Available online at www.sciencedirect.com

SCIENCE DIRECT®

Ecological Modelling 192 (2006) 188–196

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecolmodel

Improving ecological niche models by data mining large environmental datasets for surrogate models

David R.B. Stockwell [*]

San Diego Supercomputer Center, University of California, San Diego,
9500 Gilman Drive, La Jolla, CA 92093-0505, USA

Received 30 June 2004; received in revised form 6 May 2005; accepted 20 May 2005
Available online 20 July 2005

Abstract

WhyWhere is a new ecological niche modeling (ENM) algorithm for mapping and explaining the distribution of species. The algorithm uses image processing methods to efficiently sift through large amounts of data to find the few variables that best predict species occurrence. The purpose of this paper is to describe and justify the main parameterizations and to show preliminary success at rapidly providing accurate, scalable, and simple ENMs. Preliminary results for six species of plants and animals in different regions indicate a significant ($p < 0.01$) 14% increase in accuracy over the GARP algorithm using models with few, typically two, variables. The increase is attributed to access to additional data, particularly remotely sensed monthly versus annual climate averages. WhyWhere is also six times faster than GARP on large datasets. A data mining based approach with transparent access to remote data archives is a new paradigm for ENM, particularly suited to finding correlates in large databases of fine resolution surfaces. Software for WhyWhere is freely available, both as a service and in a desktop downloadable form from the web site http://biodi.sdsc.edu/ww_home.html.

Keywords: WhyWhere; Ecological niche modeling; Surrogate models; Data mining; Remote sensing

---

1. Introduction

Since the inception of ecological niche modeling, finding better methods of answering the question “where is it and why?” has been a fundamental objective of modelers (Stockwell, 1993). Choosing among the many forms of predictive models of habitat distribution in ecology has not been based on statistical performance in a single trial, but has included the objectives of the study and generality (Guisan and Zimmermann, 2000). For example, it has been shown that different forms of models may be more accurate at different sample sizes (Stockwell and Peterson, 2002) although ecological niche models (ENMs) developed by in Genetic Algorithm for Rule-set Prediction (GARP) are accurate over

---

[*] Tel.: +1 858 8220942; fax: +1 858 8223631.
E-mail address: davids@sdsc.edu.

0304-3800/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2005.05.029

a range of sample sizes (Stockwell and Peters, 1999). However, the ‘why’ question is multi-faceted and not easily inferred from ‘black-box’ complex models (Stockwell et al., 1990). A stepwise removal procedure indicates critical variables in GARP rule-sets (Peterson and Cohoon, 1999), but this approach is time-consuming and provides little additional information. Two segmented networks have been used to overcome the ‘black-box’ problem and explain how confident a neural net is in its conclusions (Werner and Obach, 2001). Generalized linear models (GLMs) and Generalized additive models (GAMs) explain by nature of their functional forms representing the uni-modal expectations of ecological niche theory (Austin, 2002).

It is not clear that agreement with ecological theory – e.g. the principle of central tendency in niche theory – is sufficient to confidently answer ‘why’. For example, although the use of linear instead of non-linear models leads to sub-optimal species distribution models (Austin et al., 1990; Austin, 2002), and climatic envelopes defined by the Mahalanobis distance are more accurate that rectilinear envelopes (Farber and Kadmon, 2003). However, highly non-Gaussian distributions are remain problematic for all parametric statistical methods (James and McCulloch, 1990). The distribution of environmental values are almost always highly skewed, e.g. temperatures are generally moderate throughout most of the landscape, with small areas of extreme cold on mountain summits. In addition, extreme non-linearities not representable by uni-modal functions are common, particularly in remote sensing data. For example, values of the percent of vegetation cover in the continuous fields data (Hansen et al., 2003) range from 20 to 80, but zero cover has value 255. These data sets would not perform well in models of central tendency due to inversion of the natural placement for the zero value. The ideal ENM method will (1) be capable of modeling a wide range of responses, (2) allow critical examination of assumptions, (3) be a simple approach that will not fit inappropriate functions, but (4) will handle extremely non-linear data, and (5) will efficiently turn an increasing flood of data from satellites, geographic information systems and climate model outputs into simple, scalable ENMs.

