Sensitivity of distributional prediction algorithms to geographic data completeness

A. Townsend Peterson

Ecological Modelling

ELSEVIER

Ecological Modelling 117 (1999) 159–164

ECOLOGICAL MODELLING

Short communication

Sensitivity of distributional prediction algorithms to geographic data completeness

A. Townsend Peterson *, Kevin P. Cohoon

Natural History Museum, The University of Kansas, Dyche Hall, Lawrence, KS 66045-2454, USA

Received 6 March 1998; accepted 8 December 1998

Abstract

The sensitivity of one algorithm for prediction of geographic distributions of species from point data to depth of geographic information was tested for three species of North American birds. Test species were chosen to represent three distinct distributional patterns—western North America (Pygmy Nuthatch *Sitta pygmaea*), eastern North America (Barred Owl *Strix varia*), and the Great Plains in the central part of the continent (Lark Bunting *Calamospiza melanocorys*). Distributional predictions were made using the expert-system algorithm Genetic Algorithm for Role-set Prediction (GARP). Depth of geographic information was manipulated by rarifying the number of coverages on which predictions were based, from the full complement of eight down to one, using a combination of jackknifing and bootstrapping. In all three species, five of the eight coverages were necessary to arrive at the asymptotic maximum predictive efficiency and to avoid broad variance in resulting predictive efficiencies. Annual mean temperature was a critical variable, in some cases more important than inclusion of additional coverages, to producing accurate distributional predictions. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Distributional prediction; Geographic coverages; Sensitivity; GARP; Biodiversity

1. Introduction

An emerging field in landscape ecology and conservation biology is that of building ecological niche models to predict geographic distributions of species (e.g. Breininger et al., 1991; Dewey et al., 1991; Pereira and Itami, 1991; Stoms et al. 1992; Herr and Queen, 1993; Mladenoff et al., 1995; Sperduto and Congalton, 1996; Bian and West, 1997; Li et al., 1997). Especially challenging has been developing such models from point data, especially data such as museum specimen locality

* Corresponding author. Tel.: +1-785-8643926; fax: +1-785-8645335.

E-mail address: town@ukans.edu (A. Townsend Peterson)

PII: S0304-3800(99)00023-X

data, which are rarely accompanied by data indicating absence or abundance at sites. Several analytical approaches have been applied to these challenges, including logistic regression, discriminant function analysis, approaches based on distance measures, and parallelipiped set-based approaches.

Although broad comparisons of efficiency have not yet been carried out, our preliminary trials (L.G. Ball et al., in preparation) indicate that an especially promising algorithm is an expert-system approach based on a genetic algorithm, developed by David Stockwell (Stockwell and Noble, 1991; Stockwell, 1993). This algorithm, called Genetic Algorithm for Rule-set Prediction (GARP) produces a model of species, niches in geographic space based on heterogeneous rule-sets (see Methods). GARP appears especially well-suited to reducing error in predicted distributions, both in the form of omission of real distributional areas and inclusion of areas not holding actual populations.

One dimension of the challenge of distributional prediction that has not seen careful testing is that of the depth or completeness of geographic information necessary for accurate predictions. The purpose of the present contribution is to analyze—for a particular geographic and taxonomic challenge—the completeness of geographic data necessary for accurate distributional prediction, and if possible, what particular thematic coverages are especially important.

2. Methods

The three species employed in this study were chosen to represent distinct distributional areas in diverse ecosystems across North America. The Pygmy Nuthatch (*Sitta pygmaea*) is distributed in coniferous forests throughout western North America south to central Mexico. The Barred Owl (*Strix varia*) is distributed in deciduous and coniferous forests almost exclusively in eastern North America, and locally south to southern Mexico. Finally, the Lark Bunting (*Calamospiza melanocorys*) is distributed in prairie habitats across the northern Great Plains of interior North America, migrating south to the southem Great Plains and northern Mexico in the winter (only summer records were used for this species in the present study). Hence, these three species present distinct distributional patterns in very different habitats that likely encompass a significant portion of the diversity of habitats of North America.

Distributional data for the three species were drawn from the records of the 1966–1995 US Fish & Wildlife Service Breeding Bird Survey (BBS). These surveys are conducted in May and June only, and are thus focused only on breeding populations. Total numbers of unique geographic localities available for each species were: 180 for Pygmy Nuthatch, 1116 for Barred Owl, and 485 for Lark Bunting. Latitude-longitude coordinates accurate to the nearest 0.1 min were provided for each locality in the BBS data set. Half of the data set was randomly chosen for rule development ('training data'), and half for rule evaluation ('test data').

