. Oikos **116**,99-110.

Nix HA. (1986). A biogeographic analysis of Australian Elapid Snakes. In *Atlas of Elapid Snakes of Australia*. Longmore R, ed. (Australian Government Publishing Service, Canberra). pp 415.

Norris JR, Jackson ST, Betancourt JL. (2006). Classification tree and minimum-volume ellipsoid analyses of the distribution of ponderosa pine in the western USA. Journal of Biogeography **33**,342-360.

Nyári Á, Ryall C, Peterson AT. (2006). Global invasive potential of the house crow Corvus splendus based on ecological niche modeling. Journal of Avian Biology **37**,306-311.

Oberhauser K, Peterson AT. (2003). Modeling current and future potential wintering distributions of eastern North American monarch butterflies. Proceedings of the National Academy of Science **100**,14063-14068.

Olding-Smee FJ, Laland KN, Feldman MW. (2003). *Niche construction: the neglected process in evolution*. (Princeton University Press, Princeton, NJ).

Ortega-Huerta MA, Peterson AT. (2004). Modelling spatial patterns of biodiversity for conservation prioritization in North-eastern Mexico. Diversity and Distributions **10**,39-54.

Palmer MW. (1993). Putting things in even better order: the advantages of canonical correspondence analysis. Ecology **74**,2215-2230.

Papes M. (2007). Ecological niche modeling approaches to conservation of endangered and threatened birds in central and eastern Europe. Biodiversity Informatics **4**,14-26.

Pearce JL, Boyce MS. (2006). Modelling distribution and abundance with presence-only data. Journal of Applied Ecology **43**,405-412.

Pearson RG, Dawson TP. (2003). Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? Global Ecology and Biogeography **12**,361-371.

Pearson RG, Raxworthy CJ, Nakamura M, Peterson AT. (2007). Predicting species distributions from small numbers of occurrence records: A test case using cryptic geckos in Madagascar. Journal of Biogeography **34**,102-117.

Pearson RG, Thuiller W, Araujo MB, Martinez-Meyer E, Brotons L, McClean C, Miles L, Segurado P, Dawson TP, Lees DC. (2006). Model-based uncertainty in species range prediction. Journal of Biogeography **33**,1704-1711.

Peppler-Lisbach C, Schräder B. (2004). Predicting the species composition of Nardus stricta communities by logistic regression modelling. Journal of Vegetation Science **15**,623-634.

Peterson AT. (2003). Predicting the geography of species' invasion via ecological niche modeling. The Quarterly Review of Biology **78**,419-433.

Peterson AT. (2005). Kansas GAP Analysis: the importance of validating distributional models before using them. The Southwestern Naturalist **50**,230-236.

Peterson AT, Nakazawa Y. (2007). Environmental data sets matter in ecological niche modelling: an example with *Solenopsis invicta* and *Solenopsis richteri*. Global Ecology and Biogeography **17**,135-144.

Peterson AT, Papes M, Kluza DA. (2003). Predicting the potential invasive distributions of four alien plant species in North America. Weed Science **51**,863-868.

Peterson AT, Vieglais DA. (2001). Predicting species invasions using ecological niche modeling: new approaches from bioinformatics attack a pressing problem. Bioscience **51**,363-371.

Peterson TA, 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, Dudik M, Elith J, Graham CH, Lehmann A, Leathwick JR, Ferrier S. (2009). Sample selection bias and presence-only models of species distributions: implications for selection of background and pseudo-absences. In press.

Phillips SJ, Anderson RP, Schapire RE. (2006). Maximum entropy modeling of species geographic distributions. Ecological Modelling **190**,231-259.

Prasad AM, Iverson LR, Liaw A. (2006). Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. Ecosystems **9**,181-199.

Prates-Clark CDC, Saatchi SS, Agosti D. (2008). Predicting geographical distribution models of high-value timber trees in the Amazon Basin using remotely sensed data. Ecological Modelling **211**,309-323.

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

Raxworthy CJ, Schneider GE, Ortega-Huerta MA, Peterson AT, Martinez-Meyer E, Horning N, Nussbaum RA. (2003). Predicting distributions of known and unknown reptile species in Madagascar. Nature **426**,837-841.

Robertson MP, Caithness N, Villet MH. (2001). A PCA-based modelling technique for predicting environmental suitability for organisms from presence records. Diversity and Distributions **7**,15-27.

Rodríguez JP, Brotons L, Bustamante J, Seoane J. (2007). The application of predictive modeling of species distribution to biodiversity conservation. Diversity and Distributions **13**,243-251.

Römermann C, Tackenberg O, Scheuerer M, May R, Poschlod P. (2007). Predicting habitat distribution and frequency from plant species co-occurrence data. Journal of Biogeography **34**,1041-1052.

Rowe RJ. (2005). Elevational gradient analyses and the use of historical museum specimens: A cautionary tale. Journal of Biogeography **32**,1883-1897.

Rushton SP, Ormerod SJ, Kerby G. (2004). New paradigms for modelling species distributions? Journal of Applied Ecology **41**,193-200.

Sattler T, Bontadina F, Hirzel AH, Arlettaz R. (2007). Ecological niche modelling of two cryptic bat species calls for a reassessment of their conservation status. Journal of Applied Ecology.

Schapire RE. (2002). The boosting approach to machine learning: an overview. In *MSRI Workshop on Nonlinear estimation and classification*. pp 1-23.

Segurado P, Araújo MB. (2004). An evaluation of methods for modelling species distributions. Journal of Biogeography **31**,1555-1568.

Stockwell D, Peters D. (1999). The GARP modelling system: problems and solutions to automated spatial prediction. International Journal of Geographical Information Science **13**,143-158.

Stockwell DRB. (2007). *Niche modeling: predictions from statistical distributions*. (Chapman & Hall/CRC, Boca Raton, FL).

Stockwell DRB, Peterson AT. (2002). Effects of sample size on accuracy of species distribution models. Ecological Modelling **148**,1-13.

Swets JA. (1988). Measuring the Accuracy of Diagnostic Systems. Science **240**,1285-1293.

ter Braak CJF. (1986). Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology **67**,1167-1179.

ter Steege H, Van Der Hout P, Daly DC, Silveira M, Phillips O, Vasquez R, Van Andel T, Duivenvoorden J, De Oliveira AA, Ek R and others. (2003). A spatial model of tree • -diversity and tree density for the Amazon. Biodiversity and Conservation **12**,2255-2277.

Thuiller W. (2003). BIOMOD - optimizing predictions of species distributions and projecting potential future shifts under global change. Global Change Biology **9**,1353-1362.

Thuiller W, Araujo MB, Lavorel S. (2003). Generalized models vs. classification tree analysis: Predicting spatial distributions of plant species at different scales. Journal of Vegetation Science **14**,669-680.

Thuiller W, Lavorel S, Ara˙jo MB, Sykes MT, Prentice IC. (2005a). Climate change threats to plant diversity in Europe. Proceedings of the National Academy of Science **102**,8245-8250.

Thuiller W, Lavorel S, Araujo MB. (2005b). Niche properties and geographical extent as predictors of species sensitivity to climate change. Global Ecology and Biogeography **14**,347-357.

Tole L. (2006). Choosing reserve sites probabilistically: a Columbian Amazon case study. Ecological Modelling **194**,344-356.

Van Mannen FT, Clark JD, Schlarbaum SE, Johnson K, Taylor G. (2002). A model to predict the occurence of surviving butternut trees in the Souther Blue Ridge Mountains. In *Predicting species occurences*. Scott JM, Heglund PJ, al. e, eds. (Island Press, Washington). pp 868.

Vaughan IP, Ormerod SJ. (2003). Improving the Quality of Distribution Models for Conservation by Addressing Shortcomings in the Field Collection of Training Data. Conservation Biology **17**,1601-1611.

Vaughan IP, Ormerod SJ. (2005). The continuing challenges of testing species distribution models. Journal of Applied Ecology **42**,720-730.

Vayssieres MP, Plant RE, Allen-Diaz BH. (2000). Classification trees: An alternative non-parametric approach for predicting species distributions. Journal of Vegetation Science **11**,679-694.

Welk E, Schubert K, Hoffmann MH. (2002). Present and potential distribution of invasive garlic mustard (Alliaria petiolata) in North America. Diversity and Distributions **8**,219-233.

Yee TW, Mitchell ND. (1991). Generalized additive models in plant ecology. Journal of Vegetation Science **2**,587-602.

Zhu L, Sun OJ, Sang W, Li Z, Ma K. (2007). Predicting the spatial distribution of an invasive plant species (*Eupatorium adenophorum*) in China. Landscape Ecology **22**,1143-1154.

Zweig MH, Cambell G. (1993). Reciever-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clinical Chemistry **39**,561-577.

Table 1. A typical confusion matrix. TP = true presence; FP = false presence; FA = false absence; TA = true absence.

		Actual	
		present	absent
Predicted	present	TP	FP
	absent	FA	TA

Table 2. A quick comparison of common modeling techniques in the current literature.

	Statistical Method	Categorical data	Assumptions	Good with small samples?	Different Weights for predictor variables?	Implementation	Response data requirements	Misc Comments
Envelope Models	transparent	no		no	no	Bioclim software	uses only presence	consistently performs poorly in model comparisons
BRT	machine learning - mysterious	yes	non-parametric	yes	yes	Package for R	presence and absence, or abundance	newer technique that is performing well
Domain	interpretable	yes		yes	no	ArcGIS tool	uses only presence	consistently performs poorly in model comparisons
CCA	transparent	yes	parametric	neutral	yes	Canoco	presence and absence, or abundance	difficult to create a predictive map from results
CART	interpretable	yes	non-parametric	yes	yes	Package for R	presence and absence, or abundance	allows interaction of predictors
ENFA	transparent	yes	assumes unimodal relationship	no	?	Package for R	uses only presence	no longer regularly used
GARP	machine learning - mysterious	yes, but hasn't been thoroughly tested		yes	?	Desktop GARP	uses only presence	consistently performs poorly in model comparisons
Maxent	machine learning - mysterious	yes		yes	yes	Stand alone software	uses only presence	newer technique that is performing well
Random Forest	machine learning - mysterious	yes	non-parametric	yes	yes	Package for R	presence and absence, or abundance	can deal with uneven prevalence
Regression - GLM	transparent	yes	parametric, linear relationships	neutral	yes	Package for R	requires presence and absence	well studied model that performs moderately well
Regression - GAM	transparent	yes	complex relationships	no	yes	Package for R	requires presence and absence	improves on GLM, but may overfit

Figure 1. A unimodal response curve of a species along an environmental gradient may appear linear when only a portion of the range is examined (dashed box).

Figure 1

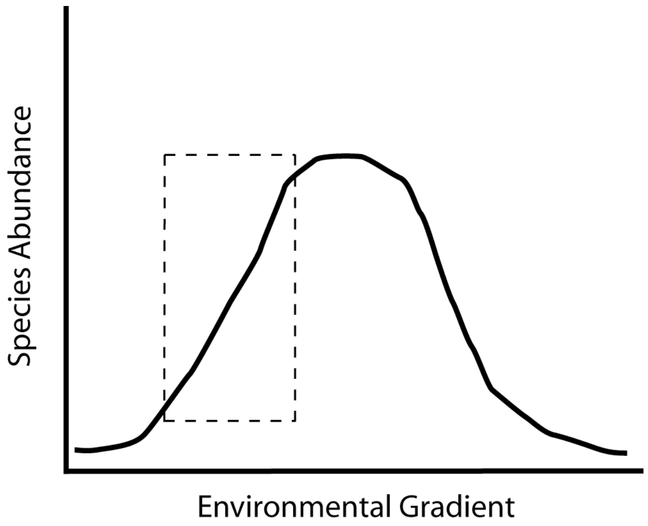
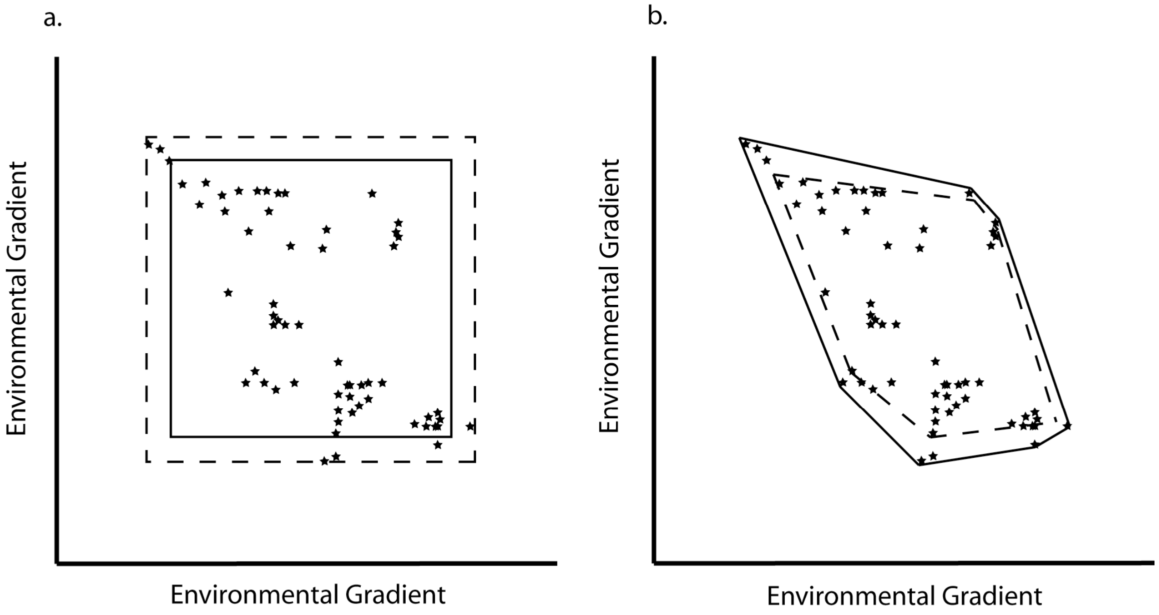


Figure 2. This example illustrates a simple two variable climate envelope model. The stars represent values of the environmental variables for individual species observations. The area within the solid box is the core environment and the dotted line is the boundary of the marginal environment. The original envelope model is constrained to a box or rectangular shape (a); however, the convex hull may be an irregular polygon (b).

Figure 2


a.
b.
Environmental Gradient
Environmental Gradient
Environmental Gradient
Environmental Gradient

CHAPTER 3

Predicting Areas of Suitable Habitat for the Endangered American Burying Beetle (*Nicrophorus americanus*) in Oklahoma.

Priscilla H. C. Crawford[*]

& Bruce W. Hoagland[*†]

[*]Oklahoma Biological Survey, University of Oklahoma,
111 E. Chesapeake St., Norman, OK 73019, USA;
[†]Department of Geography, University of Oklahoma,
Norman, OK 73019, USA

formatted for submission to Conservation Biology

INTRODUCTION

The goal of the Endangered Species Act is to improve the chances of listed species' survival by increasing population levels, as outlined in an endangered species recovery plan (US Fish and Wildlife Service 1991). If successful, this can result in a species being delisted, but in order to achieve the goal of species recovery the demography, habitat preferences, reproductive biology, and cause of the species decline must be understood. However there are disparities in the level of available knowledge for threatened and endangered species. For example, considerable information has been compiled on the status and life history of species such as the Red-cockaded Woodpecker or Mexican grey wolf, but less in known about the Soccoro springsnail or rock gnome lichen (US Fish and Wildlife Service 2009).

The American burying beetle (*Nicrophorus americanus*) was listed as an endangered species in 1989 (Federal Register 54 (133): 29652-29655). Like many threatened and endangered invertebrates, information about *N. americanus* prior to listing consisted of the taxonomic description and morphological characterization (US Fish and Wildlife Service 1991, 2009). Although 1000s of surveys across the United States conducted since listing have contributed to our knowledge of *N. americanus*'s range and populations, they focused on determining species presence and have minimally contributed to our knowledge of its habitat affinities and reproductive biology. Research conducted since its addition to the endangered species list has focused on the breeding season and over-wintering habitat preferences (Creighton et al. 1993b; Lomolino & Creighton 1996; Lomolino et al. 1994; Schnell et al. 2007), population dynamics (Bedick et al. 1999; Holloway & Schnell 1997; Peyton 2003; Raithel et al. 2006), and best survey practices (Bedick et al. 2004; Creighton et al. 1993a). However, we believe much remains to be discovered about the reproductive and over-wintering requirements of *N. americanus.*

N. americanus was once considered common throughout eastern North America (US Fish and Wildlife Service 1991), but at the time of its listing, the range had been reduced to two disjunct populations; one on an island off the coast of Rhode Island and another in eastern Oklahoma. Surveys throughout the historic range since listing have located extant populations in central Nebraska, south-central South Dakota, southeastern Kansas, western Arkansas, and northeast Texas (US Fish and Wildlife Service 1991). Populations in the historic range east of the Mississippi River have not been found.

Endangered species are generally rare for one of two reasons: they were always rare due to habitat specialization or restricted endemism or their population size was substantially reduced due to habitat loss or catastrophic events (Rosenzweig & Lomolino 1997). The cause of *N. americanus* population and range decline over the past 100 years remains uncertain. Sikes and Raithel (2002) presented the following eight possible causes for *N. americanus* decline: pesticide use, artificial lighting, pathogen, habitat loss, vegetation change (both as an old growth woodland specialist or prairie specialist), vertebrate competition, loss of ideal carrion, and congener competition. Of those, they conclude that the most plausible explanation is competition with congeners and vertebrates for carrion and a reduction in optimal prey size. Schnell et al. (2007) suggest that availability of food, in the form of a carcass, during over-wintering will significantly affect the survival of individuals.

Extensive surveys for *N. americanus* within its historic range provide much data that can be integrated into spatial models to help predict suitable habitat. Georeferenced species data can be combined with GIS layers of environmental data in habitat suitability models to generate predictions of areas of suitable habitat within the presumed range of *N. americanus*. Species distribution models (SDM, also known as habitat suitability or ecological niche models) are used to understand species' distributions (Anderson 2003; Camarero et al. 2005; Van Mannen et al. 2002), ecological requirements (Costa et al. 2007; De'ath 2002; Laurent et al. 2004; Murphy & Lovett-Doust 2007; Norris et al. 2006), locate new populations (Pearson et al. 2007; Peppler-Lisbach & Schräder 2004), plan land conservation (Buechling & Tobalske 2007; Danks & Klein 2002; Ortega-Huerta & Peterson 2004; Rodríguez et al. 2007; Tole 2006), and predict new habitats associated with climate change (Berry et al. 2002; Pearson et al. 2006). SDMs correlate species occurrence data with environmental data to produce a predictive map of a species potential distribution or suitable habitat. Different modelling techniques utilize a variety of algorithms to calculate probabilities that a species will occupy a given area. The efficacy of an algorithm to accurately predict the presence or absence varies based on the quantity and quality of species data and the specificity of its environmental requirements. The vast and growing literature on distribution modelling suggest that some techniques are generally more effective, but there is not one algorithm applicable to all species, all data sets, or all research objectives (Elith et al. 2006; Guisan et al. 2006; Pearson et al. 2006; Rushton et al. 2004).

A nearly straight north-south line bisecting the eastern third of Oklahoma demarcates the southwest edge of the range for *N. americanus* (Fig. 1). Using specific location information coupled with environmental data, we hope to delineate a less generalized map for potential *N. americanus* habitat and to understand the constraints on the range. Currently sufficient information is not available for optimal conservation planning for *N. americanus*. Modelling may clarify habitat characteristics and focus conservation efforts.

Our objective in modeling the potential distribution of *N. americanus* is to evaluate the suitability of these models for generating maps of potential habitat, thus focusing survey and recovering efforts as well is contributing to the knowledge of this species ecology. The purpose of this study is to evaluate the ability of current modelling techniques to predict suitable habitat for *N. americanus* using presence-absence data from species observations and surveys. Modelling will facilitate the location of highly suitable habitat, assist in defining and managing conservation lands for *N. americanus*, and help to assess the likely presence of the species prior to surveys. We have chosen to compare six modelling algorithms that utilize both presence and absence data. Although techniques that use absence data have been shown to perform better when absence information is available, we suspect the absence data for the beetle surveys may not truly represent habitat that is unsuitable for *N. americanus*.

METHODS

Study Area

The study area is the eastern half of Oklahoma, a state in the south-central USA. Elevation within this area ranges from 87 m to 806 m with major topographic features including the Ouachita Mountains in the southeast and the Ozark Plateau in the northeast. The natural vegetation of this region is primarily oak-hickory, oak-pine, or post oakblackjack oak forest (Hoagland 2000). The geomorphic provinces represent a variety of surface geology from resistant sandstone and limestone to soft sands, clays, and gravels (Curtis et al. 2008). Oklahoma has a strong longitudinal and latitudinal gradient in both precipitation and temperature. Average annual temperature ranges from 16.2° C in the southeastern corner of the study area to 14.4° C in the northwest with the growing season ranging from 201-222 days. The coldest month is January with an average temperature in the southeast being 4.1° C and in the northwest being 1.6° C. The warmest month for the study area is July with an average temperature in the southeast being 26.9° C and in the northwest being 27.7° C. Average annual precipitation within the study area ranges from 54.2 cm in the southeast to 33.4 cm in the northwest, with the wettest month being May for all areas (Brock et al. 1995).

Study Species and Data Set

N. americanus is the largest species (approximately 2.5-3.5 cm adult length) within the *Nicrophorus* genus, a group of beetles that bury vertebrate carcasses on which to raise their young (Lomolino et al. 1994). Both parents care for the offspring on the underground brood carcass with secretions that apparently slow decay while feeding the larvae regurgitate and protecting them from predators. The young require 48-60 days to develop and surface as teneral adults in July and August. Adults over-winter underground beginning in late September and emerge in April during spring. Adults are nocturnal and require warm nights of 15.5°C for activity (US Fish and Wildlife Service 1991).

The *N. americanus* data set was compiled from records provided by the U. S. Fish and Wildlife Service Tulsa Ecological Services Field Office and the Oklahoma

Biological Survey. The data set contained records from both opportunistic collections and standardized transect surveys gathered from 1979 to 2008. Presence of *N. americanus* may have been recorded with either method, but absence was only recorded when the species was not collected during a standardized survey. Standardized surveys are series of carrion traps along a 20 m transect that is maintained for three rainless nights with temperatures above 15.5°C [for survey details see (US Fish and Wildlife Service 1991, 2007)]. Biologists permitted by the U.S. Fish and Wildlife Service conducted the surveys, of which a majority were located in areas of road or pipeline construction.

Multiple surveys were conducted at some locations over the course several years. Surveys at one location may be both positive or negative over time. Therefore records were analyzed to determine the repeatability of the results at one site. Based on the likelihood that a site with a positive observation had subsequent positive observations in following years, a location was considered positive if any survey conducted at the site yielded a positive beetle observation. We tested for spatial autocorrelation in the *N. americanus* data set with Moran's *I* (Rangel et al. 2006).

Predictor Variables

In previous research, *N. americanus* has been found to be a generalist species (Bedick et al. 1999; Holloway & Schnell 1997; Lomolino et al. 1994), and it is unclear which environmental variables are important in determining its distribution. Therefore, we chose a variety of predictor variables that we believe are likely to affect a burrowing insect. These predictor variables fall into three major categories: topographic, vegetation and landcover, and climatic (Table 1).

Some research indicates that *N. americanus* may be found more often in certain types of habitat. Creighton et al. (1993b) found that *N. americanus* are more likely to be found in oak-hickory forest than other habitat types in eastern Oklahoma. To include vegetation type in the models we used potential natural vegetation, vegetation, forest cover, landcover, and landcover change. Also, preliminary work points to soil texture being an important factor in burying beetle habitat choice (Schnell et al. 2007; Smith 2007). Consequently we included in the models soil association obtained from the STATSGO data set (Soil Survey Staff 2005). Additionally, geologic data were included in the predictor variable set because, similar to soil type, surface and subsurface geology may affect the beetles ability to bury carrion and raise young underground.

Many insects are significantly affected by local climate variation. We included several climate variables in the models which were obtained in point format from the Oklahoma Mesonet administered by the Oklahoma Climatological Survey (Brock et al. 1995). These data were interpolated by simple kriging, except days below freezing which was determined by universal kriging 50% local. Topographic data were obtained from the Digital Elevation Model (DEM) of Oklahoma derived from 1:100,000-scale digital topographic maps. Slope and elevation, which influence microclimate of the area, were included in the models.

We attributed values for all predictor variables to each species data point. To accomplish this, all predictor variables were converted into raster format with 60 m grid cell resolution. Models were run initially with all predictor variables. However, some modelling techniques, particularly regressions, are significantly affected by correlation among the predictor variables. Therefore we ran bivariate correlations to determine which variables were highly correlated prior to a second round of model building.

Among those variables that were highly correlated, we conducted logistic regressions of each variable with the species data set to determine which variable had a greater effect on *N. americanus* occurrence. The variable within each correlated group that had the greatest effect on the species was kept for a second round of model building.

Because many modelling techniques, especially regression based techniques, are negatively affected by an unequal ratio of presence and absence data (Manel et al. 2001), we randomly removed absence data points until the number of absence and presence was approximately equal. The final species data set used for modelling contained 426 locations with 203 presence and 223 absence points.

Modelling Techniques

We used six modelling techniques to create predictive models of habitat suitable for *N. americanus*. Many researchers suggest comparing the results of several techniques because no one method has proven to be the best for all species and study areas (Elith et al. 2006; Guisan et al. 2006). We wanted to compare methods that were based on traditional statistics and machine learning; and methods that utilized absence data and generated pseudo-absence data.

Generalized Linear and Generalized Additive Models

Generalized linear models (GLMs) and generalized additive models (GAMs) are applied extensively in species distribution modelling because of their statistical power and their potential to realistically model species–environment relationships (Austin 2002; Guisan et al. 2002; Yee & Mitchell 1991). GLMs are parametric techniques that assume a linear relationship, which may not always be ecologically appropriate. However, at finer scale or at the edge of a species range a linear relationship may be the best representation of the relationship (Austin 2002). GAMs are considered more ecologically realistic then GLMs because they use non-parametric functions that are more capable of modelling complex response–predictor relationships (Guisan et al. 2002). Although GAMs may create models that fit the training data better than GLMs, there is a cost. When validated with independent evaluation data, GAMs do not perform as well because of over-fitting, which limits the transferability of the model to different areas or time periods (Randin et al. 2006). GAMs require a large training data set to produce an accurate model because of the complexity of the algorithm used to determine the shape of the species–environment relationship (Yee & Mitchell 1991). GLM and GAM models require absence data and results can be affected by an uneven ratio of presence and absence points. For our model building, it was necessary to reduce the number of absence points from the data set to achieve an appropriate presence–absence ratio. Both models were implemented in R using the BIOMOD package (Thuiller 2003).

Regression Trees

Classification and Regression Tree (CART) methods divide the training data iteratively into two sub-sets based on the environmental variable that best reduce the variance in the response variable. A tree is constructed by further divisions causing dichotomous branching for each split of the data. This continues with all new sub-sets until all occurrences have been classified. The branches of the classification tree can lead to presence or absence points based on the environmental variable used to sort the data (De'ath & Fabricius 2000). CART was implemented in R using the BIOMOD package (Thuiller 2003)

Random Forest is a form of CART that increases the power of the classification tree by generating multiple models from repeatedly sub-sampled training data sets (bootstrapping). The multiple models grow a “forest” of trees of which each tree is “grown” from a randomized subset of environmental variables. Each species data point is classified by all trees in the “forest.” The classification backed by the greatest number of trees becomes the value for the data point (Breiman 2001). Although increasing the number of trees does not appear to increase over-fitting in Random Forests (Prasad et al. 2006), it does complicate model interpretability (De'ath 2002; Prasad et al. 2006). Random Forest was implemented in R using the BIOMOD package (Thuiller 2003).

The Generalized Boosted Method (GBM, also known as Boosted Regression Trees) is an advanced CART method that incorporates the regression tree algorithm with a boosting algorithm that combines and summarizes a collection of many — 100s to 1000s — trees. In contrast, conventional regression, CART, and Random Forest methods find a single tree or model that is the best fit. Boosting works on the premise that “it is easier to find many rough rules of thumb than it is to find a single highly accurate prediction rule” (Schapire 2002). The boosting procedure builds many models then combines them to produce an average model. A basic CART method, because of its dichotomous nature does not easily represent a smooth response curve (Austin 2002), but the addition of the boosting algorithm enables the GBM models to better represent smooth species response curves by averaging many trees (Elith et al. 2008). The GBM models are also able to represent non-linear relationships and interactions between predictor variables (Elith et al. 2008). Although GBM modelling could be considered a “black-box” method, as many other machine learning methods have been labeled, it appears of model comparisons that GBM results are ecologically sensible and were accurate representations of species distributions (Elith et al. 2006). We implemented GBM using ‘gbm’ in the BIOMOD package in R (Ridgeway 2006; Thuiller 2003).

Maxent

Maximum entropy (Maxent) is a machine learning method that is able to make predictions using presence only data. Like other machine learning techniques, Maxent improves the modelling algorithm automatically through a series of trainings with the data set. Maxent estimates the spatial distribution of the presence points with maximum entropy (the most uniform or spread out distribution) given the constraints put on the distribution with respect to the point’s relationship to the environmental layers. This relationship is quantified using the empirical average of the environmental variable for all presence points (Phillips et al. 2006). Although Maxent was designed to use presence-only data, it also performs well when compared to presence–absence procedures that utilize both real and pseudo-absence data (Elith et al. 2006; Hernandez et al. 2006; Pearson et al. 2007). We chose to use Maxent because of its superior performance in model comparisons despite the availability of absence data for *N. americanus*. We implemented Maxent with stand-alone software (Phillips et al. 2006; Phillips & Dudik 2008).

Model Evaluation

We used the threshold independent method, receiver-operating characteristic curve (ROC) to evaluate all models. The area under the curve (AUC) of a ROC plot has been widely recommended to assess the predictive performance of species distribution models (Barry & Elith 2006; De'ath & Fabricius 2000; Elith et al. 2006; Ferrier & Guisan 2006; Fielding & Bell 1997; Guisan et al. 2007; Rushton et al. 2004). AUC is calculated by plotting sensitivity against 1-specificity for all possible thresholds. Sensitivity is the likelihood that a predicted presence point should actually be absent. While, specificity is the likelihood that a predicted absence point should really be classified as present. AUC values range from 0 to 1; with 0.5 being random performance and values near 1 being good predictive performance (Fielding & Bell 1997). An index has been developed for AUC values: 0.5-0.7 = low accuracy; 0.7-0.9 = potentially useful; and > 0.9 high accuracy (Swets 1988). Models were evaluated by calculating the AUC for the evaluation data set which was 25% of the species data points held out from the original species data set.

RESULTS

Species Data Set

From 1979 to 2008, 1182 surveys for *N. americanus* were conducted across the eastern third of Oklahoma with 1089 surveys conducted in the past 10 years (Fig. 1). Of those, 230 (20%) of the surveys collected at least one *N. americanus* specimen. Of the total number of surveys, 72 locations were surveyed more than once, representing 173 survey events (15%). Of the 72 locations, 29 were negative for all surveys; 28 were positive for all surveys; 15 of the locations had surveys of both negative and positive results. We considered the 15 locations with conflicting survey results as positive. Spatial autocorrelation of presence and absence was weak for neighboring data points and became 0 at a distance of 84 km (Table 2, Fig. 2).

Predictor Variables

Eight environmental variables were removed for a second round of model building due to high correlation (Table 1). Three of the categorical landcover and vegetation layers were highly correlated and two were removed. Landcover was retained. Six climatic variables were removed leaving annual temperature, days below freezing, and May precipitation.

Model Comparison

Ten of the twelve models performed within the AUC index category of “potentially useful“ with an AUC value between 0.7-0.9 (Table 3). As expected, removing correlated variables improved the performance of GLM, GBM, and GAM, and also improved the Random Forest model. The model with the best performance was Maxent using all the predictor variables (AUC 0.857). Other models with AUC values in the “useful” category were Random Forest, GBM, and Maxent - all which used the smaller set of predictor variables (Table 3).

The map of the best Maxent model indicates that *N. americanus* is more likely to be present in the northern part of the southern half of the study area (Fig. 3), with small areas in the far north and southeast. May precipitation, geology, days below freezing, annual temperature, and last day of growing season were accounted for the highest gain in AUC in the Maxent jackknife test of variable importance. Slope was the only variable responsible for reducing model performance.

Of the other model predictions, the spatial representation of CART and Random Forest appear to have the most agreement with the best Maxent model. Both CART and Random Forest predict greatest habitat suitability in the lower middle of the study area, but also indicate suitable habitat in the far north and southeastern corner. However, none of the model predictions were obviously different from the Maxent predictive map (Fig.

3).

