*Diversity and Distributions, (Diversity Distrib.)* (2006) **12**, 511–520

BIODIVERSITY RESEARCH

# Filling in the gaps: modelling native species richness and invasions using spatially incomplete data

Catherine S. Jarnevich[1]*, Thomas J. Stohlgren[1], David Barnett[2] and John Kartesz[3]

[1]*Fort Collins Science Center, US Geological Survey, 2150 Centre Ave Bldg C, Fort Collins, Colorado 80526, USA,* [2]*Natural Resource Ecology Laboratory, Colorado State University, Fort Collins, Colorado 80523, USA and* [3]*Biota of North America Program, University of North Carolina, Chapel Hill, North Carolina 27599, USA*

*Correspondence: Catherine S. Jarnevich, Fort Collins Science Center, US Geological Survey, 2150 Centre Ave Bldg C, Fort Collins, CO 80526, USA. E-mail: catherine_jarnevich@usgs.gov

## ABSTRACT

Detailed knowledge of patterns of native species richness, an important component of biodiversity, and non-native species invasions is often lacking even though this knowledge is essential to conservation efforts. However, we cannot afford to wait for complete information on the distribution and abundance of native and harmful invasive species. Using information from counties well surveyed for plants across the USA, we developed models to fill data gaps in poorly surveyed areas by estimating the density (number of species km$^{-2}$) of native and non-native plant species. Here, we show that native plant species density is non-random, predictable, and is the best predictor of non-native plant species density. We found that eastern agricultural sites and coastal areas are among the most invaded in terms of non-native plant species densities, and that the central USA appears to have the greatest ratio of non-native to native species. These large-scale models could also be applied to smaller spatial scales or other taxa to set priorities for conservation and invasion mitigation, prevention, and control efforts.

**Keywords**

**biodiversity, biological invasions, modelling, non-native, prediction.**

## INTRODUCTION

Distributions of native species have been shaped over long evolutionary periods (Ricklefs, 2004), but only recently have researchers begun to collect data on these richness patterns in the USA. Survey efforts over the last few hundred years and some more recent intensive survey efforts have provided us with no more than a rudimentary knowledge of diversity patterns. Even with intensive landscape-scale surveys, less than 1% of the landscape can be effectively sampled. For example, a 6-year floristic survey of Grand Staircase Escalante National Monument included 379 0.1 ha plots and sampled only 0.004% of the 850,000 ha Monument; and a 5-year landscape analysis of Rocky Mountain National Park included 181 0.1 ha plots and sampled only 0.0001% of the Park. These intensive sampling efforts still miss species (e.g. 390 of the 940 plants known to occur in the Monument were missed; Stohlgren *et al.*, 2005b). More problematic yet, these floristic patterns that have evolved over long time periods that we are just beginning to understand are now being affected and potentially altered by the introduction of non-native species. Non-native species invasions threaten native biodiversity (Mack *et al.*, 2000), decrease human economical wealth by impacting agricultural land, rangeland, and forests (Pimentel *et al.*, 2000), alter ecosystem functioning (Vitousek *et al.*, 1987), and threaten human health (Mack *et al.*, 2000).

Efforts to control and mitigate these negative effects of invasive, non-native species are hampered by incomplete knowledge of species distributions (Stohlgren & Schnase, 2006). Because there are limited resources to be expended, species abundance and distribution data are needed to set management priorities.

Therefore, good baseline information on both native and non-native species distribution patterns is important. Knowledge of native species richness hotspots guides conservation efforts (Myers *et al.*, 2000; Myers, 2003), while identified non-native richness hot spots can be monitored for early detection of new invasions and aid control effort prioritization. The relationship between native and non-native richness may be another important component in determining conservation priorities. A highly invaded region with high native richness may be of less value than a slightly less rich area with low invasion. Competition with or predation by non-native species is also one of the top two reasons cited for the listing of federally threatened and endangered species (Wilcove *et al.*, 1998). Thus, knowledge of non-native species richness patterns may also aid in efforts aimed at conserving native species.

The best-sampled floristic data sets tend to be in densely populated areas or near herbaria. However, even these data sets are not without bias; even the best sampled areas miss species (Crosier & Stohlgren, 2004). This fact, and our inability to exhaustively survey any landscape, highlights the need to extrapolate

© 2006 The Authors
Journal compilation © 2006 Blackwell Publishing Ltd

DOI: 10.1111/j.1366-9516.2006.00278.x
www.blackwellpublishing.com/ddi

information from the few, well-surveyed areas to many, poorly surveyed areas.

Attempts to improve species lists using various methods of prediction are not new. Palmer *et al.* (2002) described numerous quantitative methods to improve species information using multiple species lists and broad species ranges. Others created predictive models for species richness using only the counties with a resident botanist (Iverson & Prasad, 1998), or by using the number of specimens in a collection to estimate survey intensity at the state level (Palmer *et al.*, 2002). Pearson (1994) suggested criteria for choosing indicator species whose occurrence patterns are functionally related to species richness as a surrogate for species richness. In fact, Scott *et al.* (2002) provided an edited volume of articles on different ways to predict species occurrences in space and time.

These efforts generally covered a small spatial extent, at the scale of states or smaller areas, or have a spatial grain greater even than counties. Additionally, many concentrate on predicting the occurrence of a single species rather than species richness. Using these approaches, a model would need to be generated for each species and then combined to determine species richness, which could propagate errors and be very data-, time-, and labour-intensive. Often, these different methods also require intensive amounts of data that are not readily available. For example, for indicator species, criteria for indicator species must be determined to choose the particular species, and then intensive survey data for that species must be compiled. Large-scale patterns at a relatively fine spatial resolution will be necessary to set regional, national, and global priorities for conservation and invasive species management.