In a new approach to this problem, we describe the WhyWhere algorithm, which integrates a dynamic categorization procedure with a form of data mining, sifting through large amounts of data with an efficient image processing routine to discover accurate ‘surrogate’ models. A surrogate model in ENM is a one-dimensional (1D) model with a discrete categorization of the landscape, such as an aerial photographic or satellite maps of land classes based on vegetation. Surrogate, refers to the way landcover maps stand in place of species habitat (Stomes and Estes, 1993). Developers calibrate surrogates for prediction with the frequency with which survey points fall within the classes. The surrogate approach is simple, intuitive, and has figured highly in recent work on the prediction of species distributions and understanding patterns of biodiversity (e.g. Scott et al., 1996). Difficulties in obtaining adequate survey data, high-resolution vegetation maps, and concerns with the statistical validity have tended to limit their application (Stockwell and Peterson, 2003). However, in a comparative study with a large number of species, simple ‘surrogate’ maps generally equaled or exceeded accuracy of other multivariate methods (Stockwell and Peterson, 2002). WhyWhere exploits a natural analogy between a landcover map and the colors in an image, developing a surrogate model by converting between 3D ‘raw’ formats, where the intensity of each color is stored for each pixel, and a 1D ‘palette’ format (e.g. GIF), a size-limited list of colors. The color reduction algorithm called Heckbert’s median cut (Heckbert, 1982) converts a continuous color image to a reduced color-categorized image. This algorithm is widely regarded as providing a natural appearance to the human eye with high data compression. This algorithm divides the pixels in the image into equally sized categories. More categories are assigned to large gradually changing color areas (e.g. flesh tones, green vegetation) than to small areas of color (e.g. a small red ball, a mountain top). The median cut approach and discrete binning of colors in Heckbert’s algorithm inherently handles extremely skewed color distributions. Software for WhyWhere is freely available, both as a service and in a desktop downloadable form from the web site http://biodi.sdsc.edu/ww_home.html.

2. Method

Model development follows the usual stages of species distribution modeling (Stockwell and Peters, 1999). There are a number of enhancements for ease of use, however, particularly transparent access to environmental data from a large remote data archive.

2.1. Preparation of environmental data

Point data are entered as longitude latitude (x, y) pairs. A set of environmental layers is prepared by cropping and scaling global extent datasets to the same resolution and extent as the point data. The global datasets may be stored in a remote archive and accessed transparently to the user (providing they are on the internet). Alternatively, global data may be stored locally, or copied into the working directory. Regions of the image not of interest (such as water in a terrestrial analysis) are masked out. Point data are then filtered to produce one unique point per grid cell, and the geographic extent is determined from the range of the point data.

2.2. Model development and testing

At each iteration, the algorithm generates new surrogate models from the combination of each variable with the optimal variable in the previous channel(s), stopping when a new channel does not significantly exceed the accuracy of the previous channels. The most accurate optimal model is the image at the penultimate iteration, e.g.

1. find the most accurate environmental variable;
2. put this variable in the red channel of an image and find the most accurate combination with each of the environmental files in the green channel;
3. if necessary, test this image with each of the environmental layers in the blue channel.

Calibration of surrogate: In this case developing the model reduces to calibrating each category (color) for the conditional probability of a species. A simple heuristic on the number of data points auto-selects the number of colors c if required:

$$\text{if}\, n < 16, \quad \text{then}\, c = 16\, \text{else}\, c = 255$$

For a color i, where P is the proportion of each occurrence points of a color i and B_i is the proportion of absences (or background) of a specific color then the expected probability Pr is given by the rule:

$$\text{if}\,(P_i + B_i) > 0, \quad \text{then}\, Pr_i = \frac{P_i}{P_i + B_i}\ \text{else}\ Pr_i = 0$$