DISCUSSION

Even the best performing models did not fall in to the highly accurate category (AUC ≥ 0.9). Several factors may have inhibited predictive performance. The errors in model building generally fall into two categories: data deficiencies, in both species and predictors, and incorrect model specifications (Barry & Elith 2006). Let us first consider model specifications and parameterization. The variation in model output for *N. americanus* is consistent with other studies comparing these modelling techniques (Elith et al. 2006; Hernandez et al. 2006; Loiselle et al. 2003; Meynard & Quinn 2007; Muñoz & Felicísimo 2004; Pearson et al. 2006). GAM and GLM were two of the worst performing models — both techniques utilized absence data from the *N. americanus* surveys and are known to be significantly affected by spatial autocorrelation (Austin 2002; Diniz-Filho et al. 2008; Dormann et al. 2007; Guisan et al. 2006; Segurado et al. 2006). The spatial autocorrelation for the species data set was low (Table 2), but may have been high enough to affect the model algorithm. It has been suggested that when using these regression techniques that a covariate term be added to account for spatial autocorrelation (Segurado & Araújo 2004). Autoregressive techniques designed to account for spatial autocorrelation can also be used, but have mixed results with models built with presence/absence data sets as compared to those using abundance values. The addition of covariates or using autoregressive techniques do not consistently improve the results of models from binary data (Dormann et al. 2007). The use of ensemble or consensus methods may improve model predictions. By comparing, averaging, and measuring variation in the predictions of multiple modelling techniques, ensemble methods can draw out the correctly predicted areas from several models and indicate areas of uncertainty (Marmion et al. 2009). Ensemble methods have been used for other analyses, but only recently applied to SDM by a few researchers (Araújo & New 2007; Araújo & Rahbek 2006; Araújo et al. 2005; Marmion et al. 2009).

What factors in the species data set may have confounded model predictions? Absence data points from the *N. americanus* surveys may not truly represent unsuitable habitat. Habitat suitability models work on the principle that the observed occurrences of a species reflects the species ecological requirements. Most models rely on the assumption that the organism will be present in suitable habitat and absent from unsuitable habitat — that the species is in equilibrium with its environment. Unfortunately that assumption is often fallacious because organisms can be found and recorded in unsuitable habitat or not found in highly suitable habitat. The current distribution of *N. americanus* is almost certainly not at equilibrium with the environment or the species would occupy more of its historic range. Knowing the cause of the range reduction would help to choose predictor variables that directly affect the current distribution. Methods relying on these absence data will therefore have errors. Techniques that use presence and absence data usually have higher AUC values than presence only methods, but only when true absence data is available (Brotons et al. 2004; Pearson et al. 2006). However, we argue that the absence data for *N. americanus* do not represent true absence, and using it to build the models introduced error into the predictions. If false absences are suspected it is better to use a presence-only method (Chefaoui & Lobo 2008; Hirzel & Le Lay 2008; Jimenéz-Valverde et al. 2008; Pearson et al. 2006). Consequently, Maxent may have performed better because it does not rely on absence data, but uses pseudo-absences or "background" data that characterizes the environment of the entire study area (Phillips et al. 2006).

Although the majority of the data comes from standardized surveys conducted over the past twenty years, we believe there are some problematic features of the data set. The survey method relies on rotten meat to lure insects to a pit fall trap and is likely to attract *N. americanus* to suboptimal habitat. The USFWS provides trap specifications and notes that beetles within a 8 km radius could be attracted to the bait (US Fish and Wildlife Service 2007). Raithel and colleagues (2006) found that *N. americanus* traveled "considerable distances" both on their own or aided by prevailing winds. Bedick and colleagues (1999) found beetles traveled up to 6 km in a breeding season in Nebraska. For other flying invertebrates, such as butterflies, distribution model performance decreases as mobility and flight period increases (Pöyry et al. 2008). Although *N. americanus* are attracted to carrion traps, this does not necessarily signify that the trap location is suitable habitat for reproduction.

Because survey locations were not placed randomly on the landscape or in a strict grid pattern covering the entire region, some geographic biases are apparent in the data. Much of the *N. americanus* survey data was conducted in roadside or pipeline right-of-ways because it was commissioned by agencies prior to construction projects. Therefore a pronounced bias exists in the *N. americanus* data set that may affect model results. Kadmon and colleagues (2004) found that even though woody plant records in Israel had a strong roadside bias, they were able to produce accurate models from the data set. However, their models were built simply from the species data set and included only three climatic variables which they found were only weakly correlated. Our models, however, were built using topographic and landcover features that may be more highly correlated to road networks.

Species life history characteristics can affect the accuracy of a model. *N. americanus* is considered a generalist species and thus has no specialized habitat requirements (Bedick et al. 1999; Holloway & Schnell 1997; Lomolino et al. 1994). Generalist species have proven difficult to model because environmental requirements are not simply correlated to predictor variables unlike species with strong habitat or host specificity (Brotons et al. 2004; Evangelista et al. 2008; Guisan et al. 2007).

The predictive performance of our models may be reduced by not including predictors that directly affect the distribution of *N. americanus*. We used a variety of predictor variables that should influence the distribution of *N. americanus* at several ecological scales. Climatic variables are known to determine the continental or regional distribution of a species. Topographic and landcover variables often affect the species at a finer scale. However, we need to have greater emphasis on predictor variables that directly affect the organism at the sub-state scale. Derived bioclimatic variables, such as evapotranspiration, may make more ecological sense and are more appropriate to the smaller scale than precipitation or temperature considered separately.

Despite the low predictive success of our models, the work we have done suggests future avenues of research that will improve our understanding of the *N. americanus*'s biology and ecology. Maxent's test of variable importance identifies variables that were most responsible for improving the model's performance: May precipitation, geology, days below freezing, annual temperature, and last day of growing season. Number of days below freezing and last day of growing season indicate that environmental conditions during over-wintering may account for part of the species suitable habitat. Over-wintering survival has been studied with regard to habitat type, carrion availability, and depth in soil (Schnell et al. 2007), but another factor may be soil temperature. Although we were able to see a signal on a large scale, the affect of soil temperature on *N. americanus* distribution may be better studied at a smaller scale while taking into consideration the microclimate variation in small study areas. The importance of geology in contributing to model performance indicates that substrate may limit what *N. americanus* finds to be suitable habitat. Substrate will affect the insect's ability to bury carrion and successfully raise a brood. Preliminary results from Smith's (Smith 2007) research indicates that brood carcasses were most likely to be buried in loose soil with low clay content. The addition of an accurate soil texture layer, rather than soil association, may enhance future habitat models.

The model results that indicate increased habitat suitability with increased May precipitation could suggest a physiological effect with over-wintering or brooding carcass decay or may simply be a surrogate for a predictor variable that we did not use. Because of the strong southeast-northwest precipitation gradient in Oklahoma, precipitation may be a surrogate for the distribution of a competitor or prey item. Research into the direct effect of precipitation on *N. americanus* reproduction and over-wintering might prove useful in understanding the current distribution of the species and the possible reasons for the historic range collapse.

Inclusion of biotic interactions such as overlap with competitor distribution and shared resources improve model performance at small and macroscales for a variety of organisms (Araújo & Luoto 2007; Davis et al. 1998; Guisan & Thuiller 2005; Heikkinen et al. 2007; Preston et al. 2008). Indeed, Sikes and Raithel (2002) have hypothesized that competition with congeneric and other scavengers and a reduction in suitably sized carrion affects the distribution and abundance of *N. americanus*. The effect of congeneric competitors on distribution models has been demonstrated for South American pocket mice (genus *Heteromys*) (Anderson et al. 2002). Including competitors of native trees (four species of *Nothofagus*) in New Zealand also produce more accurate species distribution models (Leathwick & Austin 2001). While they indicate that more work needs to be done, they believe that the most plausible cause for *N. americanus* decline is related to a change in these biotic interactions. Habitat fragmentation may be altering the biotic interactions that have led to the decline of *N. americanus*. Holloway and Schnell (Holloway & Schnell 1997) suggest that fragmentation has caused an increase in vertebrate scavengers and a reduction in carrion supply. Bedick et al. (1999) agree, but also found that not all land-cover change is detrimental — agricultural areas can still be suitable habitat for *N. americanus*.

Another challenge for modelers is the inclusion of processes that affect the distribution of a species (Austin 2002; Guisan & Thuiller 2005). *N. americanus* may be directly affected processes ongoing on the landscape, such as: fire, dispersal, and succession. Woody plant encroachment is affecting the *N. americanus* population in the grasslands of Nebraska (Walker & Hoback 2007). Revising the 48 categories of landcover change by grouping types of change that is more likely to *N. americanus* could improve the variable importance in the models. Integrating information of fire history or intervals could not only help improve model performance, but also inform land managers of conservation practices that would increase habitat suitability.

Modelling *N. americanus* only in Oklahoma has allowed us to use a finer scale of environmental variables. We may have compromised the predictive ability of the model by looking at the species at the edge of its western range. More sophisticated algorithms have been developed recently that may be better for modelling species at the edge of the range, where habitat may be suboptimal and the species-environment relationship is skewed compared to the whole range (Braunisch et al. 2008).

CONCLUSIONS

Other researchers have repeatedly encouraged better links from ecological theory and biology of the organism to the model building process (Austin 2002; Austin 2007; Guisan et al. 2006; Guisan & Thuiller 2005). To improve model performance, we should think more carefully about the cause of *N. americanus*'s endangered status and its population shrinkage. Sikes and Raithel's (2002) review concludes that the most plausible explanation for *N. americanus*'s decline is a combination of factors associated with biotic interactions including congener and vertebrate competition and a reduction in optimally sized prey. To improve the models and consequently the recovery effort for the species, we need to take into account these important variables. Creating an accurate spatial layer of this data will be a future challenge.

Our objective was to produce a map of potentially suitable habitat for *N. americanus* that would guide conservation efforts within the state of Oklahoma. Although the model was not highly accurate, the map of suitable habitat can help to inform conservation biologists of areas that have suitable habitat for the *N. americanus*.

Overgenerous models can mislead conservation planners in thinking that more areas are highly suited to the species. It is better to be conservative and find the best areas if resources are limited for planning preserves for the species or are looking for areas of reintroduction (Loiselle et al. 2003). Therefore, we urge caution in interpreting the predictive map. We offer it as a suggestion from which additional research can be done to support or refute our suitability map.

ACKNOWLEDGEMENTS

We thank P.T. Crawford, Oklahoma Biological Survey, and H. Dikeman, U.S. Fish and Wildlife Service, for access to the species data, J. Kelly and M. Patten for important statistical advice, R. Channell and G. Schnell for helpful discussions regarding *N. americanus* and distribution models, and T. Fagin for GIS assistance. C. Vaughn, W. Elisens, and J.S. Greene made comments on earlier versions of the manuscript. D. Arndt of the Oklahoma Climatological Survey was very helpful in compiling the climate data. This work was supported by the Oklahoma Natural Heritage Inventory and the Oklahoma Natural Areas Registry programs.

LITERATURE CITED

Anderson, R. P. 2003. Real vs. artefactual absences in species distributions: tests for *Oryzomys albigularis* (Rodentia: Muridae) in Venezuela. Journal of Biogeography **30**:591-605.

Anderson, R. P., A. T. Peterson, and M. Gomez-Laverde. 2002. Using niche-based GIS modeling to test geographic predictions of competitive exclusion and competitive release in South American pocket mice. Oikos **98**:3-17.

Araújo, M. B., and M. Luoto. 2007. The importance of biotic interactions for modelling species distributions under climate change. Global Ecology and Biogeography **16**:743-753.

Araújo, M. B., and M. New. 2007. Ensemble forecasting of species distributions. Trends in Ecology & Evolution **22**:42-47.

Araújo, M. B., and C. Rahbek. 2006. How does climate change affect biodiversity? Science **313**:1396-1397.

Araújo, M. B., R. J. Whittaker, R. J. Ladle, and M. Erhard. 2005. Reducing uncertainty in projections of extinction risk from climate change. Global Ecology and Biogeography **14**:529-538.

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

Austin, M. P. 2007. Species distribution models and ecological theory: A critical assessment and some possible new approaches. Ecological Modelling **200**:1-19.

Barry, S., and J. Elith. 2006. Error and uncertainty in habitat models. Journal of Applied Ecology **43**:413-423.

Bedick, J. C., B. C. Ratcliffe, and L. G. Higley. 2004. A new sampling protocol for the endangered American burying beetle, *Nicrophorus americanus* Olivier (Coleoptera: Silphidae). Coleopterists Bulletin **58**:57-70.

Bedick, J. C., B. C. Ratcliffe, W. W. Hoback, and L. G. Higley. 1999. Distribution, ecology, and population dynamics of the American burying beetle [*Nicrophorus* *americanus* Olivier (Coleoptera, Silphidae)] in south-central Nebraksa, USA. Journal of Insect Conservation **3**:171-181.

Berry, P. M., T. P. Dawson, P. A. Harrison, and R. G. Pearson. 2002. Modelling potential impacts of climate change on the bioclimatic envelope of species in Britain and Ireland. Global Ecology and Biogeography **11**:453-462.

Braunisch, V., K. Bollmann, R. F. Graf, and A. H. Hirzel. 2008. Living on the edge—Modelling habitat suitability for species at the edge of their fundamental niche. Ecological Modelling **214**:153-167.

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

Brock, F. V., K. C. Crawford, R. L. Elliott, G. W. Cuperus, S. J. Stadler, H. J. Johnson, and M. D. Eilts. 1995. The Oklahoma Mesonet: a technical overview. Journal of Atmospheric and Oceanic Technology **12**:5-19.

Brotons, L., W. Thuiller, M. B. Araujo, and A. H. Hirzel. 2004. Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. Ecography **27**:437-448.

Buechling, A., and C. Tobalske. 2007. Habitat modeling of rare plant species in Pacific Northwest forests. Page 35. Oregon Natural Heritage Information Center, Portland, OR.

Camarero, J. J., E. Gutiérrez, M. J. Fortin, and E. Ribbens. 2005. Spatial patterns of tree recruitment in a relict population of *Pinus uncinata*: Forest expansion through stratified diffusion. Journal of Biogeography **32**:1979-1992.

Chefaoui, R. M., and J. M. Lobo. 2008. Assessing the effects of pseudo-absences on predictive distribution model performance. Ecological Modelling **210**:478-486.

Costa, G. C., C. Wolfe, D. B. Shepard, J. P. Caldwell, and L. J. Vitt. 2007. Detecting the influence of climatic variables on species distributions: a test using GIS niche-based models along a steep longitudinal environmental gradient. Journal of Biogeography **35**:637-646.

Creighton, J. C., M. V. Lomolino, and G. D. Schnell. 1993a. Survey methods for the American burying beetle (*Nicrophorus americanus*) in Oklahoma and Arkansas. Oklahoma Biological Survey, Norman, Oklahoma.

Creighton, J. C., C. C. Vaughn, and B. R. Chapman. 1993b. Habitat preferences of the endangered American burying beetle (*Nicrophorus americanus*) in Oklahoma. The Southwestern Naturalist **38**:275-306.

Curtis, N. M., W. E. Ham, and K. S. Johnson. 2008. Geomorphic provinces of Oklahoma in K. S. Johnson, and K. V. Luza, editors. Earth sciences and mineral resources of Oklahoma. Oklahoma Geological Survey, Norman, OK.

Danks, F. S., and D. R. Klein. 2002. Using GIS to predict potential wildlife habitat: A case study of muskoxen in northern Alaska. International Journal of Remote Sensing **23**:4611-4632.

Davis, A. J., L. S. Jenkinson, J. H. Lawton, B. Shorrocks, and S. Wood. 1998. Making mistakes when predicting shifts in species range in response to global warming. Nature **391**:783-786.

De'ath, G. 2002. Multivariate regression trees: A new technique for modeling species-environment relationships. Ecology **83**:1105-1117.

De'ath, G., and K. E. Fabricius. 2000. Classification and regression trees: A powerful yet simple technique for ecological data analysis. Ecology **81**:3178-3192.

Diniz-Filho, J. A. F., T. F. L. V. B. Rangel, and L. M. Bini. 2008. Model selection and information theory in geographical ecology. Global Ecology and Biogeography **17**:479-488.

Dormann, C. F., J. M. McPherson, M. B. Araújo, R. Bivand, J. Bolliger, G. Carl, R. G. Davies, A. Hirzel, W. Jetz, W. Daniel Kissling, I. Kuhn, R. R. Ohlemuller, P. Peres-Neto, B. Reineking, B. Schröder, F. M. Schurr, and R. Wilson. 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography **30**:609-628.

Elith, J., C. H. Graham, R. P. Anderson, M. Dudik, S. Ferrier, A. Guisan, R. J. Hijmans, F. Huettmann, J. R. Leathwick, A. Lehmann, J. Li, L. G. Lohmann, B. A. Loiselle, G. Manion, C. Moritz, M. Nakamura, Y. Nakazawa, J. M. Overton, A. T. Peterson, S. J. Phillips, K. Richardson, R. Scachetti-Pereira, R. E. Schapire, J.

Soberon, S. Williams, M. S. Wisz, and N. E. Zimmermann. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography **29**:129-151.

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

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

Ferrier, S., and A. Guisan. 2006. Spatial modelling of biodiversity at the community level. Journal of Applied Ecology **43**:393-404.

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

Guisan, A., T. C. Edwards, and T. Hastie. 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecological Modelling **157**:89-100.

Guisan, A., J. M. C. Overton, R. Aspinall, T. Hastie, A. Lehmann, S. Ferrier, and M. Austin. 2006. Making better biogeographical predictions of species' distributions. Journal of Applied Ecology **43**:386-392.

Guisan, A., and W. Thuiller. 2005. Predicting species distribution: offering more than simple habitat models. Ecology Letters **8**:993-1009.

Guisan, A., N. E. Zimmermann, J. Elith, C. H. Graham, S. Phillips, and A. T. Peterson. 2007. What matters for predicting the occurrences of trees: Techniques, data, or species' characteristics? Ecological Monographs **77**:615-630.

Heikkinen, R. K., M. Luoto, R. Virkkala, R. G. Pearson, and J.-H. Körber. 2007. Biotic interactions improve prediction of boreal bird distributions at macro-scales. Global Ecology and Biogeography **16**:754-763.

Hernandez, P., C. H. Graham, L. Master, and D. L. Albert. 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography **29**:773-785.

Hirzel, A. H., and G. Le Lay. 2008. Habitat suitability modelling and niche theory. Journal of Applied Ecology **45**:1372-1381.

Hoagland, B. 2000. The vegetation of Oklahoma: A classification for landscape mapping and conservation planning. Southwestern Naturalist **45**:385-420.

Holloway, A. K., and G. D. Schnell. 1997. Relationship between numbers of the endangered American burying beetle *Nicrophorus americanus* Olivier (Coleoptera: Silphidae) and available food resources. Biological Conservation **81**:145-152.

Jimenéz-Valverde, A., J. M. Lobo, and J. Hortal. 2008. Not as good as they seem: the importance of concepts in species distribution modelling. Diversity and Distributions **14**:885-890.

Kadmon, R., O. Farber, and A. Danin. 2004. Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. Ecological Applications **14**:401-413.

Laurent, J. M., R. Cheddadi, A. Bar-Hen, L. François, and M. Ghislain. 2004. Refining vegetation simulation models: From plant functional types to bioclimatic affinity groups of plants. Journal of Vegetation Science **15**:739-746.

Leathwick, J. R., and M. P. Austin. 2001. Competitive interactions between tree species in New Zealand's old-growth indigenous forests. Ecology **82**:2560-2573.

Loiselle, B. A., T. Brooks, K. G. Smith, P. H. Williams, C. A. Howell, C. H. Graham, and J. M. Goerck. 2003. Avoiding Pitfalls of Using Species Distribution Models in Conservation Planning. Conservation Biology **17**:1591-1600.

Lomolino, M. V., and J. C. Creighton. 1996. Habitat selection, breeding success and conservation of the endangered American burying beetle *Nicrophorus americanus*. Biological Conservation **77**:235-241.

Lomolino, M. V., J. C. Creighton, G. D. Schnell, and D. L. Certain. 1994. Ecology and conservation of the endangered American burying beetle (*Nicrophorus americanus*). Conservation Biology **9**:605-614.

Manel, S., H. C. Williams, and S. J. Ormerod. 2001. Evaluating presence-absence models in ecology: the need to account for prevalence. Journal of Applied Ecology **38**.

Marmion, M., M. Parviainen, M. Luoto, R. K. Heikkinen, and W. Thuiller. 2009. Evaluation of consensus methods in predictive species distribution modelling. Diversity and Distributions **15**:59-69.

Meynard, C. N., and J. F. Quinn. 2007. Predicting species distributions: a critical comparison of the most common statistical models using artificial species. Journal of Biogeography **34**:1455-1469.

Muñoz, J., and Á. M. Felicísimo. 2004. Comparison of statistical methods commonly used in predictive modelling. Journal of Vegetation Science **15**:285-292.

Murphy, H. T., and J. Lovett-Doust. 2007. Accounting for regional niche variation in habitat suitability models. Oikos **116**:99-110.

Norris, J. R., S. T. Jackson, and J. L. Betancourt. 2006. Classification tree and minimum-volume ellipsoid analyses of the distribution of ponderosa pine in the western USA. Journal of Biogeography **33**:342-360.

Ortega-Huerta, M. A., and A. T. Peterson. 2004. Modelling spatial patterns of biodiversity for conservation prioritization in North-eastern Mexico. Diversity and Distributions **10**:39-54.

Pearson, R. G., C. J. Raxworthy, M. Nakamura, and A. T. Peterson. 2007. Predicting species distributions from small numbers of occurrence records: A test case using cryptic geckos in Madagascar. Journal of Biogeography **34**:102-117.

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

Peppler-Lisbach, C., and B. Schräder. 2004. Predicting the species composition of Nardus stricta communities by logistic regression modelling. Journal of Vegetation Science **15**:623-634.

Peyton, M. M. 2003. Range and population size of the American Burying Beetle (Coleoptera: Silphidae) in the dissected hills of south-central Nebraska. Great Plains Research **13**:127-138.

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

Phillips, S. J., and M. Dudik. 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography **31**:161-175.

Pöyry, J., M. Luoto, R. K. Heikkinen, and K. Saarinen. 2008. Species traits are associated with the quality of bioclimatic models. Global Ecology and Biogeography **17**:403-414.

Prasad, A. M., L. R. Iverson, and A. Liaw. 2006. Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. Ecosystems **9**:181-199.

Preston, K. L., J. T. Rotenberry, R. A. Redak, and M. F. Allen. 2008. Habitat shifts of endangered species under altered climate conditions: importance of biotic interactions. Global Change Biology **14**:2501-2515.

Raithel, C. J., H. S. Ginsberg, and M. L. Prospero. 2006. Population trends and flight behavior of the American burying beetle, *Nicrophorus americanus* (Coleoptera: Silphidae), on Block Island, RI. Journal of Insect Conservation **10**:317-322.

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

Rangel, T., J. A. F. Diniz-Filho, and L. M. Bini. 2006. Towards an integrated computational tool for spatial analysis in macroecology and biogeography. Global Ecology and Biogeography **15**:321-327.

Ridgeway, G. 2006. Generalized boosted regression models. Documentation on the R Package 'gbm', version 1.6-3 **CRAN Package Library, http://cran.cnr.Berkeley.edu, accessed 2 January 2009**.

Rodríguez, J. P., L. Brotons, J. Bustamante, and J. Seoane. 2007. The application of predictive modeling of species distribution to biodiversity conservation. Diversity and Distributions **13**:243-251.

Rosenzweig, M. L., and M. V. Lomolino. 1997. Who gets the short bits of the broken stick? Pages 63-90 in W. E. Konin, and K. J. Gaston, editors. The biology of rarity. Chapman & Hall, London.

Rushton, S. P., S. J. Ormerod, and G. Kerby. 2004. New paradigms for modelling species distributions? Journal of Applied Ecology **41**:193-200.

Schapire, R. E. 2002. The boosting approach to machine learning: an overview. Pages 1-23. MSRI Workshop on Nonlinear estimation and classification.

Schnell, G. D., A. E. Hiott, J. C. Creighton, V. L. Smyth, and A. Komendat. 2007. Factors affecting overwinter survival of the American burying beetle, *Nicrophorus americanus* (Colelptera: Silphidae). Journal of Insect Conservation.

Segurado, P., and M. B. Araújo. 2004. An evaluation of methods for modelling species distributions. Journal of Biogeography **31**:1555-1568.

Segurado, P., M. B. Araújo, and W. E. Kunin. 2006. Consequences of spatial autocorrelation for niche-based models. Journal of Applied Ecology **43**:433-444.

Sikes, D. S., and C. J. Raithel. 2002. A review of hypotheses of decline of the endangered American burying beetle (Silphidae: *Nicrophorus americanus* Olivier). Journal of Insect Conservation **6**:103-113.

Smith, A. 2007. Camp Gruber American Burying Beetle Reproductive Study. 2007 American Burying Beetle Workshop. U.S. Fish and Wildlife Service, Tahlequah, OK.

Soil Survey Staff. 2005. State Soil Geographic database (STATSGO). U.S. Department of Agriculture, Natural Resources Conservation Service, Fort Worth, TX.

Swets, J. A. 1988. Measuring the Accuracy of Diagnostic Systems. Science **240**:1285-1293.

Thuiller, W. 2003. BIOMOD - optimizing predictions of species distributions and projecting potential future shifts under global change. Global Change Biology **9**:1353-1362.

Tole, L. 2006. Choosing reserve sites probabilistically: a Columbian Amazon case study. Ecological Modelling **194**:344-356.

US Fish and Wildlife Service. 1991. American burying beetle (*Nicrophorus americanus*) recovery plan. Page 80, Newton Corner, MA.

US Fish and Wildlife Service. 2007. American burying beetle, *Nicrophorus americanus*, survey guidance for Oklahoma in F. a. W. S. United States Department of Interior, editor.

US Fish and Wildlife Service. 2009. Endangered Species Program.

Van Mannen, F. T., J. D. Clark, S. E. Schlarbaum, K. Johnson, and G. Taylor. 2002. A model to predict the occurence of surviving butternut trees in the Souther Blue Ridge Mountains. Page 868 in J. M. Scott, P. J. Heglund, and e. al., editors. Predicting species occurences. Island Press, Washington.

Walker, T. L., and W. W. Hoback. 2007. Effects of invasive eastern red cedar on capture rates of *Nicrophorus americanus* and other Silphidae. Environmental Entomology **36**:297-307.

Yee, T. W., and N. D. Mitchell. 1991. Generalized additive models in plant ecology. Journal of Vegetation Science **2**:587-602.

Table 1. Environmental layers used as predictor variables in models of potential habitat suitability of the endangered *Nicrophorus americanus* in eastern Oklahoma.

Variable	**Range & Unit**	**Source**
Elevation	87 - 806 m	Oklahoma Digital Elevation Model (Cederstrand and Rea 1995; geo.ou.edu)
Slope	0 - 46°	Oklahoma Digital Elevation Model (Cederstrand and Rea 1995; geo.ou.edu)
Soil association	228 categories	STATSGO (Soil Survey Staff 2005; soils.usda.gov/survey/geography/statsgo)
Surface geology	133 categories	U.S. Geological Survey (Heran et al. 2003; pubs.usgs.gov/of/2003/ofr-03-247
Vegetation *	34 categories	Oklahoma Gap Project (Fisher and Gregory 2001; www.biosurvey.ou.edu/gap-ok.html)
Potential vegetation *	8 categories	Game Type Map of Oklahoma (Duck and Fletcher 1943; www.biosurvey.ou.edu/duckflt/dfhome.html)
Landcover	15 categories	National Land Cover Database (www.mrlc.gov)
Forest cover	0 - 100 %	
Landcover change	48 categories	
Annual temperature	14.4 - 16.2° C	Oklahoma Climatological Survey Oklahoma Mesonet (Brock et al. 1994; www.mesonet.org)
Number of days below freezing (0° C) *	57 - 93 days	
Number of days above 32.2° C *	56 - 85 days	
Length of growing season *	201 - 222 days	
First growing season day *	87th - 97th day of year	
Last growing season day	299th - 310th day of year	
Annual precipitation *	32.5 - 55.5 cm	
May precipitation	4.8 - 6.7 cm	
September precipitation *	3.4 - 5.6 cm	

Table 2. Analysis of spatial autocorrelation of *Nicrophorus americanus* occurrence records in Oklahoma. The average Moran's *I* is given for 16 distance classes. Values for I can range from -1 to 1; values close to 1 indicate a positive spatial autocorrelation and negative values a negative spatial autocorrelation. Spatial autocorrelation is low at the closest distances and approaches 0 at 84 km.

Average Paired Distance (km)	Moran's *I*	*I* (max)
15.4	0.23 ± 0.012 *	0.592
39.3	0.176 ± 0.013 *	0.523
55.7	0.054 ± 0.013 *	0.401
70.5	0.065 ± 0.013 *	0.371
84.1	0.01 ± 0.013	0.333
96.8	-0.001 ± 0.013	0.343
108.3	0.011 ± 0.013	0.323
119.0	-0.051 ± 0.013 *	0.324
129.9	-0.093 ± 0.013 *	0.360
141.2	-0.124 ± 0.013 *	0.391
153.1	-0.142 ± 0.013 *	0.439
167.0	-0.157 ± 0.013 *	0.456
183.7	-0.118 ± 0.013 *	0.468
204.6	-0.093 ± 0.013 *	0.486
233.7	-0.011 ± 0.012	0.500
320.7	0.206 ± 0.011 *	0.717

* $p < 0.001$

Table 3. Performance of different modelling techniques for *Nicrophorus americanus* using all available predictor variables and a reduced set of variables based on variable correlations. AUC value of 0.5-0.7 is considered low accuracy; 0.7-0.9 is considered useful; and 0.9 and above is considered high accuracy. Models were evaluated with 25% holdout data from the occurrence data set. Classification and regression tree, CART; generalized additive model, GAM; generalized boosted model, GBM; generalized linear model, GLM; maximum entropy, Maxent.

	All Predictors	Correlated Predictors Removed
CART	0.726	0.688
GAM	0.780	0.802
GBM	0.765	0.813
GLM	0.674	0.731
Maxent	0.857	0.831
Random Forest	0.792	0.834

Figure 1. Occurrence records of *Nicrophorus americanus* in Oklahoma, south-central United States, used in habitat suitability modelling. Presence records are indicated with circles, absences with small crosses (+). To the east of the black line indicates the historic range within Oklahoma.

Figure 1

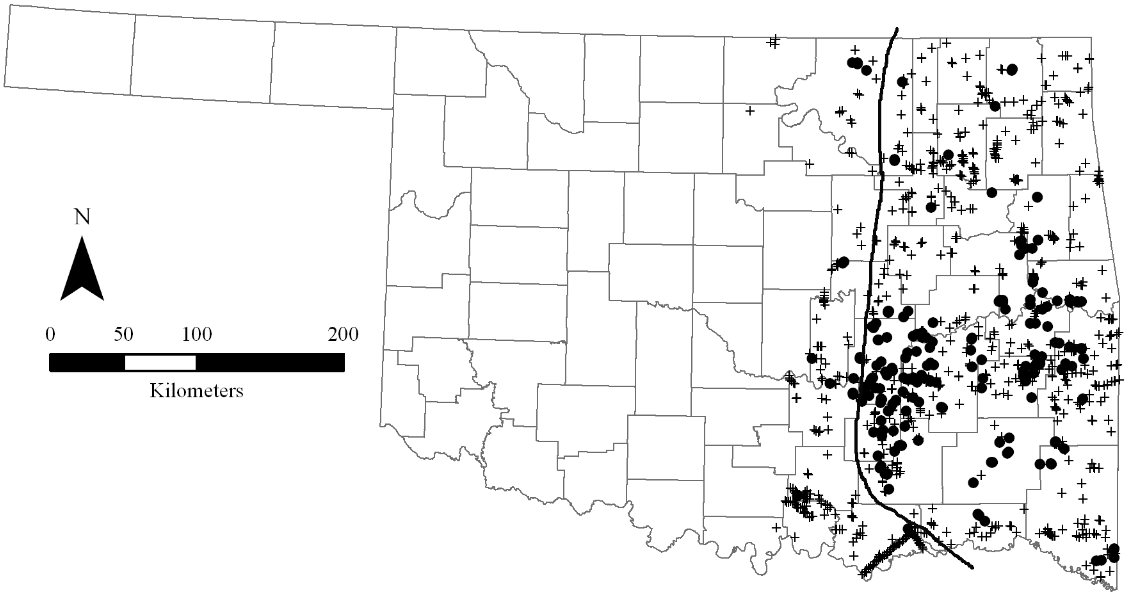


Figure 2. Spatial correlograms of *Nicrophorus americanus* occurrences in Oklahoma. Circles indicate the Moran's *I* for each distance pair. Triangles are the highest Moran's *I* value for each distance class.