Stohlgren *et al.* (2005a) explored patterns of plant diversity in the USA using data from almost all counties in the coterminous USA. Here, we take their work a step farther and show how a few well-surveyed counties can be used to improve knowledge of species richness patterns for the whole USA, including poorly sampled counties. We hypothesize that a small, well-dispersed sample of more complete species lists can be used to predict plant species richness in large, poorly sampled areas. We expect that we will find similar plant diversity patterns to those found by Stohlgren *et al.* (2005a) despite much smaller sample sizes due to predictable environmental controls on biodiversity.

## METHODS

We chose county-level plant species richness lists from the Biota of North America Program (BONAP) for the 48 coterminous states as our example data set. It has been assembled over the past 20 years and is regarded as highly accurate and relatively complete, being accepted as the standard by many different agencies and organizations. This data set is the finest resolution and best available for the entire country, as most data sets with finer resolution cover a small spatial extent, whereas data sets for large spatial extents are often at a very low resolution. The BONAP data set is based on herbarium records, giving it high accuracy, but it still has inherent bias like any data set. The bias stems from the data's origins, with more species reported for areas with a greater concentration of survey efforts (e.g. areas near herbaria or universities, Crosier & Stohlgren, 2004).

To overcome this bias, we chose the four counties from each state with the greatest native plant species richness reported in the BONAP county-level plant lists as a surrogate for the best sampled counties, creating an initial pool of 192 counties out of 3111 counties in the conterminous USA. Alaska and Hawaii were excluded due to a lack of ancillary data for predictor variables for the counties in these states. From this group of 192, we removed Georgia, Maryland, and Mississippi counties due to incomplete data sets for the entire state. We also removed Washington, D.C. and two of Delaware's three counties, resulting in a sample size of 177 counties for the analysis.

We used native species richness as the selection criteria because native species' distributions are relatively stable because of their long evolutionary history in the USA. We assumed that they were present in a county for the opportunity of being captured in a survey effort for a longer period of time than many non-natives that are still expanding their range. In general, we believe that native species lists are typically more complete than non-native species lists for these two reasons. We defined nativity based on BONAP definitions, classifying species as native if they are believed to have evolved in the USA as suggested by Pyšek *et al.* (2004). Non-native species included all species that are not believed to have evolved in the USA. We considered several metrics besides richness for selecting the best-sampled counties, such as counties with state herbariums or ones with large population centres. However, the number of counties with state herbariums would reduce the sample size by more than half. Population centres are often clustered within a state, and are highly disparate in size between states (compare New York and North Dakota urban areas). Therefore, we decided that native species richness, when stratified by state, was the best way to select the best-sampled counties. Richness was chosen rather than density, as we wanted to examine the species–area relationship before deciding on a transformation.

We stratified our sampling by state to ensure that we captured the range in environments found across the USA. Ecosystems on the west coast differ from the more species-poor central region, which differs from the comparatively species-rich east coast. We used species density (number of species $km^{-2}$) as the metric for our dependent variable because of the great variation in county size across the USA to account for the species–area relationship. We empirically chose this transformation by examining four different species–area relationships, including linear (untransformed); semilog; log-log; and the power model (nonlinear). The linear relationship, referred to as density, had the most support in the data, both for the 177 counties actually used in the analysis and for all counties together (T.J. Stohlgren, unpubl. data).

Species density values for native and non-native plant species from BONAP were extracted along with other predictor variable values for each county. These other predictor variables were assembled from a variety of Geographic Information System (GIS) sources (Appendix). These comprised 20 possible predictor variables divided into three classes that could affect richness patterns. These variables included 13 environmental and topographical

© 2006 The Authors

variables such as mean annual precipitation and elevation, four anthropogenic factors such as human population density, and three biotic factors such as Normalized Difference Vegetation Index (NDVI). Variables were selected based on previous studies examining species richness and invasion patterns (Stohlgren *et al.*, 2005a). Within each of the three groups, variables were examined for collinearity and those that were highly cross-correlated ($r \pm 0.80$; Bonferroni tests) were excluded. This resulted in seven environmental/topographical (mean annual precipitation, mean solar radiation, potential and actual evapotranspiration, elevation variance per county, mean elevation, and land cover classes count), three human (human population density, index of habitat disturbance, and percentage of crop area per county), and three biotic factors (native bird species density in a county, vegetative carbon, and NDVI) being possible predictors for the models. Vegetative carbon and NDVI were included as surrogates for plant productivity, and bird species density was included as an additional measure of habitat heterogeneity. Data were transformed when necessary.

SYSTAT version 11 (Systat Software Inc., Point Richmond, CA, USA) was used for all statistical analyses. Complete regression models were fitted for 18 candidate models. This included all possible combinations of the three classes of variables for both native and non-native species (seven each) and a linear and a nonlinear model for non-native species density using native species density as the predictor. The Akaike's Information Criteria for small samples ($AIC_c$) values were calculated using:

$$AIC_c = n \log\left(\frac{RSS}{n}\right) + 2K + \frac{2K(K+1)}{n-K-1} \quad (1)$$

where $n$ was the sample size, $RSS$ was the residual sum of squares in the model such that $RSS/n$ was the maximum likelihood estimator, and $K$ was the number of parameters in the model including the constant (Burnham & Anderson, 1998). $AIC_c$ values were used as opposed to AIC values because $n/K$ was $\ll 40$ (Burnham & Anderson, 1998, p. 66). This value is an estimate of how the model compares to the expected 'truth'. These values for different models can be compared and the model with the lowest $AIC_c$ value is said to be the one with the most support from the data. Models were assessed using both $AIC_c$ and adjusted $R^2$ values. The models with the lowest $AIC_c$ and highest $R^2$ values for both native and non-native plant species density were then applied to all counties in the conterminous USA, and used to calculate the proportion of non-native flora per county.

