Estimating tree species richness from forest inventory plot data

Ronald E. McRoberts and Dacia M. Meneguzzo
North Central Research Station, USDA Forest Service
Saint Paul, Minnesota 55108 USA

Abstract

Montréal Process Criterion 1, Conservation of Biological Diversity, expresses species diversity in terms of number of forest dependent species. Species richness, defined as the total number of species present, is a common metric for analyzing species diversity. Several model-based and non-parametric techniques have been developed to estimate tree species richness from sample data. Four sample-based approaches to estimating tree species richness were compared using data obtained from forest inventory plots in Minnesota, USA. The results indicate that approaches based on a 3-parameter exponential model and the non-parametric jackknife were superior to approaches based on the Michaelis-Menten model and the non-parametric bootstrap. Estimates of tree species richness were greater for forested areas with low housing density than for areas that were uninhabited.

Introduction

Of the international forest sustainability initiatives, the Montréal Process (1998) is geographically the largest, involving 12 countries on five continents and accounting for 90 percent of the world's temperate and boreal forests. The Montréal Process prescribes a scientifically rigorous set of criteria and indicators that have been accepted for estimating the status and trends of the condition of forested ecosystems. A criterion is a category of conditions or processes and is characterized by a set of measurable quantitative or qualitative variables called indicators which, when observed over time, demonstrate trends. The Montréal Process includes seven criteria (McRoberts et al. 2004) of which Criterion 1, Conservation of Biological Diversity, focuses on the maintenance of ecosystem, species, and genetic diversity. Of the indicators associated with Criterion 1, one of the most intuitive is Indicator 6, Number of Forest Dependent Species. When the emphasis is on the number of tree species, this indicator is characterized as tree species richness. A primary research interest is to determine the factors that affect tree species richness; e.g., do increases in housing density or forest fragmentation affect tree species richness?

Because species richness relates only to the presence or absence of species, regardless of distribution or abundance, estimation of species richness is difficult apart from a complete census. However, complete tree censuses are not practical for the naturally regenerated, mixed species, uneven aged forests that occur in much of the world. As a result, estimation of tree species richness must depend on sample data.

However, although tree species richness is an intuitive measure, it is difficult to estimate using sample data because there is no assurance that all species have been observed in the sample, particularly rare or highly clustered species.

The objectives of the study were twofold: (1) to compare two model-based and two non-parametric approaches for estimating tree species richness from forest inventory plot data, and (2) to compare estimates of tree species richness for uninhabited forest land and forest land with low levels of housing density.

Data

Forest inventory data is widely recognized as an excellent source of information for estimating the status and trends of forests in the context of the Montréal Process or the Ministerial Conference for the Protection of the Forests of Europe (McRoberts et al. 2004). The national forest inventory of the United States of America (USA) is conducted by the Forest Inventory and Analysis (FIA) program of the Forest Service, U.S. Department of Agriculture. The program collects and analyzes inventory data and reports on the status and trends of the nation's forests. The national FIA plot consists of four 7.32-m (24-ft) radius circular subplots which are configured as a central subplot and three peripheral subplots with centers located at 36.58 m (120 ft) and azimuths of 0°, 120°, and 240° from the center of the central subplot. All trees on these plots with diameters at breast height of at least 12.5 cm (5.0 in) were measured and the species identifications were recorded.

The national FIA sampling design is based on an array of 2,400-ha (6,000-ac) hexagons that tessellate the nation. This array features at least one permanent plot randomly located in each hexagon and is considered to produce an equal probability sample. The sample was systematically divided into five interpenetrating, non-overlapping panels. Panels are selected for measurement on approximate 5-, 7-, or 10-year rotating bases, depending on the region of the country, and measurement of all accessible plots in one panel is completed before measurement of plots in a subsequent panel is initiated.

The study area was in Minnesota, USA, and consisted of the geographic intersection of Bailey's ecoprovince 212 (Bailey 1995) and Mapping Zone 41 of the Multiresolution Land Characterization Consortium (Loveland and Shaw 1996) (Figure 1). Forests in the study area are generally naturally regenerated, mixed species, and uneven aged.

Using the Wildland-Urban Interface (WUI) map constructed by Radeloff et al. (2005), the study area was partitioned into four categories corresponding to combinations of levels of vegetative cover and housing density. The WUI map was constructed using information from two sources: vegetative cover information was obtained from the National Land Cover Dataset (Vogelmann et al. 2001), a 21-class land cover classification based on nominal 1991 Landsat Thematic Mapper satellite imagery and other ancillary data, and housing density information was obtained from the 2000 U.S. Census. The focus of this study was on comparing estimates of tree species richness for two WUI categories: (1) the WUI Vegetated category (VEG), defined as areas with vegetative cover but with little or no housing; and (2) the WUI Intermix category (INT),

defined as areas with 50 percent or greater vegetative cover and 0.067 or fewer houses per ha (2.47 ac). For all four WUI categories, data for 3,296 forested FIA plots were available. The VEG category consisted of approximately 47,250 km^2 (29,500 mi^2) and included 801 forested FIA plots on which 39 species were observed, and the INT category consisted of approximately 36,000 km^2 (22,500 mi^2) and included 2,373 forested FIA plots on which 53 species were observed. The differences between plot and area proportions for these two categories is due to non-forested plots which were not used for this study.

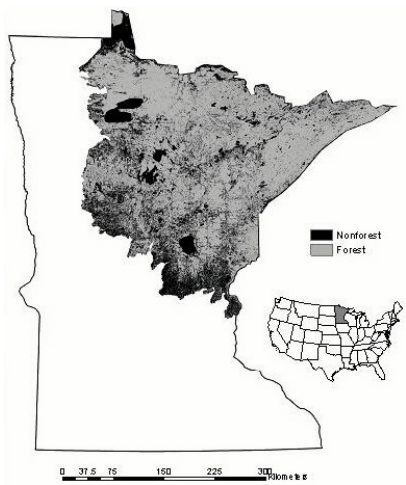

Figure 1. Minnesota, USA, study area.