Figure 2

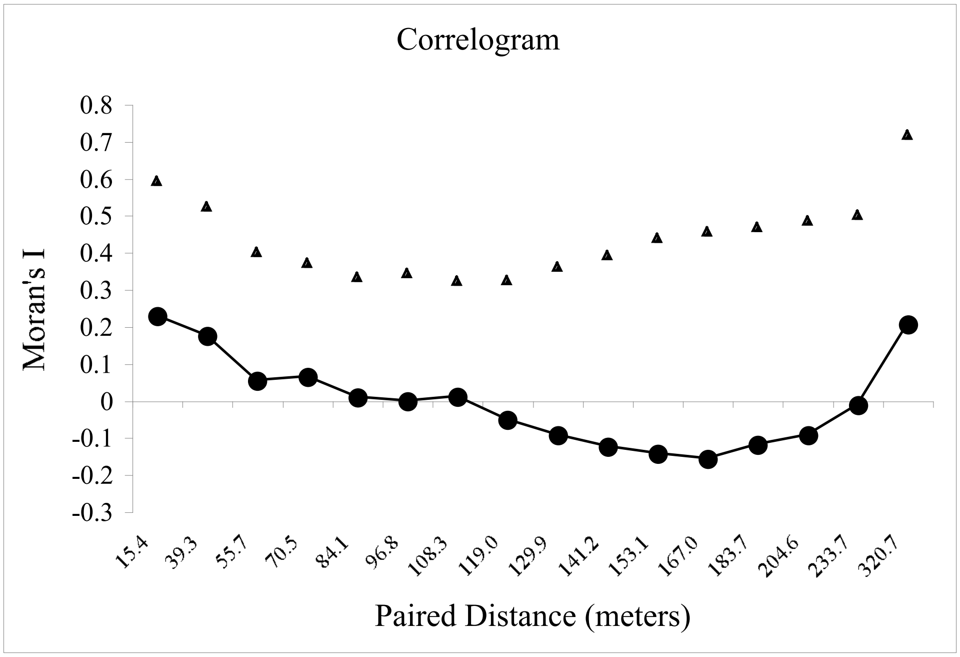
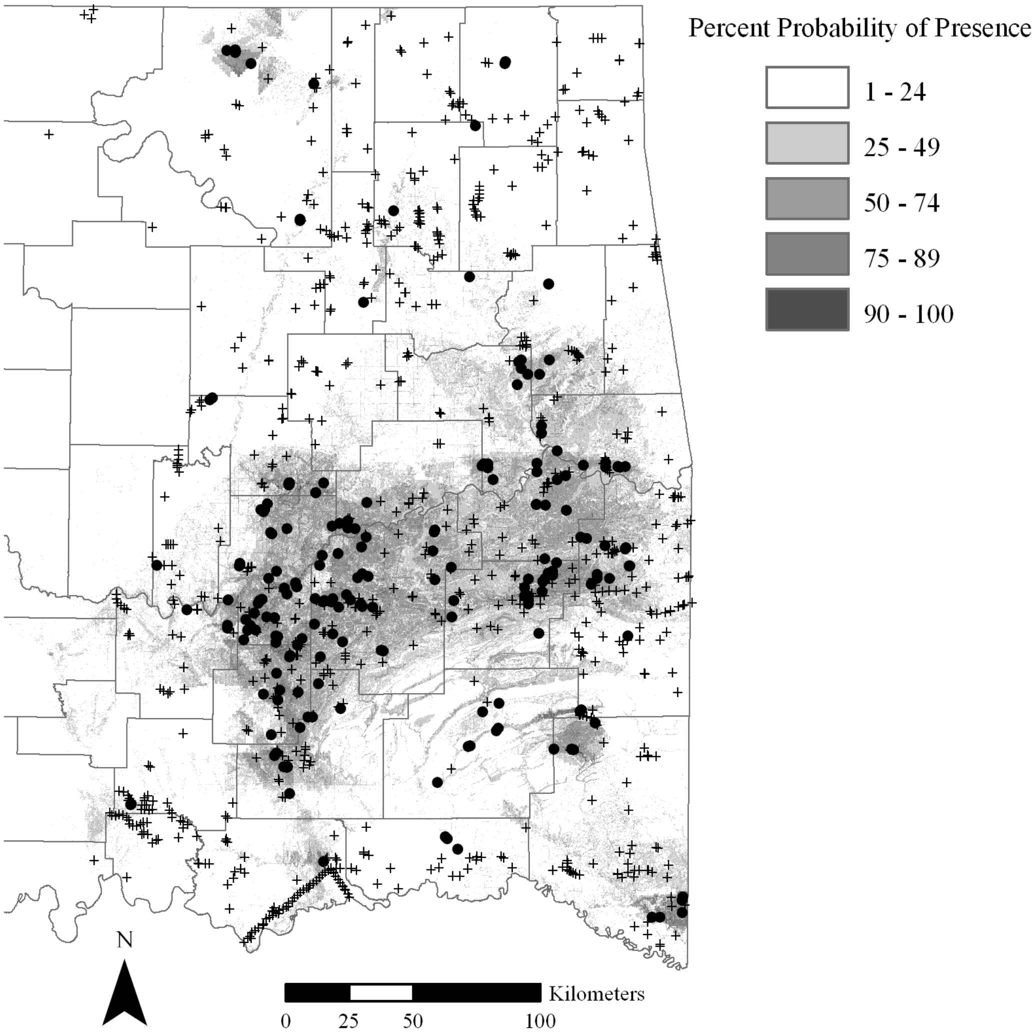

Correlogram
Moran's I
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
15.4
39.3
55.7
70.5
84.1
96.8
108.3
119.0
129.9
141.2
153.1
167.0
183.7
204.6
233.7
320.7
Paired Distance (meters)

Figure 3. Predicted habitat of *Nicrophorus americanus* in eastern Oklahoma based on the Maxent model using all predictor variables. This modelling technique produced the most accurate model of all techniques tested, with an AUC value of 0.857. Circles indicate known presence locations of *Nicrophorus americanus* and small crosses (+) indicate surveys that found no *Nicrophorus americanus*.

Figure 3


Percent Probability of Presence
1 - 24
25 - 49
50 - 74
75 - 89
90 - 100
N
Kilometers
0
25
50
100

CHAPTER 4

Models of climate suitability based on the native range of invasive plants cannot be transferred to the introduced range

Priscilla H. C. Crawford [1, *] & Bruce W. Hoagland [1, 2]

[1] Oklahoma Biological Survey, University of Oklahoma,
111 E. Chesapeake St., Norman, OK 73019, USA;
[2] Department of Geography, University of Oklahoma, Norman, OK 73019, USA

* Correspondence: Priscilla H. C. Crawford, Oklahoma Biological Survey,
University of Oklahoma, 111 E. Chesapeake St., Norman, OK 73019, USA.
E-mail: prill@ou.edu.

formatted for submission to Diversity and Distributions

INTRODUCTION

Species' ranges are controlled by environmental tolerances, biotic interactions, dispersal limitations, and historical factors. Boundaries such as mountains, rivers, oceans, and deserts limit the geographic range of many species, not because suitable habitat is limited, but barriers prevent the continued movement of species. However, humans have eliminated those barriers by accidentally or deliberately transporting species around the world (Vitousek, D'Antonio et al. 1997). If the habitat is suitable, alien species can survive and thrive in their introduced range and possibly become invasive. The term invasive has been used in a variety of ways, but we use the strict definition of invasive species to mean alien species that have spread over a considerable area after introduction from another region by humans (Richardson, Panetta et al. 2000).

The impact of invasive species is multifaceted, both from an ecological and societal perspective. Following introduction and establishment, invasive alien species can have significant ecological and economic impacts. The ecological impact of invasive species has been well documented: alteration of disturbance regimes, decline in native species abundance, nutrient cycles shifted, epidemics caused by new parasites, food web shifts, and others [for reviews see (Vitousek, D'Antonio et al. 1997; Mack, Simberloff et al. 2000)]. The economic impact of alien invasive species is best illustrated by the estimated amount expended for invasive species management each year, $137 billion in the United States alone (Pimentel, Lach et al. 2000). These expenses could be ameliorated by an effective early detection and eradication system (Hobbs and Humphries 1995; DiTomaso 2000) and early detection can be improved by identifying the potential invasive species, recognizing the likely mode of transportation into the area, and determining the potential habitat available. It is the last of these topics that we address with our research: identifying areas that are susceptible to invasion by known invasive species. Species distribution modelling can identify areas in the introduced range that have environments similar to the native range. Locating geographic areas in the introduced range that have the same fundamental niche space is the first step in mapping potential areas of invasion.

Species Distribution Modelling

Species distribution models (SDM) correlate data for known species occurrences with environmental data to produce a predictive map of the range within a study area. These models are predicated on the assumption that species populations are at equilibrium with the environment; that is, the species should be found in all suitable areas and not occupy unsuitable habitat (Hutchinson 1957) [but for a discussion on how realistic this assumption is see (Araújo and Pearson 2005)]. This assumption is problematic when attempting to model the potential range of alien, invasive species, which by definition are continuing to expand both in geographic area and abundance. Therefore, a model built with occurrence data from the introduced range may under-predict the distribution of a species that has not reached equilibrium with its environment. Thus, it has been necessary to modify distribution modelling techniques to more accurately model the potential distribution of alien species. Recent attempts to overcome the limitations of the assumption of equilibrium have been to develop distribution models based on data from the species' native range and use the result to project the potential habitat suitability onto the introduced region (Peterson 2003; Nyári, Ryall et al. 2006; López-Darias, Lobo et al. 2008). These models are generally built using coarse scale climate data from the native range and the suitable climate parameters are projected onto the introduced range.

Reciprocal Modelling

The assumption of distribution modelling of invasive species using native range occurrences is that the niche occupied in the native range will be similar to the one occupied in the introduced range (Peterson and Vieglais 2001; Pearman, Guisan et al. 2008). Comparing model predictions of an invasive species in both its native and introduced range is a test of that assumption. The ability of a model's predictions to be transferred from one region to another has been examined for species within their native range (Randin, Dirnbock et al. 2006; Barbosa, Real et al. 2009) and to introduced ranges (Mau-Crimmins, Schussman et al. 2006; Fitzpatrick, Weltzin et al. 2007). In general, these studies found that models predictions did not transfer well. When modelling the potential distribution of an alien, invasive species in a new area, many species distribution models go no further than to project the range of suitable environmental variables in the native range onto the introduced range. Reciprocal modelling, on the other hand, predicts not only the potential invaded range based on the environmental characteristics of the native range, but it also uses occurrence data from the invaded range to predict the native range. The results can then be used to evaluate habitat discrepancies or potential niche shifts (Fitzpatrick and Weltzin 2005; Fitzpatrick, Weltzin et al. 2007; Loo, Nally et al. 2007). Fitzpatrick and Weltzin (2005) proposed and demonstrated the use of reciprocal modelling as a new method of studying the prediction errors of invasive species distribution models. This modelling approach can help to determine which environmental factors within the introduced range are different from the native range and which habitats within the native range are not represented in the introduced range. Further, reciprocal modelling can reveal potential problems with occurrence data and predictor variables in both native and introduced ranges, but it also has also been used to investigate ecological phenomena, such as niche shifts of invasive species in their introduced range (Mau-Crimmins, Schussman et al. 2006; Broennimann, Treier et al. 2007; Fitzpatrick, Weltzin et al. 2007).

Objectives

We were interested in exploring the potential of species distribution models to identify areas susceptible to alien species invasion within the United States. Species distribution models are a relatively versatile and inexpensive ecological tool — species occurrence data, GIS layers of environmental data, and software to implement the models are freely available on the Internet. Species distribution models have the potential to improve our response to the threat of invasive species. Identifying potentially suitable habitat can help to focus early detection efforts and therefore reduce the resources needed to manage or eradicate the species. We used species distribution models based on native range occurrences and climate to predict the climate suitability of three invasive species. Model predictions were then projected into the introduced range to determine areas that are climatically suitable for the invasive species. To test the model’s accuracy when transferred into other regions, we compared model predictions in the introduced range to occurrence records of the alien species in the introduced range. Using reciprocal modelling and principle components analysis (PCA) we examined the differences in predicted distributions in both geographic and climate space using native and introduced occurrences.

METHODS

Study Species

We limited our investigation to three invasive wetland plant species that are considered invasive in the USA, that we assume have not reached environmental equilibrium: *Iris pseudacorus, Lythrum salicaria*, and *Saccharum ravennae*. We focused on species of European origin because of the plethora of data available on their native distribution.

Iris pseudacorus (yellow flag iris) is native to Europe and western Asia and brought to North America as an ornamental in the mid 1800s (Sutherland 1990) and is currently found throughout the USA and Canada, except the Rocky Mountains (USDA NRCS 2009). *Iris pseudacorus* is capable of forming dense monocultures from a network of rhizomes that exclude native riparian vegetation (personal observation) (Judd 1953; Raven and Thomas 1970). *Iris pseudacorus* occurs in wetlands and riparian zones and can thrive in drainage ditches. It also has been planted in sewage treatment facilities for heavy metal remediation. As with other wetland plants, *I. pseudacorus* can tolerate long periods of anoxia, but also can withstand long droughts. *Iris pseudacorus* can reproduce sexually by seed and/or asexually by rhizome fragments. Hydrochory is the typical dispersal mode via transport of floating fruits, seeds, and dislodged rhizomes (Sutherland 1990).

Lythrum salicaria (purple loosestrife) originated in Eurasia, but was well established in North America by the 1830s, thus leading John Torrey and Asa Grey to conclude it was a native species. It was probably introduced repeatedly to North America via ships ballasts, through the horticulture trade, in imported goods, and by immigrants using it as a culinary and medicinal herb (Thompson, Stuckey et al. 1987). *Lythrum salicaria* has been reported from 43 of the coterminous states in the USA and is considered a noxious weed by several (USDA NRCS 2009). As a mature, herbaceous perennial, *L. salicaria* can reach up to 2 m in height and produce over 2 million seeds per plant. Although most seeds cannot float, seedlings can and this may be the primary mode of dispersal (Thompson, Stuckey et al. 1987).

Saccharum ravennae (ravenna grass) is a large clump forming grass species native to southern Europe, northern Africa, and western Asia. It has been reported from 16 of the coterminous states (USDA NRCS 2009) and is designated as an invasive species in Arizona and Utah (Swearingen 2006). The oldest records in herbaria in the United States are from the early 1900s, but there is some speculation that the invasive genotype was introduced later (Thomsen and Meyer 2008). It is widely used as an ornamental grass in the USA and naturalized populations are presumed to be escaped from cultivation and ornamental landscaping (Utah State University 2009). Invasive populations are generally found along rivers and the grass can grow in a variety of soil types and moisture regimes.

Occurrence Data

Occurrence data sets were compiled from several data sources. A search was conducted of the Global Biodiversity Information Facility (GBIF) database for the species of interest. GBIF is an international organization that has partnered with institutions from around the world to provide biodiversity data over the Internet. A majority of the data within GBIF comes from natural history collections, including herbaria. We limited the occurrence records to the continental United States and Europe.

We removed duplicate records, records with missing location information, and records that were not georeferenced to three decimal places for latitude and longitude. It should be noted that GBIF does not guarantee the accuracy of the data provided.   Biodiversity occurrence data used for this research were provided to GBIF by institutions listed in appendix 1 (Accessed through GBIF Data Portal, www.gbif.net, 2008-12-10). Additional occurrence data for *I. pseudacorus* and *L. salicaria* within the United States were obtained from the Nonindigenous Aquatic Species Program database at the US Geological Survey (nas.er.usgs.gov). Occurrence data for all species within Oklahoma were acquired from the Oklahoma Vascular Plants Database [www.oklahomaplantdatabase.org; (Hoagland, Buthod et al. 2008)]. The occurrence data were randomly split into two data sets: model training (or building) data (75% of the data) and model evaluation data (25% of the data; also known as hold out data).

The specimen occurrence records maintained by GBIF do not represent a geographically uniform or systematic data set. Asymmetries exist in the data because collection effort is not equal for all parts of the globe nor have all natural history collections contributed data to GBIF. Precision at which a specimen was georeferenced is variable and some collections are not well georeferenced. We accounted for sampling bias in the GBIF data through use of a “bias file” in Maxent (Phillips, Anderson et al. 2006). The bias file is an additional raster file added to the modelling process that represents sampling effort. Because it is rare that sampling effort is quantified and available in spatial form, especially with large data sets such as GBIF, a bias file can be generated by using records of several common species within the same study area. This group of species should be broadly distributed within the study area and represent a variety of habitats and environmental tolerances. The distribution of this group of species is modelled using the same environmental predictor variables. Because this group of species represents a wide range of environmental variables, the distribution should not be easily predicted from the model. However, if the distribution of this group of species performs well using the environmental variables, we can infer that the model performance is being affected by geographic sampling bias and not true environmental factors (Phillips, Dudik et al. 2009). We selected 20 common herbaceous plants of Europe based on their broad European distribution and range of habitats (Appendix 2) and adequate occurrence records in GBIF. Biodiversity occurrence data used for this research were provided to GBIF by institutions listed in appendix 1 (Accessed through GBIF Data Portal, www.gbif.net, 2008-12-10). We discarded duplicate records, records with missing location information, and records that were not georeferenced to at least three decimal places for latitude and longitude. Each species contributed over 4,000 specimens to the total of over 80,000 occurrences. We randomly selected a subset of 15,000 records for model building in Maxent. Models for the bias file group of species were built with the same parameters (detailed below) as models for the invasive species. The model predictions for the bias file group of species were then used as the bias file in Maxent (Phillips and Dudik 2008).

Bioclimatic Predictor Data

Because plant species distribution is, for the most part, determined by climate at the continental scale (Woodward 1987), we used the 19 derived bioclimatic variables in 30 arc-second resolution raster grids (approximately 1 km x 1 km resolution) from WorldClim for the environmental predictor data (Table 1). Freely available over the

Internet (www.worldclim.org), the WorldClim data make up a set of fine-scale global climatic layers interpolated from a large number of weather stations and statistically enhanced with digital elevation models (Hijmans, Cameron et al. 2005). WorldClim has been used with success in species distribution models (Broennimann, Treier et al. 2007; Fitzpatrick, Weltzin et al. 2007; Pearman, Randin et al. 2008). Within ArcMap, global raster layers were clipped to rectangles surrounding the regions representing the 48 contiguous United States (from here on referred to as US) and Europe, as far east as western Russia.

Projecting models built in one region onto another region requires a similar range of values within the environmental predictor variables. To determine if the environmental variables used for model building at the continental scale have similar ranges in both the US and Europe, box plots and line graphs were used to evaluate each pair of US and European environmental variables for range of value overlap. For example, a value for each environmental variable is contained in each cell of the raster layer. The number of cells representing all possible values for each variable is tallied. The entire range of values and interquartile range of values for each variable pair is compared in 19 box plots for all bioclimatic variables. Box plot whiskers were drawn to represent the range of values and boxes were drawn to encompass values between the first and third quartile (Appendix 3). The extent of the box represents the range of values for each variable surrounding the mean for the middle 50% of cells. Another effort to visualize the data involved line graphs drawn for pairs of environmental variables (Appendix 3). The number of cells was plotted against variable value for each environmental pair. Box plots and line graphs were visually analyzed.

Maxent Modelling Algorithm

Many of the standard distribution modelling techniques, such as regression, require both presence and absence data to make accurate predictions of a species' distribution. Maximum entropy (Maxent) is a machine learning method that is able to make predictions using presence only data (Phillips, Anderson et al. 2006). Like other machine learning techniques, Maxent improves the modelling algorithm automatically through a series of trainings with the data set. The creators of the Maxent implementation for species distribution modelling state that it is able to predict a species' distribution based on "incomplete information"; meaning species observation data that do not necessarily cover the entire suitable range of environmental variables (Phillips, Anderson et al. 2006; Phillips and Dudik 2008). Maxent estimates the distribution of a species with maximum entropy (the most uniform or spread out distribution) of the known presence points given the constraints put on the distribution with respect to the point's relationship to the environmental layers. This relationship is quantified using the empirical average of the environmental variable at all presence records (Phillips, Anderson et al. 2006).

The implementation of Maxent for species distribution modelling was specifically designed for use with presence only data. In comparisons with other presence only methods, it performs significantly better. Maxent also performs well when compared to presence–absence procedures that utilize both real and pseudo-absence data (Elith, Graham et al. 2006; Hernandez, Graham et al. 2006; Pearson, Raxworthy et al. 2007).

Maxent has several features that improve the models predictive performance and interpretability: it takes into account the interaction between environmental variables; the output is the probability of distribution, which is mathematically defined; it possesses a procedure to counteract over-fitting of the model; it employs a relaxation that allows the estimated distribution to go beyond the empirical average within the error bounds, a smoothing procedure called regularization, can potentially correct for small sample size (Phillips, Anderson et al. 2006; Phillips and Dudik 2008). Maxent is also simple to implement given the user-friendly interface developed by Phillips et al. (2008). Not only does the software compute distribution models, but it also performs validation statistics, jackknifes to calculate variable importance, and produces a potential distribution map of the model results. And the software can also project the model results onto another set of environmental variables in a different region which is especially useful for modelling alien species in their introduced region (Phillips, Anderson et al. 2006; Phillips and Dudik 2008).

Reciprocal Models

Once data were corrected and a bias file generated, occurrence and predictor data were loaded into Maxent version 3.2.19 (Phillips, Schapire et al. 2008). To test the model, 25% of the occurrence data was withheld from model building and used in evaluations (Table 2). Although not a truly independent data set, withholding occurrence data from analysis for evaluation is a common and useful technique for model evaluation (Araújo, Pearson et al. 2005). The regularization multiplier affects how well the model can be applied to independent data. As the multiplier value is decreased, the model fit to the training data improves, but the risk of over-fitting the model increases. The regularization multiplier is adjusted if the model evaluation results indicate a large difference in the performance of the model for the training and testing data. Unless there is a large discrepancy between the test statistic for the training and evaluation data, the default regularization multiplier value of 1 is recommended (Phillips, Anderson et al. 2006). A maximum of 500 iterations for each model was run and a convergence threshold of 0.0001 was used. Convergence threshold of 0.0001 is the default and is considered a conservative estimate allowing the algorithm to approach convergence (Phillips, Anderson et al. 2006). Ten thousand "background" points were randomly chosen from the extent of the environmental layers as a representation of the range of values for all environmental variables across the region. Multiple occurrence points falling within one grid cell of the environmental variables were reduced to one point for both model building and evaluation.

Models were built in two stages. First, predictions from native range data were mapped in both the native range (Europe) and introduced range (US). Then a full set of reciprocal models were built in the introduced range (US) and projected into the native range (Europe). Models projected into a different region were checked for environmental variables that were restricted because of range of values encountered during training was limited [termed clamping in Maxent software (Phillips, Schapire et al. 2008)]. Predictor variable values in the new region that are outside the range used during model building will likely have an effect on predicted suitability. Models built with all variables were compared to models built with the reduced set of variables that had good range overlap between US and Europe.

Model Evaluation

We used the threshold independent method, receiver-operating characteristic curve (ROC) to evaluate all models. The area under the curve (AUC) of a ROC has been widely used to assess the predictive performance of species distribution models (Hirzel, Le Lay et al. 2006; Wisz, Hijmans et al. 2008). AUC is calculated by plotting sensitivity against (1-specificity) for all possible thresholds. AUC values range from 0 to 1; with 0.5 being random performance and values near 1 being good predictive performance (Pearce and Ferrier 2000). The Maxent model calculates AUC using 25% holdout data for presence points and the 10,000 background points as absence points (Phillips, Schapire et al. 2008). Native range models projected into the introduced range were evaluated using US occurrences; introduced range models projected into native range were evaluated using Europe occurrences. AUC values for models were compared using a Wilcoxon signed-rank test.

PCA

In addition to evaluating the distribution of the invasive species geographically, we evaluated the distribution of populations from both the native and introduced range in environmental space. A comparison of results in both geographic and environmental space can help us to evaluate discrepancies between models made in the native and introduced range. Using principle component analysis (PCA), we compared the position of the species occurrences in climate space for both the native and introduced range (McCune and Mefford 2006). Because of the large occurrence data set for *I. pseudacorus* and *L. salicaria* in Europe, we randomly selected 1,500 occurrences for each species to calculate the principle components.

RESULTS

Occurrence Data and Accounting for Sample Bias

The number of occurrence records available per species in each region ranged from 24 to 14877 (Table 2). *Iris pseudacorus* and *L. salicaria* have been collected extensively and are widely distributed throughout central and northern Europe (Table 2; Fig. 1a, 2a). *Saccharum ravennae* has significantly fewer occurrence records and is found primarily in southern Europe (Table 2; Fig. 3a). It is clear from the map of occurrence points in Europe that there is a sampling bias related to political boundaries (Fig. 1a, 2a, 3a).

The AUC value from the bias file model was high (AUC 0.881 +/- 0.01), indicating that GBIF data for Europe are not uniformly distributed in geographic space and the distribution of this group of species can be erroneously predicted with climatic variables. The resulting predictions from the target group were used as the bias file in the Maxent modeling of the invasive species.

Accounting for Difference in Range of Bioclimatic Variables

Based on the comparison of line graphs and box plots of US and Europe bioclimatic variable raster layers, 5 of the 19 variables appeared to have a large difference in value range and interquartile range (Appendix 3). These variables were excluded from the final model building (Table 1).

Although there were significant differences in the performance of the model sets, (models using all 19 bioclimatic variables versus models using the reduced set of bioclimatic variables) there was not a consistent pattern related to number of predictor variables (Table 3). Among the models built with the Europe occurrence points, either modeled in the native or introduced regions, only the *S. ravennae* model projected into the US had a significant difference between all or reduced predictor variables and contrary to expectations, the model with fewer variables performed better. Therefore, in subsequent analyses we focused only on the results from models with the reduced predictor data set due to the slight advantage or no difference between these model pairs. Focusing on this data set also allowed us to moderate the errors caused by variables with dissimilar ranges in Europe and US.

Reciprocal Models

Species distribution models were highly accurate when applied to the region in which they were built (Table 4, Figure 4b, 5b, 6b). The results from the three study species supports the general assumption that plant species distribution is governed by climate at the continental scale. However, species distribution models built in one region and projected into another region performed poorly (Table 4, Figure 4a, 5a, 6a). For all three species, models built using Europe occurrences and applied to Europe performed well, with AUC values above 0.92. Even the small data set of *S. ravennae*, with only 18 training points, still performed well (AUC = 0.959). Models created using US occurrences and applied to the US also performed well (AUC range = 0.895 to 0.922). At best, the models projected into a different region had moderate AUC values (0.759 and 0.744), but several models performed no better than random (near 0.5). There was no consistent pattern of performance for models built in the native range and projected into the introduced range or vice versa (Table 4). The model of *S. ravennae* built with Europe occurrences and projected into the US performed better than the model using US occurrences and projected into Europe. However, the model of *I. pseudacorus* built with Europe occurrences and projected into the US performed worse.

PCA

For all three species, the first three principle components accounted for over 75% of the total variation in the data (Table 5). For *I. pseudacorus*, the first principle component (PC-1) was related to temperature (especially Bio6), PC-2 was related to precipitation (particularly periods of wettest precipitation), and PC-3 was related to temperature extremes (Bio10 and 11) (Figure 7). Occurrences in Europe and US appear to separate based precipitation. PCA were similar for *L. salicaria*, except temperature extremes were more important in PC-1 (Figure 8). The Europe occurrences exhibit a variety of precipitation tolerance, and temperature seems to separate the Europe and US groups. For *S. ravennae*, PC-1 was related to precipitation, while PC-2 and PC-3 were related to temperature (Figure 9). The distribution of *S. ravennae* US occurrences appears to be more influenced by temperature and by precipitation in Europe. All three analyses illustrate a separation in environmental space for the native and introduced occurrences.

DISCUSSION

All distribution models in this study performed well when built with occurrence and climate data from the same region, but did not perform well when projected, or transferred, to a different region. Transferability of model predictions to other ranges have been examined both within native ranges and to introduced ranges. Some studies have found that models built using data from a portion of the native range are not necessarily transferable to other parts of the native range (Randin, Dirnbock et al. 2006; Barbosa, Real et al. 2009); and invasive species models projected into their introduced range and evaluated with introduced occurrences also show poor performance (Mau-Crimmins, Schussman et al. 2006; Fitzpatrick, Weltzin et al. 2007; Loo, Nally et al. 2007).

Recently postulated hypotheses regarding factors that contribute to invasive species may explain the discrepancy of ranges of the three invasive study species within Europe and US, such as escape from natural enemies, evolution in new environment, better competitors due to novel biochemicals, pre-adapted to disturbed environment, and repeated introduction with high propagule pressure [for review see (Hierro, Maron et al. 2005)]. The characteristics that make a species invasive may be the same characteristics that cause the species' environmental range to be different in the native and introduced regions. Not only do the model predictions from one continent to another illustrate a difference in climate preference, the PCA results indicate a difference in the climate space occupied by the native and alien occurrences. This difference in occupied habitats by one species after introduction to a new region can be interpreted as a niche shift.

Species distribution models assume that the species being modeled is at equilibrium with the environment, whether in the native or introduced area (Araújo and Pearson 2005); but this assumption certainly violated when modeling alien species, which may still be spreading into suitable areas. In fact, that is the point of our invasive modelling research: to find areas of suitable habitat that the species has not yet dispersed into, for whatever reason. Therefore, in this study, it is assumed that the invasive species is not at equilibrium in the introduced range. Thus, it would be expected that models built with introduced range data under-predict the native distribution and the models built with native range data to over-predict the introduced distribution. Our results are, unfortunately, not that simple. The appeal of using native range occurrences to build a model is to represent the species' environment when it is at equilibrium. However, there is some debate as to how many species are truly at equilibrium within their native environment (Araújo and Pearson 2005).

The assumption of modelling the distribution of invasive species using native range occurrences is that the niche occupied in the native range will be similar to the one occupied in the introduced range (Peterson and Vieglais 2001; Pearman, Guisan et al. 2008). Evidence is accumulating that invalidates that assumption. Fire ants (*Solenopsis invicta*) are not occupying the same climatic space in their native and introduced regions (Fitzpatrick, Weltzin et al. 2007; Fitzpatrick, Dunn et al. 2008). Spotted knapweed (*Centaurea maculosa*) occupies areas that are climatically different in Europe and North America (Broennimann, Treier et al. 2007). A lovegrass (*Eragrostis lehmanniana*) from South Africa has invaded a different environmental niche in the southwestern United States (Mau-Crimmins, Schussman et al. 2006).

The niche and niche shift concepts affect the interpretation of the model results. Observations of the distribution of a species native region only consider the species realized niche: the combination of suitable environmental conditions that is adjusted by history and biotic interactions. When considering the distribution of an alien species in its introduced region, what may be revealed is a new realized niche. Researchers using reciprocal models such as ours have demonstrated that the realized niche differs in native and introduced ranges. The niches of introduced and native ranges, as represented in climatic or environmental space predicted by the models, may overlap in part. However, nonoverlapping areas that represent different environmental space or realized niches are biologically and ecologically interesting.

The difference between the suitable habitat in the native and introduced ranges can be due to genetic differences caused either by evolution or adaptation after introduction or the introduction of a particular phenotype that has thrived in the introduced region (Dietz and Edwards 2006; Richardson and Pyšek 2006). Fitzpatrick et al. (2007, 2008) support the hypothesis that fire ants with a specific phenotype were introduced into the United States. They also suggest the fire ants' niche continued to shift due to adaptation to the introduced region's environment (Fitzpatrick, Weltzin et al. 2007; Fitzpatrick, Dunn et al. 2008). The core area of invasion for spotted knapweed (*Centaurea maculosa*) is outside of the climatic niche of the native distribution. Broennimann and colleagues (2008) argue that the niche of spotted knapweed has shifted after introduction and the difference is not due to the introduction of a specific genotype. They suggest that the niche shift could be caused by a change in realized niche (competitor release or other change in biotic interaction) or a change in the fundamental niche (evolution or adaptation of an increased competitive ability). Mau-Crimmins and colleagues (2006) found that the variety of Lehmann's lovegrass (*Eragrostis lehmanniana*) introduced in the United States was highly selected by agronomists and its environmental tolerances within the introduced range did not reflect the entire native range. Therefore models trained on occurrence data from only the introduced range performed better than models using native range information. The researchers conclude that introduced taxa that represent a genetically distinct group within a species are best modelled with only introduced occurrences because native range information would include environmental tolerances outside the narrow tolerances of the introduced taxa.

Although a species' distribution within its native region may be readily described by climate variables, that prediction may not be transferable to another location when based on climate alone. Biotic interactions also limit a species distribution. Model predictions based on the native range may under-predict the potential distribution in the introduced range if biotic interactions, such as competition or parasitism, are removed when an alien species enters a new region. For example, a competitor may be the limiting factor at the northern edge of a species range. However, that northern range edge may easily be represented by temperature. If temperature is used as a surrogate by the model and predictions based on temperature are then projected into the introduced region, the model will fail to accurately predict the distribution because competition is the true limiter. Situations such as these have lead many ecologists and modellers to call for the incorporation of biotic interactions in species distribution models (Davis, Jenkinson et al. 1998; Araújo and Luoto 2007; Guisan, Zimmermann et al. 2007). But accurately predicting areas of invasion in the introduced range may never truly incorporate the influence of biotic interactions because the introduced species are no longer affected by their native biotic interactions and are subject to another suite of species in the introduced range with which it may form new biotic interactions that are currently indescribable.

Poor model transfer may be a result of causes that are not related to ecology. Non-overlapping range of values for the predictor variables in Europe and US may still be affecting the model predictions despite removing five of the 19 variables from model building. There may continue to be error in the models caused by one or two variables that are important to the model predictions, but also have value ranges that do not match in both the introduced and native regions. Knowledge of the environmental tolerances across the entire range of the species is necessary to find all suitable habitats (Murphy and Lovett-Doust 2007). Unfortunately, we cannot evaluate this likelihood for *I. pseudacorus, L. salicaria,* and *S. ravennae* because GBIF data are highly skewed to western European and North America, but under represent their native range in eastern Europe, northern Africa, and/or the Middle East. The climatic environment of these areas within the native range was not represented in our models. Also, more complex models have recently emerged and been applied to alien species distribution. For example, fuzzy envelope models proved successful in predicting the distribution of alien species in South Africa (Robertson, Villet et al. 2004). The ongoing development of new algorithms and techniques may improve the predictions of invasive species potential distribution.