Because we were trying to predict undersampled areas using models generated from well-surveyed areas, we compared the models applied to the undersampled counties (2935 counties) to the observed values for the well-surveyed counties (177 counties), and examined the models applied to all counties together. When examining the best models, the ones using native species density to predict non-native species density were ignored for several reasons. First, we wanted to compare models of native and non-native species density, and use one to predict the other would convolute the results. Second, native species are typically harder to predict and we did not wish to propagate errors. Third, while using the transformed variable of density, the relationship between species and area did not appear to be completely absent despite this transformation having the most support in the data. Maps of native plant species density, non-native species plant density, and proportion of non-native flora per county were generated using ESRI ArcGIS, version 9.0. Maps of both predicted and observed county richness and proportion non-native were created by dividing the richness and proportion values into 10 quartiles based on the variable of interest (e.g. predicted native species density).

Models were validated by examining residual values of the observed data, including the four counties with the most native species in each state, residual values from the four counties with the next greatest richness of native plant species in the BONAP data (rank 5 to 8), and the four counties with the fewest native plant species per state in the original BONAP data. Four states (Delaware, Rhode Island, Connecticut, and New Hampshire) had fewer than 12 counties and were included in this part of the analysis as numbers permitted, with the first four counties in the top-four classification, and so on. The values for these three groups of county data were compared to each other.

## RESULTS

We found that native and non-native plant species densities were highly predictable. Of the seven candidate regression models that were composed of all possible combinations of environmental/topographical, biotic, and human variables, the model including all variables (environmental/topographical, biotic, and human) best predicted native plant species richness ($AIC_c = -1096.6$; $w_i = 0.555$, adjusted $R^2 = 0.89$; Table 1). The most influential variables based on standardized partial regression coefficients ($S_b$) were native bird species density and human population density ($S_b = 0.611$ and $S_b = 0.265$, respectively). This model is the one used in subsequent analyses. The combination of biotic and human independent variables provided the second best model for the prediction of native plant species density ($AIC_c = -1077.6$; $w_i = 0.0001$; adjusted $R^2 = 0.872$; Table 1). Despite the very similar adjusted $R^2$ values, this model and all the others were discounted because the information theoretical approach dictates that models with an $AIC_c$ difference greater than 10 are essentially not supported by the data (Burnham & Anderson, 1998).

A nonlinear regression model of native plant species density using values from the best model described above had the most support for predicting non-native plant species density ($AIC_c = -1376.8$; $w_i = 1$; adjusted $R^2 = 0.879$), followed by the model with human and biotic factors ($AIC_c = -1329.7$; $w_i = 3.4e^{-19}$; adjusted $R^2 = 0.846$; Table 2). The variables with most influence in the human and biotic variable model were native bird species density and human population density ($S_b = 0.624$ and $S_b = 0.537$, respectively). However, these $\Delta i$ values greatly exceed 10 and, therefore, the models should be discounted. This model is the one used in subsequent analyses as we did not wish to use the model where native density predicted non-native density as described in the methods. The $R^2$ values for the top two models were similar, and if the $\Delta i$ values were standardized as other statistics are, they would not be discounted.

© 2006 The Authors

Table 1 Native species-density model results including $AIC_c$, $\Delta i$ (AIC difference), RSS (residual sum of squares), adjusted $R^2$, $K$ (number of parameters), and the rank of the model compared with the others

| Model | $AIC_c$ | $\Delta i$ | RSS | Adjusted $R^2$ | $K$ | Rank |
|---|---|---|---|---|---|---|
| Topographical/environmental | −839.38 | 257.19 | 1.403 | 0.511 | 8 | 6 |
| Human | −834.65 | 261.92 | 1.512 | 0.485 | 4 | 7 |
| Biotic | −1035.65 | 60.92 | 0.489 | 0.835 | 4 | 4 |
| Topographical/environmental + human | −901.15 | 195.42 | 0.95 | 0.662 | 11 | 5 |
| Topographical/environmental + biotic | −1059.21 | 37.37 | 0.39 | 0.862 | 11 | 3 |
| Human + biotic | −1077.6 | 18.97 | 0.367 | 0.872 | 7 | 2 |
| Topographical/environmental + human + biotic | −1096.57 | 0 | 0.30 | 0.89 | 14 | 1 |

Table 2 Non-native species density model results

| Model | $AIC_c$ | $\Delta i$ | RSS | Adjusted $R^2$ | $K$ | Rank |
|---|---|---|---|---|---|---|
| Topographical/environmental | −1067.46 | 309.3 | 0.39 | 0.328 | 8 | 9 |
| Human | −1153 | 223.75 | 0.25 | 0.575 | 4 | 8 |
| Biotic | −1242 | 134.76 | 0.15 | 0.743 | 4 | 5 |
| Topographic/environmental + human | −1168.22 | 208.52 | 0.21 | 0.62753 | 11 | 7 |
| Topographic/environmental + biotic | −1231.33 | 145.42 | 0.15 | 0.739 | 11 | 6 |
| Human + biotic | −1329.72 | 47.03 | 0.089 | 0.846 | 7 | 2 |
| Topographical/environmental + human + biotic | −1324.07 | 52.69 | 0.08 | 0.849 | 14 | 3 |
| Native species density model (linear) | −1287.09 | 89.67 | 0.12 | 0.798 | 2 | 4 |
| Native species density model (nonlinear) | −1376.75 | 0 | 0.07 | 0.879 | 3 | 1 |

The greatest observed plant species densities existed in major metropolitan areas. Arlington County, Virginia, had the greatest native species density (13.4 native species $km^{-2}$) and Bronx County, New York, contained the greatest non-native species density (4.1 species $km^{-2}$). Examination of the best models of native and non-native plant density indicated that seven of the 2935 undersampled counties had predicted values outside the observed range of variation. These density values ranged from very large (111.7 species $km^{-2}$ for New York, New York) to much closer to the observed range (14.9 species $km^{-2}$ for Richmond, New York). These counties had some of the highest values for the top two predictor variables, native bird density (the six greatest values for all counties), and human population density (five counties ranked in the top 10). When examining the 177 well-surveyed counties, there were no counties with values greater than 14 species $km^{-2}$.

