A TEST OF A REGRESSION-TREE MODEL OF SPECIES DISTRIBUTION

Authors: O'Connor, Raymond J., and Wagner, Tansy L.

Source: The Auk, 121(2) : 604-609

Published By: American Ornithological Society

URL: https://doi.org/10.1642/0004-8038(2004)121[0604:ATOARM]2.0.CO;2

Commentary

The Auk 121(2):604–609, 2004

A TEST OF A REGRESSION-TREE MODEL OF SPECIES DISTRIBUTION

Raymond J. O'Connor[1] and Tansy L. Wagner

Department of Wildlife Ecology, University of Maine, 5755 Nutting Hall, Orono, Maine 04469, USA

Regression-tree analysis has increasingly found applications in ornithology, in modeling of behavior (Grubb and King 1991) and of species distributions (Rodenhouse et al. 1993; O'Connor et al. 1996, 1999; Hahn and O'Connor 2002). The method has proved particularly useful in modeling species occurrence over the large areas necessary for macroecology (Brown and Maurer 1989, Brown 1995, De'ath and Fabricius 2000, Iverson and Prasad 2002) and in accommodating the effects of anthropogenic stressors (Grubb and King 1991, Wickham et al. 1997, Allen and O'Connor 2000). O'Connor and Jones (1997) used the technique to estimate losses of bird populations to anthropogenic stressors in the United States and concluded that some 15% of the Breeding Bird Survey (BBS) routes in the conterminous United States had lost, on average, 17 or more breeding bird species. That would mean—given an average of 81 species on the typical BBS route (O'Connor et al. 1996)—that 21% of species in the bird communities in those areas were locally extinct, which is equivalent to a 3.1% average extinction rate over the conterminous United States! However, those conclusions depend on the validity of the underlying statistical model, a regression-tree analysis that involves subjective decisions that might have influenced the estimates. Here, we report the results of a test—with an entirely independent data set—that supports the ability of the O'Connor-Jones model to correctly predict local bird-species richness.

Regression-tree analysis (Breiman et al. 1984, Clark and Pregibon 1992) proceeds by recursive binary splitting of the original sample. In O'Connor and Jones's (1997) model, the predicted variable was species richness estimated from Breeding Bird Survey (BBS) data, and the predictor variables were a number of climate and land-cover variables (described in O'Connor et al. 1996). Each split in a regression-tree analysis is made by considering every available predictor (in turn) as a potential splitting variable with which to divide the sample into two subsets; data are ordered according to the values of the predictor under consideration, and each value is considered as a possible binary splitting value. At each split point, the choice is made that yields the maximum difference between the dependent-variable values in the two subsets. As more splits are incorporated, an inverted "tree" of proliferating binary divisions develops. Because some branches cease splitting before others, the final tree contains some finite number of end nodes ("leaves") within which all cases simultaneously satisfy the conditions at all of the split points in the branch from that leaf back to the root node. Each end node has a single predicted value of the dependent variable, in essence the value predicted under the conditions specified by that chain of conditions. Because the two subsets at each split point may split on different independent variables, this process allows detection of contingent effects and of interactions, without those having to be specified *a priori*. That is a major advantage over multiple linear regression, which (unless expressly configured with interaction terms) requires constancy of relationships over the entire domain of the sample, a condition unlikely to hold over regional and continental extent.

On the other hand, regression trees are prone to overfitting: in an extreme instance, the recursion could proceed until every end node (terminal set of cases) contained either a single case or multiple cases with a common value of the response variable. Breiman et al. (1984) showed that the technique's propensity to overfit is best

[1]E-mail: oconnor@umenfa.maine.edu

handled by fitting an excessively large tree and then "pruning" it back to an optimal size. Pruning is done by cross-validation with a cost-complexity function that penalizes predictions made with an excessively large tree. However, choice of penalty and basis of cross-validation are open to subjective decision by the analyst; there is no consensus among statisticians as to what constitutes the ideal approach to cross-validation (Miller 1994, Ribic and Miller 1998, J. Sifneos et al. unpubl. data). Hence, "optimal tree size" contains a degree of subjectivity, and one can always argue for a smaller or larger tree, relative to that found by cross-validation. Moreover, because cross-validation is a subset sampling technique, repetition of the cross-validation yields a slightly different cross-validation curve, which can occasionally yield a notably different estimate for the optimum size of the final tree. That different estimate may result in different predictions for a given location when used, for example, in modeling bird distributions and stressor effects.