Models of the potential distribution of invasive species have been informative at the global scale (Peterson and Vieglais 2001; Peterson, Pape• et al. 2003; Nyári, Ryall et al. 2006). For example, Thuiller and colleagues (2005) produced a global map of potential areas of invasion by 96 South African plants species. Their models performed well due to the inclusion of a generalized biome variable with bioclimatic layers and thorough documentation of the range of these South African endemics. Models utilizing the introduced occurrences have been successful in predicting new areas of invasion at the local level. In fact, because of the difficulty obtaining data from native ranges, some modelers rely only on introduced region information. Researchers have examined the local invasion potential of alien sea squirts (the tunicate *Didemnum vexillum*) off the coast of British Columbia (Herborg, O'Hara et al. 2009), invasive trees and grasses in an

American national park (Evangelista, Kumar et al. 2008), and invasive plants in China (Zhu, Sun et al. 2007). At the local scale, better performing models have included variables that are important to the distribution of a species. In addition to climate variables, modelers have included topography and landcover. Recent advancement in distribution models have incorporated dispersal vectors (del Barrio, Harrison et al. 2006; Herborg, O'Hara et al. 2009), anthropogenic influence (Lippitt, Rogan et al. 2008), and remotely sensed habitat information (Thuiller, Richardson et al. 2005; Anderson, Peterson et al. 2006).

Relatively easily obtained data and user-friendly modelling software make building models an inexpensive tool for conservation biologists. Being able to create a distribution map with little prior knowledge of the species' ecology and biology is tempting and possible with species distribution models. Models can help generate hypotheses regarding the environmental and physiological tolerances of species. It is apparent from our results and others that native region models based on climate alone are of little use in locating suitable invasive species habitat. It is not surprising that many others have come to the conclusion that species distribution models coupled with sound ecological understanding will produce the best results (Wilson, Westphal et al. 2005; Barry and Elith 2006; Guisan, Overton et al. 2006; Austin 2007).

Our goal was to create predictive models of alien plant invasion based on native range information with the intention to inform conservation efforts such as early detection and eradication programs. model areas that are suitable for invasion by specific alien species using the native climate habitat. Our model results indicate that the climate space occupied by the species are inconsistent between the native and introduced ranges.

Therefore our model predictions are not useful in determining areas of habitat suitability in the introduced range.

ACKNOWLEDGEMENTS

We thank P. Fuller with the USGS Nonindigenous Aquatic Species Program for providing US occurrence data of *I. pseudacorus* and *L. salicaria.* We thank J. Kelly, M. Patten, S. Phillips, and R. Pearson for helpful discussions of species distribution modelling. C. Vaughn, W. Elisens, and J.S. Greene made helpful comments on earlier versions of the manuscript.

LITERATURE CITED

Anderson, R. P., A. T. Peterson, et al. (2006). Vegetation-index models predict areas vulnerable to purple loosestrife (*Lythrum salicaria*) invasion in Kansas. *The Southwestern Naturalist* **51**, 471-480.

Araújo, M. B. & M. Luoto (2007). The importance of biotic interactions for modelling species distributions under climate change. *Global Ecology and Biogeography* **16**, 743-753.

Araújo, M. B. & R. G. Pearson (2005). Equilibrium of species' distributions with climate. *Ecography* **28**, 693-695.

Araújo, M. B., R. G. Pearson, et al. (2005). Validation of species-climate impact models under climate change. *Global Change Biology* **11**, 1504-1513.

Austin, M. P. (2007). Species distribution models and ecological theory: A critical assessment and some possible new approaches. *Ecological Modelling* **200**, 1-19.

Barbosa, A. M., R. Real, et al. (2009). Transferability of environmental favourability models in geographic space: The case of the Iberian desman (Galemys pyrenaicus) in Portugal and Spain. *Ecological Modelling* **220**, 747-754.

Barry, S. & J. Elith (2006). Error and uncertainty in habitat models. *Journal of Applied Ecology* **43**, 413-423.

Broennimann, O. & A. Guisan (2008). Predicting current and future biological invasions: both native and invaded ranges matter. *Biology Letters* **4**, 585-589.

Broennimann, O., U. A. Treier, et al. (2007). Evidence of climatic niche shift during biological invasion. *Ecology Letters* **10**, 701-709.

Davis, A. J., L. S. Jenkinson, et al. (1998). Making mistakes when predicting shifts in species range in response to global warming. *Nature* **391**, 783-786.

del Barrio, G., P. A. Harrison, et al. (2006). Integrating multiple modelling approaches to predict the potential impacts of climate change on species' distributions in contrasting regions: comparison and implications for policy. *Environmental Science and Policy* **9**, 129-147.

Dietz, H. & P. J. Edwards (2006). Recognition that causal processes change during plant invasion helps explain conflicts in evidence. *Ecology* **87**, 1359-1367.

DiTomaso, J. M. (2000). Invasive weeds in rangelands: species, impacts, and management. *Weed Science* **48**, 255-265.

Elith, J., C. H. Graham, et al. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography* **29**, 129-151.

Evangelista, P. H., S. Kumar, et al. (2008). Modelling invasion for a habitat generalist and a specialist plant species. *Diversity and Distributions* **14**, 808-817.

Fitzpatrick, M. C., R. R. Dunn, et al. (2008). Data sets matter, but so do evolution and ecology. *Global Ecology and Biogeography* **17**, 562-565.

Fitzpatrick, M. C. & J. F. Weltzin (2005). Ecological niche models and the geography of biological invasions: a review and a novel application. *Invasive plants: ecological and agricultural aspects*. (ed. by Inderjit), pp. 45-60. Birkhäuser Basel.

Fitzpatrick, M. C., J. F. Weltzin, et al. (2007). The biogeography of prediction error: why does the introduced range of the fire ant over-predict its native range? *Global Ecology and Biogeography* **16**, 24-33.

Guisan, A., J. M. C. Overton, et al. (2006). Making better biogeographical predictions of species' distributions. *Journal of Applied Ecology* **43**, 386-392.

Guisan, A., N. E. Zimmermann, et al. (2007). What matters for predicting the occurrences of trees: Techniques, data, or species' characteristics? *Ecological Monographs* **77**, 615-630.

Herborg, L.-M., P. O'Hara, et al. (2009). Forecasting the potential distribution of the invasive tunicate *Didemnum vexillum*. *Journal of Applied Ecology* **46**, 64-72.

Hernandez, P., C. H. Graham, et al. (2006). The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography* **29**, 773-785.

Hierro, J. L., J. L. Maron, et al. (2005). A biogeographical approach to plant invasions: the importance of studying exotics in their introduced and native range. *Journal of Ecology* **93**, 5-15.

Hijmans, R. J., S. E. Cameron, et al. (2005). Very high resoluation interpolated climate surfaces for global land areas. *International Journal of Climatology* **25**, 1965-1978.

Hirzel, A. H., G. Le Lay, et al. (2006). Evaluating the ability of habitat suitability models to predict species presences. *Ecological Modelling* **199**, 142-152.

Hoagland, B. W., A. K. Buthod, et al. (2008). Oklahoma Vascular Plants Database). Oklahoma Biological Survey, University of Oklahoma, Norman, OK.

Hobbs, R. J. & S. E. Humphries (1995). An integrated approach to the ecology and management of plant invasions. *Conservation Biology* **9**, 761-770.

Judd, W. W. (1953). *Iris pseudacorus* L. established in the vicinity of London, Ontario. *Rhodora* **55**, 244.

Lippitt, C. D., J. Rogan, et al. (2008). Incorporating anthropogenic variables into a species distribution model to map gypsy moth risk. *Ecological Modelling* **210**, 339-350.

Loo, S. E., R. M. Nally, et al. (2007). Forecasting New Zealand mudsnail invasion range: model comparisons using native and invaded ranges. *Ecological Applications* **17**, 181-189.

López-Darias, M., J. Lobo, et al. (2008). Predicting potential distributions of invasive species: the exotic Barbary ground squirrel in the Canarian archipelago and the west Mediterranean region. *Biological Invasions* **10**, 1027-1040.

Mack, R. N., D. Simberloff, et al. (2000). Biotic invasions: causes, epidemiology, global consequences and control. *Issues in Ecology* **5**, 1-22.

Mau-Crimmins, T. M., H. R. Schussman, et al. (2006). Can the invaded range of a species be predicted sufficiently using only native-range data?: Lehmann lovegrass (*Eragrostis lehmanniana*) in the southwestern United States. *Ecological Modelling* **193**, 736-746.

McCune, B. & M. J. Mefford (2006). PC-ORD. Multivariate Analysis of Ecological Data). MjM Software, Gleneden Beach, OR.

Murphy, H. T. & J. Lovett-Doust (2007). Accounting for regional niche variation in habitat suitability models. *Oikos* **116**, 99-110.

Nyári, Á., C. Ryall, et al. (2006). Global invasive potential of the house crow Corvus splendus based on ecological niche modeling. *Journal of Avian Biology* **37**, 306-311.

Pearce, J. L. & S. Ferrier (2000). Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling* **133**.

Pearman, P. B., A. Guisan, et al. (2008). Niche dynamics in space and time. *Trends in Ecology & Evolution* **23**, 149-158.

Pearman, P. B., C. F. Randin, et al. (2008). Prediction of plant species distributions across six millennia. *Ecology Letters* **11**, 357-369.

Pearson, R. G., C. J. Raxworthy, et al. (2007). Predicting species distributions from small numbers of occurrence records: A test case using cryptic geckos in Madagascar. *Journal of Biogeography* **34**, 102-117.

Peterson, A. T. (2003). Predicting the geography of species' invasion via ecological niche modeling. *The Quarterly Review of Biology* **78**, 419-433.

Peterson, A. T., M. Pape•, et al. (2003). Predicting the potential invasive distributions of four alien plant species in North America. *Weed Science* **51**, 863-868.

Peterson, A. T. & D. A. Vieglais (2001). Predicting species invasions using ecological niche modeling: new approaches from bioinformatics attack a pressing problem. *Bioscience* **51**, 363-371.

Phillips, S., M. Dudik, et al. (2009). Sample selection bias and presence-only models of species distributions: implications for selection of background and pseudo-absences. *In press*.

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

Phillips, S. J. & M. Dudik (2008). Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography* **31**, 161-175.

Phillips, S. J., R. E. Schapire, et al. (2008). Maxent software for species distribution modeling).

Pimentel, D., L. Lach, et al. (2000). Environmental and economic costs of nonindigenous species in the United States. *Bioscience* **50**, 53-65.

Randin, C. F., T. Dirnbock, et al. (2006). Are niche-based species distribution models transferable in space? *Journal of Biogeography* **33**, 1689-1703.

Raven, P. H. & J. H. Thomas (1970). *Iris pseudacorus* in western North America. *Madrono* **20**, 390-391.

Richardson, D. M., F. D. Panetta, et al. (2000). Naturalization and invasion of alien plants: Concepts and definitions. *Diversity and Distributions* **6**, 93-107.

Richardson, D. M. & P. Pyšek (2006). Plant invasions: merging the concepts of species invasiveness and community invasibility. *Progress in Physical Geography* **30**, 409-431.

Robertson, M. P., M. H. Villet, et al. (2004). A fuzzy classification technique for predicting species' distributions: applications using invasive alien plants and indigenous insects. *Diversity and Distributions* **10**, 461-474.

Sutherland, W. J. (1990). Biological flora of the British Isles. *Iris pseudacorus* L. *Journal of Ecology* **78**, 833-848.

Swearingen, J. (2006). WeedUS: Database of plants invading natural areas in the United States). Plant Conservation Alliance, Alien Plant Working Group.

Thompson, D. Q., R. L. Stuckey, et al. (1987). Spread, impact, and control of purple loosestrife (*Lythrum salicaria*) in North American wetlands. (ed. by U. S. F. a. W. Service). Northern Prairie Wildlife Research Center Online, Jamestown, ND.

Thomsen, C. & T. Meyer (2008). Ravennagrass: a major wildland weed along Cache Creek. *Cal-IPC News*), pp. 4-5, 16. California Invasive Plants Council.

Thuiller, W., D. M. Richardson, et al. (2005). Niche-based modelling as a tool for predicting the risk of alien plant invasions at a global scale. *Global Change Biology* **11**, 2234-2250.

USDA NRCS (2009). The Plants Database). National Plant Data Center, Baton Rouge, LA.

Utah State University (2009). Manual of the Grasses of the United States. (ed. by M. E. Barkworth).

Vitousek, P. M., C. M. D'Antonio, et al. (1997). Introduced species: a significant component of human-caused global change. *New Zealand Journal of Ecology* **21**, 1-16.

Wilson, K. A., M. I. Westphal, et al. (2005). Sensitivity of conservation planning to different approaches to using predicted species distribution data. *Biological Conservation* **122**, 99-112.

Wisz, M. S., R. J. Hijmans, et al. (2008). Effects of sample size on the performance of species distribution models. *Diversity and Distributions* **14**, 763-773.

Woodward, F. I. (1987). *Climate and Plant Distribution,* Cambridge, Cambridge University Press.

Zhu, L., O. J. Sun, et al. (2007). Predicting the spatial distribution of an invasive plant species (*Eupatorium adenophorum*) in China. *Landscape Ecology* **22**, 1143-1154.

Table 1. Bioclimatic variables from WorldClim (www.worldclim.org) used in Maxent models to predict the distribution of three invasive species in US.

BIO1	Annual Mean Temperature
BIO2*	Mean Diurnal Range (Mean of monthly (max temp - min temp))
BIO3	Isothermality (BIO2/BIO7) (* 100)
BIO4*	Temperature Seasonality (standard deviation *100)
BIO5	Max Temperature of Warmest Month
BIO6	Min Temperature of Coldest Month
BIO7*	Temperature Annual Range
BIO8*	Mean Temperature of Wettest Quarter
BIO9	Mean Temperature of Driest Quarter
BIO10	Mean Temperature of Warmest Quarter
BIO11	Mean Temperature of Coldest Quarter
BIO12	Annual Precipitation
BIO13	Precipitation of Wettest Month
BIO14	Precipitation of Driest Month
BIO15	Precipitation Seasonality (Coefficient of Variation)
BIO16	Precipitation of Wettest Quarter
BIO17	Precipitation of Driest Quarter
BIO18	Precipitation of Warmest Quarter
BIO19*	Precipitation of Coldest Quarter

* Variable removed from modelling due to differences in range between Europe and US.

Table 2. Number of occurrence points used for model building and evaluation for each region.

	Native (Europe)		Introduced (US)	
	Training	Evaluation	Training	Evaluation
Iris pseudacorus	11158	3719	528	176
Lythrum salicaria	9847	3282	1216	405
Saccharum ravennae	18	6	30	10

Table 3. Comparison of model performance using AUC values calculated using 25% of occurrence data held out. Standard deviations indicate the variability of model accuracy through 500 iterations. The difference in accuracy between models built with all variables and models built with five variables removed was tested with a Wilcox signed-rank test (n.s., not significant; *, p < 0.01).
** individual of a model pair that performed better in comparison.

		Models built with US occurrence points			Models built with Europe occurrence points		
		all variables		reduced variables	all variables		reduced variables
Iris pseudacorus	predicted into US	0.884 +/- 0.01	*	0.909 +/- 0.01 **	0.613 +/- 0.01	n.s.	0.648 +/- 0.01
	predicted into Europe	0.79 +/- 0.01	n.s.	0.759 +/- 0.01	0.933 +/- 0.01	n.s.	0.922 +/- 0.01
Lythrum salicaria	predicted into US	0.918 +/- 0.01	n.s.	0.922 +/- 0.01	0.436 +/- 0.01	n.s.	0.447 +/- 0.01
	predicted into Europe	0.396 +/- 0.01	*	0.457 +/- 0.01 **	0.924 +/- 0.01	n.s.	0.917 +/- 0.01
Saccharum ravennae	predicted into US	0.943 +/- 0.04 **	*	0.895 +/- 0.04	0.668 +/- 0.03	*	0.736 +/- 0.03 **
	predicted into Europe	0.452 +/- 0.06	*	0.535 +/- 0.07 **	0.956 +/- 0.02	n.s.	0.959 +/- 0.01

Table 4. Comparison of model performance using AUC values calculated using 25% of occurrence data held out. Standard deviations indicate the variability of model accuracy through 500 iterations. The difference in accuracy between models built with all variables and models built with five variables removed was tested with a Wilcox signed-rank test (n.s., not significant; *, p < 0.01).

	Models built and applied to same region				Models projected into new region		
	within native		within introduced	difference between model sets	from native to introduced		from introduced to native
Iris pseudacorus	0.922 +/- 0.01	*	0.909 +/- 0.01	*	0.648 +/- 0.01	*	0.759 +/- 0.01
Lythrum salicaria	0.921 +/- 0.01	n.s.	0.922 +/- 0.01	*	0.447 +/- 0.01	n.s.	0.457 +/- 0.01
Saccharum ravennae	0.959 +/- 0.01	*	0.859 +/- 0.04	*	0.736 +/- 0.03	*	0.535 +/- 0.07

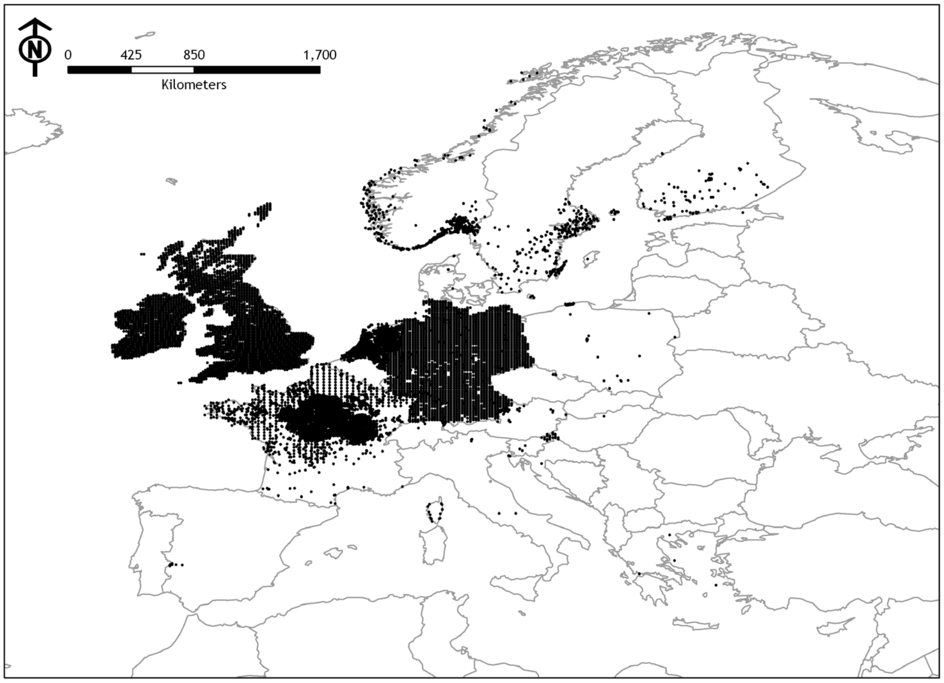
Table 5. Principle components analysis (PCA) of environmental variables associated with occurrence points for the species modelled.

	Iris pseudacorus		*Lythrum salicaria*		*Saccharum ravennae*	
	Eigenvalue	% Variance	Eigenvalue	% Variance	Eigenvalue	% Variance
PC-1	5.752	30.28	6.97	36.7	7.95	41.86
PC-2	5.119	26.96	4.42	23.25	5.29	27.82
PC-3	3.689	19.42	3.18	16.73	2.91	15.3
Total for first 3		76.63		76.68		84.98

Figure 1. Occurrences points used in models for *Iris pseudacorus* in Europe (a) and US (b). Data are from GBIF (see appendix 1 for contributing institutions), USGS Nonindigenous Aquatic Species Program, and Oklahoma Vascular Plants Database.

Figure 1

(a)



(b)

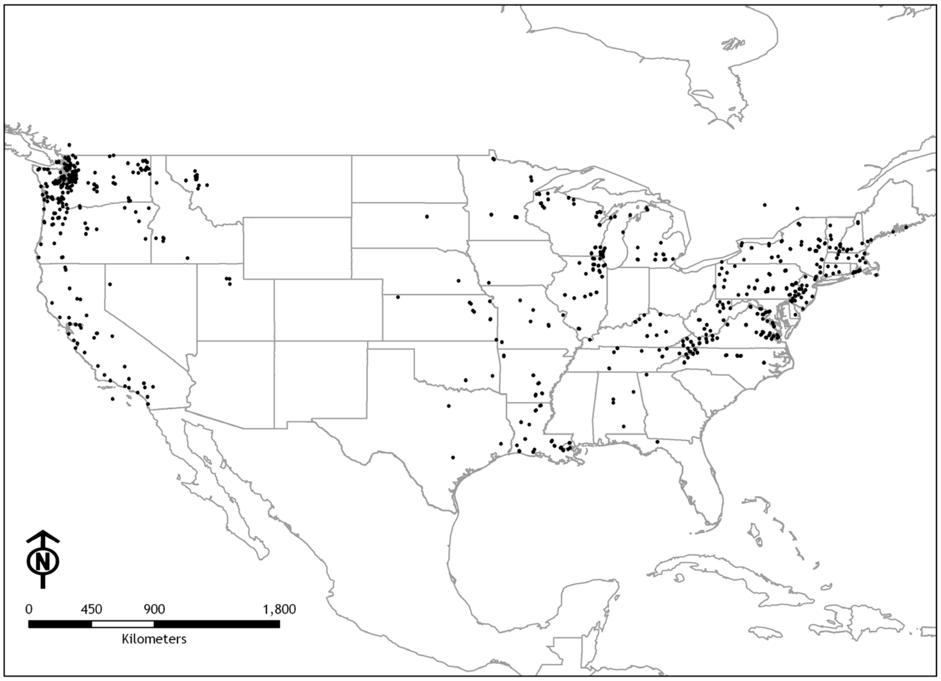
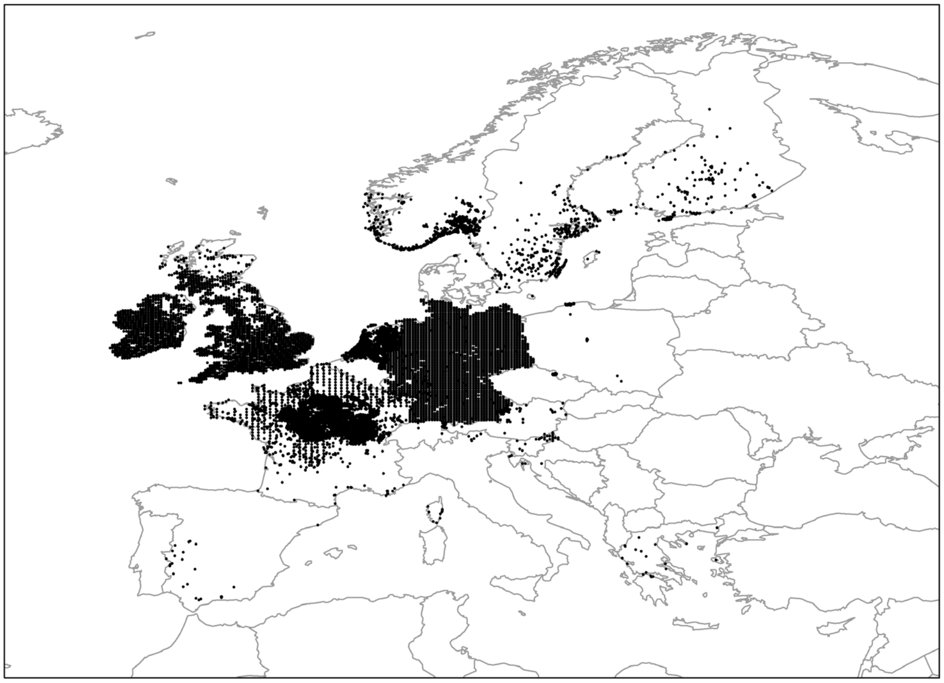


Figure 2. Occurrences points used in models for *Lythrum salicaria* in Europe (a) and US (b). Data are from GBIF (see appendix 1 for contributing institutions), USGS Nonindigenous Aquatic Species Program, and Oklahoma Vascular Plants Database.

Figure 2

(a)



(b)

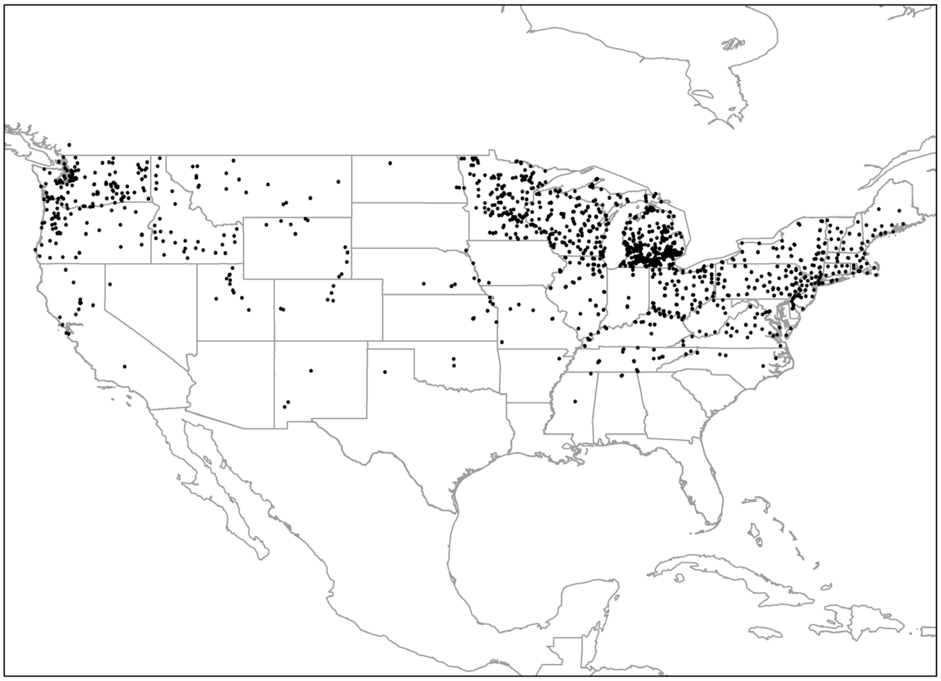
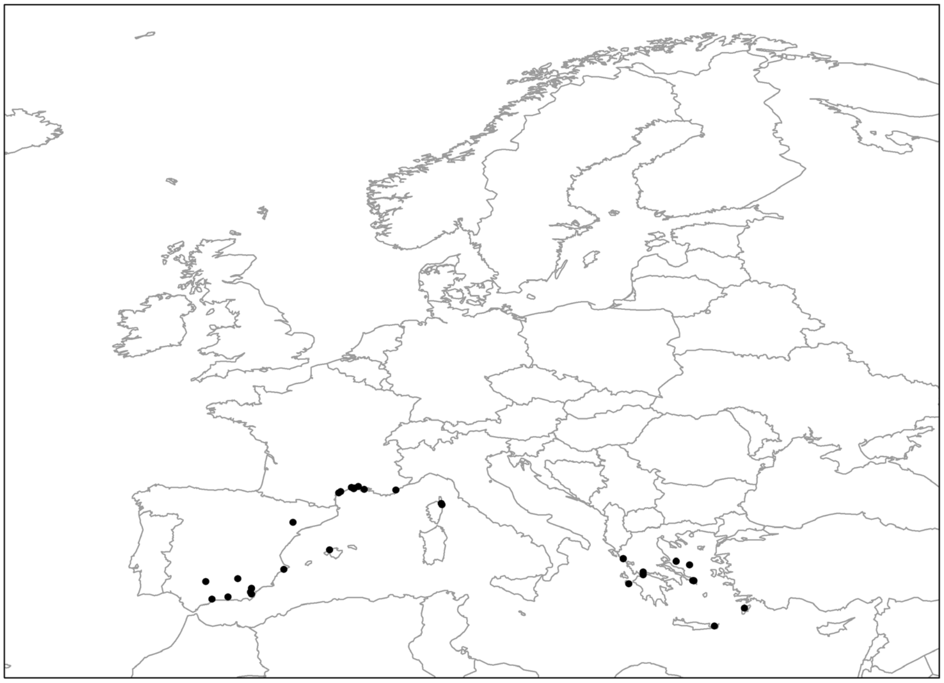


Figure 3. Occurrences points used in models for *Saccharum ravennae* in Europe (a) and US (b). Data are from GBIF (see appendix 1 for contributing institutions), USGS Nonindigenous Aquatic Species Program, and Oklahoma Vascular Plants Database.

Figure 3

(a)



(b)

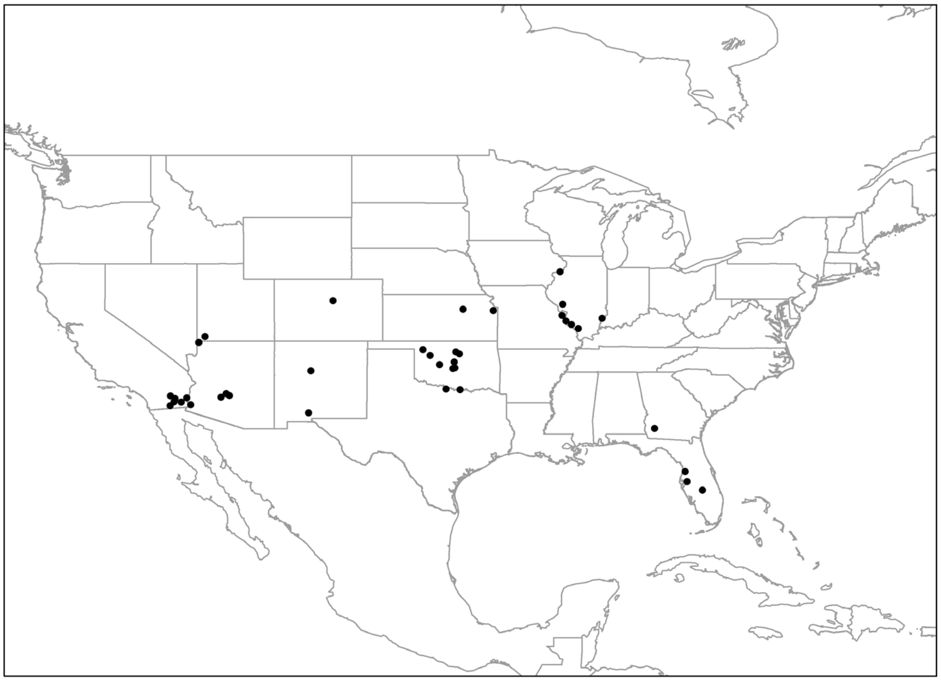
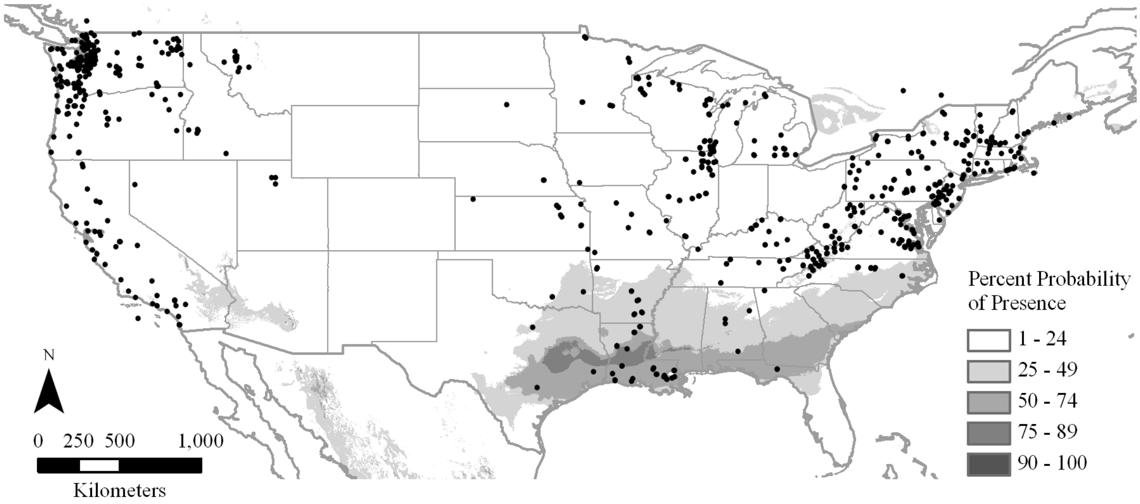


Figure 4. Potential distribution of *Iris pseudacorus* in US based on models built from native range occurrences (a) and introduced occurrences (b). Actual occurrences of the species are indicated with black dots.