The models predicted negative species densities for some counties (77 counties for non-native predictions and 3 for native predictions). When native density was used to predict non-native plant density, however, the number dropped to one county. There were 417 counties, including the negative predictions, in 14 states that had predicted values less than the observed value with an average decrease of 0.0057 species. These were scattered mainly in remote western plains and mountains in Montana, Texas, New Mexico, and Nevada.

When the models were applied to the 2935 counties not included in generating the models (the undersampled counties), native plant species density increased from an observed average of 0.48 species $km^{-2}$ to a predicted average of 0.85 species $km^{-2}$ (Table 3). This average was similar to the observed average of 0.70 native plant species/$km^2$ for the 177 well-surveyed counties used to generate the model. Examining the model for non-native plant species density, we found that density increased from an observed average of 0.08 species $km^{-2}$ to a predicted average of 0.14 species $km^{-2}$. Again, this value is very similar to the observed average non-native plant species richness of 0.15 species $km^{-2}$ for the well-sampled counties.

The calculated proportion of non-native to native plant species density was 0.16, compared to an observed average proportion of 0.14 (Table 2). This analysis used the best model for each, excluding the nonlinear and linear species density models. The predicted ratio of non-native to native plant species density was slightly higher for all counties (well-surveyed and undersurveyed) together. Thus, based on the models, the proportion of non-native plant species per county was greater than the observed values.

Examining the maps, the 10 quartiles changed between the three observed and predicted surfaces, reflecting the average increase in species density. In general, the greatest proportion of non-native plant species shifted east with the modelled data. The Great Basin had a high proportion of non-native to native plant species densities in comparison to the north-east coast (Fig. 1). This is an important distinction as the north-eastern coast is highly invaded, but also has many native plant species. Coastal counties in the north-eastern USA and California also have a

© 2006 The Authors

Table 3 Results divided into the three categories of native plant species density, non-native plant species density, and the proportion non-native and analysed by observed and predicted values based on the models for (1) counties used to generate the models (well-sampled), (2) counties not used to generate the model (undersampled), and (3) all counties together. Values in parenthesis are calculated excluding values outside the original range in variation (e.g. counties with > 13.4 species km$^{-2}$ predicted), which only occurred for seven under-sampled counties for native plant species

| | | Well-sampled Observed | Well-sampled Predicted | Undersampled Observed | Undersampled Predicted | All counties Observed | All counties Predicted |
|---|---|---|---|---|---|---|---|
| Native density (species km$^{-2}$) | Average | 0.7 | 0.7 | 0.48 | 0.85 (0.75) | 0.5 | 0.84 (0.74) |
| | Maximum | 4.3 | 6.5 | 13.4 | 111.7 (9.88) | 13.42 | 111.7 (9.88) |
| Non-native density (species km$^{-2}$) | Average | 0.15 | 0.14 | 0.08 | 0.14 | 0.08 | 0.14 |
| | Maximum | 1.93 | 1.35 | 4.06 | 12.43 | 4.06 | 12.43 |
| Proportion non-native | Average | 0.17 | 0.18 | 0.14 | 0.16 | 0.14 | 0.16 |
| | Maximum | 0.45 | 1.95 | 0.75 | 0.64 | 0.75 | 1.95 |

high proportion of non-native plant species density. However, unlike other areas, these also have both high native and non-native plant species densities. The eastern USA remains the most invaded area when comparing the observed and predicted non-native plant species density maps (Fig. 1), though the highest proportion shifted with modelled data.

Analysis of the residuals from the modelled surfaces indicated that the models performed well (Fig. 2). The observed values used to generate the model (i.e. the four with most dense native plant species counties per state) had a greater range of variation in their residuals than did either counties with native plant density ranks from 5 to 8 or the four counties per state with the lowest recorded native plant species density. When applied to the undersurveyed counties, the models do not seem to generate atypical values. Also, the net increase in species density was greater than the net decrease in the modelled surface, which achieved the objective of improving plant species density information for poorly surveyed counties in the USA.

## DISCUSSION

Patterns of the establishment of non-native plant species seem highly predictable when examining the high $R^2$ values obtained in our models, and patterns of native species density seem only slightly harder to predict. On average, modelled values of native and non-native plant species density were greater than observed values (natives: average of 0.5 to average of 3.55 species km$^{-2}$; non-natives: average of 0.09 to average of 0.39 species km$^{-2}$). These modelled values provided much more realistic values for many poorly surveyed counties. For example, many small counties in Virginia (< 20 km$^2$) are completely surrounded by a better-surveyed county. These small counties originally reported less than 20 native plant species (all but three less than four) and zero non-native plant species, resulting in native plant density values averaging 1.3 species km$^{-2}$. These low numbers are highly unlikely, especially when compared to surrounding county densities of at least 0.4 species km$^{-2}$ and averaging 0.8 species km$^{-2}$. The modelled density values for these small counties (all > 0.4 species km$^{-2}$ with an average of 0.7 species km$^{-2}$) were more similar to the surrounding counties with similar habitats and more complete survey records. Similarly, modelling improved richness and density estimates in Texas and Maryland where many counties had reported less than 100 native species.

The models we developed performed well, even though a few counties had predicted density values outside the range of densities found in the observed data set when the models were applied to all 3111 counties in the conterminous USA (Fig. 1). The large values were minimal, especially for the undersampled counties we were trying to predict, and these values were found in major urban centres such as New York City. Humans tend to settle in areas with fertile soil, near water, and along coasts, conditions that would predispose dense urban areas to high densities of species (Stohlgren *et al.*, 2005a, 2006). Additionally, these urban counties are atypical when examining human population density, which was one of the strongest predictors in the models.