The probabilities in array Pr can predict the distribution using a cut x in a characteristic function $\boldsymbol{X}{:}\boldsymbol{I} \to \{0, 1\}$:

$$\boldsymbol{X}(i) : \text{if}\, Pr_i > x, \quad \text{then 1 else 0}$$

The accuracy A is the sum over each n colors color divided by two (as P_i and B_i sum to one).

$$A = \sum_{i=0}^{n} \frac{\boldsymbol{X}(i)P_i + (1 - \boldsymbol{X}(i))B_i}{2}$$

The difference in accuracy between models provides a measure of significance, and can be calculated using the Z statistic, assuming the binomial distribution, generated from the independent sampling of each point in the populations of presences and absences. While this assumption may be violated by spatial autocorrelation, the algorithm has attempted to reduce this by eliminating duplicate occurrences and random sampling of data points. The Z statistic for the difference in accuracy A of models where n is the number of points is:

$$Z = \frac{|A_1 - A_2|}{\sqrt{\frac{A_1 A_2}{n}}}$$

2.3. Prediction to new areas

For extrapolative predictions where a model developed in a specific environmental dataset is applied in different dataset, either a different location (invasive species) or with different base layers (migration, climate change) the list of colors, is used to categorize a new multi-channel image into the original set of colors. This operation can be performed in *netpbm* tools with any of *pnmcolormap*, *pnmremap*, or *pnmquant*.

2.4. The explanation stage

Allows study of the response surface and conversion to a variety of formats for visualization and mapping. The species distribution is mapped by changing the palette colors in an image according to the probability, e.g. red is highly probable, green is low, and blue is zero. The response surface of the points of presence and absence can be shown in a histogram which is two-

Table 1
Examples of terrestrial datasets in WhyWhere

Name	Description	Resolution (°)	No. of vars
treecover	Continuous field data—treecover	0.01	1
wrzsoil	Webb et al. soil particle size properties Zobler soil types	1.0	1
wrtext	Webb et al. texture-based potential storage of water (mm)	1.0	1
wrsoil	Webb et al. soil profile thickness (cm)	1.0	1
wrroot	Webb et al. potential storage of water in root zone (mm)	1.0	1
wrprof	Webb et al. potential storage of water in soil profile (mm)	1.0	1
wrmodii	Webb et al. Model II soil water (mm)	1.0	1
wrcont	Webb et al. continent codes from the FAO/UNESCO soil map of the world	1.0	1
whcov1	Wilson & Henderson—sellers primary land cover classes	1.0	1
srztext	Staub and Rosenzweig Zobler near-surface soil texture	1.0	1
owe14dr	Resolution codes for OWE1.4D	0.5	1
owe14d	Olson World Ecosystem Classes Version 1.4D	0.5	1
owe13a	Olson World Ecosystems Version 1.3A	0.5	1
mgvc488	1988 MGV PCA components 1–4	0.5	4
mgv0001-12	Average January–December Generalized Global Vegetation Index	0.5	12
mfwwet	Matthews and Fung wetland type	1.0	1
mfwveg	Matthews and Fung vegetation type	1.0	1
mfwsrc	Matthews and Fung wetland data source	1.0	1
mfwsol	FAO soil types of Matthews and Fung wetland locations	1.0	1
mfwfrin	Matthews and Fung fractional inundation	1.0	1
maveg	Matthews vegetation types	1.0	1
malbwn	Matthews winter albedo (% × 100)	1.0	1
malbsp	Matthews spring albedo (% × 100)	1.0	1
malbsm	Matthews summer albedo (% × 100)	1.0	1
malbfa	Matthews fall albedo (% × 100)	1.0	1
macult	Matthews cultivation intensity	1.0	1
Lwtsd01-12	Legates & Willmott January–December temperature (S.D.)	0.5	12
lwtsd00	Legates & Willmott annual temperature (S.D.)	0.5	1
Lwtmp01-12	Legates & Willmott January–December temperature (0.1 °C)	0.5	12
lwtmp00	Legates & Willmott annual temperature (0.1 °C)	0.5	1
Lwmsd01-12	Legates & Willmott January–December measured precipitation (S.D.)	0.5	12
lwmsd00	Legates & Willmott annual measured precipitation (S.D.)	0.5	1
Lwmpr01-12	Legates & Willmott January–December measured precipitation (mm/month)	0.5	12
lwmpr00	Legates & Willmott annual measured precipitation (mm/year)	0.5	1
Lwcsd01-12	Legates & Willmott January–December corrected precipitation (S.D.)	0.5	12
lwcsd00	Legates & Willmott annual corrected precipitation (S.D.)	0.5	1
Lwcpr01-12	Legates & Willmott January–December corrected precipitation (mm/month)	0.5	12
lwcpr00	Legates & Willmott annual corrected precipitation (mm/year)	0.5	1
lmfmeth	Lerner et al. annual methane emission (kg/km^2)	1.0	1
lholdag	Leemans' Holdridge life zones aggregated classification	0.5	1
lhold	Leemans' Holdridge life zones classification	0.5	1
Lctmp01-12	Leemans and Cramer January–December temperature (0.1 °C)	0.5	12
Lcprc01-12	Leemans and Cramer January–December precipitation (mm/month)	0.5	12
Lccld01-12	Leemans and Cramer January–December cloudiness (% sunshine)	0.5	12
fnocwat	Navy Terrain Data—percent water cover	0.1667	1
fnocurb	Navy Terrain Data—percent urban cover	0.1667	1
fnocst	Navy Terrain Data—secondary surface type codes	0.1667	1
fnocrdg	Navy Terrain Data—number of significant ridges	0.1667	1
fnocpt	Navy Terrain Data—primary surface type codes	0.1667	1
fnocmod	Navy Terrain Data—modal elevation (m)	0.1667	1
fnocmin	Navy Terrain Data—minimum elevation (m)	0.1667	1
fnocmax	Navy Terrain Data—maximum elevation (m)	0.1667	1
fnocazm	Navy Terrain Data—direction of ridges (degrees × 10)	0.1667	1
etopo2	Etopo elevation	0.0333	1