Figure 4

(a)


Percent Probability of Presence
1 - 24
25 - 49
50 - 74
75 - 89
90 - 100
N
0 250 500 1,000
Kilometers

(b)

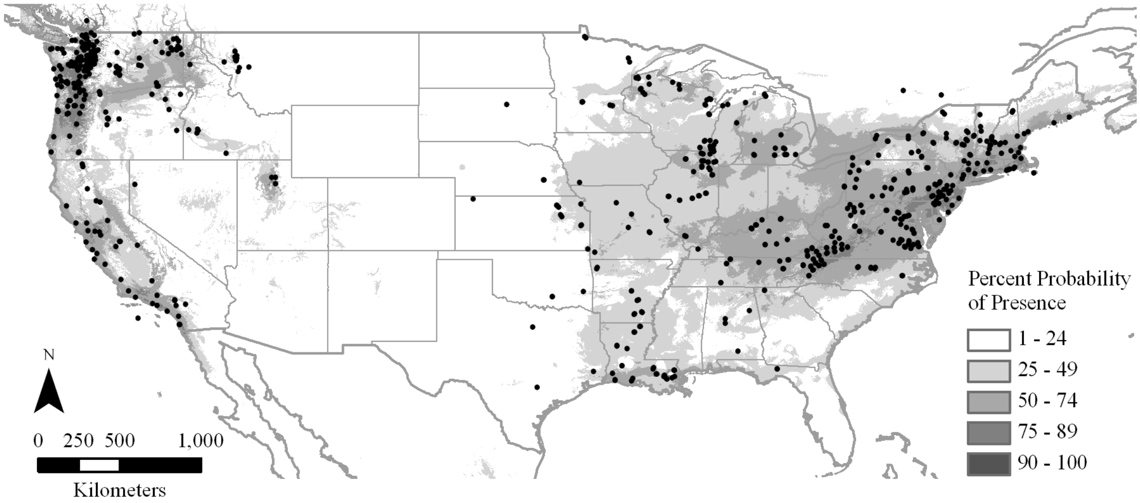
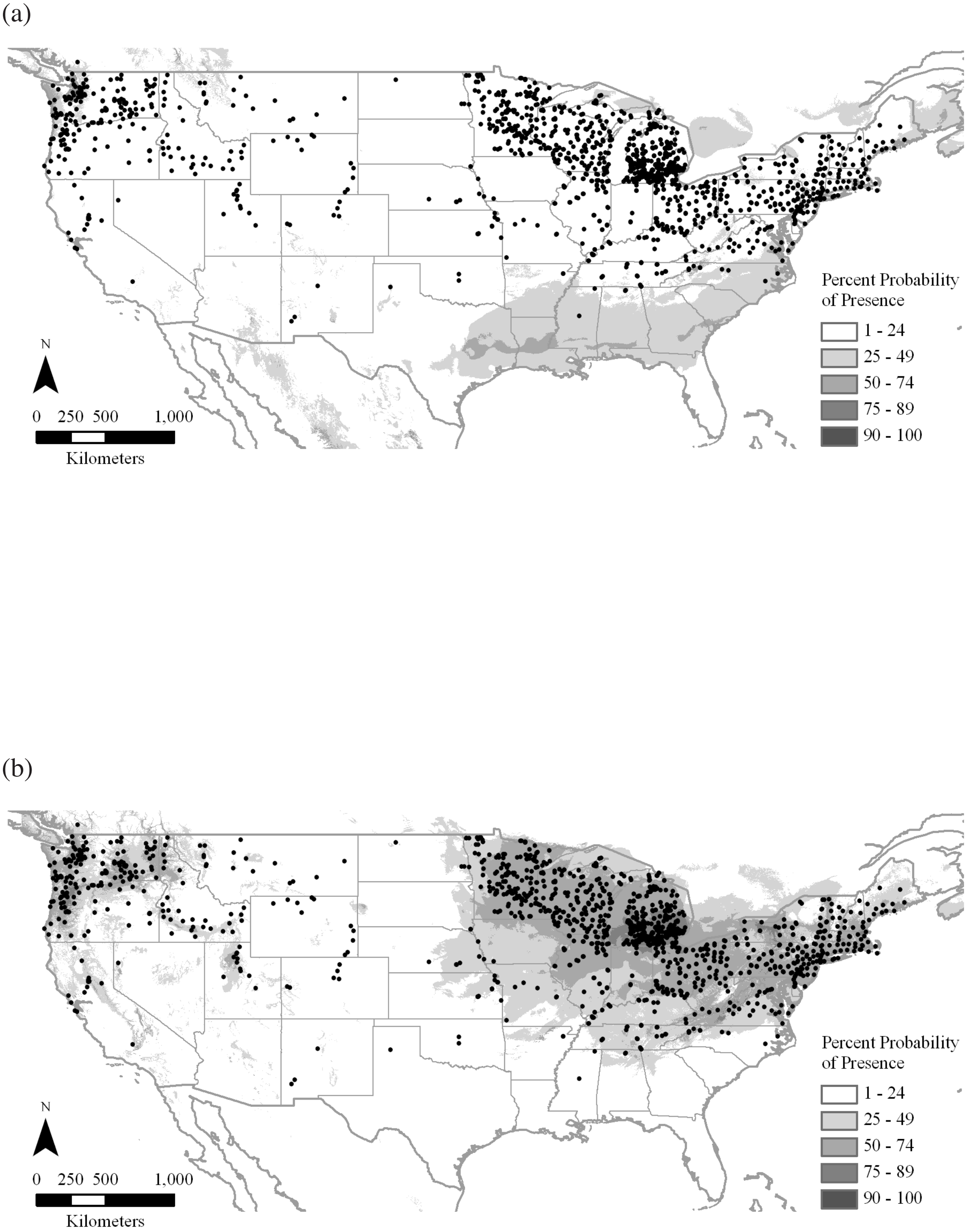

Percent Probability of Presence
1 - 24
25 - 49
50 - 74
75 - 89
90 - 100
N
0 250 500 1,000
Kilometers

Figure 5. Potential distribution of *Lythrum salicaria* in US based on models built from native range occurrences (a) and introduced occurrences (b). Actual occurrences of the species are indicated with black dots.

Figure 5



(a)

Percent Probability of Presence
1 - 24
25 - 49
50 - 74
75 - 89
90 - 100
N
0 250 500 1,000
Kilometers

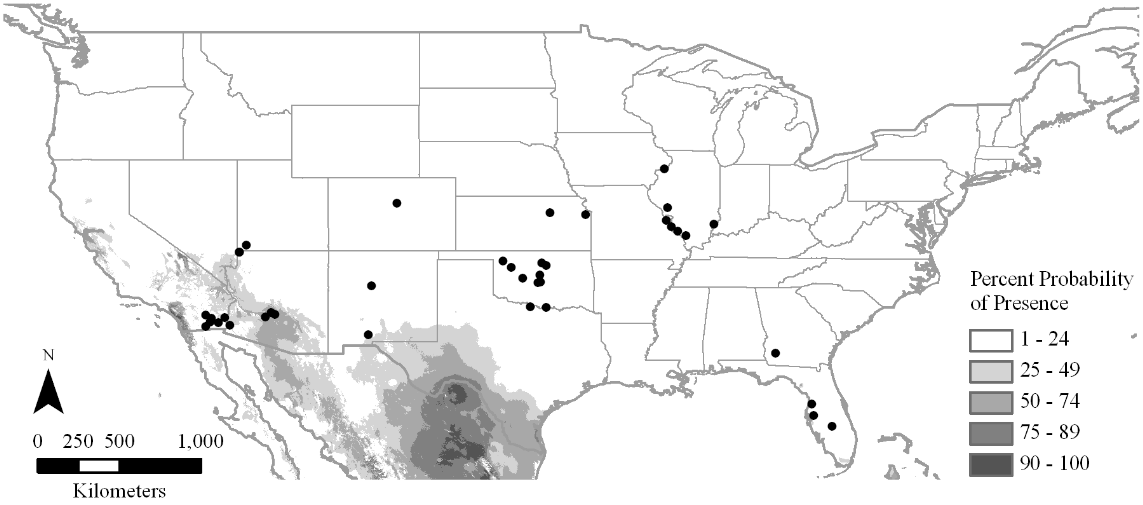
(b)

Percent Probability of Presence
1 - 24
25 - 49
50 - 74
75 - 89
90 - 100
N
0 250 500 1,000
Kilometers

Figure 6. Potential distribution of *Saccharum ravennae* in US based on models built from native range occurrences (a) and introduced occurrences (b). Actual occurrences of the species are indicated with black dots.

Figure 6

(a)


Percent Probability of Presence
1 - 24
25 - 49
50 - 74
75 - 89
90 - 100
N
0 250 500 1,000
Kilometers

(b)

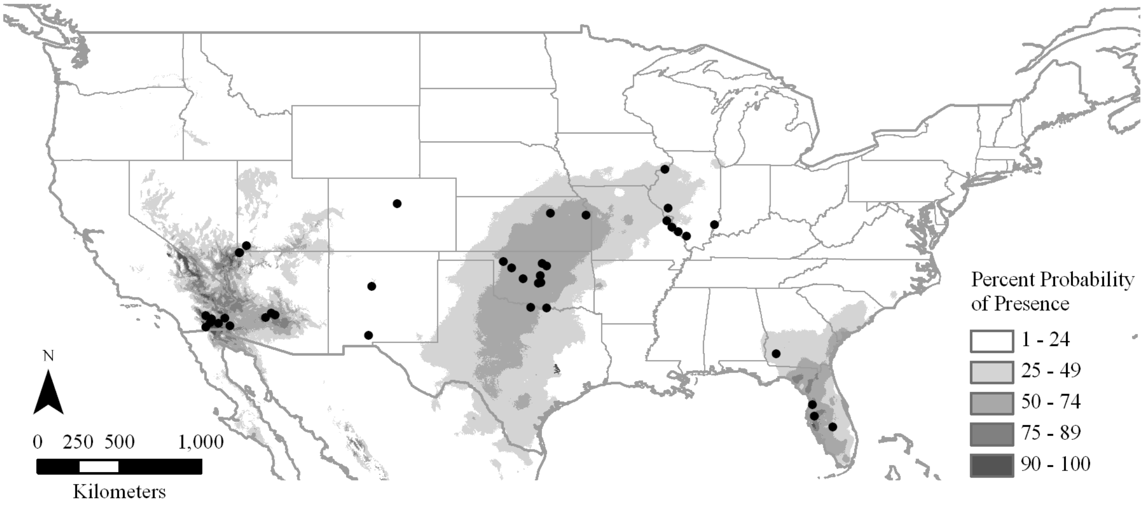

Percent Probability of Presence
1 - 24
25 - 49
50 - 74
75 - 89
90 - 100
N
0 250 500 1,000
Kilometers

Figure 7. PCA of *Iris pseudacorus* occurrences points based on values of bioclimatic variables. Open circles (o) represent the Europe records, crosses (+) represent the US records. Contribution of bioclimatic variables to the distribution of the occurrence points is indicated with the abbreviation of variable, see Table 1.

Figure 7

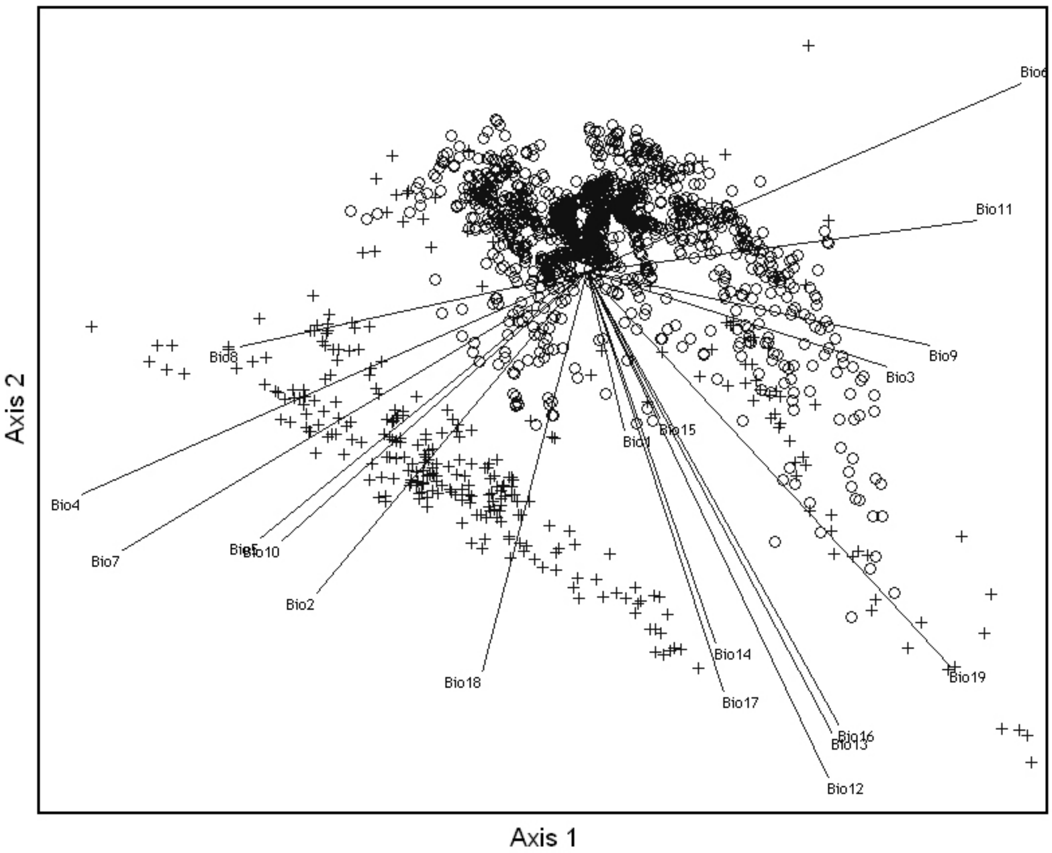

Bio6
Bio11
Bio8
Bio9
Bio3
Bio15
Bio1
Bio4
Bio7
Bio5
Bio10
Bio2
Bio18
Bio14
Bio19
Bio17
Bio16
Bio13
Bio12
Axis 2
Axis 1

Figure 8. PCA of *Lythrum salicaria* occurrences points based on values of bioclimatic variables. Open circles (o) represent the Europe records, crosses (+) represent the US records. Contribution of bioclimatic variables to the distribution of the occurrence points is indicated with the abbreviation of variable, see Table 1.

Figure 8

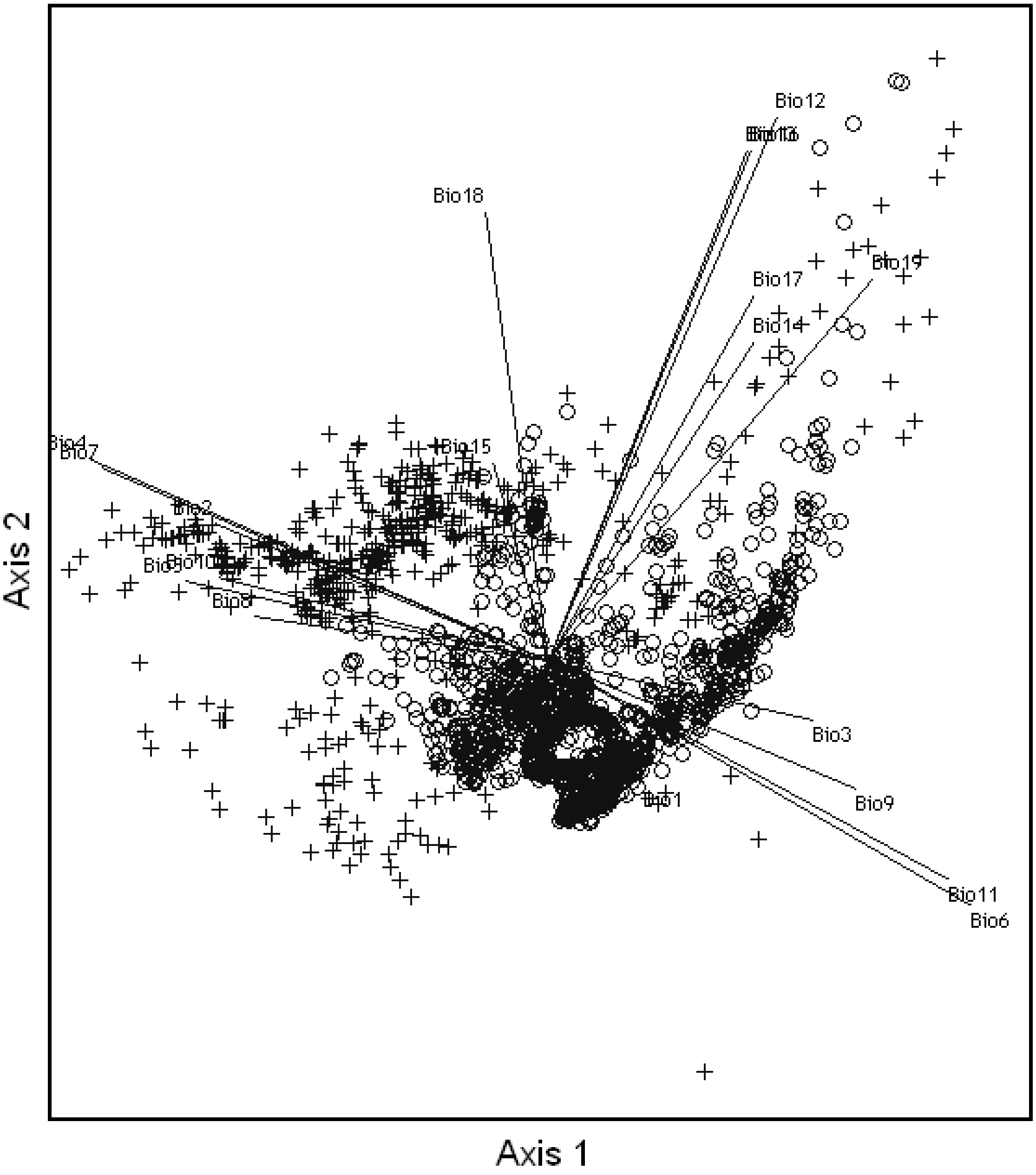
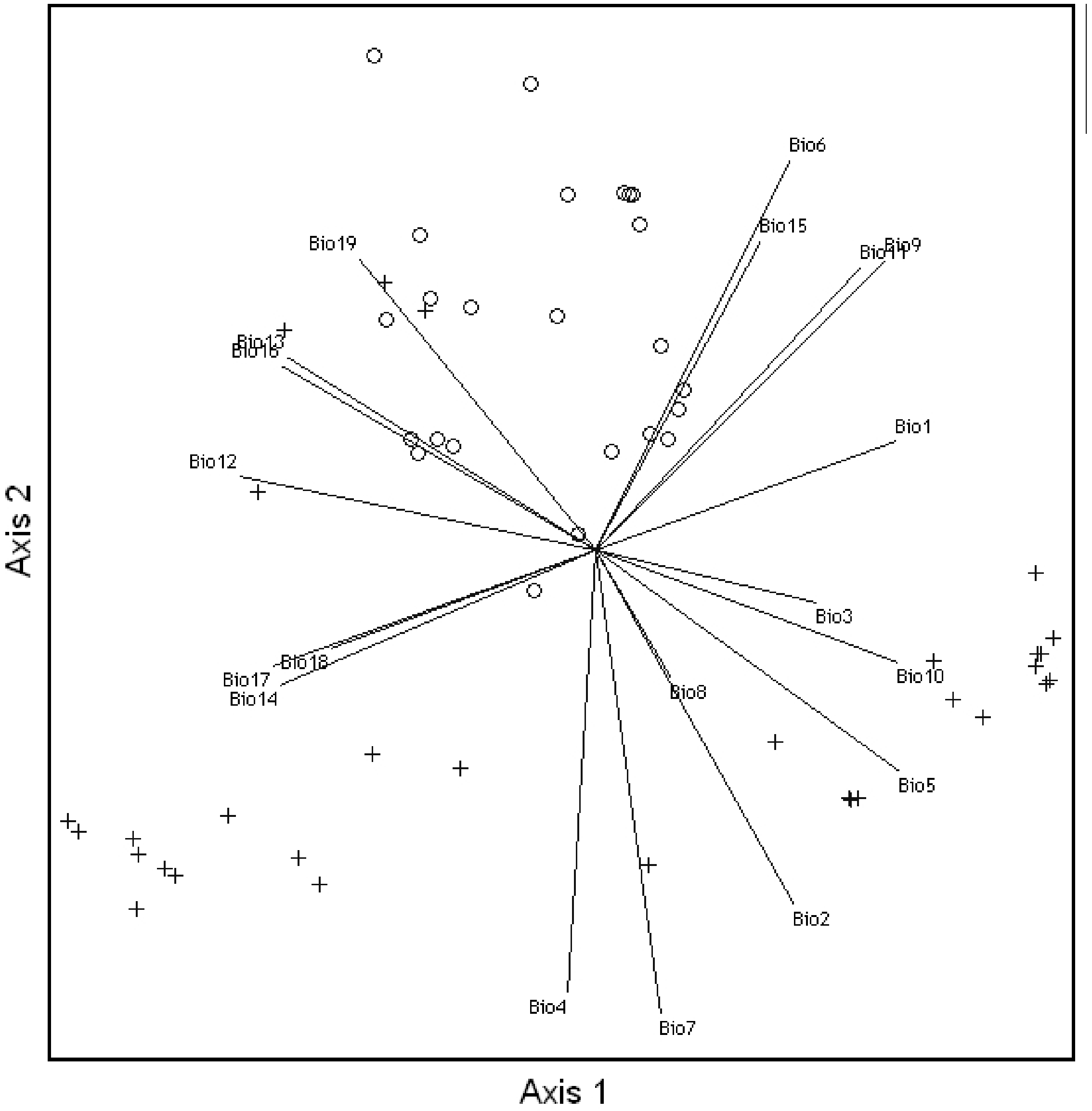

Bio12
Bio13
Bio16
Bio18
Bio17
Bio19
Bio14
Bio4
Bio7
Bio15
Bio2
Bio10
Bio5
Bio8
Bio3
Bio9
Bio1
Bio11
Bio6
Axis 2
Axis 1

Figure 9. PCA of *Saccharum ravennae* occurrences points based on values of bioclimatic variables. Open circles (o) represent the Europe records, crosses (+) represent the US records. Contribution of bioclimatic variables to the distribution of the occurrence points is indicated with the abbreviation of variable, see Table 1.

Figure 9


Bio6
Bio15
Bio11
Bio9
Bio19
Bio13
Bio16
Bio12
Bio1
Bio3
Bio10
Bio18
Bio17
Bio14
Bio8
Bio5
Bio2
Bio4
Bio7
Axis 2
Axis 1

Appendix 1. Data for analyses were obtained from the GBIF data portal from the following institutions:

10. GEO - Tag der Artenvielfalt 2008 - LSG Pfarrhübel Chemnitz (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3381 11/12/2008)

20 Jahre Naturschutzgebiet Dreienberg (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2729 09/12/2008)

3. Tag der Artenvielfalt Hockenheim (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2825 11/12/2008)

4. GEO-Tag in Eberbach (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2736 11/12/2008)

4. Tag der Artenvielfalt, Naturschutzgebiet Hockenheimer Rheinbogen (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2847 11/12/2008)

Ahrschleife bei Altenahr (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3522 11/12/2008)

AKG-Gelände (Bensheim) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2639 09/12/2008)

Angiosperm specimens of Shoji Sasamura of Iwate Prefectural Museum (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1800 09/12/2008)

Arizona State University Vascular Plant Herbarium (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/676 09/12/2008)

Artenfülle um das Schalkenmehrener Maar (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2722 11/12/2008)

Artenvielfalt auf der Weide - GEO-Hauptveranstaltung in Crawinkel (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2697 09/12/2008)

Artenvielfalt der Nordsee - Bremerhaven (Dorum-Neufeld) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2716 11/12/2008)

Artenvielfalt in der Stadt: Botanischer Garten Wuppertal und Hardt (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3385 09/12/2008)

Artenvielfalt Kreis Gießen (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2972 11/12/2008)

Außengelände KITA Mäuseburg Waldkirchen (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3074 09/12/2008)

Australian National Herbarium (CANB) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/47 09/12/2008)

Bäche, Quellen und Teiche im FFH-Gebiet Mühlhauser Halde (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3160 11/12/2008)

Bammentaler Duft- und Heilkräutergarten (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3115 09/12/2008)

Bannwald Burghauser Forst (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3379 09/12/2008)

BDBCV - III Semana de la Biodiversidad (Alicante, Spain), 2008 (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/7926 11/12/2008)

Bergbaufolgelandschaft am Muldestausee (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2751 11/12/2008)

Bergkamen- Bergehalde Großes Holz (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2797 09/12/2008)

Berkel (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/7871 11/12/2008)

Bernhardsthal (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3398 09/12/2008)

Beweidungsprojekt an der Nesse (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2938 09/12/2008)

Binsenwiesen (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3113 11/12/2008)

Biodiversidad de Costa Rica (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/333 11/12/2008)

Biologiezentrum Linz (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1104 11/12/2008)

Biologische Station im Kreis Wesel (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2703 11/12/2008)

Biosphärenpark Wienerwald - Wiener Steinhofgründe (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3392 11/12/2008)

Biotop Kohlbeke (Berlin-Marzahn) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2954 11/12/2008)

Biotop Binsenwiesen (Wehrheim/Taunus) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2761 11/12/2008)

Biotop Binsenwiesen und Ernst-Reiter-Wiese (Wehrheim/Taunus) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3062 11/12/2008)

Bishop Museum Natural History Specimen Data (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/54 11/12/2008)

Bizzenbach-Aue im Bizzenbachtal (Wehrheim/Taunus) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2835 11/12/2008)

Bizzenbachtal (Wehrheim/Taunus) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2809 09/12/2008)

Bodenseeufer Radolfzell (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2991 11/12/2008)

Bodenteicher Seewiesen (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3515 11/12/2008)

Bolzplatz (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3031 11/12/2008)

Borkhart (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2933 11/12/2008)

Borstgrasrasen um die Burg Baldenau im Oberen Dhrontal (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3107 09/12/2008)

Botanic Garden of the Finnish Museum of Natural History (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2406 09/12/2008)

Botánica, Universidad de León: LEB-Cormo (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/260 09/12/2008)

Botanical Garden Collection (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/64 11/12/2008)

Botanical Garden Yoshkar-Ola (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1392 11/12/2008)

Botanical Museum, Copenhagen. Database of type specimens (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/716 11/12/2008)

Botanical Society of the British Isles - Vascular plant data for Scottish Vice-counties (VCs 80, 84, 103 & 104) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1887 09/12/2008)

Botanical Society of the British Isles - Vascular Plants Database (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/839 09/12/2008)

Botanical specimens database of Mr. Jiro Ito collection, Shizuoka Prefecture Museum of Natural History (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1811 11/12/2008)

Botanischer Garten (Saarbrücken) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2641 09/12/2008)

Botanischer Garten Bochum (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1366 09/12/2008)

Botanischer Garten Bonn (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1374 09/12/2008)

Botanischer Garten Darmstadt (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1361 09/12/2008)

Botanischer Garten der Christian-Albrechts-Universitat zu Kiel (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1378 11/12/2008)

Botanischer Garten Frankfurt (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1372 09/12/2008)

Botanischer Garten Gie?en (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1362 09/12/2008)

Botanischer Garten Graz (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1390 09/12/2008)

Botanischer Garten Jena (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1381 09/12/2008)

Botanischer Garten Krefeld (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1379 09/12/2008)

Botanischer Garten Marburg (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1364 09/12/2008)

Botanischer Garten Munster (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1383 09/12/2008)

Botanischer Garten Osnabruck (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1382 09/12/2008)

Botanischer Garten Rostock (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1369 09/12/2008)

Botanischer Garten Saarbrucken (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1376 09/12/2008)

Botanischer Garten TU Dresden (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1363 09/12/2008)

Botanischer Garten Ulm (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1359 09/12/2008)

Botany (UPS) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1045 09/12/2008)

Botany registration database by Danish botanists (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/703 09/12/2008)

Breitkopfbecken (Berlin-Reinickendorf) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3096 11/12/2008)

Bronx River Bioblitz (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/733 11/12/2008)

BÜG (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2628 09/12/2008)

BUND - Dassower See (Lübeck/Dassow) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2707 09/12/2008)

Bundesamt fuer Naturschutz / Netzwerk Phytodiversitaet Deutschland (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1098 09/12/2008)

California State University, Chico (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/737 09/12/2008)

Canadian Museum of Nature Herbarium (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/123 11/12/2008)

Civico Orto Botanico Trieste (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1389 11/12/2008)

CONN GBIF data (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/7857 09/12/2008)

Cuxhavener Küstenheiden (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2695 11/12/2008)

Danielsberg (Mölltal, Kärnten) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2636 09/12/2008)

Danisco-Wiese (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2698 09/12/2008)

Database Schema for UC Davis [Herbarium Labels] (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/734 09/12/2008)

Departamento de Biolog. Veg. II, Facultad de Farmacia, Universidad Complutense, Madrid: MAF (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/249 09/12/2008)

Deponie Klausdorf (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2976 09/12/2008)

Die Wuhle (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3011 11/12/2008)

Dierloch, nördlicher Mooswald (Freiburg-Hochdorf) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2952 09/12/2008)

Dirección General de Investigación, Desarrollo Tecnológico e Innovación de la Junta de Extremadura(DGIDTI): HSS (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/291 09/12/2008)

Döchtbühlwald (Bad Waldsee) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2967 11/12/2008)

Dorset Environmental Records Centre - Bryophyte Survey of the Poole Basin Mires - NBN South West Pilot Project Case Studies (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/835 11/12/2008)

Dpto de Botánica, Ecología y Fisiología Vegetal (herbario_cofc).Facultad de Ciencias.Universidad de Córdoba (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/292 09/12/2008)

Draubiotop Lavamünd (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3245 11/12/2008)

E.C. Smith Herbarium (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1829 09/12/2008)

East Ayrshire Countryside Ranger Service - East Ayrshire Species Database (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1717 11/12/2008)

Ehmkendorf (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2944 11/12/2008)

EKY_Darwincore (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/7894 11/12/2008)

Entdeckertour am Muldestausee (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2709 09/12/2008)

Entomology Department Collections, ZMUC (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/711 11/12/2008)

Environment and Heritage Service - EHS Species Datasets (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/940 09/12/2008)

Eppingen und Umgebung (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2816 09/12/2008)

Erlengraben/Lipp-Tal (Östringen) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2675 09/12/2008)

EUNIS (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/198 09/12/2008)

EURISCO, The European Genetic Resources Search Catalogue (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1905 09/12/2008)

Fairchild Tropical Botanic Garden Virtual Herbarium Darwin Core format (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/202 11/12/2008)

Fern specimens collected by Mr. Hisaya Manago, Aquatic plant specimen database of Dr. Shigeru Miki collection (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/608 09/12/2008)

Feuchtbiotop, Wildtier- und Artenschutzstation Sachsenhagen, Sielmanns Natur-Ranger (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3226 11/12/2008)

Feuchtwiese am Nationalpark-Haus Neuwerk (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3590 11/12/2008)

Feuchtwiese Grüne Mitte, Klasse 5a (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3588 11/12/2008)

FFH-Gebiet Ahrbachtal (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2640 09/12/2008)

FFH-Gebiet Paartal (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3558 11/12/2008)

Fledermaus (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2948 11/12/2008)

Flora of Slowinski National Park, Poland (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2022 09/12/2008)

FloVegSI - Floristical and fitocenological database of ZRC SAZU (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2585 09/12/2008)

Föhrenried (Fronreute und Baindt) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2970 09/12/2008)

Forstbotanischer Garten Tharandt (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1370 09/12/2008)

Frauenholz (Holzmaden) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2668 11/12/2008)

Freiburger Netzwerk Artenvielfalt (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/7866 11/12/2008)

Freiburger Tag der Artenvielfalt (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2669 11/12/2008)

Freigelände Naturschutzscheune Reinheimer Teich (Kreis Darmstadt-Dieburg) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2845 11/12/2008)

Frohlinder Mühlenbach (Dortmund-Kirchlinde) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2803 11/12/2008)

Fruit and seed collection database (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1093 11/12/2008)

Fuldaaue (Stadtgebiet Fulda) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2790 11/12/2008)

Garten J. Scherrer (Lachen-Speyerdorf) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3069 11/12/2008)

Geführte Wanderung im Eselsbachtal (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3561 11/12/2008)

Gelände der Lahntalschule Biedenkopf und Lahnauen (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2982 11/12/2008)

Gelände des Schulzentrums am Himmelsbarg (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3136 11/12/2008)

Gemeinde Sursee (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2652 09/12/2008)

Gemeindegebiet Weikendorf (Marchfeld) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2765 09/12/2008)

GEO Biodiversity Day (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1094 11/12/2008)

GEO Hauptveranstaltung Tirol (Innsbruck) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2662 11/12/2008)

GEO-Hauptveranstaltung (Duisburg) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2705 09/12/2008)

GEO-Hauptveranstaltung (Insel Vilm) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2704 11/12/2008)

GEO-Hauptveranstaltung (NLP Harz / Hochharz) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2643 11/12/2008)

GEO-Hauptveranstaltung im Nationalpark Bayerischer Wald (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3378 09/12/2008)

Georgs-Padd (Wangerooge) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3298 11/12/2008)

Geo-Tag der Artenvielfalt Süßen Hornwiesen-Grundschule (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2783 11/12/2008)

Gesamtartenliste Bremerhaven, Helgoland und Sylt (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2689 11/12/2008)

Geschützter Landschaftsbestandteil - GLB Troppach (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3014 09/12/2008)

Gewann Krampf (Heilbronn) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2653 09/12/2008)

Gewässer des Wartbergparks Stuttgart (bei der Ökostation der VHS Stuttgart) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3124 11/12/2008)

Gronau - auf der Suche nach dem Neunauge (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3490 11/12/2008)

Gruga-Park Essen (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1384 09/12/2008)

Gurgltal (Tarrenz) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2727 09/12/2008)

Gymnicher Mühle (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/7906 11/12/2008)

Hainhoop - Tonkuhle - Bullenmoor (Arpke) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2951 11/12/2008)

Hamberger Brücke / Würmtal (Pforzheim) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2644 09/12/2008)

Harvard University Herbaria (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1827 11/12/2008)

Hatikka Observation Data Gateway (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2401 09/12/2008)

Hatikka Observation Data Gateway (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2401 11/12/2008)

Haus der Natur Salzburg
(accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1488 11/12/2008)

Heinersdorfer Sumpfwiese (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2734 09/12/2008)

