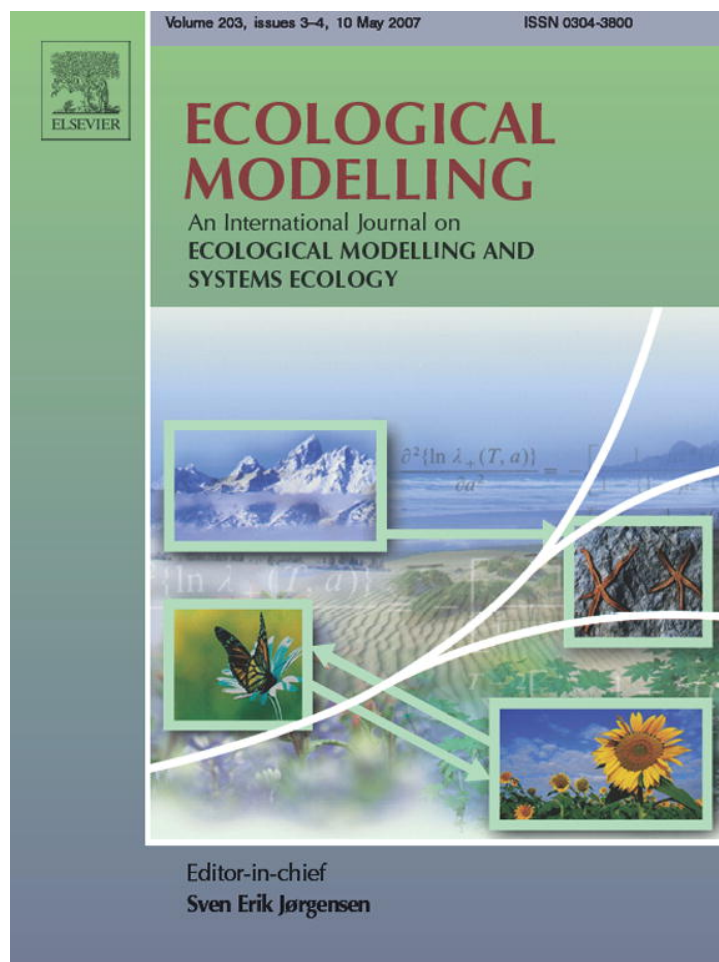


Provided for non-commercial research and educational use only.  
Not for reproduction or distribution or commercial use.



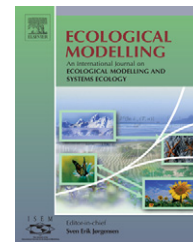
This article was originally published in a journal published by Elsevier, and the attached copy is provided by Elsevier for the author's benefit and for the benefit of the author's institution, for non-commercial research and educational use including without limitation use in instruction at your institution, sending it to specific colleagues that you know, and providing a copy to your institution's administrator.

All other uses, reproduction and distribution, including without limitation commercial reprints, selling or licensing copies or access, or posting on open internet sites, your personal or institution's website or repository, are prohibited. For exceptions, permission may be sought for such use through Elsevier's permissions site at:

<http://www.elsevier.com/locate/permissionusematerial>



ELSEVIER

available at [www.sciencedirect.com](http://www.sciencedirect.com)journal homepage: [www.elsevier.com/locate/ecolmodel](http://www.elsevier.com/locate/ecolmodel)

## Short communication

# Why not WhyWhere: The need for more complex models of simpler environmental spaces

A. Townsend Peterson

Natural History Museum and Biodiversity Research Center, The University of Kansas,  
Lawrence, Kansas 66045, USA

### ARTICLE INFO

#### Article history:

Received 24 April 2006  
Received in revised form  
2 December 2006  
Accepted 12 December 2006  
Published on line 30 January 2007

#### Keywords:

WhyWhere  
Ecological niche modeling  
Geographic distribution  
Overfitting

### ABSTRACT

WhyWhere has recently been introduced by Stockwell as a new tool for ecological niche modeling or species distribution modeling. I address two features of WhyWhere and its presentation: (1) the assertion that hundreds of environmental data layers are necessary to summarize ecological variation across environments, and (2) the idea that species' ecological needs can be summarized in just 2 or 3 dimensions. I present evidence that neither idea is valid, effectively arguing that environmental spaces are simpler than Stockwell envisages, but that more complex and dimensional models are necessary to describe species' ecological niches. Beyond these conceptual problems, WhyWhere suffers from several more practical problems, which render it little useful for any practical application.

