Ecological Modelling 227 (2012) 29–33

Contents lists available at SciVerse ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel

Sample sizes and model comparison metrics for species distribution models

B.B. Hanberry[a,*], H.S. He[a], D.C. Dey[b]

[a] University of Missouri, 203 Natural Resources Building, Columbia, MO 65211, USA
[b] USDA Forest Service, Northern Research Station, University of Missouri, 202 Natural Resources Building, Columbia, MO 65211 USA

ARTICLE INFO

Article history:
Received 22 April 2011
Accepted 5 December 2011
Available online 9 January 2012

Keywords:
Correlation
Forest inventory and analysis
Intraclass correlation coefficient
Kappa

ABSTRACT

Species distribution models use small samples to produce continuous distribution maps. The question of how small a sample can be to produce an accurate model generally has been answered based on comparisons to maximum sample sizes of 200 observations or fewer. In addition, model comparisons often are made with the kappa statistic, which has become controversial. Therefore, we used sample sizes ranging from 30 to 2500 individuals to model 16 tree species or species groups in Minnesota's Laurentian Mixed Forest. We compared all smaller sample sizes to models for 2500 records and then 1000 records using Cohen's kappa, Pearson's *r*, Cronbach's alpha, and two intraclass correlation coefficients. We then began confirmation of our findings by repeating the process using a smaller extent in a different area, a portion of Missouri's Central Hardwoods. Although there are disadvantages to using the kappa statistic and intraclass correlation coefficients, due to conversion to categories or computation limitations respectively, the model comparison metrics produced similar results. Comparison values depend on the maximum sample size, and at sample sizes roughly around 10–20% of the maximum sample size, values will begin to decrease more rapidly. Models may not be very accurate below a sample size of 200, for our study areas, extents, and grains. Nonetheless, models based on small sample sizes still may provide information for rare species. We recommend using the full sample available for modeling, after using a partial sample for accuracy assessment. Future research is needed to confirm our findings for different areas, extents, grains, and species.

1. Introduction

Species distribution models use small samples from point locations to predict species occurrence probability for a continuous spatial extent. The accuracy of species distribution models may vary by species, statistical method, explanatory variables, and study extent among other factors, although the unique ecological characteristics of species, and consequent diverse distribution patterns, may explain the greatest variance (Guisan et al., 2007; Syphard and Franklin, 2010). Nevertheless, the accuracy of species distribution models also depends on both sample size and the method for comparison of models.

Sample size is an important consideration for modeling accuracy, particularly for rare species where there are few samples. Small sample sizes that produce inaccurate models may provide some information, but uncertainties associated with these models are high. Although there is much research that compares ever-changing statistical methods, establishing an appropriate sample size as a base for appropriate comparisons has not been common. Studies that have focused on sample size also have used small maximum sample sizes for comparison (e.g. Stockwell and Peterson, 2002; Kadmon et al., 2003; Hernandez et al., 2006; Wisz et al., 2008). Even though 100–200 individuals may be more records than available, models for 100 individuals may not be the best standard.

To measure the agreement among species distribution maps, Cohen's kappa commonly is used (Cohen, 1960). Although the kappa statistic is meant to account for chance agreement, the definition of chance is uncertain (Vaughan and Omerod, 2005). The kappa statistic also behaves paradoxically due to prevalence (number of present cases) and location of species distributions (McPherson et al., 2004; Jiménez-Valverde et al., 2008). Therefore, other metrics to measure accuracy may be preferable.

One option for measurement of model agreement is the familiar interclass correlation coefficient. Interclass correlation coefficients, such as the commonly used Pearson's *r* or more rare Cronbach's alpha, are used to correlate different variables (such as height and weight), and consequently, different variance. For Pearson's *r* and Cronbach's alpha, the magnitude of difference between variables does not matter. For example, pairwise values of 0.1 and 0.8, 0.2 and 0.9, and 0.3 and 1.0, would be correlated and yet are very different values for species distribution maps.

Another option is intraclass correlation coefficients, which measure the relationship between the same variable from different

* Corresponding author. Tel.: +1 573 875 5341x230; fax: +573 882 1977.
E-mail address: hanberryb@missouri.edu (B.B. Hanberry).

0304-3800/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2011.12.001

sources (Shrout and Fleiss, 1979). This option may be more suited for model comparison, where the same variable (predicted probability) is compared and variation arises from methodological choices, for example, when comparing predicted probabilities from different statistical methods (i.e. different sources). Two intraclass correlation coefficient values may be calculated: an absolute agreement metric and a consistency metric. The absolute agreement metric incorporates the different sources in the comparison, whereas the consistency metric excludes variance from the source (i.e. different methods), which is desirable when the magnitude of difference is irrelevant. The consistency metric is similar to Pearson's correlation, but based on an additive rather than a linear transformation to relate lower and higher values (McGraw and Wong, 1996).

Due to the importance of sample size and the choice of accuracy assessment metric, we had two objectives. First, we used much larger sample sizes than previous studies to evaluate adequate minimum sample sizes. We examined sample sizes ranging from 30 to 2500 individuals of 16 tree species or species groups in a roughly 5 million hectare area of Minnesota. Secondly, in order to compare species distribution models developed under various sample sizes, we needed to determine if the current option, the kappa statistic, was as reliable as alternative options. We compared all smaller sample sizes to models based on 2500 records and then 1000 records using Pearson's correlation, Cronbach's alpha, Cohen's kappa, and two intraclass correlation coefficients. We then evaluated the comparison metrics for differences to identify the strengths and weaknesses of each metrics. To strengthen our findings, we repeated the process for a smaller extent in the Central Hardwoods of Missouri. Our work will provide guidance in selection of appropriate sample size and metrics for species distribution models.