Heinersdorfer Sumpfwiese (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2734 11/12/2008)
herbario (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/566 11/12/2008)
Herbario de la Universidad de Arizona, EUA (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2479 11/12/2008)
Herbario de la Universidad de Salamanca: SALA (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/239 09/12/2008)
Herbario de la Universidad de Sevilla, SEV (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/283 09/12/2008)
Herbario de la Universidad de Sevilla, SEV-Historico (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/284 09/12/2008)
Herbario del Instituto de Ecología, A.C., México (IE-BAJIO) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1595 11/12/2008)
Herbarium (AMNH) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/232 11/12/2008)
Herbarium (ICEL) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/231 11/12/2008)
Herbarium (UNA) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/775 11/12/2008)
Herbarium des Staatlichen Museums für Naturkunde Görlitz (GLM) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1105 09/12/2008)
Herbarium Faeroense (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/713 11/12/2008)
Herbarium GJO (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1484 09/12/2008)
Herbarium GZU (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1491 11/12/2008)
Herbarium of Kitakyushu Museum of Natural History and Human History (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/606 09/12/2008)
Herbarium of National Centre for Plant Genetic Reosurces (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/227 09/12/2008)
Herbarium of Oskarshamn (OHN) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1024 09/12/2008)
Herbarium of the Bia_owie_a Geobotanical Station (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1470 09/12/2008)
Herbarium Senckenbergianum (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1654 09/12/2008)
Herbarium Specimens of Museum of Nature and Human Activities, Hyogo Pref., Japan (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/589 09/12/2008)
Herbarium Specimens of Tokushima Prefectural Museum, Japan (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/600 11/12/2008)
Herbarium Universitat Ulm (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1224 09/12/2008)
Herbarium W (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1479 09/12/2008)
Herbarium Willing (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1096 09/12/2008)
Herbarium WU (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1496 09/12/2008)
Herbier de la Guyane (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1436 11/12/2008)
Herbier de Strasbourg (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1849 09/12/2008)
Herrensee-Gebiet (Fischbachtal im Odenwald) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3055 09/12/2008)
Hintere Halde (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2830 09/12/2008)
Hortus Botanicus Sollerensis Herbarium (FBonafè) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/300 11/12/2008)
Ibaraki Nature Museum, Dr.Masatomo Suzuki collection:Vascular Plants (1) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1813 09/12/2008)
inatura - Erlebnis Naturschau Dornbirn (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1866 11/12/2008)
Institut Botanic de Barcelona, BC (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/299 09/12/2008)
Institut d'Ecologia Litoral: IEL_Plantae (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/263 11/12/2008)
Internation Botanical Collections (S) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1983 09/12/2008)
Inventaire national du Patrimoine naturel (INPN) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2620 09/12/2008)
IPK Genebank (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1851 11/12/2008)
Israel Nature and Parks Authority (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1431 09/12/2008)
Issumer Fleuth (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3252 09/12/2008)
Jardín Botánico de Córdoba: Herbarium COA (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/247 09/12/2008)
Jardin Botanique de la Ville Lyon (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1388 09/12/2008)
Joint Nature Conservation Committee - Vegetation surveys of coastal shingle in Great Britain (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/849 09/12/2008)
KARSTLANDSCHAFT SÜDHARZ - VOM GIPSABBAU BEDROHT (Grenzstreifen am Röseberg) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2726 09/12/2008)
Kiesbagger (Mittelhausen) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2760 11/12/2008)
Kiesgruben Wemb (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2963 09/12/2008)
Kinderbauernhof Pinke-Panke (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3192 11/12/2008)
Klasse 3a (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2929 11/12/2008)
Klutensee (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2631 11/12/2008)
Knechtweide (Kohlfurth) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2742 09/12/2008)
Königstetten (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2667 09/12/2008)
Korean Ethnobotany Database (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/111 11/12/2008)
Kremmer Luch (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2937 11/12/2008)
Kurashiki Museum of Natural History (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/599 09/12/2008)

Küste Wismar-Wendorf bis Hoben (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2818 11/12/2008)

Küstenschutzwald (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2934 11/12/2008)

LaBoOb02 (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2629 09/12/2008)

Landschaftspark St.Leonhard-Deisendorf (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3161 11/12/2008)

Landschaftspflegehof (Berlin) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2656 11/12/2008)

Landschaftsschutzgebiet Holmer Sandberge (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3040 11/12/2008)

Landschaftsschutzgebiet Schmutterwald (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3375 09/12/2008)

Langenberger Forst am Ochsenweg/ Niebüll-Leck (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2658 11/12/2008)

Langes Tannen (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2682 11/12/2008)

Langes Tannen (Uetersen) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2671 11/12/2008)

Laubenheimer Bodenheimer Ried - von Stromtalwiesen und Flutrasen (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3501 09/12/2008)

Leben im Finkensteiner Moor (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3154 11/12/2008)

Lebensraum Fluß/Zwickauer Mulde in Wolkenburg (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2973 11/12/2008)

Lebensraum Gesamtschule (Langerwehe) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2767 09/12/2008)

Leiner-Herbar Konstanz (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1473 09/12/2008)

Liether Kalkgrube (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3507 11/12/2008)

Liether Park 1 (LMS), Klasse 5b (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3530 11/12/2008)

Liether Park 2 (LMS), Klasse 6c (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3492 11/12/2008)

Limnodata (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1466 09/12/2008)

Lindau im Bodensee (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2801 11/12/2008)

LK 11 im Mönchspark (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3396 11/12/2008)

Lothian Wildlife Information Centre - Lothian Wildlife Information Centre Secret Garden Survey (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/856 11/12/2008)

Luch Niederlehme, Schüler der Klasse 7 (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2719 11/12/2008)

Lund Botanical Museum (LD) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1028 09/12/2008)

Lustadter Wald . (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/7904 09/12/2008)

Lustbach-Umland (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3494 09/12/2008)

Magnoliophyta- Taiwan Biodiversity Data for GBIF (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/727 11/12/2008)

Mainufer (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3043 11/12/2008)

MEXU/Plantas Vasculares (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/780 11/12/2008)

MISS_DC_01MAR2006 (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/7895 09/12/2008)

Mißmahlsche Anlage (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2852 11/12/2008)

Missouri Botanical Garden (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/621 09/12/2008)

Mooswald (Freiburg) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2651 11/12/2008)

Müritz-Nationalpark (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3384 11/12/2008)

Museum of Natural History, Wroclaw University, Flora of the Sto_owe Mts. (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1456 09/12/2008)

NABU Naturschutzhof Netttetal (Sassenfeld) e.V. (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2759 09/12/2008)

NABU-Auerochsenweide (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3118 11/12/2008)

NABUGEO1 (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3140 09/12/2008)

NABU-Projekt (Osterode am Harz) Südharzer Gipskarst (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2821 09/12/2008)

Nationaal Herbarium Nederland (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1211 09/12/2008)

National System of Proetcted Areas (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1791 09/12/2008)

National Vegetation Data bank (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2471 09/12/2008)

Natur aus zweiter Hand am Muldestausee (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2770 11/12/2008)

Natural History Museum Rotterdam (NMR) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/693 09/12/2008)

Natur-Erlebnis-Kindergarten Waldkirchen/Erzgebirge (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3090 11/12/2008)

Naturerlebnisraum Koppelsberg (Plön) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3132 11/12/2008)

NatureServe Network Species Occurrence Data (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/607 09/12/2008)

Naturgarten Langenholtensen (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2857 09/12/2008)

Naturgrundstück (Eutin) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2961 11/12/2008)

Naturnahes Tal in Siena (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/7909 11/12/2008)

Naturparadies in Gräfenhausen am Trifels (bei Annweiler) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3093 09/12/2008)

Naturpark Drömling (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/7864 09/12/2008)

Naturschutzgebiet Gellener Torfmöörte (Landkreis Wesermarsch) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3338 11/12/2008)

Naturschutzgebiet Bausenberg (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2657 09/12/2008)

Naturschutzgebiet Börstig bei Hallstadt (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3485 09/12/2008)

Naturschutzgebiet Kochertgraben (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3233 11/12/2008)

Naturschutzgebiet Lippeaue (Marl) - Pfadis in Sickingmühle (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3087 11/12/2008)
Naturschutzgebiet Lochbusch-Königswiesen (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3094 11/12/2008)
nazza (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2699 11/12/2008)
Neanderthal (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3131 11/12/2008)
Neckartalsüdhang (Horb) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2680 11/12/2008)
Neuer Botanischer Garten Gottingen (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1373 09/12/2008)
New Mexico Biodiversity Collections Consortium database (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3607 09/12/2008)
New Zealand Biodiversity Recording Network (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/7910 11/12/2008)
New Zealand National Plant Herbarium (CHR) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/474 09/12/2008)
NMNH Botany Collections (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1874 09/12/2008)
Nordic Herbarium (S) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1025 09/12/2008)
NSG Haunestausee, Hauneteiche (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/7876 09/12/2008)
NSG Karwendel (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2678 11/12/2008)
NSG Leist bei Ziegenhain (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3097 11/12/2008)
NSW herbarium collection (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/968 09/12/2008)
NW-Innenhof Gesamtschule Herten 7.6.2001 (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3321 11/12/2008)
Observational database of Icelandic plants (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/233 11/12/2008)
Observations du Conservatoire botanique national du Bassin parisien. (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1103 09/12/2008)
Oklahoma Vascular Plants Database Provider (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2558 09/12/2008)
Okologisch Botanischer Garten Bayreuth (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1360 09/12/2008)
Ökostation (Freiburg) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2750 09/12/2008)
Orto Botanico di Pisa (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1386 09/12/2008)
Paleobiology Database (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/563 11/12/2008)
Panke und Ufer am Kinderbauernhof Pinke-Panke (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2995 11/12/2008)
Perchtoldsdorfer Heide (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/7863 09/12/2008)
Phanerogamie (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1506 09/12/2008)
Phanerogamie (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1506 11/12/2008)
Philosophenwald und Wieseckaue in Gießen (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2690 09/12/2008)
Phragmites of Canada (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/526 11/12/2008)
Pilstingermoos (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2721 09/12/2008)
Plant (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/469 09/12/2008)
Plant Observation Records of Japan (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2547 09/12/2008)
Plant observations from Bia_owie_a National Park (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1861 09/12/2008)
Plant specimens depodited in Osaka Museum of Natural History, Japan. (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1973 09/12/2008)
Plant Systematics Laboratory, Ajou University, Korea (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2469 11/12/2008)
Plants (GBIF-SE:Artdatabanken) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1034 09/12/2008)
Please cite this data as follows:
Pottundkopp (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2741 11/12/2008)
Priest Pot species list, Cumbria, Britain (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/717 09/12/2008)
privater Garten (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3016 11/12/2008)
Promberg1 (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2702 09/12/2008)
Prophetensee Quickborn (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2826 11/12/2008)
Quarrendorfer Landschaftsschutzgebiet (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2778 09/12/2008)
Real Jardin Botanico (Madrid), Vascular Plant Herbarium (MA) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/240 09/12/2008)
Regenrückhaltebecken (Zeulenroda) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2974 11/12/2008)
Regenwasserabfangsbecken (Erlenbach) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3133 11/12/2008)
Regionalpark(Hattersheim) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2753 09/12/2008)
renaturierter Main (Kemmern bei Bamberg) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2823 11/12/2008)
Renaturierung Werse (Innenbereich Beckum) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2795 11/12/2008)
Repatriación de datos del Herbario de Arizona (ARIZ) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2480 11/12/2008)
Ried und Sand - Artenvielfalt durch Beweidung (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3023 09/12/2008)
Riedensee (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2724 11/12/2008)

Rohrmeistereiplateau und angrenzendes Gebiet (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3382 11/12/2008)

Rosarium (LMS), Klasse 6a (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3580 11/12/2008)

Royal Botanic Gardens, Kew (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/629 11/12/2008)

Royal Botanical Gardens Herbarium (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/512 09/12/2008)

Royal Museum of Central Africa - Metafro-Infosys - Xylarium (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/95 09/12/2008)

Rund um das LUGY (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3022 11/12/2008)

Rund um den Eichwald,Schulhof Friedrich Fröbel Gymnasium- Bad Blankenburg (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2684 11/12/2008)

Rund ums Cani (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3128 11/12/2008)

Salzwiese Diekskiel (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3293 11/12/2008)

SANT herbarium vascular plants collection (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/222 09/12/2008)

Schatzinsel Wangerooge (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3493 11/12/2008)

Schlern - (Bozen) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2661 09/12/2008)

Schüler erforschen die Helme-Aue (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3577 11/12/2008)

Schulgarten der Volksschule (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3511 11/12/2008)

Schulgarten Hans-Carossa-Oberschule (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3027 11/12/2008)

Schulgarten mit Klasse 8a (Essen) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2966 11/12/2008)

Schulgarten-St.-Georg-Gymnasium (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3248 11/12/2008)

Schulgelände Ceciliengymnasium (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3224 11/12/2008)

Schulgelände Schule auf der Aue, Münster (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2771 11/12/2008)

Schulhof der Astrid-Lindgren-Schule Elmshorn (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3092 11/12/2008)

Schulprojekt (Bremen) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2789 11/12/2008)

Schulteich Freie Waldorfschule Darmstadt (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3335 11/12/2008)

Schulteich Heinrich-Mann-Schule (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3253 11/12/2008)

Schulumfeld Albert-Einstein-Gymnasium (Sankt Augustin) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2764 09/12/2008)

Schulzentrum Parc Hosingen (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3394 11/12/2008)

Schussenaue (Weingarten) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2833 09/12/2008)

Schussenaue bei Berg (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3020 09/12/2008)

Schwanheimer Wald (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/7865 09/12/2008)

Schwanseepark (87645 Schwangau) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3058 11/12/2008)

Scottish Borders Biological Records Centre - SWT Scottish Borders Local Wildlife Site Survey data 1996-2000 - species information (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/848 09/12/2008)

Selz-Renaturierung (Hahnheim/Sörgenloch) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3255 11/12/2008)

Selztal bei Friesenheim (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3091 09/12/2008)

Siegen/ Gymnasium Am Löhrtor (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2932 11/12/2008)

Silbertor + Wasserbachtal (Rutesheim / Renningen) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2677 11/12/2008)

Sonnentaugemeinschaft (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2686 11/12/2008)

Spandau HBO (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2840 11/12/2008)

Specimen Database of Colorado Vascular Plants (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1832 09/12/2008)

Spießwoogtal / Königsbruch (Fischbach) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3049 11/12/2008)

Spreewaldfließe und Feuchtwiese bei Lübbenau (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3246 09/12/2008)

Staatliches Museum für Naturkunde Stuttgart, Herbarium (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1100 09/12/2008)

Stadtgebiet (Dannenberg) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2792 11/12/2008)

Stadtpark Herzberg (Elster) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2979 09/12/2008)

Stadtpark Sulzbach-Rosenberg (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2800 11/12/2008)

Stausee (Oberdigisheim/Meßstetten) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2673 11/12/2008)

Steinbruch Mainz-Weisenau (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2994 11/12/2008)

Stever (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3030 09/12/2008)

Streuobstwiese Stedar (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3506 11/12/2008)

Streuobstwiesengelände St.Meinrad Gymnasium (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3065 09/12/2008)

Sudeniederung (Amt Neuhaus) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3260 11/12/2008)

Sudeniederung (Amt Neuhaus), Landkreis Lüneburg (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2715 11/12/2008)

Sürther Aue (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3512 11/12/2008)

SysTax (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1875 09/12/2008)

Tag der Artenvielfalt (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2861 11/12/2008)

Tag der Artenvielfalt in Heidelberg (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3486 11/12/2008)

Tag der Artenvielfalt mit SchülerInnen des Europa-Gymnasiums in Wörth am Rhein (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/7872 11/12/2008)

Tag der Artenvielfalt mit SchülerInnen des Leibniz-Gymnasiums in Neustadt a.d.W. (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/7873 11/12/2008)

Tage der Artenvielfalt rund um die Naturschutzstation Molsberg (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/7868 11/12/2008)

Take a Pride in Fife Environmental Information Centre - Records for Fife from TAPIF EIC (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/927 09/12/2008)

Taxa (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/7903 09/12/2008)

The AAU Herbarium Database (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/224 11/12/2008)

The Deaver Herbarium, Northern Arizona University (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/678 11/12/2008)

The Shimane Nature Museum of Mt. Sanbe (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1978 09/12/2008)

Tiere und Pflanzen am Pfannenbach (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3355 09/12/2008)

Tiergarten Straubing (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2806 09/12/2008)

Tiroler Landesmuseum Ferdinandeum (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1509 09/12/2008)

Tongrube bei Hettstedt (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3488 11/12/2008)

Tornoer Teich (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3502 11/12/2008)

Triebesbach (Zeulenroda-Triebes) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2996 09/12/2008)

Tümpel Schulbiologiezentrum (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3285 11/12/2008)

Type herbarium, Gottingen (GOET) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1494 11/12/2008)

UA Herbarium (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/7900 11/12/2008)

UAM Botany Specimens (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/975 09/12/2008)

Umgebung der Elsa-Brändström-Schule (Krückaupark) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2781 11/12/2008)

Umgebung der Gesamtschule Hamburg-Winterhude (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2681 11/12/2008)

Umgebung der Gesamtschule Winterhude (Hamburg) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2766 11/12/2008)

Umgebung der Grundschule Oderberg (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3009 11/12/2008)

Umgebung von Schorndorf (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2696 11/12/2008)

United States National Plant Germplasm System Collection (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1429 11/12/2008)

Universidad de Almería, HUAL (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/244 09/12/2008)

Universidad de Costa Rica (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1184 09/12/2008)

Universidad de Extremadura, UNEX (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/255 09/12/2008)

Universidad de Granada, Herbario: GDA (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1741 11/12/2008)

Universidad de Málaga: MGC-Cormof (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/259 09/12/2008)

Universidad de Oviedo. Departamento de Biología de Organismos y Sistemas: FCO (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/245 09/12/2008)

Universidad Politécnica de Madrid, Dpto. Biología Vegetal, Banco de Germoplasma (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1521 11/12/2008)

University and Jepson Herbaria DiGIR provider (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1413 09/12/2008)

University Museums of Norway (MUSIT) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1996 09/12/2008)

University of California Botanical Garden DiGIR provider (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1412 11/12/2008)

Unna-Mühlhausen, Wiesen (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2865 09/12/2008)

Unser kleines Rasenstück/ Dürer-Gymnasium Nürnberg (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2810 09/12/2008)

Unterbrucker Weiher (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2824 11/12/2008)

USDA PLANTS Database (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1066 09/12/2008)

USU-UTC Specimen Database (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1508 09/12/2008)

Utah Valley State College Herbarium (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1013 09/12/2008)

Vascular Plant Collection - University of Washington Herbarium (WTU) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/126 09/12/2008)

Vascular Plant Collection (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/622 09/12/2008)

Vascular plant collection of Jyvaskyla University Museum (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/462 09/12/2008)

Vascular Plant Herbarium, Oslo (O) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1078 09/12/2008)

Vascular Plants Collection of Sagamihara City Museum (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1809 11/12/2008)

Vascular plants of south-central China (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1828 11/12/2008)

Vascular Plants, Field notes, Oslo (O) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1079 09/12/2008)

VegetWeb: zentrale Datenbank der Arbeitsgemeinschaft Vegetationsdatenbanken; Teil des Netzwerks für Phytodiversität Deutschland (NetPhyD) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1081 09/12/2008)

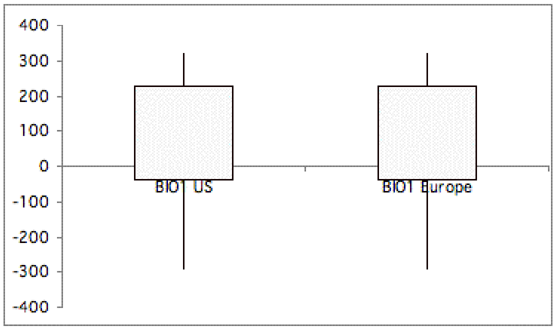
verschiedene Kleingewässer um Oldenburg/Holstein (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3000 11/12/2008)

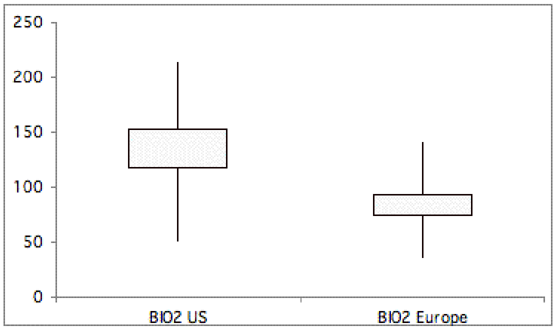
Verwilderter Hausgarten mit angrenzendem Gelände (Laufenburg-Hochsal) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2986 09/12/2008)
VFD-H: Rheingau: Pferdeweide Loock (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2928 09/12/2008)
VFD-RP: Hunsrück: Pferdeweide Kucher (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3503 09/12/2008)
VFD-RP: Taunus: Kirchenweide Köpplers (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3399 09/12/2008)
VFD-RP: Taunus: Ponykoppel Thurner (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3125 09/12/2008)
Wald am Schloss Wittgenstein Bad Laasphe (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2747 09/12/2008)
Wald und Wiese am Buchwald (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2676 11/12/2008)
Waldhusener Moor (Lübeck-Kücknitz) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2969 11/12/2008)
Waldi-Weiher (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3346 09/12/2008)
Waldränder der Frankenhöhe (Rothenburg ob der Tauber) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2647 11/12/2008)
Walldorf-Wiesloch: Natur über den Gleisen (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2850 09/12/2008)
Wanderweg am Windebyer Noor (bei Eckernförde) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2706 11/12/2008)
Warnowtal (Rostock) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3086 11/12/2008)
Wassermann (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3034 11/12/2008)
Wedeler Au (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2990 11/12/2008)
Weide am Ostufer des Zotzensees, Müritz-Nationalpark (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3111 11/12/2008)
Weide an der Mosselde / Dortmund-Kirchlinde/Westerfilde (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3219 09/12/2008)
Weinberg Reichersdorf (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3401 11/12/2008)
Westerwälder Umwelt- und Naturschutztag Limesgemeinde Hillscheid (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3017 11/12/2008)
Wiese am Waldrand (Gurtweil) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2784 09/12/2008)
Wiese und Bach am Kleinen Eutiner See (Eutin) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2773 11/12/2008)
Wiesen-Wälder-Wasser um Dansenberg, Biosphärenreservat Pfälzerwald (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3500 11/12/2008)
Wildes Bremer Leben im Park (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2708 09/12/2008)
WildesMoor bei Schwabstedt (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/3109 11/12/2008)
Wildkräuter (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2745 11/12/2008)
Wismar Bucht coast-watching (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2786 11/12/2008)
Zukünftiges NSG Höftland/Bockholmwik (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2665 11/12/2008)
Zwei Flüsse - eine Stadt (Villingen-Schwenningen) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/2829 11/12/2008)

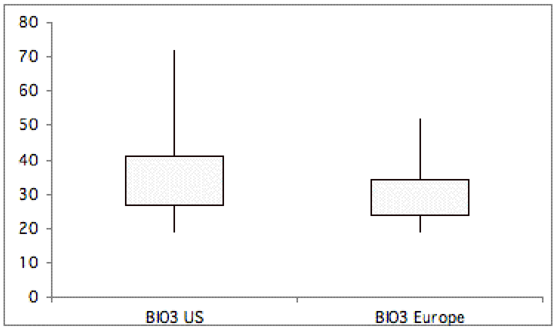
Appendix 2. Common, widespread species used for target group to create bias file used in Maxent model. These 20 species are found throughout Europe and represent a variety of habitats (Fitter et al. 1996).

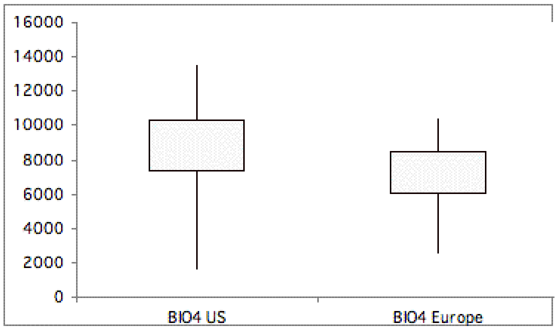
Scientific Name	Habitat
Achillea millefolium	Grassland, banks and waysides, often a weed in lawns
Ajuga reptans	Damp woods, hedge banks, meadows
Alisma plantago-aquatica	In or beside ponds, ditches, canals, slowmoving rivers
Anemone nemorosa	Woodlands, old hedge banks, upland meadows
Arum maculatum	Hedgerows, woodland, brown earth soils
Calluna vulgaris	Heaths, moors, rocky places, bogs, open woodland, mainly on sandy/peaty soils
Digitalis purpurea	Open spaces, woodland clearings, heaths, mountainsides
Dipsacus fullonum	Open woods, stream banks, roadsides, rough ground, grassland, marginal habitats, railway banks
Filipendula ulmaria	Wet, damp places of all kinds
Galium aparine	Cultivated and arable land, waste-ground, woodland, beaches, scrub, open ground, gardens
Geranium pratense	Meadows, roadsides, grasslands, open woods, dunes
Hyacinthoides non-scripta	Wide distribution except mountains and fens, but mainly woodlands
Lamium purpureum	Arable and waste ground, hedgerows, roadsides, garden weed
Leucanthemum vulgare	Grassy areas, especially nutrient-rich soils
Lotus corniculatus	Well-drained grassland, roadsides except on very acid soils
Papaver rhoeas	Arable, waste ground, field edges, roadsides
Plantago coronopus	Common near sea, on rocks, cliffs; dry sandy gravelly grasslands; inland commons, paths and roadsides
Potentilla anserina	Wasteland, pastures, waysides, sand dunes, especially damper places
Primula veris	Open woods, grassy places, meadows, roadside banks
Urtica dioica	Wasteland, woods, fens, roadsides, hedge banks. Favours phosphate rich soils

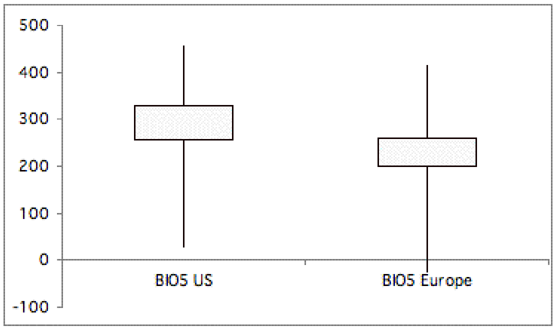
Appendix 3. Box plots and line graphs comparing the overall range and interquartile range of the 19 bioclimatic predictor variables available from WorldClim. Based on visual interpretation of these graphs, we eliminated 5 variables that had large differences in ranges, see Table 1 for abbreviations and variables that were eliminated from modelling.