Despite using the best model of the species–area relationship, we were unable to completely remove the trend. Some small counties still appear to have slightly inflated densities, while some larger counties appear to have lower values, but all of the candidate models for the transformation had some residual effect. Additionally, some of these differences may reflect real patterns. County boundaries are politically defined and not ecologically defined. Neighbouring counties of different sizes in homogenous areas may have similar species richness, as new habitat types and resources are not being added with additional area. In these cases, the larger counties would have relatively smaller density values.

Several interesting patterns can be observed in the model results that inform theories of native and non-native plant diversity. Primarily, native plant species density was selected as the best predictor of non-native plant species density. These findings support the results of several large-scale observational studies in which native and non-native species richness were positively correlated at multiple scales (Wiser *et al.*, 1998; Levine & D'Antonio, 1999; Lonsdale, 1999; Smith & Knapp, 1999; Stohlgren *et al.*, 1999, 2003, 2005a, 2006; Levine, 2000; Richardson *et al.*, 2005).

© 2006 The Authors

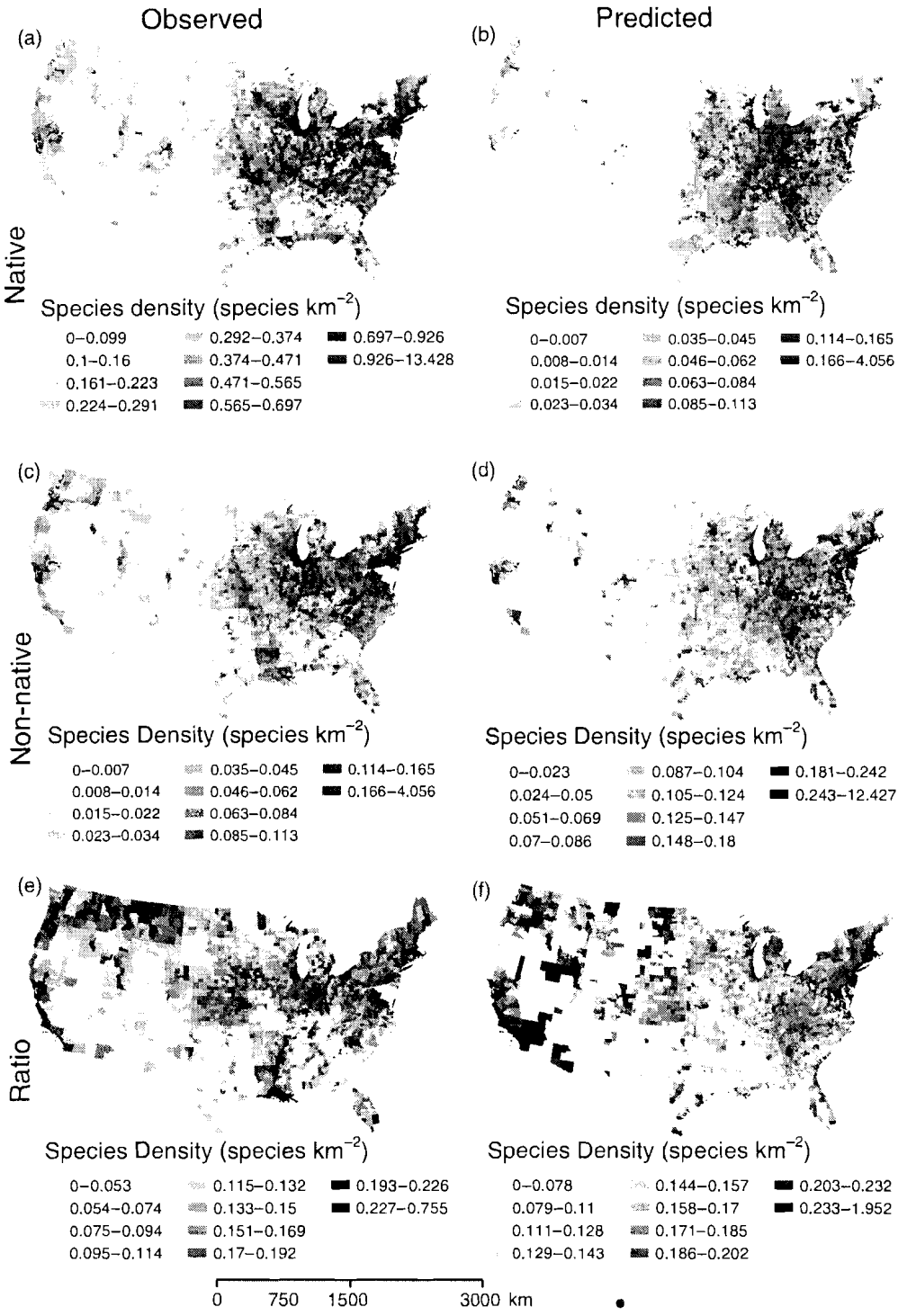


**Figure 1** Maps for observed and predicted plant species density (species km$^{-2}$) per US county by species origin and by the proportion of non-native species in the flora. (a) Observed native species density. (b) Predicted native species density from model with environmental/topographical, biotic, and human variables (global model). (c) Observed non-native species density. (d) Predicted non-native species density from model with biotic and human variables. (e) Observed proportion of non-native to native species. (f) Predicted proportion of non-native to native species from b and d. The map projection is USA Contiguous Albers Equal Area Conic USGS version, datum NAD83.