© 2007 Elsevier B.V. All rights reserved.

The field that is termed 'ecological niche modeling' or 'species distribution modeling' is seeing an impressive swell of activity, with numerous tests, new applications, and reviews (Guisan and Zimmermann, 2000; Soberón and Peterson, 2004; Thomas et al., 2004; Wiens and Graham, 2005). Concurrent with this activity has been careful attention to the algorithms and methods used, including the development of new methods (Nix, 1986; Carpenter et al., 1993; Austin and Meyers, 1996; Lehmann et al., 2002; Pearson et al., 2002; Phillips et al., 2006) and tests and comparisons of the relative performance of different methods (Manel et al., 1999a, 1999b; Elith and Burgman, 2002; Elith et al., 2006). Most recently, a paper in this journal has introduced the algorithm WhyWhere, a new approach that purports to offer significant advances over existing approaches (Stockwell, 2006)—this method has 'been around' for a while, but has not yet

seen broad application; the purpose of this note is to provide an independent evaluation of this new method from both conceptual and practical standpoints, correcting several points in the original description that will otherwise prove misleading.

Stockwell makes two assertions that constitute the essence of the WhyWhere idea: (1) that ecological spaces (i.e., the ecological variation across landscapes) are describable in large numbers of environmental dimensions, numbering in the hundreds to thousands (Stockwell, n.d.) and (2) that very accurate models of species' ecological distributions can be developed based on "few, typically two, variables" (Stockwell, 2006). I here dispute both claims based on objective evidence and analyses, and show that these key assumptions in the WhyWhere approach are not valid.

E-mail address: [town@ku.edu](mailto:town@ku.edu).

0304-3800/\$ – see front matter © 2007 Elsevier B.V. All rights reserved.  
doi:10.1016/j.ecolmodel.2006.12.023

## 1. Complexity of ecological landscapes

Ecological landscapes can be divided at the outset into those that are physical characteristics of environments (“scenopoetic variables”) versus those that are shaped and modified by biotic factors (Hutchinson, 1978)—both Stockwell and I refer to the former, and reserve the latter from consideration, at least for the present. Stockwell, however, puts considerable weight on the fact that WhyWhere surveys hundreds of data layers that summarize ecological variation across landscapes, and as such should be ‘better’ than other algorithms (Stockwell, 2006). One certainly can – and Stockwell has – assemble hundreds or thousands of such data layers.

In reality, however, ecological characteristics of landscapes that are relevant to species’ distributions in general probably distill down to a relative few real dimensions—energy, time, space, substrate, nutrients, etc. These dimensions are the real, proximate factors that likely determine the bulk of the distribution and ecology of species in real-world landscapes. Most of the data layers that we have and use in ecological niche models are approximations to one or more of these essential quantities (Whittaker et al., 1973; Hutchinson, 1978). I would assert, however, that ‘hundreds’ of such dimensions do not exist; rather, it is possible to include many highly redundant data layers in analyses, which leads to numerous problems with singularity, odd weighting, and misleading results.

The intercorrelations observable in such large data compilations indeed demonstrate considerable redundancy. For example, in a recent study of climatic variation (as relates to blackbird niches) across the Americas (M. Eaton et al. in preparation),  $\geq 99\%$  of the variation in nine variables describing climates across the region (Hijmans et al., 2005) could be summarized in just five principal component dimensions. Similarly, in analyses of 23 topographic and climatic variables (USGS, 2001; Hijmans et al., 2005) across the United States and Canada (Peterson et al. in preparation), three suites of variables showed high intercorrelations and considerable lack of independence; here again,  $\geq 99\%$  of the variation could be summarized in just 11 principal component dimensions. In both of these examples, the redundancy of the ‘hundreds’ of data layers listed in Stockwell’s (2006) Table 1 becomes eminently clear—the ‘ecological world’ is considerably simpler than his 800 data layers might suggest.

## 2. Simplicity of ecological niches

On the other side of the coin, considerably more serious is WhyWhere’s assumption that the ecological niches of species can be summarized in very few, “typically two,” environmental dimensions. This viewpoint flies in the face of decades of ecological research—indeed, dating back to the very beginning of the field, researchers have appreciated a more complex, multivariate nature of factors limiting species’ geographic and ecological distributions (Grinnell, 1917, 1924; Hutchinson, 1957, 1978; Austin et al., 1990).