dimensional in the case of one variable, and three-dimensional in the case of models with more than one variable.

The datasets consisted of publicly available, raster formatted environmental datasets in geographic (lat. long.) projection, with global coverage drawn from the EPA Global Ecosystems Dataset v2.0 (NOAA-EPA, 1992). Additional satellite data, such as the NDVI continuous fields data with a 1 km resolution, were obtained from the Global Land Cover Facility (Hansen et al., 2003). Table 1 contains a selection to demonstrate the current range of datasets and their resolutions. Occurrence points for six species on six species of birds and plants (Cerulean Warbler, Swainson's Flycatcher, Eared Trogon, Rubber Vine, and two user supplied mammal datasets) in six regions of the world (North America, South America, Mexico, Australia, Veracruz, and Brazil), respectively, test and evaluate parameter settings for the algorithm. External accuracy is the only reliable measure of accuracy, as models with high internal accuracy may be 'overfitting' the data, being too specialized and consequently performing poorly on new data (Verbyla and Litvaitis, 1989). In this study, 80% of data were used to develop the model and 20% were held back for evaluating external accuracy.

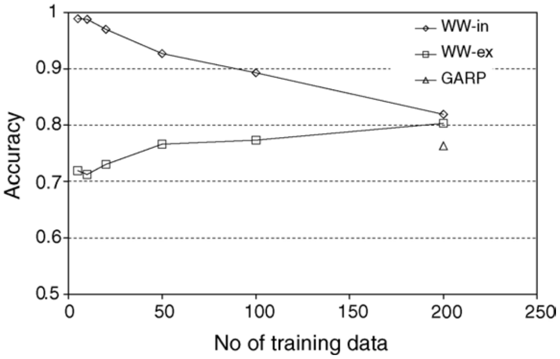

Fig. 1. Accuracy on Cerulean Warbler data with number of species data. WW, WhyWhere internal and external; GARP, GARP external accuracy.