Geographic distributional predictions were developed based on algorithms designed to evaluate correlations between distributional occurrences and environmental characteristics. The Biodiversity Species Workshop facility developed by David Stockwell (http://biodi.sdsc.edu) provides an implementation of the GARP (Stockwell and Noble, 1991). The GARP algorithm works in an iterative process of rule selection, evaluation, testing, and incorporation or rejection: first, a method is chosen from a set of possibilities (logistic regression, BIOCLIM rules (Nix, 1986), etc.), applied to the data, and a rule developed. Then, predictive accuracy is evaluated based on 1250 points resampled from the test data set and 1250 points sampled randomly from the study region as a whole (the lower 48 United States), calculated as the sum of points actually present predicted as present and actually absent predicted as absent, divided by the total number of points in the map (Stockwell and Noble, 1991). The change in predictive accuracy from one iteration to the next is used to evaluate whether a particular rule should be incorporated into the model. The algorithm runs 1000 iterations, or until addition of rules has no appreciable effect on the accuracy measure. Complete details and documentation of the algorithm are available at http://biodi.sdsc.edu/.

Geograph
drawn fro
ject (http:
html); Lif
perature,
Vegetatior
tems. Eac
resolution

Geogra
follows. T
coverages
Then, eacl
(a jackkni
possible e
ages each;
sample-siz
systematic
tween, be
analyses w
quartets,
replaceme
times for
each speci
was one
seven and
through si

Results
presenting
sence of e
and overa
rations of
niques we
approach–
coverages
als of the
yielded ma
to this inst
cal techn
avoids the
approache
models, an
pendent v
fects (Che
1996). Thi
important
varying nu
ticular ana

summer
e present
t distinct
habitats
on of the

cies were
1995 US
d Survey
May and
breeding
eographic
: 180 for
, and 485
ordinates
vided for
f the data
velopment
tion ('test

were de-
evaluate
currences
Biodiver-
loped by
provides
kwell and
orks in an
ation, test-
: first, a
ties (logis-
986), etc.),
ed. Then,
l on 1250
and 1250
y region as
lculated as
edicted as
as absent,
n the map
ge in pre-
the next is
ule should
algorithm
of rules has
y measure.
of the al-
i.sdsc.edu/.

Geographic themes consisted of eight coverages drawn from the Global Ecosystems Database Project (http://www.ngdc.noaa.gov/seg/fliers/se-2006.html): Life Zones, Soil Class, Annual Mean Temperature, Annual Precipitation, Vegetation Class, Vegetation Type, Wetlands, and World Ecosystems. Each coverage was generalized to a pixel resolution of 50 km.

Geographic data density was manipulated as follows. The full complement of eight thematic coverages was run as a baseline for comparisons. Then, each coverage was eliminated systematically (a jackknife manipulation) to provide the eight possible examples of analyses with seven coverages each; similarly, at the opposite extreme of the sample-size spectrum, each coverage was included systematically in single-coverage analyses. In between, because numbers of combinations of analyses would have been prohibitive, pairs, trios, quartets, etc., of coverages were sampled with replacement (a bootstrapping manipulation) ten times for each number of coverages. Hence, for each species, the total set of analyses conducted was one with eight coverages, eight each with seven and one coverages, and ten each for two through six coverages, for a total of 67 analyses.

Results of analyses were summarized in tables presenting number of coverages, presence or absence of each coverage in a particular analysis, and overall predictive accuracy. Initial explorations of the data with multiple regression techniques were disturbingly sensitive to analytical approach—for example, inclusion of number of coverages as a covariate versus analysis of residuals of the accuracy–coverage number relationship yielded markedly different results. As a solution to this instability, we used a more robust statistical technique, hierarchical partitioning, that avoids the path-dependence of multiple regression approaches by calculating all possible regression models, and decomposing the effects of each independent variable into independent and joint effects (Chevan and Sutherland, 1991; MacNally, 1996). This technique identified coverages most important to accurate distributional prediction at varying numbers of coverages included in a particular analysis.

3. Results

Predictive accuracy was clearly related to the number of coverages, with low accuracy associated with few coverages, and highest accuracy generally achieved with the full or near-full complement of coverages. Fig. 1 illustrates the highly accurate prediction based on all eight coverages, compared with an example prediction based on a single coverage, for Pygmy Nuthatches. In each case, the decline in correspondence of the predicted distributional area with known distributional points in the rarified prediction is clear.