Quantitative studies of the contributions of different environmental data layers to model quality have invariably indicated that species’ ecological niches are quite complex

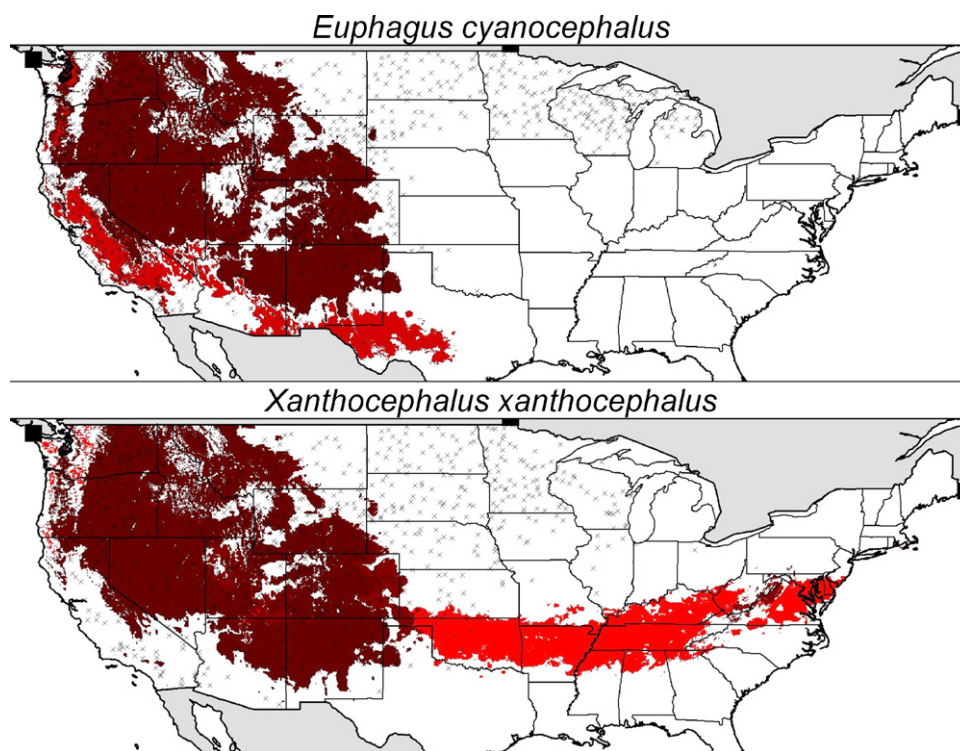
(i.e., defined in a highly dimensional space). For example, an early study using GARP jackknifed the inclusion of data layers in predictions of bird species in North America, and found that at least five data layers were required to achieve maximum predictivity—models developed with only two input layers were about 25% less accurate than those developed with larger complements of environmental data (Peterson and Cohoon, 1999). Large numbers of additional studies by numerous investigators (examples too numerous to cite, see, e.g., recent comparative analyses (Elith et al., 2006)), all conscious of the need to keep models simple, have nonetheless all seen the need to include the information from many environmental variables in order to achieve highly predictive models of species’ ecological niches.

From where does this black-and-white difference of opinion come? Stockwell appears to have been attracted by the idea of applying the tri-dimensional red–green–blue color scheme to niche modeling, which allows use of very efficient image-processing algorithms (Stockwell, 2006). What he has missed, however, is that *the devil is in the details*—that is, those third, fourth, and succeeding data layers may provide small amounts of information to the final model, but those tiny details are precisely what makes the difference between bad and good models. WhyWhere is, simply put, handicapped from the outset, and cannot marshal enough information to produce good models.

## 3. Other problems

As a means of exploring WhyWhere further, I attempted to apply it to several data sets that I know well. The results of these explorations were not satisfying, and led me to appreciate several additional problems with the program—Fig. 1 presents two representative examples of WhyWhere output when challenged with reconstructing a species’ geographic distribution based on ample species’ occurrence data. I believe that the most fundamental problems are the conceptual points listed above. However, the additional problems noted are also relevant, and so are described briefly below.