3. Results

Fig. 1 shows the internal and external accuracies with Cerulean Warbler with different numbers of data, showing the typical convergence of internal and external accuracies as the number of data points increases, as found in all modeling algorithms (Stockwell and Peterson, 2002). Fig. 2 shows the response surface

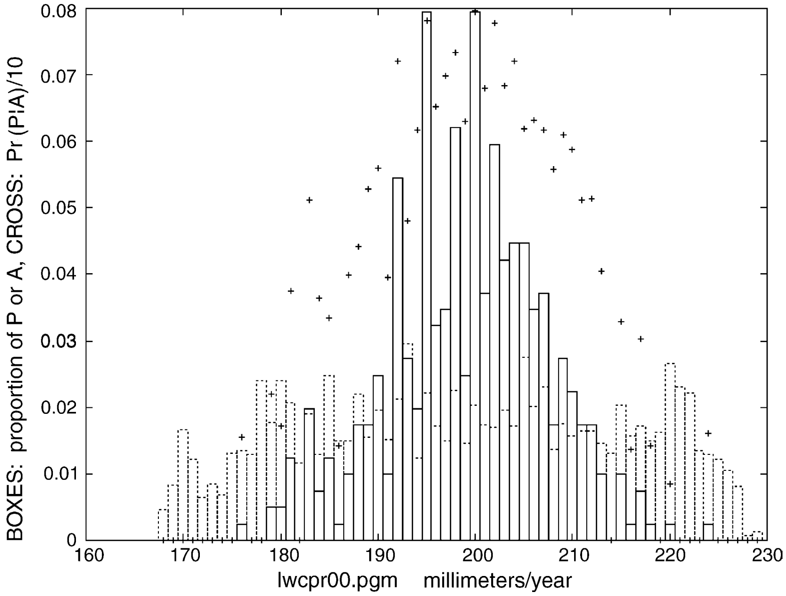

Fig. 2. The response of Cerulean Warbler to the most predictive variable. Solid line boxes are the proportion of occurrence points and dashed line boxes are the proportion background points at each color intensity (class). The crosses are the probability of presence given a color class, showing a typical univariate response to the temperature range.

for Cerulean Warbler from WhyWhere: the number of points at each color, covering the range of average annual precipitation (lwcpr00). The response surface illustrates critical features of an ecological niche model: the distribution of the presence points is restricted to a range; the background points cover the whole range, and presence of an outlier and the non-linearity of the response.

3.1. *Channels*

Is the restriction to three variables (red, green, and blue channels) adequate to optimize accuracy? Fig. 3 shows the internal and external accuracies of models at each iteration of the algorithm on three test species. Both the internal and external accuracies generally maximize in two iterations. Thus, the simpler strategy of selecting two variables using internal accuracy may be adequate.

3.2. *Categories*

Fig. 4 shows the accuracy on a range of different size Cerulean Warbler datasets for 255, 128, and 64 categories. The 255 categories produced higher external accuracies for all datasets except for the smallest. Thus, 255 categories are adequate, but optimal choice of numbers of categories is important on small numbers of data.

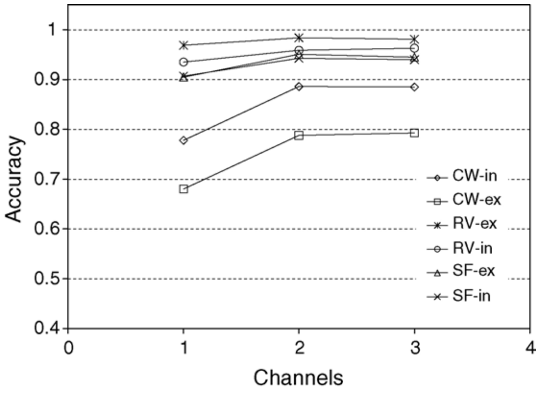

Fig. 3. Increase in internal and external accuracies with number of layers in surrogate model (or iterations of algorithm) for three species and continents; CW, Cerulean Warbler (N. America); RV, Rubber Vine (Australia); SF, Swainson's Flycatcher (S. America).