2. Methods

2.1. Study area

The primary study area covers about half of the 9.3 million hectare Laurentian Mixed Forest province in northeastern Minnesota (Fig. 1; National Hierarchical Framework of Ecological Units; ECOMAP, 1993). In the Laurentian Mixed Forest province, landforms (e.g. moraines and wetlands) were created by glaciers (Albert, 1995). Annual precipitation increases from about 55 cm in the west to 80 cm in the east and long, cold winters prevail (mean annual temperature about 2 °C).

2.2. Tree surveys

The USDA Forest Service Forest Inventory and Analysis (FIA) surveys fixed plots (each composed of four subplots that are a total of 0.065 ha) during a five year cycle. The latest complete cycle was during 2004–2008 for Minnesota's Laurentian Mixed Forest (Fig. 1). The USDA Forest Service joined our predictor variables to plots (in a table but based on accurate spatial locations) for modeling and prediction because the available FIA plot locations are fuzzed (i.e. location moved) and swapped to protect landowner privacy.

We selected tree species that had at least 2500 individuals. The species were American Basswood (*Tilia Americana*), balsam fir (*Abies balsamea*), balsam poplar (*Populus balsamifera*), black ash (*Fraxinus nigra*); black spruce (*Picea mariana*), bur oak (*Quercus macrocarpa*), jack pine (*Pinus banksiana*), northern white cedar (*Thuja occidentalis*), paper birch (*Betula papyrifera*), red maple (*Acer rubrum*), red pine (*Pinus resinosa*), sugar maple (*A. saccharum*), tamarack (*Larix laricina*), and quaking aspen (*Populus tremuloides*). We also created two mixed species groups by genus, aspens (*Populus tremuloides*, *P. balsamifera*) and maples (*Acer rubrum*, *A. saccharum*).

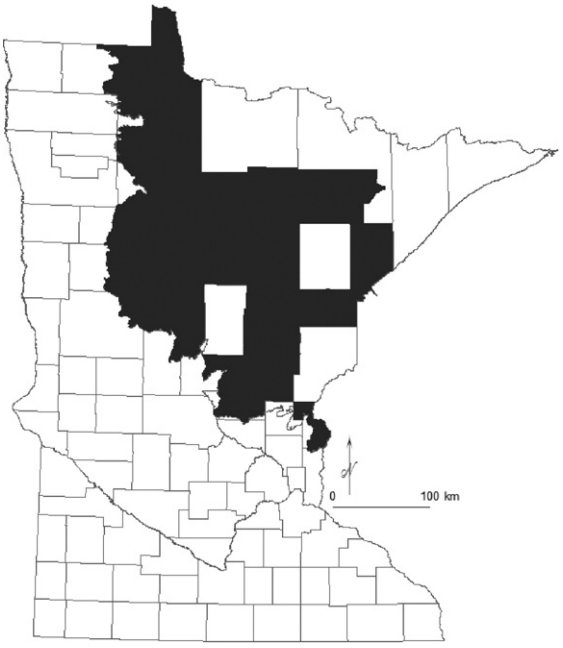

Fig. 1. Primary study area (shaded black), about 5 million ha in the Laurentian Mixed Forest of Minnesota.

2.3. Spatial units and environmental variables

Our spatial units were Soil Survey Geographic (SSURGO) Database (Natural Resources Conservation Service; http://soildatamart.nrcs.usda.gov) polygons. Soil surveys have not been completed in Cook, Crow Wing, Isanti, Koochiching, Lake, Pine, and St. Louis counties, leaving a study extent of about 4,895,238 ha (Fig. 1). After removal of polygons that were water or otherwise miscellaneous areas e.g. mines, pits, dumps), there were 310,000 soil polygons.

We used sixteen predictor variables that are important for tree presence. For soil variables, we determined values based on polygons with similar characteristics by county (map units; 2364 map units total). Soil variables were (1) drainage class (very poorly drained to excessively drained), (2) hydric soil presence class, (3) water holding capacity (cm/cm), (4) pH, (5) organic matter (%), (6) clay (%), and (7) sand (%). We intersected two more categorical variables to each soil polygon: (8) ecological subsection, which is an ecological classification (ECOMAP, 1993), and (9) bedrock geology. From a 30 m DEM (digital elevation model), we determined mean values of terrain variables by a unique unit of map unit, land type association (an ecological classification), and bedrock geology, which contained spatially distinct soil polygons that averaged about 210 ha, and became our unit for predicted probabilities. Terrain variables were (10) elevation (m), (11) slope (%), (12) transformed aspect (1 + sin(aspect/180/3.14 + 0.79; Beers et al., 1966) (13) solar radiation (0700–1900 in 4 h intervals on summer solstice for re-sampled 60 m DEM), (14) topographic roughness (Sappington et al., 2007), (15) wetness convergence, and (16) topographic position index (T. Dilts; http://arcscripts.esri.com).

2.4. Sample sizes and statistical analysis

We randomly selected 2500, 1250, 1000, 5000, 200, 100, 50 and 30 polygons for each tree species or tree species group for modeling. We reserved the rest of the present samples for accuracy