© 2006 The Authors

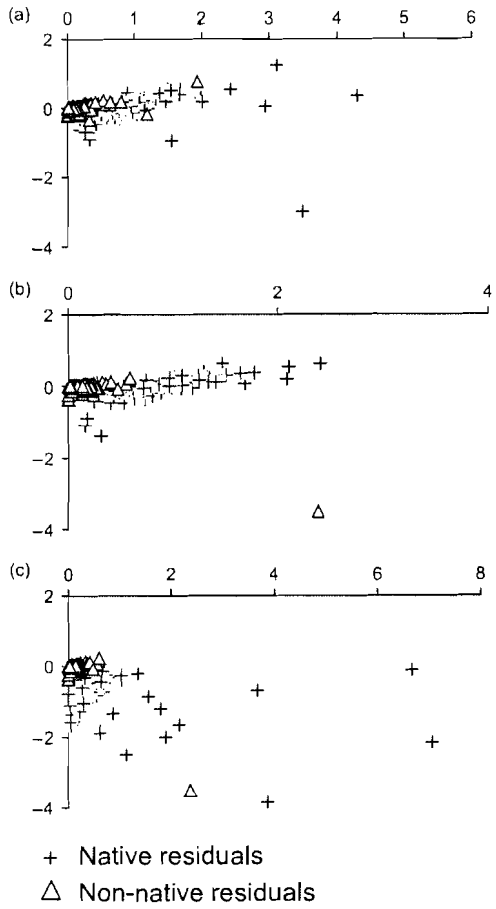


Figure 2 Observed vs. residual values for native and non-native plant species density. (a) Values from the top four counties per state for native plant species density used to generate the models. (b) Values for counties ranked five to eight per state for observed native plant species density. (c) Values for the four counties per state with the lowest observed native plant species densities.

However, several small-scale experimental studies have found the opposite result (Levine, 2000; Naeem *et al.*, 2000; Hector *et al.*, 2001; Lyons & Schwartz, 2001; Kennedy *et al.*, 2002; Prieur-Richard *et al.*, 2002; Troumbis *et al.*, 2002; Tilman, 2004). Our results support the observational studies' findings where, at least at the county-level scale used here, non-native plant species density is highly predictable with knowledge of native plant species density patterns. This supports the theory that the same environmental factors affect native and non-native plant species richness (Richardson *et al.*, 2005; Stohlgren *et al.*, 2005b).

This theory is also supported when comparing the ranking of the different models for native and non-native plant species, excluding the density models that used natives to predict non-natives. The two best models for both native and non-native species were the human and biotic factors model and the global model. In all four models, native bird species density and human population density were the most important predictors. Native bird species density is a surrogate for habitat heterogeneity, acting as a species indicator group similar to an indicator species. Other studies have shown that native bird species density is highly correlated with native plant density (Stohlgren *et al.*, 2006). These biotic variables are not assumed to be causal, but represent the complex factors driving species richness and density. Biotic substitutes such as these that are correlated with these driving variables provide good predictors. Given that the same variables were most important and that the same models were selected for both native and non-native species density, we suggest that the same factors affect the species richness, and, therefore, diversity, of both native and non-native species.

Our effort to understand undersampled counties can be evaluated by comparing results to a similar analysis with the same data set. Stohlgren *et al.* (2006) used the same BONAP data set and the same independent variables but included all counties with more than 100 native species (3000 counties) to examine patterns of plant species richness. A primary difference was that models using the 177 best-sampled counties described much more of the variability than did the other group's models using the 3000 counties (native models adjusted $R^2 = 0.89$ and 0.69; non-native models adjusted $R^2 = 0.88$ and 0.86, respectively). However, the best predictor of non-native species richness was the same in both studies, a nonlinear model of native plant species richness, with both models explaining more than 90% of the variation (adjusted $R^2 > 0.9$). The model for native species density with the most support, the global model, was also the same. Model rank was also very similar, with this study and the Stohlgren study having the same model for each of the top five ranks.

In general, the models with biotic factors had more support than the other models, further strengthening the conclusions of Stohlgren *et al.* (2006) that biotic factors are the most important in predicting patterns of diversity. The ranking for the non-native models was even stronger than the native models for biotic variables. All three models that included the biotic variables ranked in the top three when the models using native density are ignored. These comparisons indicate that limiting the dependent variable to well-surveyed areas did not change the important predictor variables. Thus, for examining correlations of variables with diversity, using all available data or a subset of well-surveyed areas does not change the results. The coefficients for the predictor variables, however, did differ between the two data sets. So, if one is interested in spatial patterns of diversity (see Fig. 1), then the methods described in this paper are important for accurate results because they correct for poorly sampled areas.

These methods provide a strategy to determine areas most in need of sampling in the USA. For example, some areas such as Maryland, other counties were removed from the model calibration dataset due to poor sampling, appeared easily predictable. Other counties, such as those that had a modelled value less than the observed, may be important ones to target for future survey efforts as they appear more difficult to model. Also, hotspots identified in the models may be important areas to target survey efforts as they may be more biologically important. So, despite the fact that these models may not be completely accurate, they are the best that can be developed with available data, and may prove highly valuable in guiding future survey efforts with limited

© 2006 The Authors

resources. As new data are obtained, the models may be rerun and re-examined in an iterative approach.

A leading concept to set priorities for conservation is known as the biodiversity hotspots thesis (Myers *et al.*, 2000; Myers, 2003). According to Myers *et al.* (2003), a few small areas containing most of the world's endemic species and facing the most rapid habitat loss should be protected. However, invasions are potentially another serious threat that should be considered in developing these priorities (Rouget *et al.*, 2003; Stohlgren *et al.*, 2005b). An area high in native species richness that is being heavily invaded may not be an ideal location for targeting long-term conservation without very active management of harmful invaders. A hotspot of biodiversity that has not yet been invaded may be a better choice, with steps taken to try and prevent large-scale invasion. The increase in the proportion of non-native flora between the observed and predicted maps here demonstrates that there are potentially more non-native species relative to natives than are currently observed. In areas like the Great Basin where there are few native species, the addition of only a few non-natives can drastically change the proportion of the flora that is non-native. These changes stand out as alarming when compared to areas with a lot of native species and non-native species. Areas such as the Great Basin may be important, unique areas to concentrate conservation efforts, especially since a few non-natives may have a greater impact given their high proportion in the total flora.