One way to assess the validity of a regression-tree model is to test its predictions against observations from an entirely independent data set. Models developed by O'Connor et al. (1996) and O'Connor and Jones (1997) used species richness estimates obtained from the national BBS, a roadside-count scheme based on 3-min, 40-km radius counts of all bird species detected by a volunteer observer at each of 50 stops at 0.8-km intervals along a survey route 40 km long. Route locations are chosen in a stratified random sampling with physiographic regions as strata. Because the breeding-species tally for a route typically increases over time as a result of limited census efficiency, O'Connor et al. (1996) used data only from routes with high-quality surveys for 7 or more years between 1981 and 1990; tallies from 7 to 9 years of surveys were adjusted to a 10-year total (shown from data for long-running sites to be a good estimate of long-term species richness on the route). Those tallies were related, by means of regression-tree analysis, to land-cover and climate data in their home cells on a hexagonal grid—the U.S. Environmental Protection Agency's Environmental Monitoring and Assessment Program (EMAP) grid (White et al. 1992), which has average hexagon area of ~635 km^2 and centroid-to-centroid distance of ~27 km. Species richness estimates and predictor data were available for 1,198 hexagons and yielded a regression tree with 33 end nodes. Because the independent (climate, land cover, etc.) data were available for all 12,600 hexagons in the conterminous United States, and because a variety of checks confirm that the BBS routes used were representative of national land-cover and climate distribution (J. J. Lawler and R. J. O'Connor unpubl. data), every hexagon in the country could be assigned one of the corresponding 33 values of species richness as a predicted value. Consequently, results of an independent survey assessing bird species richness at any location in the conterminous United States can be compared against the O'Connor and Jones (1997) prediction for the hexagon in which the survey was conducted.

Here, we used estimates of species richness on Breeding Bird Census (BBC) plots as an independent source of test data. Data for 67 BBC plots surveyed in both 1989 and 1990 (Van Velzen 1990, Anonymous 1991) were obtained from the U.S. Geological Service, Biological Resources Division, at their Patuxent Wildlife Research Center's web page (see Acknowledgments). The hexagon in which each BBC plot was located was determined with a point-in-polygon ARCINFO routine, and the empirical species richness of the plot was compared against the value predicted for that hexagon by the regression tree. The BBC uses a "spot mapping" method to identify territories within a study plot by plotting registrations from multiple visits onto a map. Such mapping greatly reduces the chance of overlooking breeding species actually present (O'Connor 1981) but does not completely eliminate it. In addition, some species may occur regularly on a plot but not occur there every year. We therefore adjusted the single-year (1990) species tally for each plot by using the SPECRICH2 program of Boulinier et al. (1998). That program treats multiple censuses at a site as analogous to mark–recapture data from a population, using differences in species lists between years to estimate census efficiency, then using that estimate to correct to a true species tally. We used data from 1989 and 1990 to obtain an adjusted species total for each BBC plot; if the assumptions of Boulinier et al. (1998) are correct, we thus avoided the problem of differences in time-span of observations between the two schemes. All other statistical calculations were done using the SYSTAT statistical package

(version 8.0; SPSS Inc., Chicago, Illinois). For all calculations involving regression trees, we used the S-PLUS statistical analysis software (Insightful Corporation, Seattle, Washington).