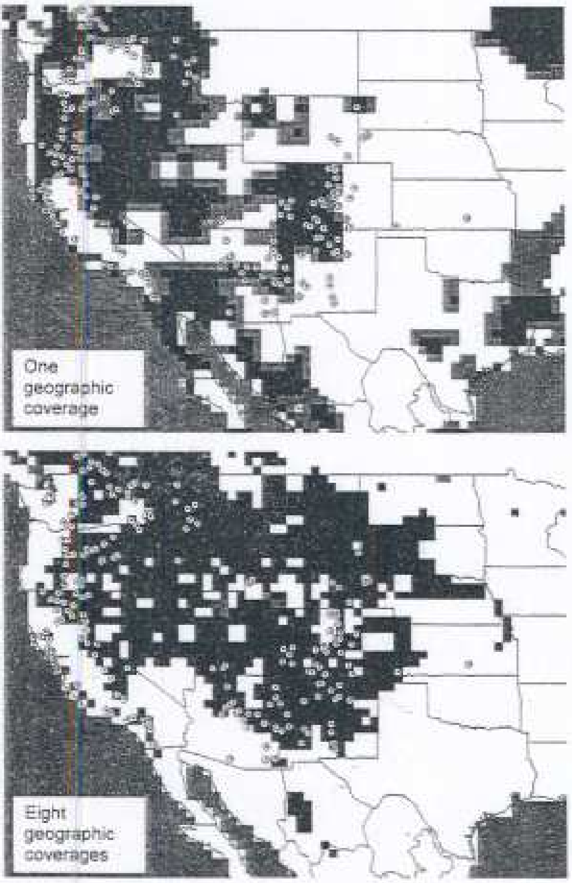


Fig. 1. Images resulting from predictive algorithms for low (soil class only, top) and high (all eight coverages, bottom) geographic infommation density for Pygmy Nuthatch. Black indicates predicted presence, white predicted absence, and grey areas not predicted; known occurrence points are shown as dotted circles.

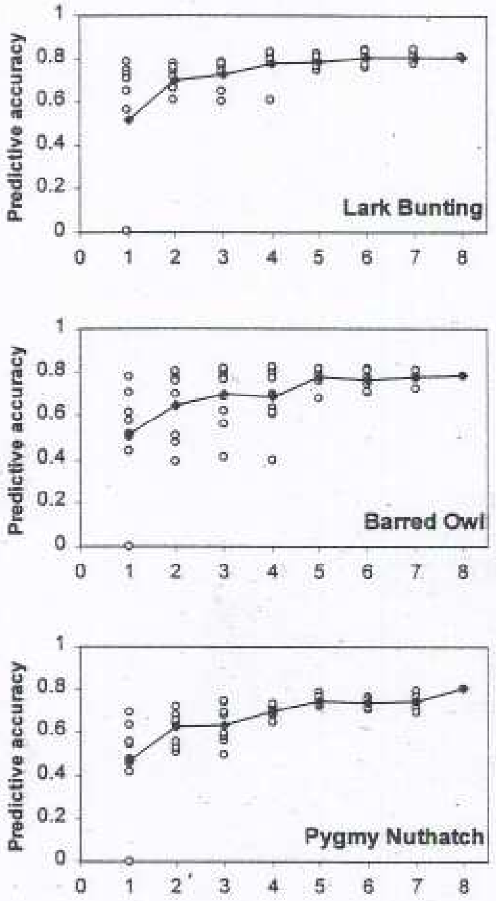


Fig. 2. Relationship of predictive accuracy to number of coverages included in the analysis. Shown as filled diamonds connected by lines are averages at each sample-size, and in open circles are raw data points.

In all three species analyzed, a sill was reached in the curve between four and five coverages—this point thus represents a number of coverages beyond which accuracy is not improved by adding more geographic data (Fig. 2). The spread of individual analyses around the mean also declined with increasing numbers of coverages, with variation tightly clustered around the mean value by five coverages as well. Hence, with five coverages, all analyses were extremely close to maximum accuracy.

The hierarchical partitioning analyses summarized the relationship between each independent variable and predictive accuracy (Fig. 3). In each case, the contribution of number of coverages to predictive accuracy was high. Among individual coverages, however, one was consistently closely related to high accuracy: annual mean temperature, which in all three species was the most important of the coverages. Interestingly, for two of the three species, presence of annual mean temperature in an analysis was more important than the number of coverages in determining overall predictive accuracy, indicating that that single coverage was especially relevant to the challenge of distributional prediction.