The modelled locations of hotspots of invasion should also be targeted in an early detection/rapid response program because non-native plant density has again been shown to track native-rich areas. Areas that show a high proportion of non-native species in our models that are not yet well surveyed should also be targeted for inventory and monitoring. Using the predicted surfaces in Fig. 1 (b,d,f) rather than the surfaces illustrating current knowledge (fig. 1 a,c,e) may be important for targeting invasions as these models fill in data gaps.

The potential of the techniques described here to improve information on species richness patterns is supported by our results. These same methods could be applied to other taxa at different spatial scales and have the potential to be modified for a single species. Reliable tools such as these are urgently needed to determine patterns of biodiversity for conservation efforts and to evaluate patterns of invasion at various spatial scales. Large-scale predictive maps to fill in gaps in knowledge are an important step towards understanding distribution patterns of species, improving effectiveness of conservation efforts, and making progress in managing the invasive species problem.

## ACKNOWLEDGEMENTS

We thank Alycia Crall for comments on the manuscript; Curtis Flather for supplying data for the independent variables used in the modelling process; the US Geological Survey, NASA, and the National Biological Information Infrastructure for funding; and the US Geological Survey Fort Collins Science Center and the Natural Resource Ecology Laboratory at Colorado State University for logistical support.

## REFERENCES

Burnham, K.P. & Anderson, D.R. (1998) *Model selection and inference: a practical information-theoretic approach.* Springer, New York.

Crosier, C.S. & Stohlgren, T.J. (2004) Improving biodiversity knowledge with data set synergy: a case study of nonnative plants in Colorado. *Weed Technology*, **18**, 1441–1444.

Hector, A., Dobson, K., Minns, A., Bazeley-White, E. & Lawton, J.H. (2001) Community diversity and invasion resistance: an experimental test in a grassland ecosystem and a review of comparable studies. *Ecological Research*, **16**, 819–831.

Hof, J., Flather, C., Baltic, T. & King, R. (2004) Forest and rangeland ecosystem condition indicators: identifying national areas of opportunity using data envelopment analysis. *Forest Science*, **50**, 473–494.

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

Kennedy, T.A., Naeem, S., Howe, K.M., Knops, J.M.H., Tilman, D. & Reich, P. (2002) Biodiversity as a barrier to ecological invasion. *Nature*, **417**, 636–638.

Levine, J.M. (2000) Species diversity and biological invasions: relating local process to community pattern. *Science*, **288**, 852–854.

Levine, J.M. & D'Antonio, C. (1999) Elton revisited: a review of evidence linking diversity and invasibility. *Oikos*, **87**, 15–26.

Lonsdale, W.M. (1999) Global patterns of plant invasions and the concept of invasibility. *Ecology*, **80**, 1522–1536.

Lyons, K.G. & Schwartz, M.W. (2001) Rare species loss alters ecosystem function — invasion resistance. *Ecology Letters*, **4**, 358–365.

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

Myers, N. (2003) Biodiversity hotspots revisited. *Biological Science*, **53**, 916–917.

Myers, N., Mittermeier, R.A., Mittermeier, C.G., da Fonseca, G.A.B. & Kent, J. (2000) Biodiversity hotspots for conservation priorities. *Nature*, **403**, 853–858.

Naeem, S., Knops, J.M.H., Tilman, D., Howe, K.M., Kennedy, T. & Gale, S. (2000) Plant diversity increases resistance to invasion in the absence of covarying extrinsic factors. *Oikos*, **91**, 97–108.

Palmer, M.W., Earls, P.G., Hoagland, B.W., White, P.S. & Wohlgemuth, T. (2002) Quantitative tools for perfecting species lists. *Environmetrics*, **13**, 121–137.

Pearson, D.L. (1994) Selecting indicator taxa for the quantitative assessment of biodiversity. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences*, **345**, 75–79.

Pimentel, D., Lach, L., Zuniga, R. & Morrison, D. (2000) Environmental and economic costs of nonindigenous species in the United States. *Biological Science*, **50**, 53–65.

Prieur-Richard, A.H., Lavorel, S., Linhart, Y.B. & Dos Santos, A. (2002) Plant diversity, herbivory and resistance of a plant

© 2006 The Authors

community to invasion in Mediterranean annual communities. *Oecologia*, **130**, 96–104.

Pyšek, P., Richardson, D.M., Rejmánek, M., Webster, G.L., Williamson, M. & Kirschner, J. (2004) Alien plants in checklists and floras: towards better communication between taxonomists and ecologists. *Taxon*, **53**, 131–143.

Richardson, D.M., Rouget, M., Ralston, S.J., Cowling, R.M., van Rensburg, B.J. & Thuiller, W. (2005) Species richness of alien plants in South Africa: environmental correlates and the relationship with indigenous plant species richness. *Ecoscience*, **12**, 391–402.

Ricklefs, R.E. (2004) A comprehensive framework for global patterns in biodiversity. *Ecology Letters*, **7**, 1–15.

Rouget, M., Richardson, D.M., Cowling, R.M., Lloyd, J.W. & Lombard, A.T. (2003) Current patterns of habitat transformation and future threats to biodiversity in the Cape Floristic Region, South Africa. *Biological Conservation*, **112**, 63–85.

Scott, J.M., Heglund, P.J., Morrison, M., Raphael, M., Haufler, J. & Wall, B. (2002) *Predicting species occurrences: issues of scale and accuracy*. Island Press, Washington, D.C.

Smith, M.D. & Knapp, A.K. (1999) Exotic plant species in a C-4-dominated grassland: invasibility, disturbance, and community structure. *Oecologia*, **120**, 605–612.