- *Documentation*—The documentation provided (Stockwell, n.d.) is fragmentary at best, and numerous features of the program are not treated at all in the document.
- *Resolution of environmental data sets*—It is well known that error related to meshing distinct spatial scales and resolutions can be propagated dramatically in GIS operations (Heuvelink, 1998), yet Stockwell states that data sets in the program range from 0.0333° to 1.0° resolution (Stockwell, 2006). Clearly, these very distinct resolutions will cause problems in any result from the program.
- *Projection of the environmental data sets and results*—When the images resulting from WhyWhere analyses are imported into GIS programs, it is clear that the program does not control for the diverse projections in which its input data are stored. That is to say, images output from different individual analyses come out in different projections, with spatial errors of even hundreds of kilometers. These images can, of course, be georectified to correct the problem, but the lack of attention to this detail is worrying.



**Fig. 1 – Two example WhyWhere predictions, developed using default program settings and the USA.0.0416667 environmental data sets, with a working resolution of 0.1° and a border beyond the data range of 10°. Darker shading indicates higher level of prediction by WhyWhere. The Xs represent known occurrences of species in the North American Breeding Bird Survey data set. The large numbers of Xs falling in white areas (i.e., areas predicted absent) represent model failures in anticipating the geographic range of the species.**

- *Path-dependence*—Most multivariate statistical approaches have appreciated the danger of path-dependence in step-wise variable selection in model building (MacNally, 1996). WhyWhere, nonetheless, is highly path-dependent—a first variable is chosen, and then “the algorithm generates new surrogate models from the combination of each (remaining) variable with the optimal variable in the previous channel(s)” (Stockwell, 2006). This feature of the algorithm clearly can lead to problems with entrapment on local optima.

#### 4. Conclusion

Given the challenge of sorting through the impressive diversity of geospatial data now available for ecological niche modeling, Stockwell should be congratulated on his development of a data-mining-based approach to this challenge. The idea of integrating data mining and modeling is a useful addition to the literature. What is more, WhyWhere’s remote data access features seem also to be a useful development, although I was not able to download or upload data sets as Stockwell had promised.

Regarding the main purpose of WhyWhere’s development, however, the point of this note is quite simple—silver bullets do not exist. That is to say, Stockwell assembled a massive stash of environmental data, and an algorithm that searches

across that stash to find two or three data layers that best serve as surrogates in ‘modeling’ a species’ geographic and ecological distribution. This is akin to the belief that a silver bullet must exist that will somehow, magically, solve all of one’s problems regarding a particular species (Ho and Pepyne, 2002). That belief (and several other elements of the conceptual design of WhyWhere as well) is simply wrong—ecological characteristics of landscapes are not as diverse and complex as Stockwell believes, and species’ ecological niches are more diverse and complex than Stockwell believes. These two assumptions yield a fatal combination of overly simple WhyWhere models that have little or no hope of providing useful insights.

#### Acknowledgements

Many thanks to David Stockwell for discussion of these ideas, to Jorge Soberón for a careful read of the manuscript, and to two anonymous readers for insightful comments.

#### REFERENCES

Austin, M.P., Meyers, J.A., 1996. Current approaches to modelling the environmental niche of eucalypts: implications for management of forest biodiversity. *Forest Ecol. Manag.* 85, 95–106.