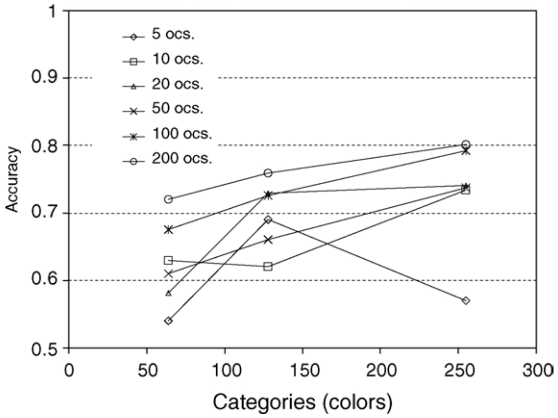

Fig. 4. Variation in external accuracy of Cerulean Warbler by number of categories and training set sizes.

3.3. *Accuracy*

We compared WhyWhere with GARP under typical usage conditions. GARP used annual average climate, elevation, and vegetation datasets described in Stockwell and Peterson (2003), and WhyWhere had access to a suite of datasets from Table 1. The sampling protocol was identical, with the data selected to an initial prior probability of 0.5, and testing carried out on random samples with replacement. The accuracy of GARP was 0.77 ± 0.03, and WhyWhere has 0.88 ± 0.03, a highly significant ($p < 0.01$) 14% increase.

3.4. *Speed*

At small map sizes (coarse resolution), WhyWhere is slowed by constant aspects of the algorithm but at finer resolutions and larger datasets WhyWhere was 583% faster than GARP making models with higher resolution of distributions feasible (Fig. 5).

3.5. *Resolution*

The Cerulean Warbler was predicted at spatial resolutions from 1° to 0.05° (Table 2). Three variables were chosen repeatedly for the first channel: mgv0009 (Average September Generalized Global Vegetation Index), lcprc04 (Leemans and Cramer April Precipitation), or lwcpr04 (Legates & Willmott April Corrected Precipitation) and the second channel varied.

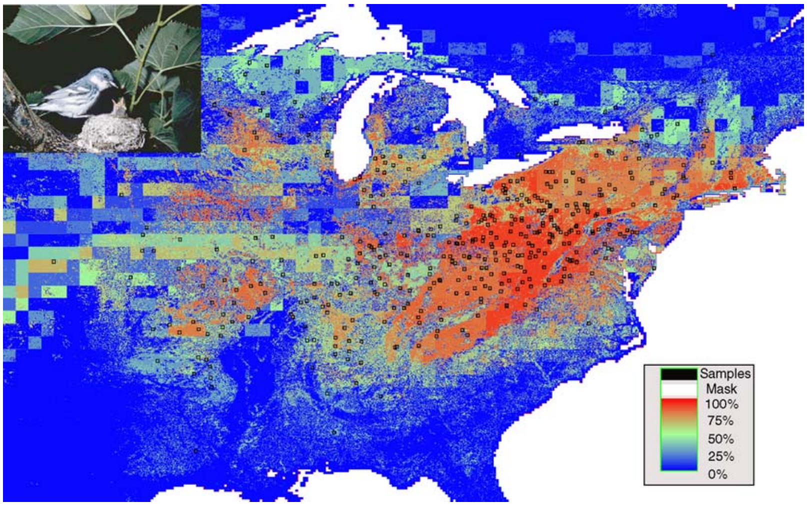

Fig. 5. WhyWhere generated image of the predicted distribution of the Cerulean Warbler.

Table 2
The accuracy and variables selected in the models of the Cerulean Warbler at a range of spatial resolutions

Resolution	In	Ex	Var 1	Var 2
1	0.868	0.868	lcprc04	lmfdcow
0.5	0.861	0.855	lwcpr04	mgvc288
0.2	0.878	0.875	mgv0009	lccld01
0.1	0.895	0.884	mgv0009	lmfpig
0.1	0.89	0.886	lcprc04	lwtmp02
0.05	0.91	0.898	mgv0009	lwtmp04

The accuracy increased slightly with increasing resolution (0.87–0.89) but was not significant.