The regression-tree model of O'Connor and Jones (1997) gave 33 distinct predicted values (one for each of the various combinations of environmental and land-use constraints found within the tree), covering all 12,600 hexagons in the country. However, BBC plots are distributed opportunistically rather than with a representative sampling, and only 19 of the 33 prediction zones contained BBC plots. In contrast to the spatially extensive BBS routes used to characterize the avifauna of the surrounding hexagon, the BBC plots were very small, averaging only 15.4 ha in our 1990 sample. It is well known from island biogeography theory (MacArthur and Wilson 1967) that species richness increases with area sampled, both for statistical reasons and because of greater heterogeneity of habitat in a larger area. Pairing the BBS and BBC estimates of species richness for each hexagon yielded mean values of 85.7 ± 9.7 (SD) and 29.3 ± 13.6 species, respectively, with a within-hexagon (paired) difference of 56.4 ± 17.0 species. Thus, the larger areas of the BBS routes resulted in much larger species-richness estimates. That precluded comparison of absolute species-richness values as a test of the regression-tree predictions but left open the use of correlation and regression analysis. That is, even though the absolute number of species could not be predicted for a BBC plot because of the difference in area surveyed by BBS and BBC, two BBC plots of equal size that differed by (say) 50% in species tally should be located in hexagons for which the two predictions also differed by 50%. In that way, species tallies in hexagon and BBC plot should be correlated across locations. (The alternative—construction of a species-area curve from the BBC species tally and plot-area data—would have involved major [at least 40-fold] extrapolation from a small range of plot areas to the area surveyed on a BBS route.)

Bird species richness differed significantly between BBC plots of different habitat types (Table 1), though we were unable to discriminate any single habitat as significantly different in richness from the others (general linear model of richness against the five habitat categories, percent of variation explained (r^2) = 12.4%, NS).

Table 1. Mean species richness by habitat within the Breeding Bird Census plots (Kruskal-Wallis nonparametric ANOVA = 30.8, $P < 0.001$).

Type of land cover	Number of plots	Mean number of species	Standard deviation
Deciduous forest	25	37.3	10.0
Mixed forest	10	33.8	16.0
Wetland	3	32.0	13.0
Coniferous forest	10	27.4	11.2
Open habitats	18	16.2	7.0

Hence, habitats could not validly be pooled for analysis, and evaluation of the regression-tree prediction against the BBC plot tally needed to consider the plot habitat in relation to hexagon land-cover. That is, the bird community of a small woodlot or wetland used as a BBC plot would not be representative of the bird species present in a surrounding hexagon dominated by farmland, but a wooded BBC plot should be representative of a forested hexagon. In fact, species-richness estimates across all BBC plots were uncorrelated with hexagon-specific predictions if habitat was ignored (Pearson $r = –0.046$, NS). However, when only the 25 BBC deciduous plots embedded in hexagons with at least 40% of pixels classified as deciduous forest were considered, the correlation increased (Pearson $r = 0.46$, $P < 0.02$). The regression equation obtained (and interpreted below) was

$$y = 69.68\ (\pm 5.18) + 0.33\ (\pm 0.13)\ x$$

where y is BBC species tally, x is predicted species-richness for the embedding hexagon, and standard errors of the coefficients are shown in parentheses. If the level of matching between BBC plot and hexagon was increased to 50% (i.e. at least half of the hexagon was forested), only nine plots were available but the correlation rose still further ($r = 0.70$, $P < 0.05$). The regression equation obtained was

$$y = 62.00\ (\pm 9.05) + 0.54\ (\pm 0.21)\ x$$

Note that the intercept in each of these equations was large relative to the mean species tally (Table 1), again reflecting the large area of the hexagon, relative to the BBC plot area. Note also that the slope in this and in the previous regression is less than unity, implying that a gain of one species in moving to a more speciose BBC plot was accompanied by a gain of two or three species in the embedding hexagon. That is

as expected within hexagons that are far larger than BBC plots. Linear regression is not strictly appropriate here, because the *x*-axis values are themselves subject to error by virtue of being predictions, whereas linear regression strictly requires zero error in the independent variable (Zar 1998). However, because their relative errors (based on rather large samples) are very much less than those in the BBC estimates, the error involved in using regression must be quite small.