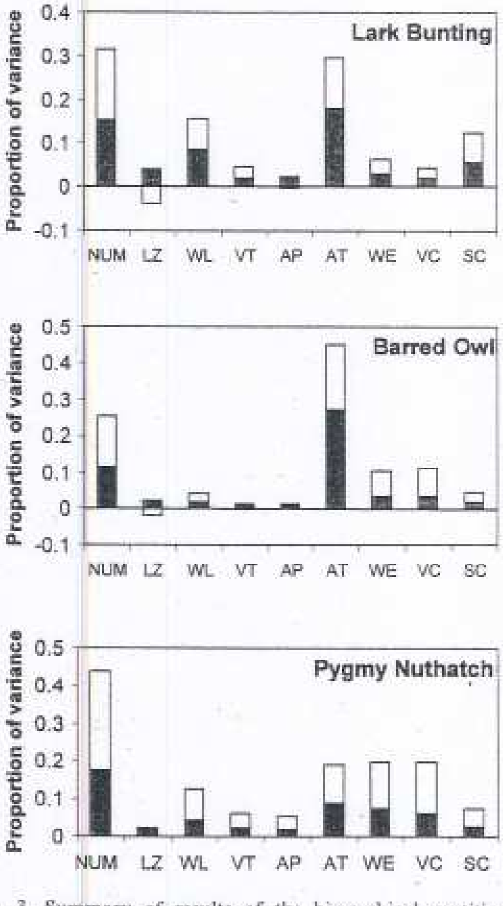


Fig. 3. Summary of results of the hierarchical partitioning analyses, indicating independent (gray) and joint (white) contributions of total number of coverages and presence or absence of each coverage to the total predictive accuracy. NUM, number of coverages; LZ, life zones; WL, wetlands; VT, vegetation types; AP, annual mean precipitation; AT, annual mean temperature; WE, world ecosystems; VC, vegetation classes; and SC, soil class.

4. Discussio

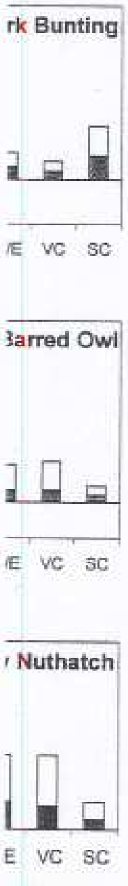
4.1. Methoc

This prel
geographic
distribution
simple in t
ages, one
temperatur
curate prec
ages and
resulting a
pends to s
and on th
ages availa
tests of th
geographic
lite imager

An imp
itive, resul
observed
tion. An c
BIOCLIM
graphic ir
point, ove
personal
equivalent
ses, in wh
reality is
decline wa
densities,
to such pr
at which
yond the

A furth
ing the ir
further pr
nature of
tization o
agery), na
be critical
entire spe
pling pre
was fairly
tion (67
alternativ
we sugge

f coverages to
ɔng individual
stently closely
ıean tempera-
vas the most
ingly, for two
annual mean
ore important
ı determining
ing that that
nt to the chal-



ıical partitioning
oint (white) con-
presence or ab-
accuracy. NUM,
, wetlands; VT,
ion; AT, annual
VC, vegetation

4. Discussion

4.1. Methodology

This preliminary study assessed the amount of geographic information necessary for accurate distributional predictions. The conclusion was simple in the case of North America; five coverages, one of which being that of annual mean temperature, were necessary for consistently accurate prediction. Inclusion of additional coverages and themes did not greatly affect the resulting accuracy. This result, of course, depends to some degree on the particular species, and on the array of geographic thematic coverages available in a particular study—planned are tests of the effects of addition of other types of geographic data, such as information from satellite imagery.

An important, though perhaps counterintuitive, result is that no decline in accuracy was observed with addition of geographic information. An observed quality of algorithms such as BIOCLIM (Nix, 1986) is that addition of geographic information to an analysis, at some point, over-restricts the prediction (D. Stockwell, personal communication). This effect is the equivalent of overprediction in regression analyses, in which points are overfit, and biological reality is lost. In the present case, because no decline was observed at highest geographic data densities, either the GARP algorithm is immune to such problems of overprediction, or the point at which such problems are encountered is beyond the dimensions assessed in this study.