4. Discussion

The WhyWhere system shows many desirable characteristics. Firstly, there is a significant increase in accuracy and speed over GARP. Secondly, the models are typically composed of few (typically two) variables. Thirdly, the approach has desirable features for a modeling method in a generic, analytical information infrastructure. The capacity to crop and scale environmental data from large remote archive removes the need to develop region-specific base layers. If required the user can use their own datasets, or contribute data to the archive. The expansion of potential datasets will increase the of applications beyond species distributions to other types of occurrences with other potential environmental correlates, e.g. geomorphology, such as landslides, or social, such as mapping crime frequency.

Fewer variables gave greater accuracy. If we observe Occams' razor, we should prefer the simpler model, as fewer variables implies better explanations than complex models with many variables, such as those produced by GARP or neural nets. In addition, the simple model of the response of the species shown in Fig. 2 illustrates the classical univariate response also seen in GLMs and GAMs, and the potential to represent more extreme non-linearities, such as outliers, out of order values, and skewed and multi-modal responses. Improvements in both prediction and explanation have implications for one of the most vexing questions in ecological modeling: the difficulty of simultaneously developing models that both predict and explain (Loehle, 1983; Stockwell, 1993). Thus, in the future, with the robust non-linearity of the analytical approach, WhyWhere could become a general-purpose predictive and explanatory system to enable new research and development directions.

A larger evaluation study found surrogate, logistic regression, and GARP methods gave similar accuracy standardized protocol (Stockwell and Peterson, 2002), suggesting the increase in accuracy is due to access to a greater range of data rather than inherent in the model itself. Examination of the variables selected by the system for the six species (Table 3) shows *treecover* is

Table 3
The variables selected for models of each of six species at a resolution of 0.05°

Name	Description	Native resolution	Resolution 0.05
treecover	Continuous field data—treecover	0.0083	3
mgvc188	1988 MGV PCA component 1	0.167	2
lwmpr04	Legates & Willmott April measured precipitation (mm/month)	0.5	2
lctmp12	Leemans and Cramer December temperature (0.1 °C)	0.5	2
lwmsd08	Legates & Willmott August measured precipitation (S.D.)	0.5	1
owe13a	Olson World Ecosystems Version 1.3A	0.5	1
mgvc388	1988 MGV PCA component 3	0.5	1
lwmsd09	Legates & Willmott September measured precipitation (S.D.)	0.5	1
lwmpr00	Legates & Willmott annual measured precipitation (mm/year)	0.5	1
lwmpr10	Legates & Willmott October measured precipitation (mm/month)	0.5	1
lwtsd08	Legates & Willmott August temperature (S.D.)	0.5	1
lwmpr06	Legates & Willmott June measured precipitation (mm/month)	0.5	1
lcprc05	Leemans and Cramer May precipitation (mm/month)	0.5	1
fnocrdg	Navy Terrain Data—number of significant ridges	0.167	1
macult	Matthews cultivation intensity	1	1
lccld06	Leemans and Cramer June cloudiness (% sunshine)	0.5	1
lccld08	Leemans and Cramer August cloudiness (% sunshine)	0.5	1
lwcpr02	Legates & Willmott February corrected precipitation (mm/month)	0.5	1
mgv0001	Average January Generalized Global Vegetation Index	0.167	1
lwcpr05	Legates & Willmott May corrected precipitation (mm/month)	0.5	1

selected most frequently (three times), while mgvc188, the first eigenvector of vegetation, is selected next most frequently (two times) and the rest were monthly climate variables. As the datasets used in GARP were annual averages of climate and vegetation, improvements in the accuracy of WhyWhere could be attributed to the remotely sensed and monthly climate datasets (i.e. greater temporal resolution). One area for improvement is in the heuristic for binning environmental data. The number of categories affects accuracy of surrogate models (Stockwell and Peterson, 2002). Optimizing this choice is important, as one of the main sources of generalization in the method is the choice of categories, controlling overfitting and subsequent accuracy on independent test data.