The analysis here provides a limited but critical confirmation that regression-tree predictions of the continental distribution of species richness on the basis of climate and remotely sensed land-use are reliable. Although the BBC data were able to support analysis only of forested hexagons, that is the habitat for which independent confirmation of regression-tree predictions is most valuable. Forest fragmentation was a major stressor in O'Connor and Jones's (1997) model of population losses. Considerable concern has focused on the fragmentation of forests in the United States, because Neotropical migrants nesting in small forest patches are disproportionately affected by predation (Wilcove 1985, Robinson et al. 1995). The systematic decline in abundance of many of those species in recent years has been attributed to increase in such fragmentation (Robbins et al. 1989), a conclusion supported by regional reversal of decline where re-afforestation has progressed (Askins 1993). Moreover, extension of the O'Connor and Jones (1997) model yields estimates of individual species losses to fragmentation as high as 36% of the national population (R. J. O'Connor and L. Hayes unpubl. data). Although the models were subjected to stringent checking within the regression-tree algorithms, lack of consensus among statisticians as to the optimal cross-validation criterion to use (Miller 1994, Ribic and Miller 1998) and the subjectivity of the penalty used in the cost-complexity function of regression-tree analysis (Clark and Pregibon 1992) could call into question the validity of those estimates of bird population losses. Therefore, agreement between the regression-tree predictions and independent data from the BBC for forested areas powerfully supports the validity of regression-tree predictions within ornithological contexts. It is worth remarking that the correlation was obtained despite very large differences in relative area of spatial units between BBS and BBC analyses, despite their very different census methodologies, and particularly despite the typical BBC plot occupying but a tiny fraction of the hexagon.

Because the cross-validation process is indifferent to the identity of variables in the tree it optimizes, confirmation of the predictions for forested nodes by our test with independent data argues that the predictions for nonforested nodes are also likely to be valid. Thus, the increasing application of regression-tree analysis to ecological problems is probably sufficiently protected against the risk of overfitting in the resultant models by the cross-validation procedure recommended by Breiman et al. (1984). The combination of regression-tree analysis and remotely sensed land-cover data for forest thus appears to be a powerful tool for expanding the approaches of Weber and Theberge (1977) and Flather and Sauer (1996) in that it frees such studies from the need for *a priori* specification of regional characteristics. But that freedom comes at a price. It enables identification of a restricted subset of model relationships that can have considerable predictive power within their domain of application but that have weak statistical inference to a larger universe. Had the sample been drawn from a restricted, regional domain, inference of avian environmental correlates on a continental scale would require considerable caution. In the context of species-distribution modeling across the conterminous United States, however, the continental scale of our sample already defines the domain of interest, the weakness of any further statistical inference is somewhat moot, and we have strong predictive power where it is needed. In particular, regression-tree analysis coupled with remotely sensed land-cover data allows identification of spatial occurrence of stressors (such as habitat fragmentation) on a continental scale and permits modeling of their effects on bird species.

Acknowledgments

We thank W. Glanz, W. Halteman, C. Slavin, and S. Cohn for comments on an initial draft; and B. W. Brook and K. G. Beal for helpful comments on the manuscript. We appreciate the assistance of J. Bartlett in providing the ARCINFO analysis. The research reported here was supported by National Science Foundation (NSF) Award 9711623 to R.J.O'C. and D. M. Mageean. Data for the Breeding Bird Census plots are available at http://www.im.nbs.gov/birds/bbc.html.

Literature Cited

Allen, A. P., and R. J. O'Connor. 2000. Interacting effects of land use and other factors on regional bird distributions. Journal of Biogeography 27: 889–900.

Anonymous. 1991. Fifty-fourth breeding bird census. American Birds 45:58–60.

Askins, R. A. 1993. Population trends in grassland, shrubland, and forest birds in eastern North America. Current Ornithology 11:1–34.

Boulinier, T., J. D. Nichols, J. R. Sauer, J. E. Hines, and K. H. Pollock. 1998. Estimating species richness: The importance of heterogeneity in species detectability. Ecology 79:1018–1028.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. 1984. Classification and Regression Trees. Wadsworth, Belmont, California.

Brown, J. H. 1995. Macroecology. University of Chicago Press, Chicago.

Brown, J. H., and B. A. Maurer. 1989. Macroecology: The division of food and space among species on continents. Science 243: 1145–1150.

Clark, L. A., and D. Pregibon. 1992. Tree-based models. Pages 377–419 *in* Statistical Models in S (J. M. Chambers and T. J. Hastie, Eds.). Wadsworth and Brooks, Pacific Grove, California.