A further methodological result is that of testing the importance of individual coverages for further prediction applications. Given the costly nature of generation of some coverages (e.g. digitization of paper maps, purchase of satellite imagery), narrowing selections of data themes can be critical. The approach used herein crossed the entire spectrum of sample-sizes, randomly sampling presences of particular coverages, which was fairly time-expensive in terms of computation (67 analyses per species). As a time-efficient alternative for identification of critical coverages, we suggest using the jackknife manipulation at 1 or $N-1$ of the N coverages as a useful approach to this question. In the present case, in the single coverage analyses, annual mean temperature had the highest predictive accuracy of the eight coverages in all three species (Lark Bunting: 0.736 compared with mean of 0.485 for remaining coverages; Barred Owl: 0.780 compared with mean of 0.479; Pygmy Nuthatch: 0.696 compared with mean of 0.441). The $N-1$ jackknife indicated annual mean temperature as the most important coverage (i.e. strongest negative effect on accuracy) in two of the three species (Lark Bunting: 0.768 compared with mean of 0.812 upon elimination of other coverages; Barred Owl: 0.726 compared with 0.785; Pygmy Nuthatch: 0.696 compared with 0.756). Hence, tendencies indicated by the jackknife approach agree well with the results of our multivariate statistical approach, suggesting that the jackknife is a viable, time-efficient manner to identify critical coverages for further study.

4.2. Biology

Although this study was primarily methodological in nature, its results have interesting biological implications. In all three species examined, annual mean temperature was the critical variable in determining quality of distributional predictions, suggesting that this dimension is particularly important in determining distributional limits in these taxa. This idea has been suggested in a number of studies developed from other methological perspectives (e.g. Root, 1988), and hence results encountered herein constitute an interesting independent confirmation.

More generally, these biological insights suggest additional applications of algorithms for distributional predictions. In addition to simply providing a prediction of geographic distributions of taxa, these approaches can reflect interesting dimensions of biology and natural history. Characteristics of component models, or in the present case, characteristics of the decline of the models as data are rarified, can indicate interesting dimensions of underlying biological factors. Further explorations of these applications and approaches are in preparation.

Acknowledgements

We thank Dr David Stockwell for his generous and capable assistance in all phases of this study. Thanks also to Dr Bruce Peterjohn, of the Breeding Bird Survey, for providing critical avian distributional data. Finally, this study was supported financially by the National Geographic Society, the National Science Foundation, the Kansas EPSCor, and the University of Kansas Honors Program.

References

Bian, L., West, E., 1997. GIS modeling of elk calving habitat in a prairie environment with statistics. Photogr. Eng. Remote Sens. 63, 161–167.

Breininger, D.R., Provancha, M.J., Smith, R.B., 1991. Mapping Florida Scrub Jay habitat for purposes of land-use management. Photogr. Eng. Remote Sens. 57, 1467–1474.

Chevan, A., Sutherland, M., 1991. Hierarchical partitioning. Am. Stat. 45, 90–96.

Dewey, S.A., Price, K.P., Ramsey, D., 1991. Satellite remote sensing to predict potential distribution of dyers woad (*Isatis tinctoria*). Weed Tech. 5, 479–484.

Herr, A.M., Queen, L.P., 1993. Crane habitat evaluation using GIS and remote sensing. Photogr. Eng. Remote Sens. 59, 1531–1538.

Li, W., Wang, Z., Ma, Z., Tang, H., 1997. A regression model for the spatial distribution of red-crown crane in Yancheng Biosphere Reserve, China. Ecol. Model. 103, 115–121.

MacNally, R., 1996. Hierarchical partitioning as an interpretative tool in multivariate inference. Aust. J. Ecol. 21, 224–228.

Mladenoff, D.J., Sickley, T.A., Haight, R.G., Wydeven, A.P., 1995. A regional landscape analysis and prediction of favorable grey wolf habitat in the northern great lakes region. Conserv. Biol. 9, 279–294.

Nix, H.A. 1986. A biogeographic analysis of Australian elapid snakes. In: Atlas of Australian elapid snakes. pp. 4–15. (Bureau of Flora and Fauna, Canberra).

Pereira, J.M., Itami, R.M., 1991. GIS-based habitat modeling using logistic multiple regression: A study of the Mt. Graham red squirrel. Photogr. Eng. Remote Sens. 57, 1475–1486.

Root, T., 1988. Atlas of wintering North American birds: An analysis of Christmas Bird Count data. University Chicago Press, Chicago.

Sperduto, M.B., Congalton, R.G., 1996. Predicting rare orchid habitat (small whorled pogonia) habitat using GIS. Photogr. Eng. Remote Sens. 62, 1269–1279.

Stockwell, D.R.B., 1993. LBS: Bayesian Learning System for rapid expert system development. Expert Syst. Appl. 6, 137–147.

Stockwell, D.R.B., Noble, I.R., 1991. Induction of sets of rules from animal distribution data: A robust and informative method of data analysis. Math. Comp. Simul. 32, 249–254.

Stoms, D.M., Davis, F.W., Cogan, C.B., 1992. Sensitivity of wildlife habitat models to uncertainties in GIS data. Photogr. Eng. Remote Sens. 58, 843–850.