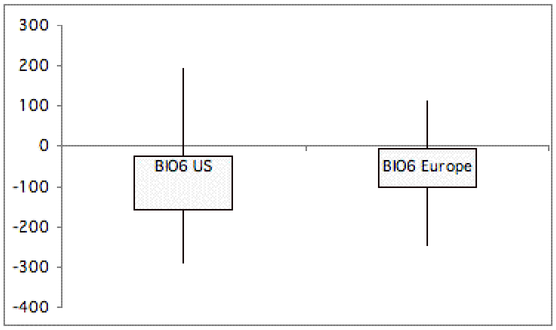

400
300
200
100
0
-100
-200
-300
-400
BIO1 US
BIO1 Europe

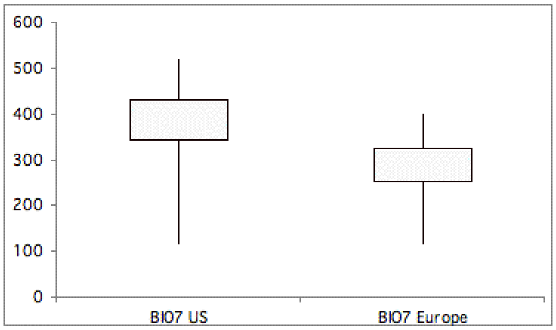

250
200
150
100
50
0
BIO2 US
BIO2 Europe

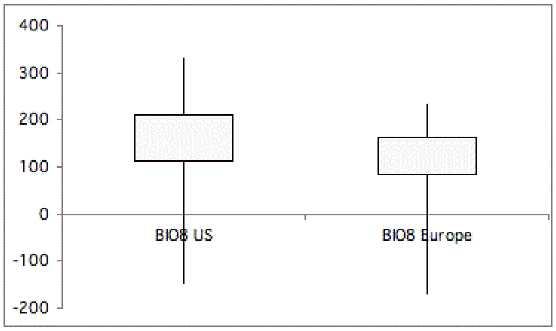

80
70
60
50
40
30
20
10
0
BIO3 US
BIO3 Europe

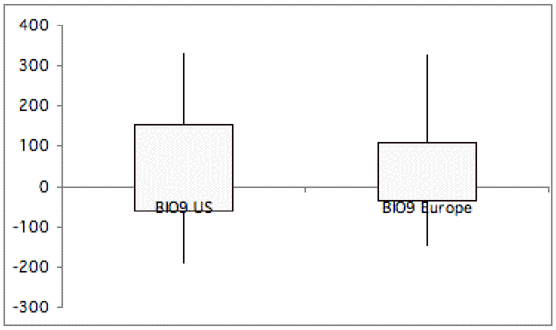

16000
14000
12000
10000
8000
6000
4000
2000
0
BIO4 US
BIO4 Europe

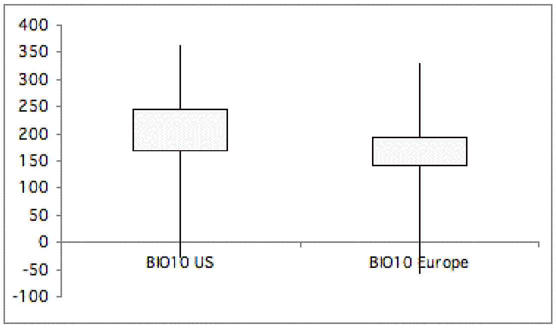

500
400
300
200
100
0
-100
BIO5 US
BIO5 Europe

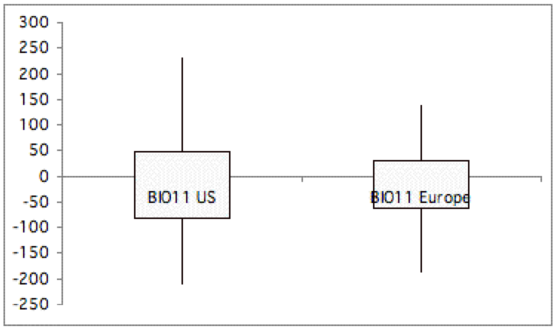

300
200
100
0
-100
-200
-300
-400
BIO6 US
BIO6 Europe

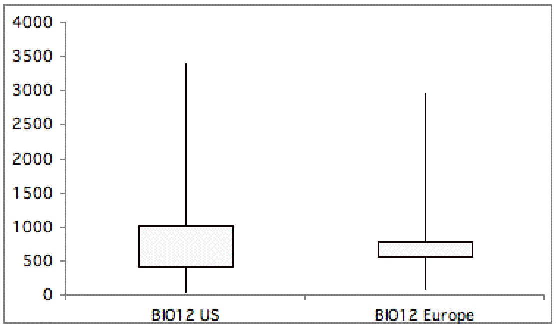

600
500
400
300
200
100
0
BIO7 US
BIO7 Europe

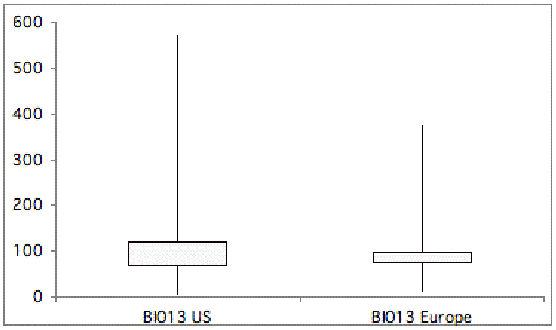

400
300
200
100
0
-100
-200
BIO8 US
BIO8 Europe

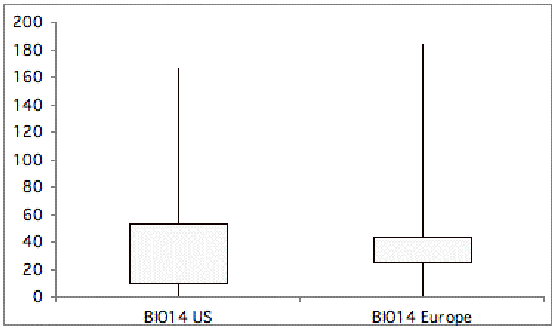

400
300
200
100
0
-100
-200
-300
BIO9 US
BIO9 Europe

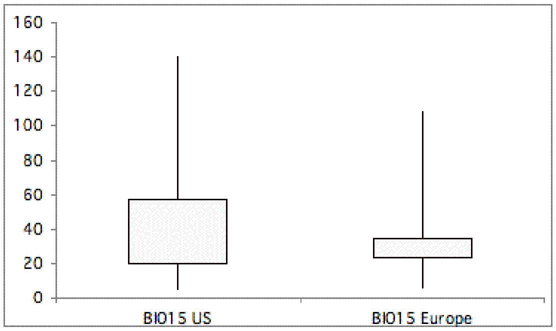

400
350
300
250
200
150
100
50
0
-50
-100
BIO10 US
BIO10 Europe

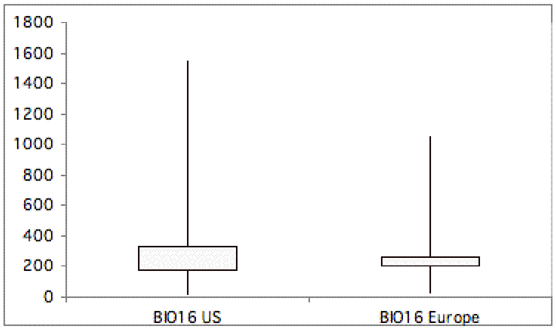

300
250
200
150
100
50
0
-50
-100
-150
-200
-250
BIO11 US
BIO11 Europe

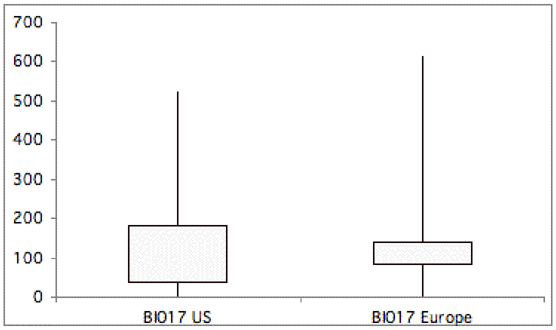

4000
3500
3000
2500
2000
1500
1000
500
0
BIO12 US
BIO12 Europe

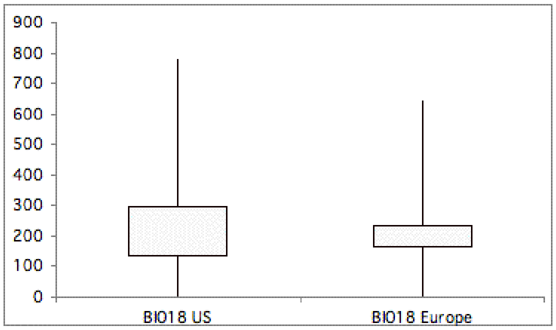

600
500
400
300
200
100
0
BIO13 US
BIO13 Europe

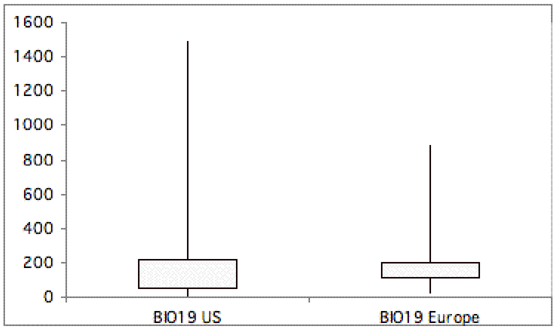

200
180
160
140
120
100
80
60
40
20
0
BIO14 US
BIO14 Europe

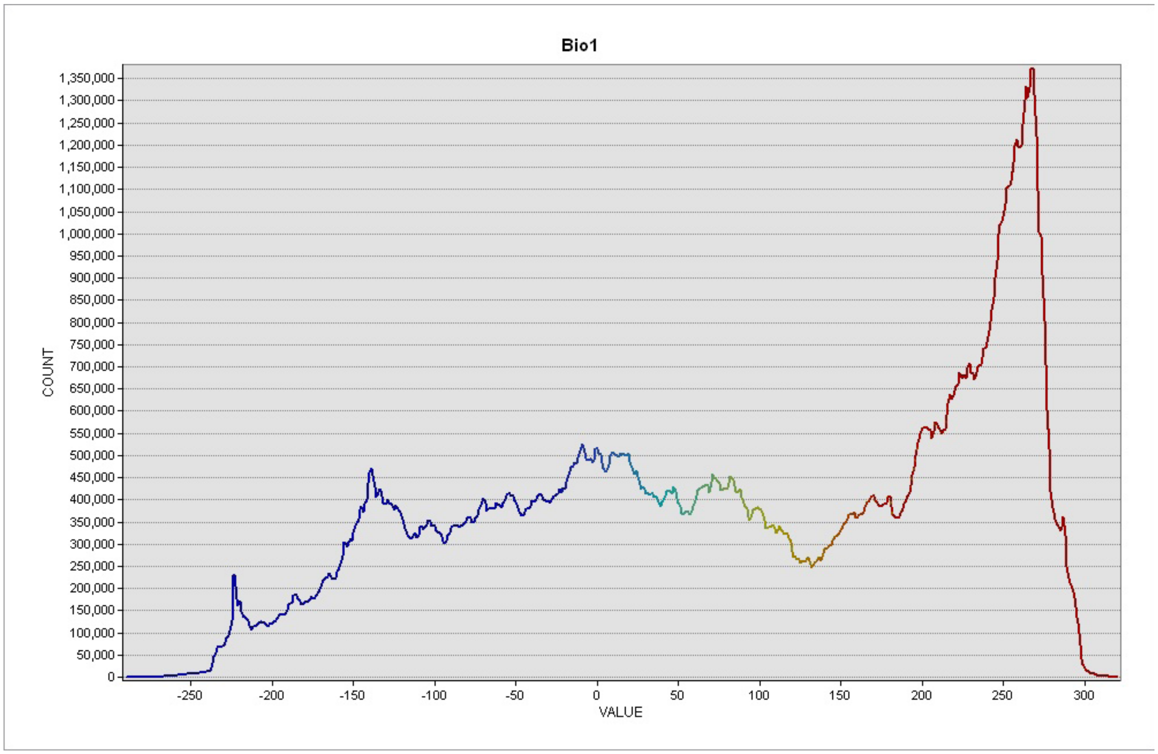

160
140
120
100
80
60
40
20
0
BIO15 US
BIO15 Europe

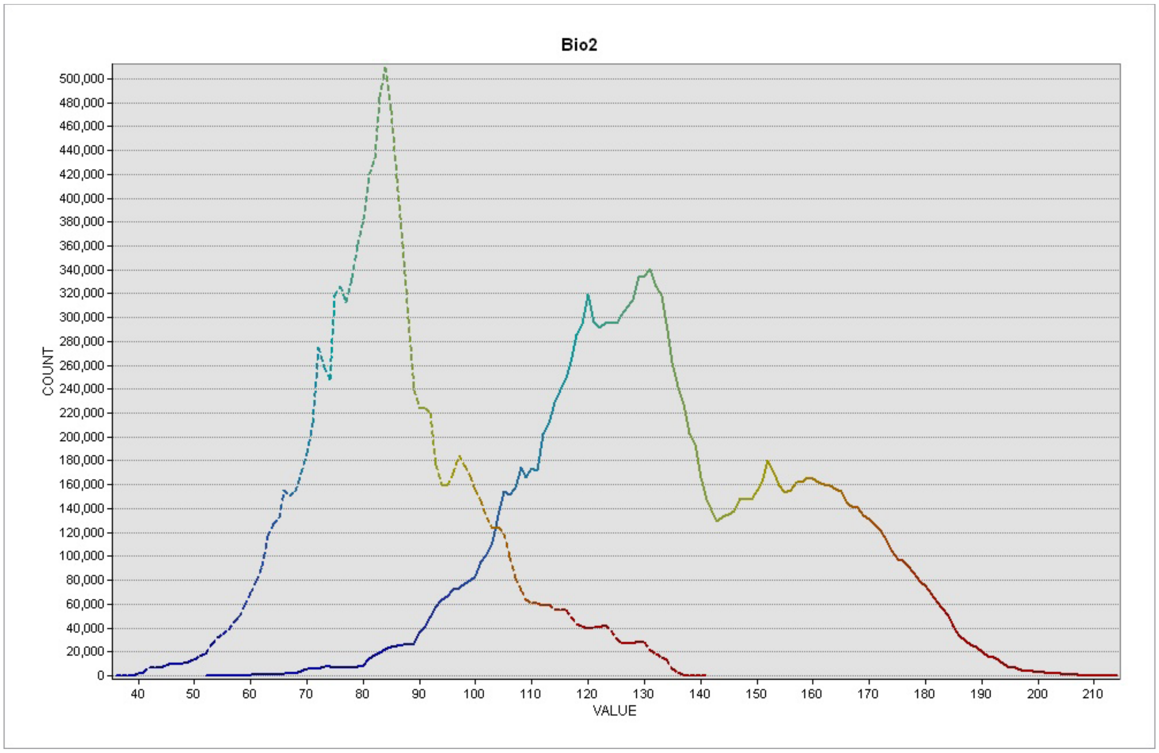

1800
1600
1400
1200
1000
800
600
400
200
0
BIO16 US
BIO16 Europe

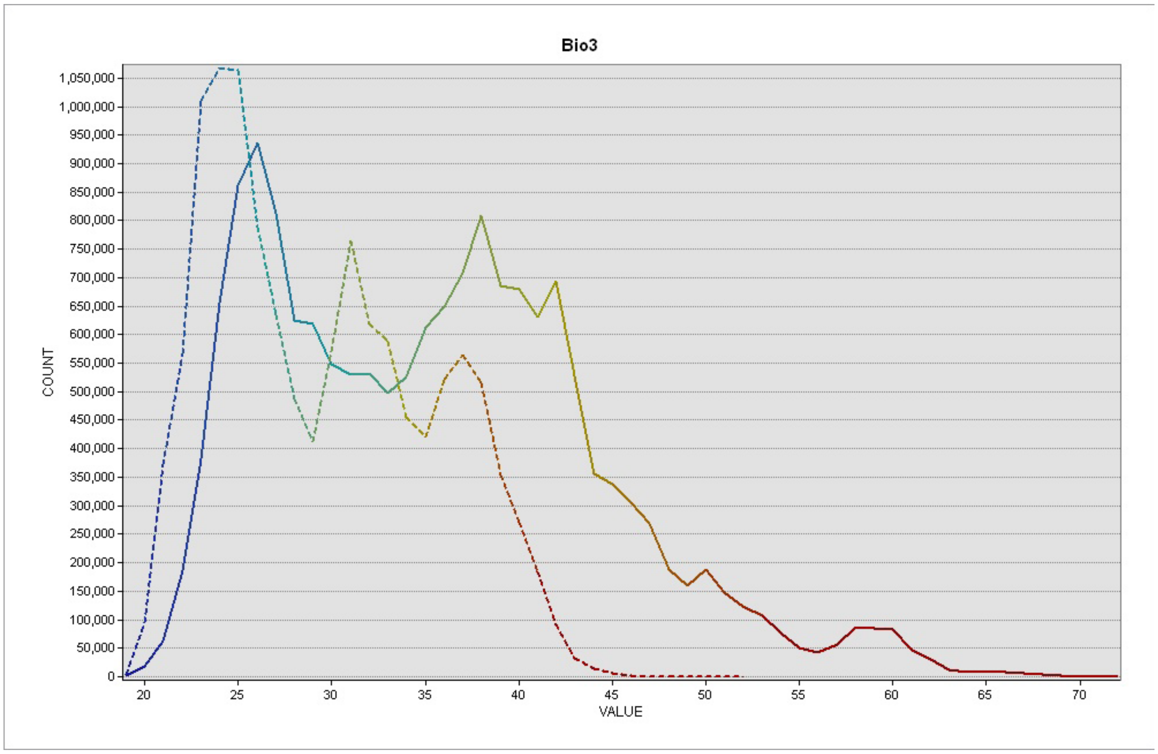

700
600
500
400
300
200
100
0
BIO17 US
BIO17 Europe

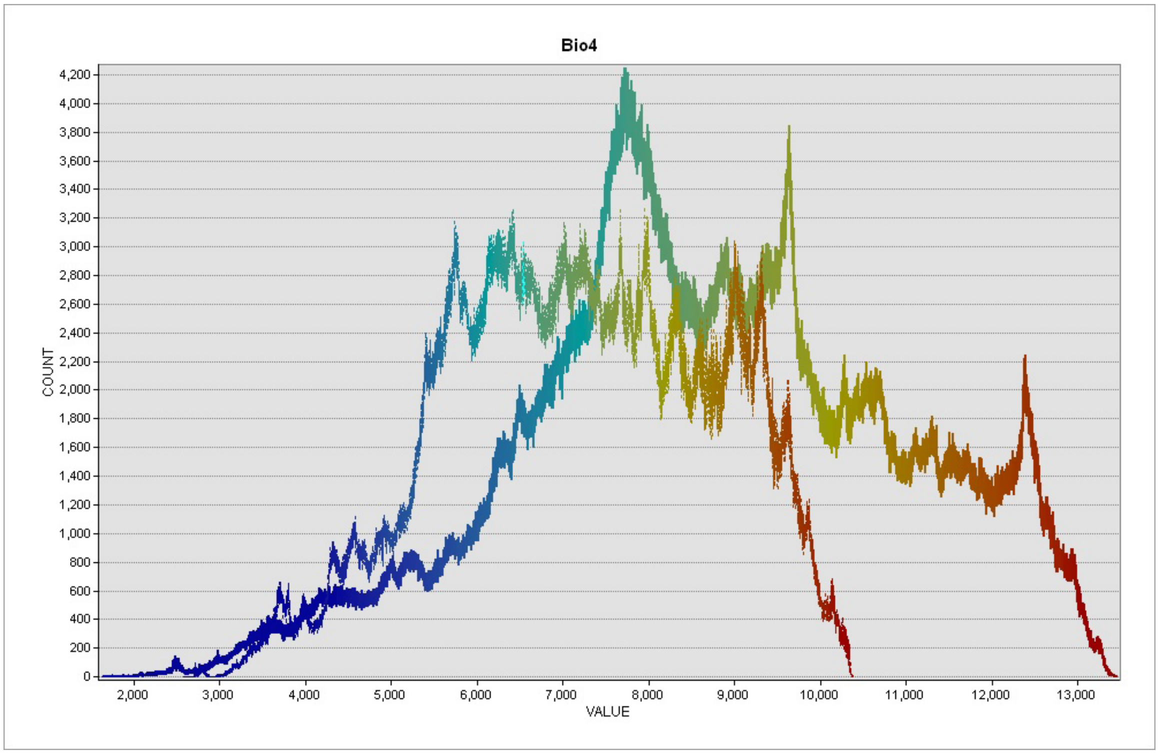

900
800
700
600
500
400
300
200
100
0
BIO18 US
BIO18 Europe

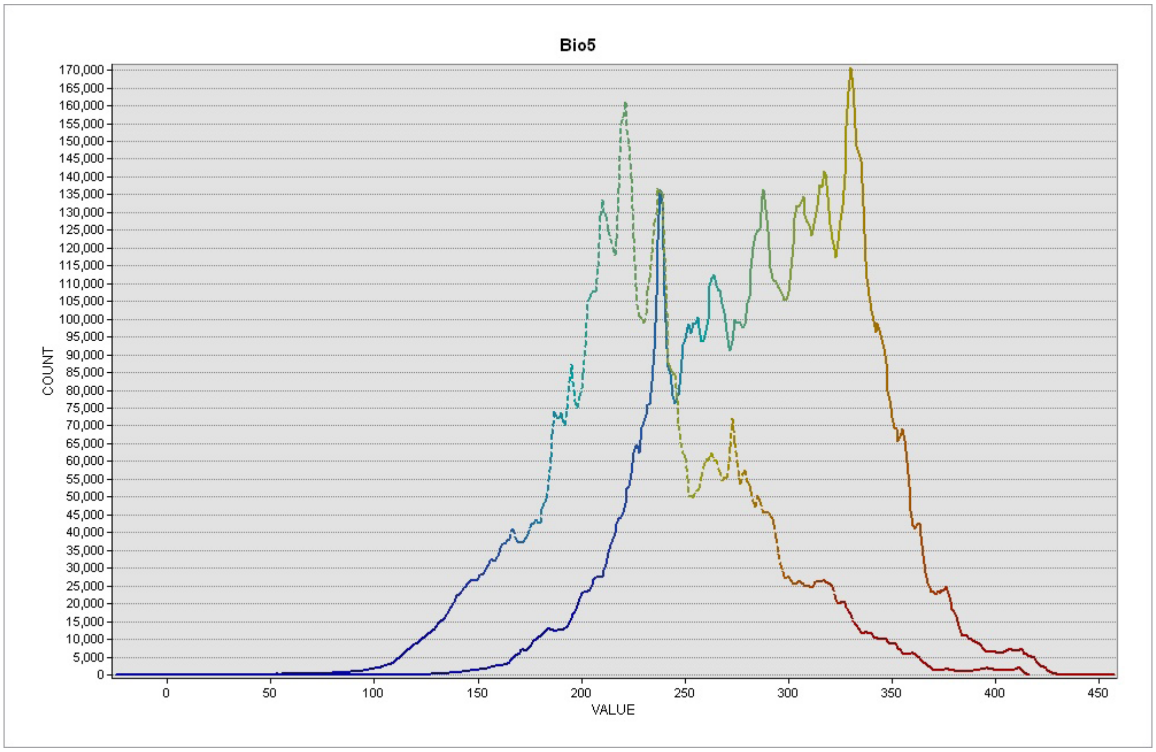

1600
1400
1200
1000
800
600
400
200
0
BIO19 US
BIO19 Europe

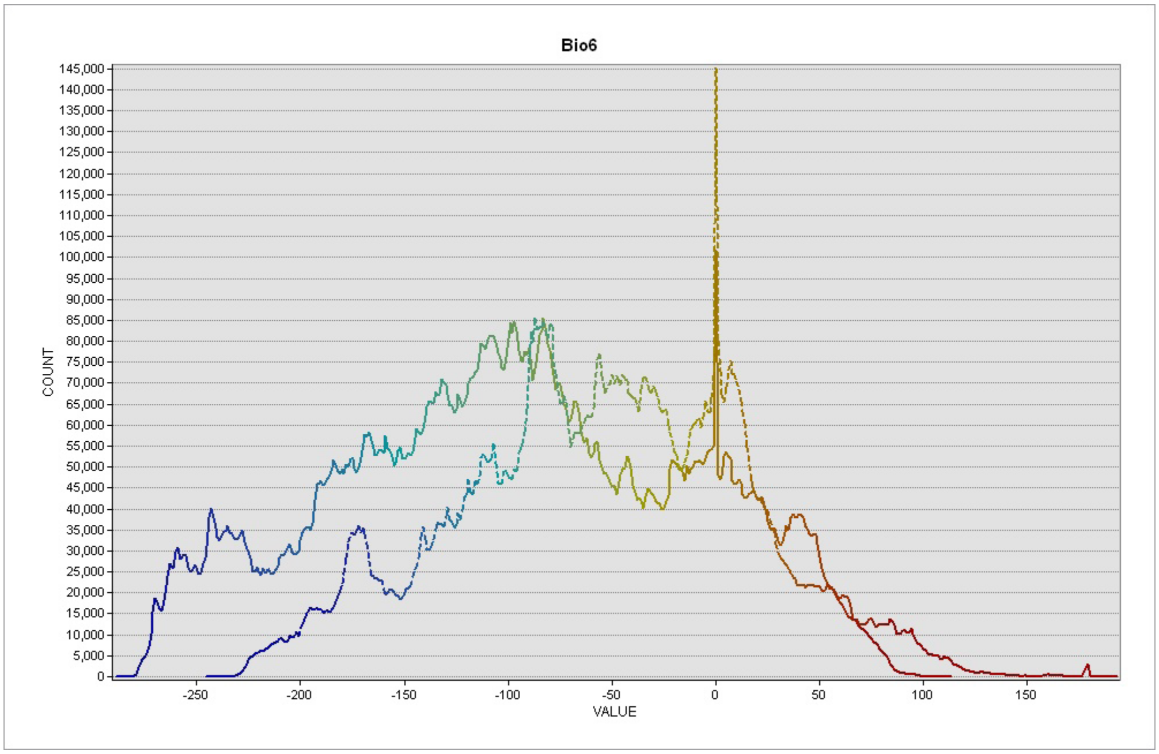

Bio1
COUNT
1,350,000
1,300,000
1,250,000
1,200,000
1,150,000
1,100,000
1,050,000
1,000,000
950,000
900,000
850,000
800,000
750,000
700,000
650,000
600,000
550,000
500,000
450,000
400,000
350,000
300,000
250,000
200,000
150,000
100,000
50,000
0
-250
-200
-150
-100
-50
0
50
100
150
200
250
300
VALUE

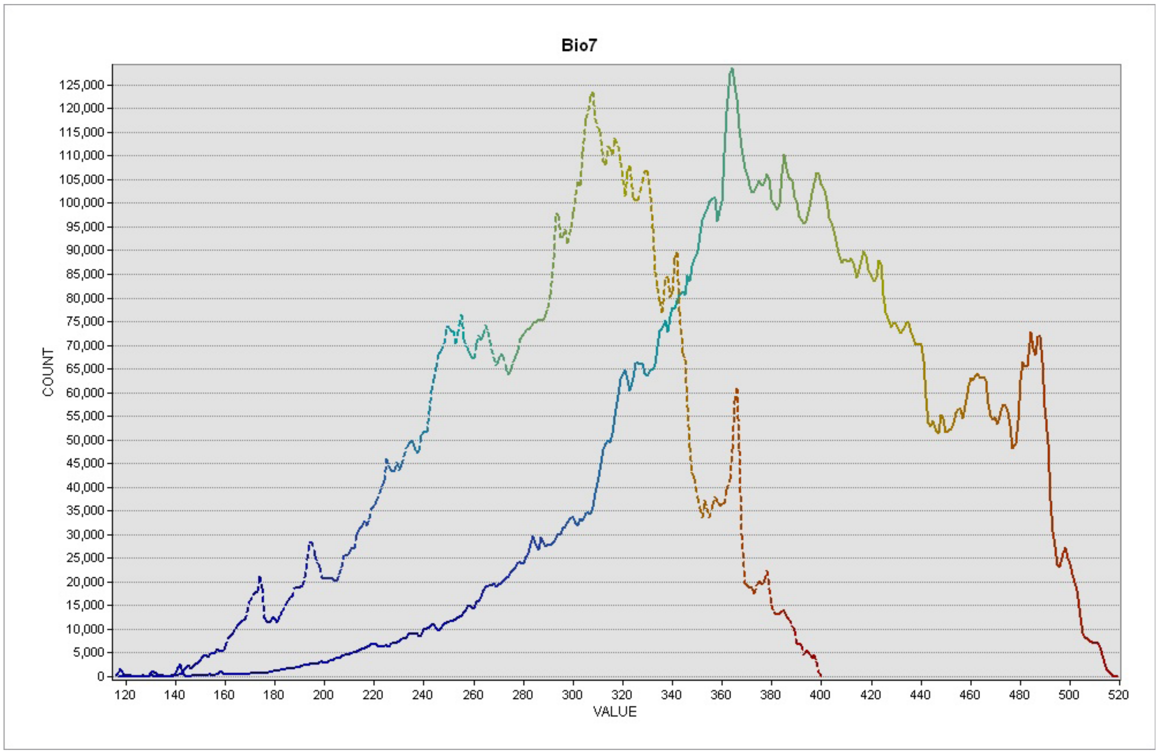

Bio2
COUNT
500,000
480,000
460,000
440,000
420,000
400,000
380,000
360,000
340,000
320,000
300,000
280,000
260,000
240,000
220,000
200,000
180,000
160,000
140,000
120,000
100,000
80,000
60,000
40,000
20,000
0
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
210
VALUE

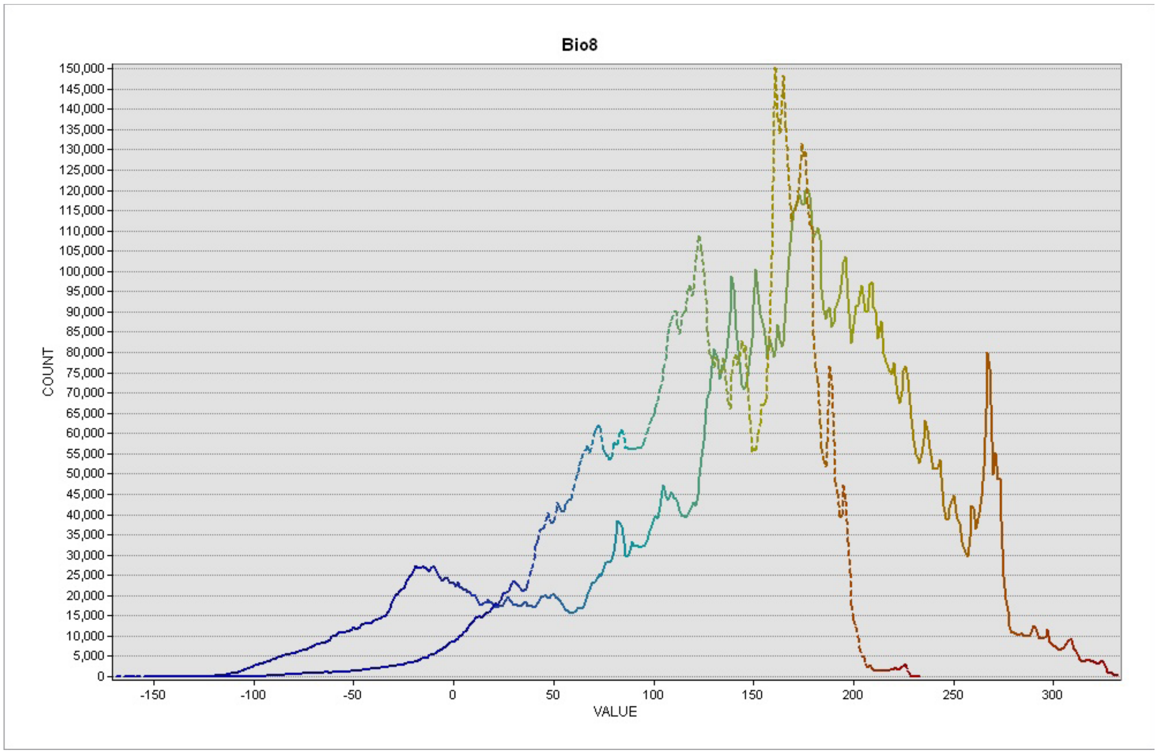

Bio3
COUNT
1,050,000
1,000,000
950,000
900,000
850,000
800,000
750,000
700,000
650,000
600,000
550,000
500,000
450,000
400,000
350,000
300,000
250,000
200,000
150,000
100,000
50,000
0
20
25
30
35
40
45
50
55
60
65
70
VALUE

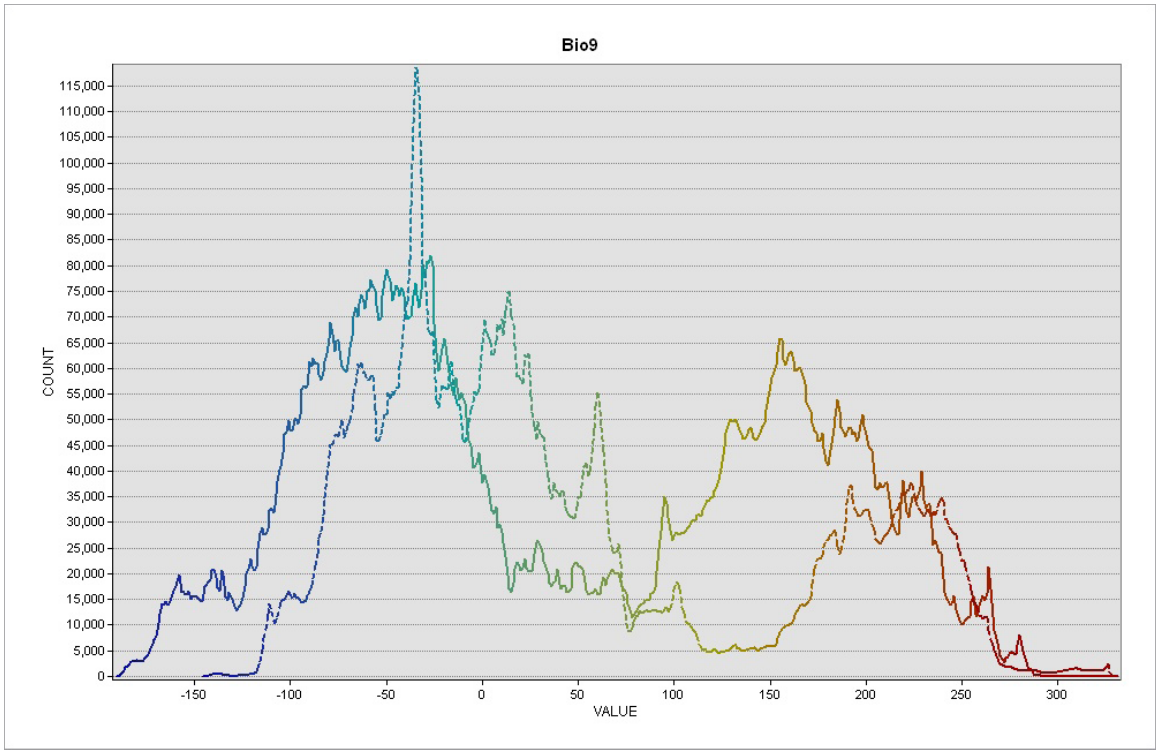

Bio4
COUNT
VALUE
4,200
4,000
3,800
3,600
3,400
3,200
3,000
2,800
2,600
2,400
2,200
2,000
1,800
1,600
1,400
1,200
1,000
800
600
400
200
0
2,000
3,000
4,000
5,000
6,000
7,000
8,000
9,000
10,000
11,000
12,000
13,000

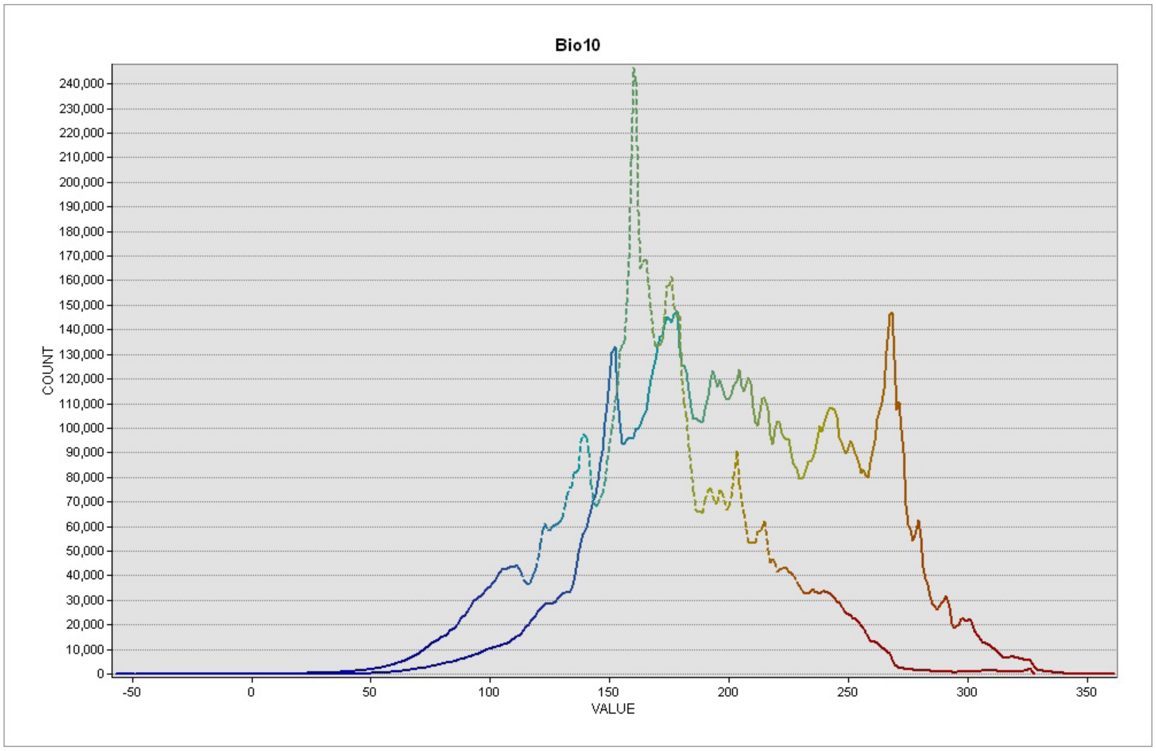

Bio5
COUNT
VALUE
170,000
165,000
160,000
155,000
150,000
145,000
140,000
135,000
130,000
125,000
120,000
115,000
110,000
105,000
100,000
95,000
90,000
85,000
80,000
75,000
70,000
65,000
60,000
55,000
50,000
45,000
40,000
35,000
30,000
25,000
20,000
15,000
10,000
5,000
0
0
50
100
150
200
250
300
350
400
450