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

Flather, C.H., and J. R. Sauer 1996. Using landscape ecology to test hypotheses about large-scale abundance patterns in migratory birds. Ecology 77:28–35.

Grubb, T. G., and R. M. King. 1991. Assessing human disturbance of breeding Bald Eagles with classification tree models. Journal of Wildlife Management 55:500–511.

Hahn, D. C., and R. J. O'Connor. 2002. Contrasting determinants of the abundance of an invasive species in its ancestral and colonized ranges. Pages 219–228 *in* Predicting Species Occurrences: Issues of Scale and Accuracy (J. M. Scott, P. J. Heglund, M. L. Morrison, J. B. Haufler, M. G. Raphael, W. A. Wall, and F. B. Samson, Eds.). Island Press, Washington, D.C.

Iverson, L. R., and A. M. Prasad. 2002. Potential tree species shifts with five climate change scenarios in the eastern United States. Forest Ecology and Management 155:205–222.

MacArthur, R. H., and E. O. Wilson. 1967. The Theory of Island Biogeography. Princeton University Press, Princeton, New Jersey.

Miller, T. W. 1994. Model selection in tree-structured regression. Pages 158–163 *in* 1994 Proceedings of the Statistical Computing Section. American Statistical Association, Alexandria, Virginia.

O'Connor, R. J. 1981. The influence of observer and analyst efficiency in mapping method censuses. Studies in Avian Biology 6:372–376.

O'Connor, R. J., R. B. Boone, M. T. Jones, and T. B. Lauber. 1999. Linking continental climate and land use patterns with grassland bird distribution in the conterminous United States. Studies in Avian Biology 19:45–59.

O'Connor, R. J., and M. T. Jones. 1997. Using hierarchical models to assess the ecological health of the nation. Transactions of the 62nd North American Wildlife and Natural Resources Conference 62:501–508.

O'Connor, R. J., M. T. Jones, D. White, C. T. Hunsaker, T. Loveland, B. Jones, and E. Preston. 1996. Spatial partitioning of the environmental correlates of avian biodiversity in the lower United States. Biodiversity Letters 3: 97–110.

Ribic, C.A., and T. W. Miller. 1998. Evaluation of alternative model selection criteria in the analysis of unimodal response curves using CART. Journal of Applied Statistics 25:685–698.

Robbins, C. S., J. R. Sauer, R. S. Greenberg, and S. Droege. 1989. Population declines in North American birds that migrate to the Neotropics. Proceedings of the National Academy of Sciences USA 86:7658–7662.

Robinson, S. K., F. R. Thompson III, T. M. Donovan, D. R. Whitehead, and J. Faaborg. 1995. Regional forest fragmentation and the nesting success of migratory birds. Science 67:1987–1990.

Rodenhouse, N. L., L. B. Best, R. J. O'Connor, and E. K. Bollinger. 1993. Effects of temperate agriculture on Neotropical migrant landbirds. Pages 280–295 *in* Status and Management of Neotropical Migratory Birds (D. M. Finch and P. W. Stangel, Eds.). U.S. Department of Agriculture, Forest Service, General Technical Report RM-229.

Van Velzen, W. T. 1990. Fifty-third breeding bird census. American Birds 44:170–172.

Weber, W. C., and J. B. Theberge. 1977. Breeding bird survey counts as related to habitat and date. Wilson Bulletin 89:543–561.

White, D., A. J. Kimmerling, and W. S. Overton. 1992. Cartographic and geometric components of a global sampling design for environmental monitoring. Cartography and Geographic Information Systems 19:5–22.

Wickham, J. D., J. Wu, and D. F. Bradford. 1997. A conceptual framework for selecting and analyzing stressor data to study species richness at large spatial scales. Environmental Management 21:247–257.

Wilcove, D. S. 1985. Nest predation in forest tracts and the decline of migratory songbirds. Ecology 66:1211–1214.

Zar, J. H. 1998. Biostatistical Analysis, 4th ed. Prentice Hall, Upper Saddle River, New Jersey.

Received 12 March 2003, accepted 14 January 2004.
Associate Editor: N. S. Sodhi