- Austin, M.P., Nicholls, A.O., Margules, C.R., 1990. Measurement of the realized qualitative niche: environmental niches of five *Eucalyptus* species. *Ecol. Monogr.* 60, 161–177.
- Carpenter, G., Gillison, A.N., Winter, J., 1993. DOMAIN: a flexible modeling procedure for mapping potential distributions of animals and plants. *Biodiversity Conservation* 2, 667–680.
- Elith, J., Burgman, M., 2002. Predictions and their validation: rare plants in the Central Highlands, Victoria. In: Scott, J.M., Heglund, P.J., Morrison, M.L. (Eds.), *Predicting Species Occurrences: Issues of Scale and Accuracy*. Island Press, Washington, DC, pp. 303–313.
- Elith, J., Graham, C.H., Anderson, R.P., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettman, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.M., Peterson, A.T., Phillips, S.J., Richardson, K., Scachetti-Pereira, R., Schapire, R.E., Soberón, J., Williams, S.E., Wisz, M.S., Zimmermann, N.E., 2006. Novel methods improve prediction of species' distributions from occurrence data. *Ecography* 29, 129–151.
- Grinnell, J., 1917. Field tests of theories concerning distributional control. *Am. Nat.* 51, 115–128.
- Grinnell, J., 1924. Geography and Evolution. *Ecology* 5, 225–229.
- Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. *Ecol. Model.* 135, 147–186.
- Heuvelink, G.B.M., 1998. *Error Propagation in Environmental Modelling with GIS*. Taylor & Francis, London.
- Hijmans, R.J., Cameron, S., Parra, J., 2005. WorldClim, Version 1.3, University of California, Berkeley.  
<http://biogeography.berkeley.edu/worldclim.htm>.
- Ho, Y.C., Pepyne, D.L., 2002. Simple explanation of the no-free-lunch theorem and its implications. *J. Optimization Theory Appl.* 115, 549–570.
- Hutchinson, G.E., 1957. Concluding remarks. *Cold Spring Harb. Symp. Quant. Biol.* 22, 415–427.
- Hutchinson, G.E., 1978. *An Introduction to Population Ecology*. Yale University Press, New Haven.
- Lehmann, A., Overton, J.M., Leathwick, J.R., 2002. GRASP: generalized regression analysis and spatial prediction. *Ecol. Model.* 157, 189–207.
- MacNally, R., 1996. Hierarchical partitioning as an interpretative tool in multivariate inference. *Aust. J. Ecol.* 21, 224–228.
- Manel, S., Dias, J.M., Buckton, S.T., Ormerod, S.J., 1999a. Alternative methods for predicting species distribution: an illustration with Himalayan river birds. *J. Appl. Ecol.* 36, 734–747.
- Manel, S., Dias, J.M., Ormerod, S.J., 1999b. Comparing discriminant analysis, neural networks, and logistic regression for predicting species distributions: a case study with a Himalayan river bird. *Ecol. Model.* 120, 337–347.
- Nix, H.A., 1986. A biogeographic analysis of Australian elapid snakes. In: Longmore, R. (Ed.), *Atlas of Elapid Snakes of Australia*. Australian Government Publishing Service, Canberra, pp. 4–15.
- Pearson, R.G., Dawson, T.P., Berry, P.M., Harrison, P.A., 2002. SPECIES: a spatial evaluation of climate impact on the envelope of species. *Ecol. Model.* 154, 289–300.
- Peterson, A.T., Cohoon, K.C., 1999. Sensitivity of distributional prediction algorithms to geographic data completeness. *Ecol. Model.* 117, 159–164.
- Phillips, S.J., Anderson, R.P., Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. *Ecol. Model.* 190, 231–259.
- Soberón, J., Peterson, A.T., 2004. Biodiversity informatics: managing and applying primary biodiversity data. *Philos. Trans. R. Soc. Lond. B* 359, 689–698.
- Stockwell, D.R.B., 2006. Improving ecological niche models by data mining large environmental datasets for surrogate models. *Ecol. Model.* 192, 188–196.
- Stockwell, D.R.B., n.d. WhyWhere—answering the question “Where is it and why?” on a global scale.  
<http://biodi.sdsc.edu/Doc/WhyWhere/tutorial.html>.
- Thomas, C.D., Cameron, A., Green, R.E., Bakkenes, M., Beaumont, L.J., Collingham, Y.C., Erasmus, B.F.N., Ferreira de Siqueira, M., Grainger, A., Hannah, L., Hughes, L., Huntley, B., Van Jaarsveld, A.S., Midgely, G.E., Miles, L., Ortega-Huerta, M.A., Peterson, A.T., Phillips, O.L., Williams, S.E., 2004. Extinction risk from climate change. *Nature* 427, 145–148.
- USGS, 2001. HYDRO1k Elevation Derivative Database, U.S. Geological Survey, Washington, DC.  
<http://edcdaac.usgs.gov/gtopo30/hydro/>.
- Whittaker, R.H., Levin, S.A., Root, R.B., 1973. Niche, habitat, and ecotope. *Am. Nat.* 107, 321–388.
- Wiens, J.J., Graham, C.H., 2005. Niche conservatism: integrating evolution, ecology, and conservation biology. *Annu. Rev. Ecol. Syst.* 36, 519–539.