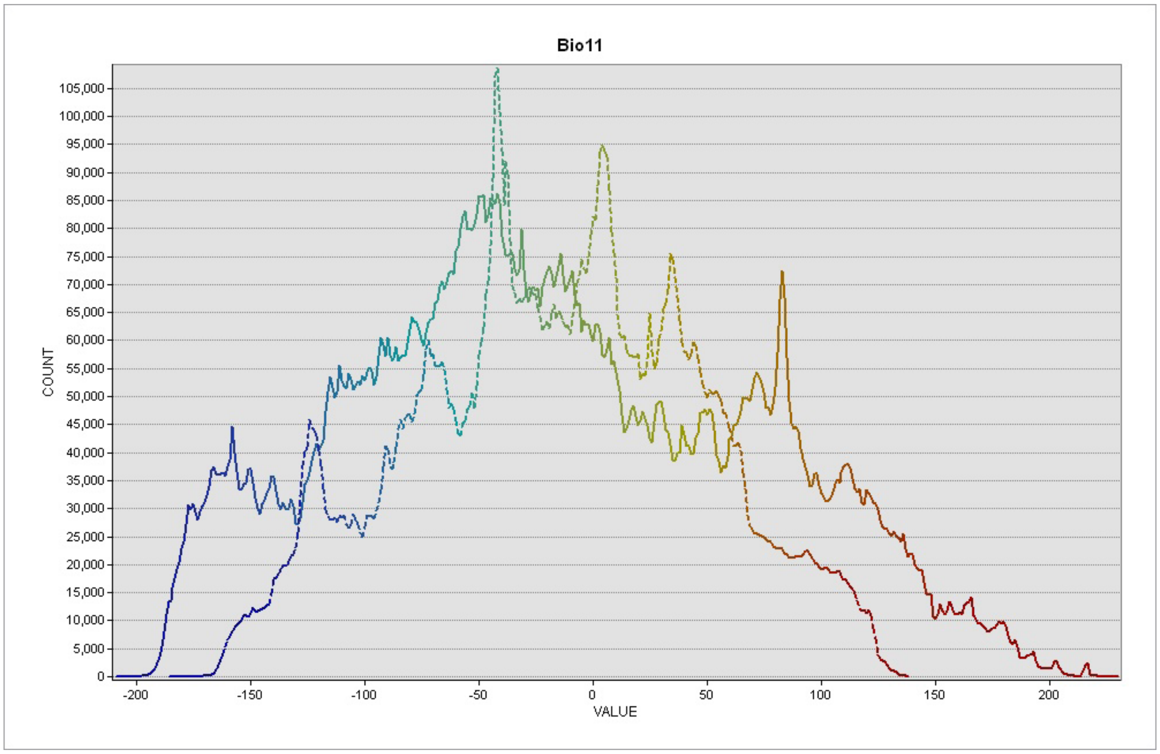

Bio6
COUNT
145,000
140,000
135,000
130,000
125,000
120,000
115,000
110,000
105,000
100,000
95,000
90,000
85,000
80,000
75,000
70,000
65,000
60,000
55,000
50,000
45,000
40,000
35,000
30,000
25,000
20,000
15,000
10,000
5,000
0
-250
-200
-150
-100
-50
0
50
100
150
VALUE

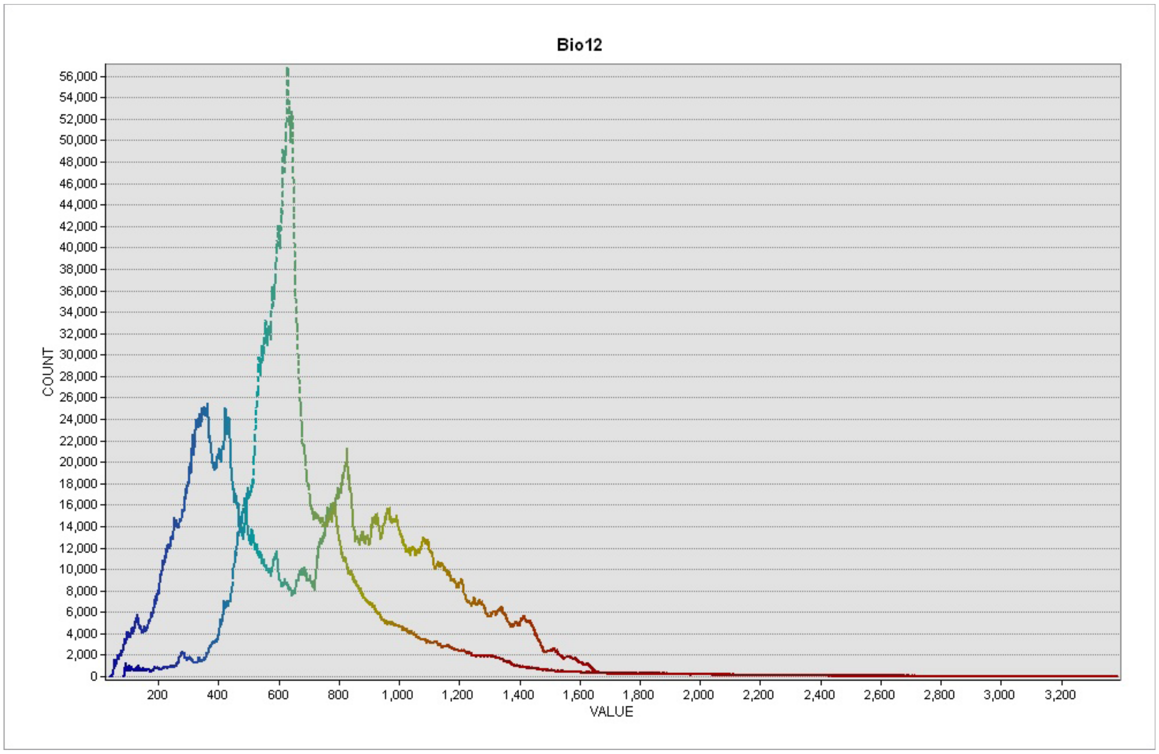

Bio7
COUNT
125,000
120,000
115,000
110,000
105,000
100,000
95,000
90,000
85,000
80,000
75,000
70,000
65,000
60,000
55,000
50,000
45,000
40,000
35,000
30,000
25,000
20,000
15,000
10,000
5,000
0
120
140
160
180
200
220
240
260
280
300
320
340
360
380
400
420
440
460
480
500
520
VALUE

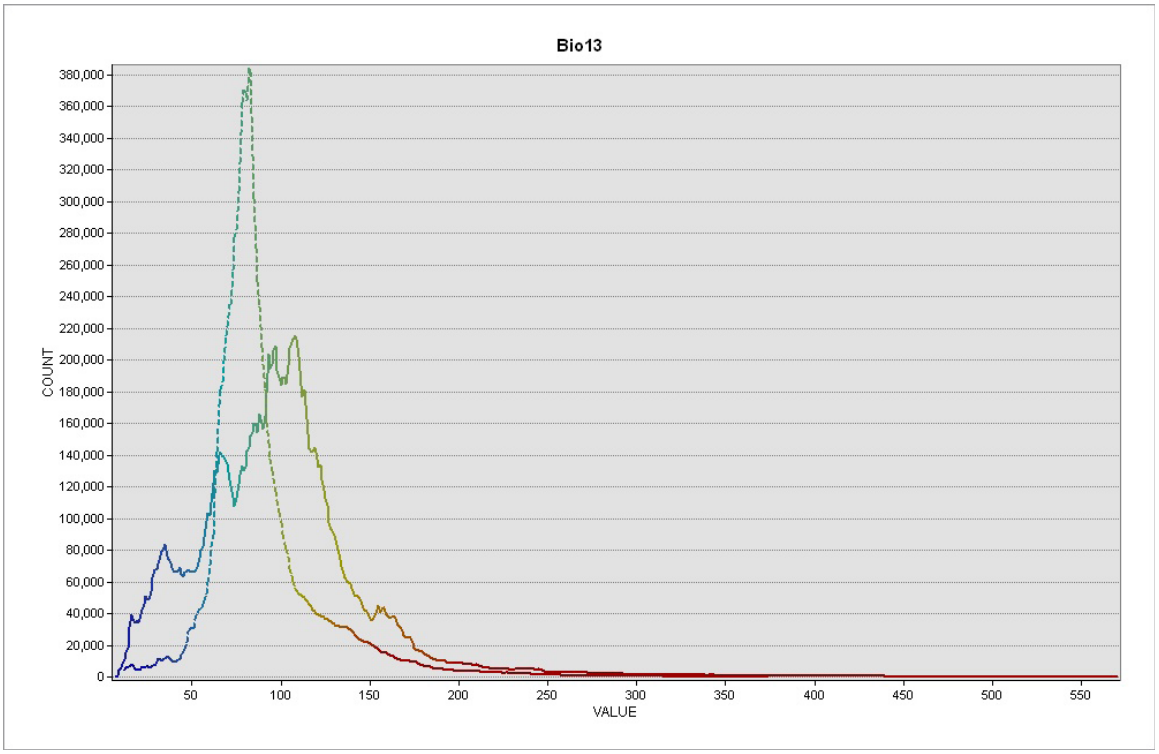

Bio8
COUNT
VALUE
150,000
145,000
140,000
135,000
130,000
125,000
120,000
115,000
110,000
105,000
100,000
95,000
90,000
85,000
80,000
75,000
70,000
65,000
60,000
55,000
50,000
45,000
40,000
35,000
30,000
25,000
20,000
15,000
10,000
5,000
0
-150
-100
-50
0
50
100
150
200
250
300

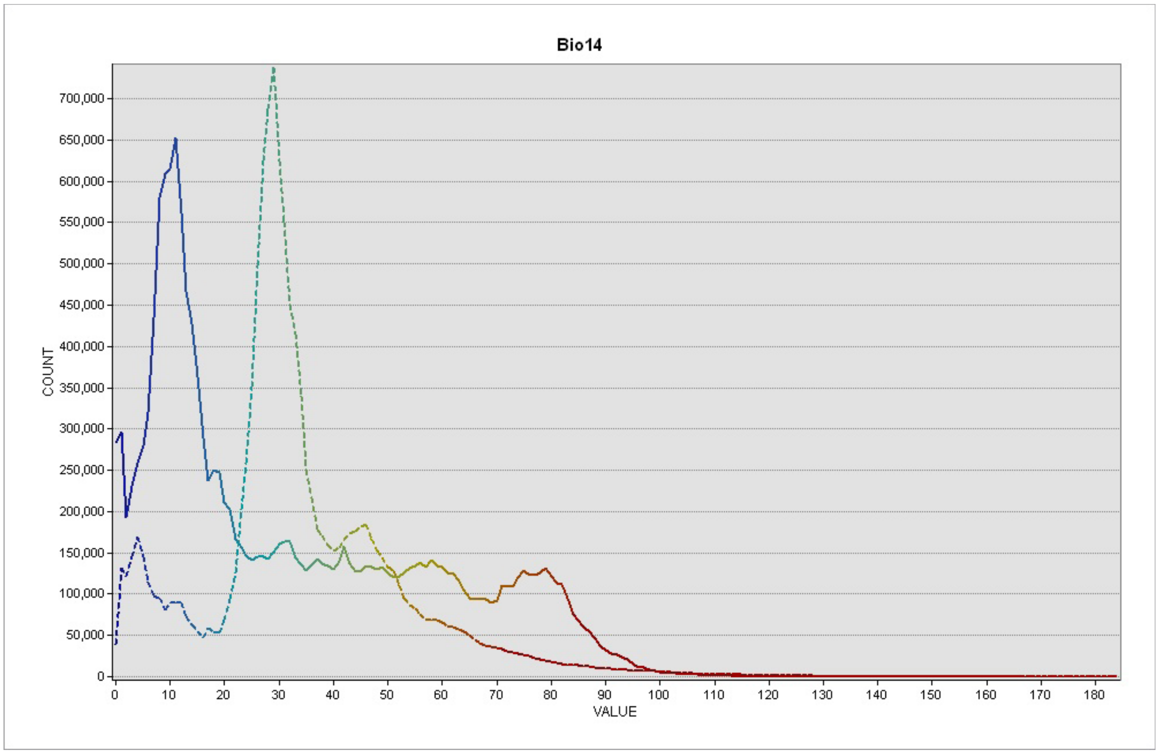

Bio9
COUNT
VALUE
115,000
110,000
105,000
100,000
95,000
90,000
85,000
80,000
75,000
70,000
65,000
60,000
55,000
50,000
45,000
40,000
35,000
30,000
25,000
20,000
15,000
10,000
5,000
0
-150
-100
-50
0
50
100
150
200
250
300

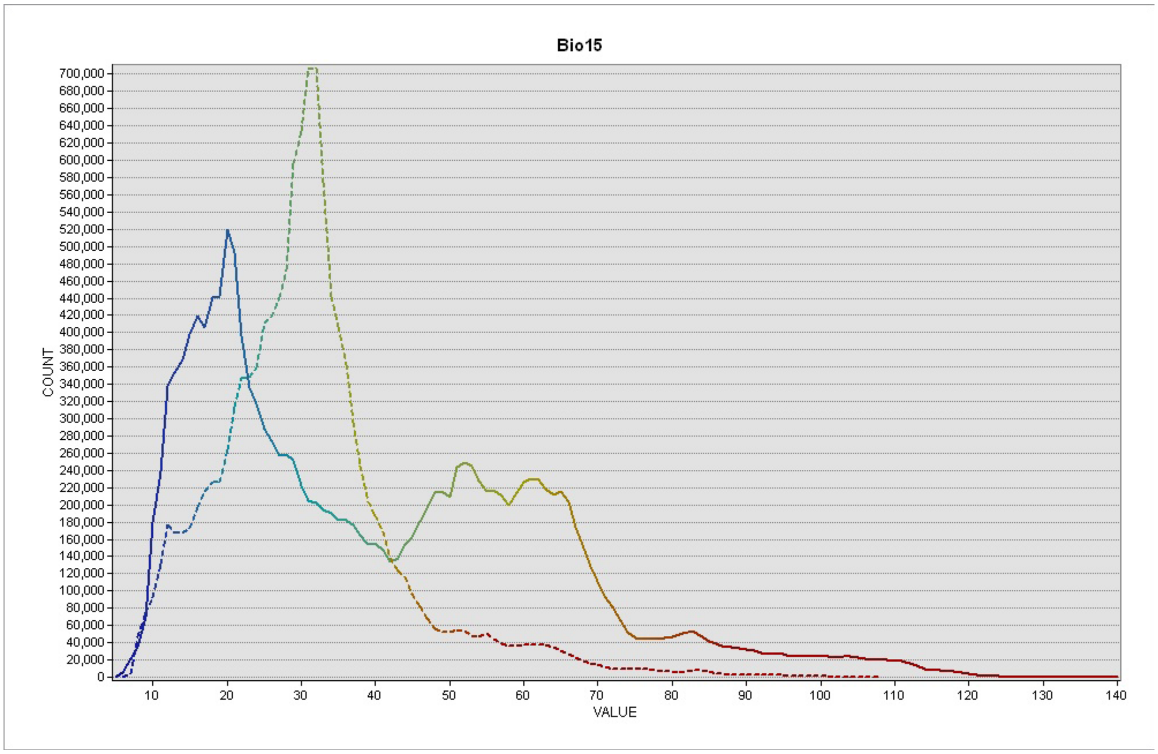

Bio10
COUNT
240,000
230,000
220,000
210,000
200,000
190,000
180,000
170,000
160,000
150,000
140,000
130,000
120,000
110,000
100,000
90,000
80,000
70,000
60,000
50,000
40,000
30,000
20,000
10,000
0
-50
0
50
100
150
200
250
300
350
VALUE

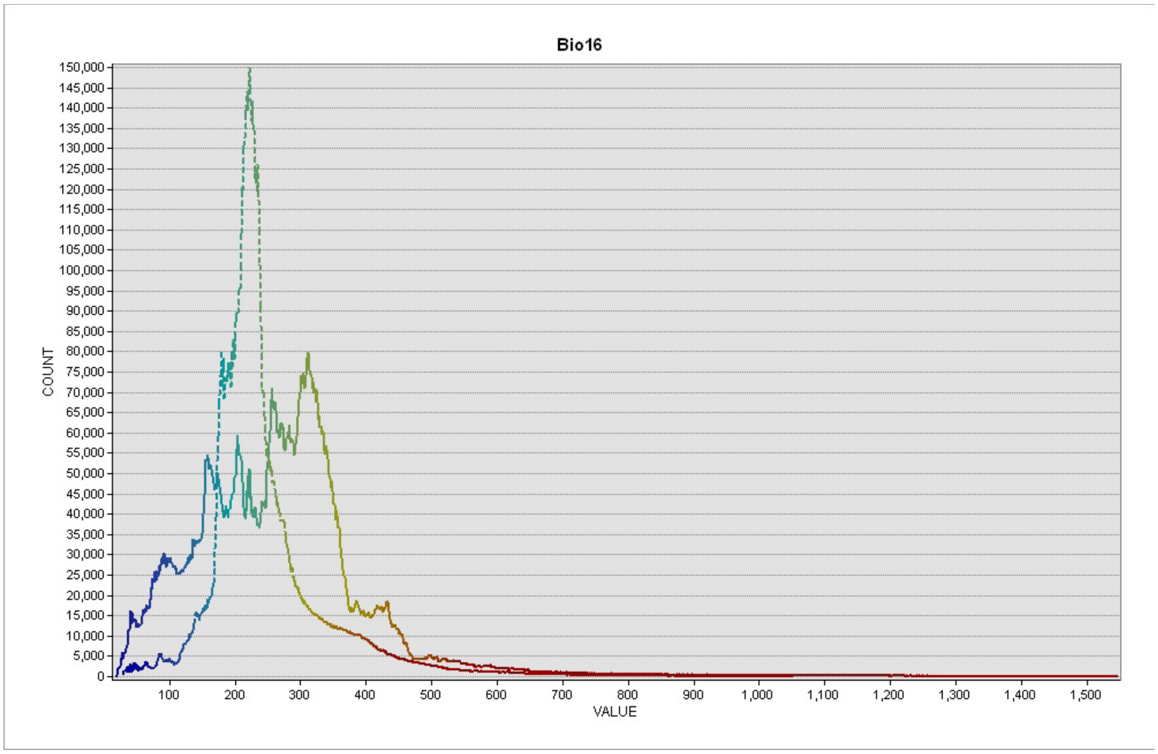

Bio11
COUNT
105,000
100,000
95,000
90,000
85,000
80,000
75,000
70,000
65,000
60,000
55,000
50,000
45,000
40,000
35,000
30,000
25,000
20,000
15,000
10,000
5,000
0
-200
-150
-100
-50
0
50
100
150
200
VALUE

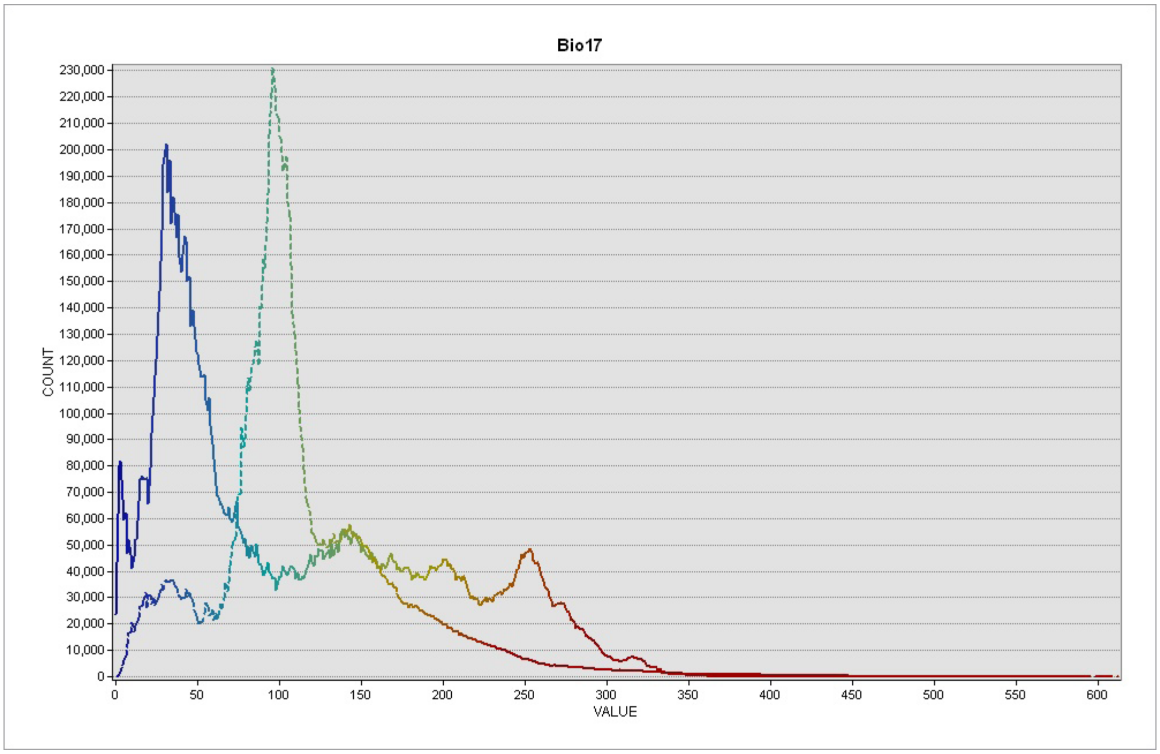

Bio12
COUNT
56,000
54,000
52,000
50,000
48,000
46,000
44,000
42,000
40,000
38,000
36,000
34,000
32,000
30,000
28,000
26,000
24,000
22,000
20,000
18,000
16,000
14,000
12,000
10,000
8,000
6,000
4,000
2,000
0
200
400
600
800
1,000
1,200
1,400
1,600
1,800
2,000
2,200
2,400
2,600
2,800
3,000
3,200
VALUE

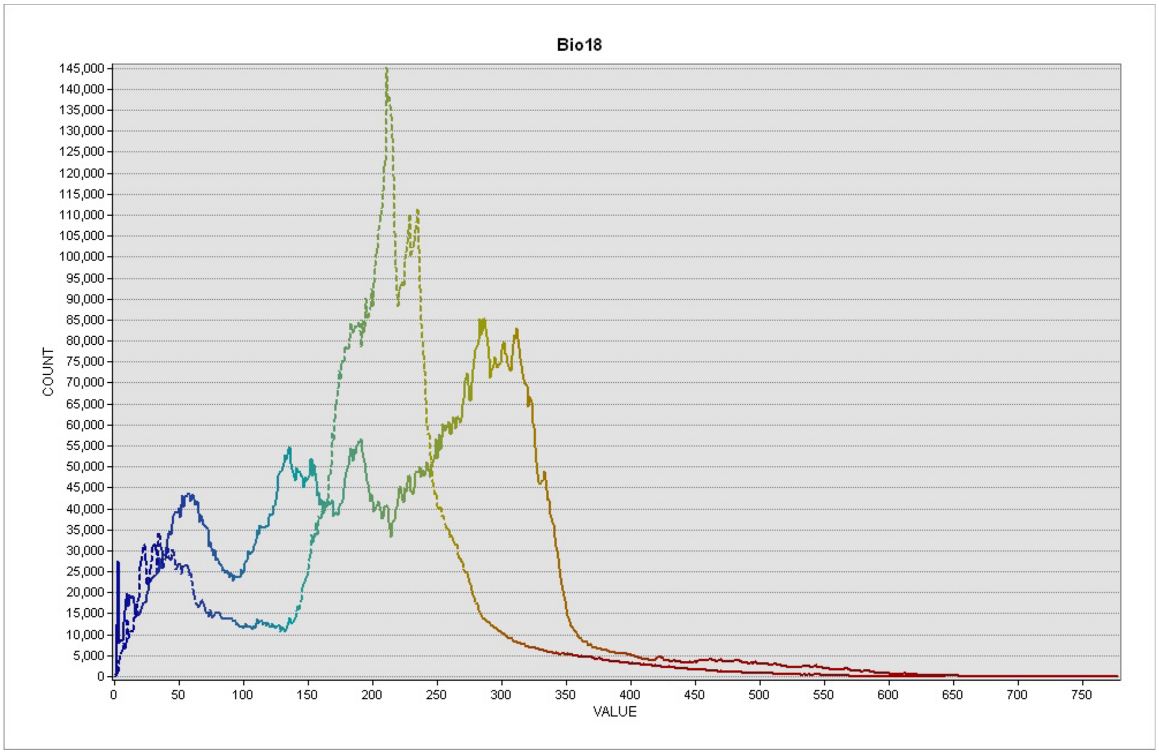

Bio13
COUNT
380,000
360,000
340,000
320,000
300,000
280,000
260,000
240,000
220,000
200,000
180,000
160,000
140,000
120,000
100,000
80,000
60,000
40,000
20,000
0
50
100
150
200
250
300
350
400
450
500
550
VALUE

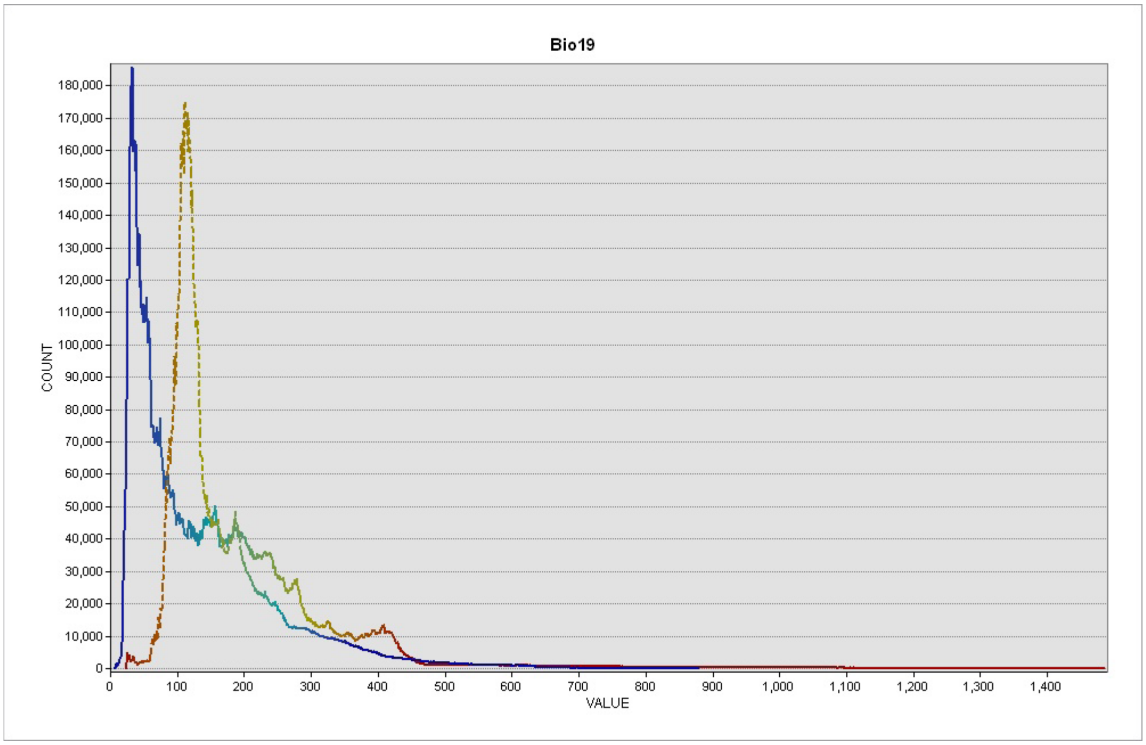

Bio14
COUNT
700,000
650,000
600,000
550,000
500,000
450,000
400,000
350,000
300,000
250,000
200,000
150,000
100,000
50,000
0
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
VALUE


Bio15
COUNT
700,000
680,000
660,000
640,000
620,000
600,000
580,000
560,000
540,000
520,000
500,000
480,000
460,000
440,000
420,000
400,000
380,000
360,000
340,000
320,000
300,000
280,000
260,000
240,000
220,000
200,000
180,000
160,000
140,000
120,000
100,000
80,000
60,000
40,000
20,000
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
VALUE


Bio16
COUNT
150,000
145,000
140,000
135,000
130,000
125,000
120,000
115,000
110,000
105,000
100,000
95,000
90,000
85,000
80,000
75,000
70,000
65,000
60,000
55,000
50,000
45,000
40,000
35,000
30,000
25,000
20,000
15,000
10,000
5,000
0
100
200
300
400
500
600
700
800
900
1,000
1,100
1,200
1,300
1,400
1,500
VALUE


Bio17
COUNT
230,000
220,000
210,000
200,000
190,000
180,000
170,000
160,000
150,000
140,000
130,000
120,000
110,000
100,000
90,000
80,000
70,000
60,000
50,000
40,000
30,000
20,000
10,000
0
0
50
100
150
200
250
300
350
400
450
500
550
600
VALUE


Bio18
COUNT
145,000
140,000
135,000
130,000
125,000
120,000
115,000
110,000
105,000
100,000
95,000
90,000
85,000
80,000
75,000
70,000
65,000
60,000
55,000
50,000
45,000
40,000
35,000
30,000
25,000
20,000
15,000
10,000
5,000
0
0
50
100
150
200
250
300
350
400
450
500
550
600
650
700
750
VALUE


Bio19
COUNT
180,000
170,000
160,000
150,000
140,000
130,000
120,000
110,000
100,000
90,000
80,000
70,000
60,000
50,000
40,000
30,000
20,000
10,000
0
0
100
200
300
400
500
600
700
800
900
1,000
1,100
1,200
1,300
1,400
VALUE

SUMMARY

I began my research with an interest in utilizing data collected in the herbaria of Oklahoma, now digitized as the Oklahoma Vascular Plants Database. Biogeographic research has benefited from the digitizing of large databases derived from natural history collections and biological surveys. These resources made available via the Internet can be accessed by biogeographers around the world to address a multitude of ecological and geographic questions. Utilizing this data taps into hundreds of years of study and countless hours of research conducted by biologists across the globe. This dissertation could not have been completed without the availability of data collected by legions of researchers from museums, herbaria, and government agencies. By taking advantage of data collected by others, I was able to work at a geographic scale that would have been impossible had I needed to gather all my own data. I was able to explore biogeographic questions at the continental and state level by mining the data collected by biologists over the past 100+ years.

My interest in the ecological conundrum of invasive plants led me to chose my first dissertation topic — Can we use herbarium data to map the historic invasion of plants? And can map the expansion of native "weedy" species in response to land use change? In chapter one, I used herbarium data to describe the temporal and spatial patterns of invasive and expansive species for the entire state of Oklahoma. I found that patterns of species invasion and expansion in Oklahoma could be detected using these techniques which were developed for regions with longer collecting plant histories. However, the expansion of native "weedy" species were not so easily documented. The information found in herbaria may not be sensitive enough to detect the increase of abundance of native species.

One of the greatest caveats associated with modelling is the biased nature of opportunistically collected data. Few studies take into consideration the biased nature of natural history collections such as: unequal sampling effort over time, non-random geographic representation, poor location information, incorrect identification, and disproportionately represented taxa. Therefore, methods must be developed to remove such biases to reveal the true pattern of invasion. Researchers must make the effort to reduce the power of these biases to control the results of analyses. The research in chapter one addressed temporal sampling bias using methods developed by researchers in Europe and Canada.

Having explored the historic spread of invasive species, I was interested to see if we could predict the future distribution of invasive species that have not yet become well established in Oklahoma. A recently developed and growing sub-field of biogeography - species distribution modelling - became an excellent tool to study the potential distribution of new invasive species. Species distribution modelling (SDM) is currently the trendy line of research and the literature is extensive and rapidly growing. Because of its relatively new status, there were few texts or articles that compile and review the literature when I began my research into SDM. I conducted a review of the literature for my own use to better understand the background and proper use and interpretation of the models produced by these techniques (chapter two).

During the course of researching and writing the literature review, it became clear that these techniques were complicated and involved many assumptions. To introduce myself to SDM, I modelled the distribution of the American burying beetle using a smaller data set that contained both presence and absence records. Survey data for the beetle were available and a model of its habitat preference would be useful for conservation efforts within Oklahoma for this endangered species. By modelling this species at the sub-state level, I was able to make some predictions of the species habitat preference. Although, as a generalist species, these results were less than ideal. Model performance could be improved by incorporating information on the cause of the beetle's endangered status and its population shrinkage. To improve the models and consequently the recovery effort for the species, the models need to include biotic interactions, such as congener and vertebrate competition and a reduction in optimally sized prey.   Creating an accurate spatial layer of this type data will be a future challenge.

In chapter four, I returned to the invasive species theme by addressing the question of whether the introduced distribution of invasive species can be predicted from its native range. I modelled the potential distribution within the United States of three alien invasive species native to Europe using the Maxent modelling technique. Using occurrence data from both the native (Europe) and introduced (US) ranges, I used reciprocal modelling to evaluate habitat discrepancies between the introduced and native ranges. The native occurrences in Europe accurately predicted the distribution within Europe; and introduced occurrences in the US accurately predicted the US distribution. However, the reciprocal models did not perform well. My model results indicate that the occupied niches are too inconsistent between the native and introduced ranges to make models useful at the scale at which early invasive species detection can occur.

The role of biotic interactions will need to play a bigger role in species distribution modelling if they are to be ecologically meaningful. Inclusion of biotic interactions such as overlap with competitor distribution and shared resources will improve model performance. Model predictions based on the native range may under-predict the potential distribution in the introduced range if biotic interactions, such as competition or parasitism, are removed when an alien species enters a new region. But accurately predicting areas of invasion in the introduced range may never truly incorporate the influence of biotic interactions because the introduced species are no longer affected by their native biotic interactions and are subject to another suite of species in the introduced range with which it may form new biotic interactions that are currently indescribable.

Another interesting avenue of research that will significantly improve the modelling of invasive species is the inclusion of mechanistic variables. Instead of relying on correlations with the environment to predict the environmental preferences of a species, a mechanistic model uses information from detailed physiological tolerance experiments to model the fundamental niche of a species. I would expect to accurately model the potential of invasive species is to model the fundamental niche and project that information onto the introduced range. This will not necessarily mean the species will be able to thrive in those locations, because new biotic interactions will be in place to limit the species range.

The methods explored in this dissertation illustrate the potential of natural history collections and survey data have in contributing to modern biogeographical research. Although the data is not perfect and the techniques do not perfectly represent the ecology, we can still take advantage of the newly digitized historical data to answer new and fundamental questions concerning biogeography. Advances in bias reduction will no doubt occur in the next several years. Improvements will be made to modelling algorithms to better represent ecological processes. Predictor data will be enhanced by including biological meaningful and derived variables. Using any technique to model species distribution should be done with care. Too often in the literature it is apparent that the researchers plugged their data into a model, the model drew a map, and the researchers presented the map as truth. This is done with little thought to proper evaluation and noted accuracy. Researchers should understand their goal when they model and verify that their approach is appropriate for that outcome.