Stohlgren, T., Barnett, D.T., Flather, C., Fuller, P., Peterjohn, B., Kartesz, J. & Master, L.L. (2006) Species richness and patterns of invasion in plants, birds, and fishes in the United States. *Biological Invasions*, **8**, 427–447.

Stohlgren, T.J., Schell, L.D., Bull, K.A., Otsuki, Y., Newman, G., Bashkin, M., Son, Y., Binkley, D., Chong, G.W. & Kalkhan, M.A. (1999) Exotic plant species invade hot spots of native plant diversity. *Ecological Monographs*, **69**, 25–46.

Stohlgren, T.J., Barnett, D.T. & Kartesz, J.T. (2003) The rich get richer: patterns of plant invasions in the United States. *Frontiers in Ecology and the Environment*, **1**, 11–14.

Stohlgren, T.J., Barnett, D.T., Flather, C., Kartesz, J. & Peterjohn, B. (2005a) Plant species invasions along the latitudinal gradient in the United States. *Ecology*, **86**, 2298–2309.

Stohlgren, T.J., Guenther, D., Evangelista, P. & Alley, N. (2005b) Patterns of plant rarity, endemism, and uniqueness in an arid landscape. *Ecological Applications*, **15**, 715–725.

Stohlgren, T.J. & Schnase, J.L. (2006) Biological hazards: what we need to know about invasive species. *Risk Assessment Journal*, **26**, 163–173.

Thornthwaite, C.W. & Mather, J.R. (1955) The water balance. *Publications in Climatology*, **8**, 1–86.

Tilman, D. (2004) Niche tradeoffs, neutrality, and community structure: a stochastic theory of resource competition, invasion, and community assembly. *Proceedings of the National Academy of Sciences USA*, **101**, 10854–10861.

Troumbis, A.Y., Galanidis, A. & Kokkoris, G.D. (2002) Components of short-term invasibility in experimental Mediterranean grasslands. *Oikos*, **98**, 239–250.

Vitousek, P.M., Walker, L.R., Whiteaker, L.D., Mueller-Dombois, D. & Matson, P.A. (1987) Biological invasion by *Myrica faya* alters ecosystem development in Hawaii. *Science*, **238**, 802–804.

Wilcove, D.S., Rothstein, D., Dubow, J., Phillips, A. & Losos, E. (1998) Quantifying threats to imperiled species in the United States. *Biological Science*, **48**, 607–615.

Wiser, S.K., Allen, R.B., Clinton, P.W. & Platt, K.H. (1998) Community structure and forest invasion by an exotic herb over 23 years. *Ecology*, **79**, 2071–2081.

© 2006 The Authors

**Appendix** Source of data for the independent variables used in the analysis

| Data set | Description | Source |
|---|---|---|
| Native and non-native plant species richness | Number of native and non-native plant species per county (species $km^{-2}$) | Biota of North America Program, John Kartesz, University of North Carolina Chapel Hill |
| County area | Size of county ($km^2$) | Environmental Systems Research Institute (ESRI, ArcView 3.2) |
| Minimum temperature | Mean daily minimum temperature (°C) | National Climatic Data Center, Climate Maps of the United States database |
| Mean temperature | Mean daily average temperature (°C) | National Climatic Data Center, Climate Maps of the United States database |
| Precipitation | Mean total precipitation (mm) | National Climatic Data Center, Climate Maps of the United States database |
| Mean elevation* | Counties were defined as zones, and zonal means of gridded elevation data were calculated (m) | Oregon Climate Service, PRISM digital elevation model (DEM), 1996. |
| Variation in elevation* | Counties were defined as zones, and zonal standard deviations of gridded elevation data were calculated (m) | Oregon Climate Service, PRISM climate digital data, 1996. |
| Potential evapotranspiration (PET)* | Thornthwaite's formula (mm; Thornthwaite & Mather, 1955) | Curtis Flather, USDA Forest Service |
| Actual evapotranspiration (AET)* | Thornthwaite's formula (mm; Thornthwaite & Mather, 1955) | Curtis Flather, USDA Forest Service |
| Solar radiation | 18-year annual average (1980–97) of daily shortwave radiation (MJ $m^{-2}$ $day^{-1}$) | DAYMET US Data Center |
| Percentage of crop area | Percentage of each county that was cropland in 1987 (%) | Environmental Research Systems Institute (ESRI, ArcView 3.2) |
| Human population | Number of people per county from the 2000 census (people $km^{-2}$) | Census 2000, US Census Bureau. |
| Land cover class count | Count of NLCD land cover classes per county | National Land Cover Data (NLCD) developed from 30 m Landsat Thematic Mapper (TM) data by The Multi-resolution Land Characterization (MRLC) Consortium, Version 09-06-2000 |
| Native and non-native bird species richness | Number of native and non-indigenous bird species per county data (species $km^{-2}$) | Bruce Peterjohn, US Geological Survey |
| Vegetation carbon* | Total vegetation carbon (potential — no land use effects). Kriged 30-year annual average (1961–90) at 3168 lat/long locations (gC $m^{-2}$). | National Center for Atmospheric Research (NCAR). VEMAP2 DATA, 2000. |
| Habitat disturbance* | Index of habitat disturbance. Ratio of area of disturbed land (developed, herbaceous planted/cultivated, non-natural woody vegetation (e.g. orchards, vineyards), surface mines (e.g. quarries, strip mines, gravel pits)) to total area in a county. Land cover classes as defined in the National Land Cover Data (NLCD). | National Land Cover Data (NLCD) developed from 30-m Landsat Thematic Mapper (TM) data by The Multi-resolution Land Characterization (MRLC) Consortium, version 09-06-2000. |

*2004 Resource Interactions Database, John Hoff, Curtis Flather, and Tony Baltic. USDA Forest Service, Rocky Mountain Research Station (Hoff *et al.*, 2004).


© 2006 The Authors