Acknowledgements

Point data for this study were kindly made available Townsend Peterson, Leo Joseph, Karen Stocks, and Denis Filer. Versions of the software for distribution were developed by Haowei Liu. Joseph Kirkebride provided suggestions on a later draft. This work was partially funded by the National Science Foundation primarily through the National Science Foundation grants SEEK: Science Environment for Ecological Knowledge (DBI0225674) and ITR: Collaborative Research: Building the Tree of Life—A National Resource for Phyloinformatics and Computational Phylogenetics (EF0331648).

References

Austin, M., Nicholls, A., Margules, C., 1990. Measurement of the realized qualitative niche—environmental niches of 5 Eucalypt species. Ecol. Monogr. 60, 161–177.

Austin, M.P., 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modeling. Ecol. Modell. 157, 101–118.

Farber, O., Kadmon, R., 2003. Assessment of alternative approaches for bioclimatic modeling with special emphasis on the Mahalanobis distance. Ecol. Modell. 160, 115–130.

Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecol. Modell. 135, 147–186.

Heckbert, P., 1982. Color image quantization for frame buffer display. In: SIGGRAPH'82 Proceedings, p. 297.

James, F.C., McCulloch, C.E., 1990. Multivariate analysis in ecology and systematics: panacea or Pandora's box? Annu. Rev. Ecol. Syst. 21, 129–166.

Hansen, M., DeFries, R., Townshend, J.R., Carroll, M., Dimiceli, C., Sohlberg, R., 2003. 500m MODIS Vegetation Continuous Fields. The Global Land Cover Facility, College Park, Maryland.

Loehle, C., 1983. Evaluation of theories and calculation tools in ecology. Ecol. Modell. 19, 239–247.

NOAA-EPA Global Ecosystems Database Project, 1992. Global Ecosystems Database Version 1.0. User's Guide, Documentation, Reprints, and Digital Data on CD-ROM. USDOC/NOAA National Geophysical Data Center, Boulder, CO.

Peterson, A.T., Cohoon, K.P., 1999. Sensitivity of distributional prediction algorithms to geographic data completeness. Ecol. Modell. 117, 159–164.

Scott, J.M., Tear, T.H., Davis, F.W., 1996. Gap Analysis: A Landscape Approach to Biodiversity Planning. American Society for Photogrammetry and Remote Sensing, Bethesda, Maryland, ISBN 1-57083-03603.

Stockwell, D.R.B., Davey, S.M., Davis, J.R., Noble, I.R., 1990. Using induction of decision trees to predict Greater Glider density. A. I. Appl. Nat. Resour. Manage. 4, 33–43.

Stockwell, D.R.B., 1993. Machine learning and the problem of predictions and explanation in ecology. Ph.D. Thesis, Australian National University.

Stockwell, D.R.B., Peters, D., 1999. The GARP Modeling System: problems and solutions to automated spatial prediction. Int. J. Geographical Inf. Sci. 13, 143–158.

Stockwell, D.R.B., Peterson, A.T., 2002. Effects of sample size on accuracy of species distribution models. Ecol. Modell. 148, 1–13.

Stockwell, D.R.B., Peterson, A.T., 2003. Comparison of resolution of methods for mapping biodiversity patterns from point-occurrence data. Ecol. Indicators 3, 213–221.

Stomes, D.M., Estes, J.E., 1993. A remote sensing research agenda for mapping and monitoring biodiversity. Int. J. Remote Sensing, 1839–1860.

Verbyla, D.L., Litvaitis, J.A., 1989. Resampling methods for evaluating class accuracy of wildlife habitat models. Environ. Manage. 13, 783–787.

Werner, H., Obach, M., 2001. New neural network types estimating the accuracy of response for ecological modelling. Ecol. Modell. 146, 289–298.