Methods

Estimation of tree species richness is difficult apart from a complete census, because rare or highly clustered species may easily be missed using sample-based approaches. Several model-based and non-parametric approaches have been proposed for estimating tree species richness from sample data. All these approaches extrapolate information from the distribution of the species observed in the sample, S_o, to estimate the total number of species, S_t . The model-based approaches are based on species accumulation curves (Soberón and Llorente 1993) and feature nonlinear statistical models with horizontal asymptotes whose estimates are considered estimates of S_t. Two nonlinear models were considered, the Michaelis-Menten model and a 3-parameter exponential model. Two non-parametric approaches were also considered, the bootstrap and the jackknife.

Michaelis-Menten model

Because relationships between S_o and n, the number of plots measured, vary depending on the order in which plots are measured, estimates of parameters of model relationships will also vary by the same plot ordering. To avoid this artificial feature, the plots were randomly re-ordered 1,000 times, and the mean of S_o over the

1,000 replications for each n was used as the dependent variable when fitting the models.

The Michaelis-Menton model (Raaijmakers 1987) has been used previously for estimating species richness (e.g., Clench 1979) and is mathematically formulated as,

$$E(S_o) = \frac{\beta_1 n}{\beta_2 + n}, \qquad [1]$$

where E(.) is statistical expectation, S_o is the number of species observed, n is the number of plots, and the βs are parameters to be estimated. The estimate of the asymptote, β_1, provides an estimate of S_t. The covariance matrix of the model parameter estimates is estimated as,

$$\hat{V} = \hat{\sigma}_\varepsilon^2 (Z'Z)^{-1}, \qquad [2]$$

where $\hat{\sigma}_\varepsilon^2$ is the residual variance estimated by the mean squared error, the elements of the Z matrix are $z_{ij} = \frac{\partial f}{\partial \beta_j}(n_i)$, and f is the statistical expectation function of the model.

<u>Exponential model</u>

An exponential model of the mathematical form,

$$E(S_o) = \beta_1 \left[1 - \exp\left(\beta_2 n^{\beta_3} \right) \right] \qquad [3]$$

was fit in the same manner as was the Michaelis-Menten model. With this model, β_1 corresponds to the asymptote, and its estimate provides an estimate of S_t.

<u>Bootstrap</u>

Following the derivations of Smith and van Belle (1984), the bootstrap procedure (Efron 1979) may be described using five steps:

1. Construct the empirical cumulative probability function with density n^{-1} at each of the n plot observations.

2. Draw a sample of size n with replacement from the empirical cumulative probability function.

3. Define

$$I_j^i = \begin{cases} 0 \text{ if the } j^{th} \text{ species is not observed in the } i^{th} \text{ sample drawn in Step 2} \\ 1 \text{ if the } j^{th} \text{ species is observed in the } i^{th} \text{ sample drawn in Step 2} \end{cases}$$

and calculate the ith estimate of S_o as $\hat{S}_o^{Bi} = \sum_{j=1}^{S_o} I_j^i$. The statistical expectation of $\hat{S}_o^{Bi}$ is

$$E\left(\hat{S}_o^{Bi}\right) = \sum_{j=1}^{S_o} E\left(I_j^i\right) = \sum_{j=1}^{S_o}\left[1-\left(1-\frac{Y_j}{n}\right)^n\right] = S_o - \sum_{j=1}^{S_o}\left(1-\frac{Y_j}{n}\right)^n ,$$

where Y_j is the number of plots in the bootstrap sample from Step 2 for which the jth species is present. The bias in $\hat{S}_o^{Bi}$ is,

$$E\left(\hat{S}_o^{Bi} - S_o\right) = -\sum_{j=1}^{S_o}\left(1-\frac{Y_j}{n}\right)^n ,$$

so that the bootstrap estimate of S_t is

$$\hat{S}_t^{Bi} = S_o + \sum_{j=1}^{S_o}\left(1-\frac{Y_j}{n}\right)^n .$$

4. Repeat Steps 2-3 N times.

5. Calculate the bootstrap estimate of S_t as

$$\hat{S}_t^{B} = \frac{1}{N}\sum_{i=1}^{N}\hat{S}_t^{Bi} . \qquad [4]$$

The variance of $\hat{S}_t^B$ is given by Smith and van Belle (1984) as,

$$Var\left(\hat{S}_t^B\right) = \sum_{j=1}^{S_o}\left(1-\frac{Y_j}{n}\right)^n\left[1-\left(1-\frac{Y_j}{n}\right)^n\right] + \sum_{j \neq k}\sum\left[\left(\frac{Z_{jk}}{n}\right)^n - \left(1-\frac{Y_j}{n}\right)^n\left(1-\frac{Y_k}{n}\right)^n\right], \qquad [5]$$

where, for the original sample, Y_j is the number of plots for which the jth species is observed and Z_{jk} is the number of plots for which the jth and k^{th} species are jointly absent.

<u>Jackknife</u>

The Jackknife estimate of S_t may be obtained in five steps (Smith and van Belle 1984):

1. Remove the observations corresponding to the ith plot, and let r_i be the number of species that were observed only on the ith plot.

2. Using only observations from the remaining plots, calculate the ith jackknife estimate of S_o as,

$$\hat{S}_o^{Ji} = S_o - r_i .$$

3. Calculate the pseudovalue, $\theta_i = nS_o - (n-1)\hat{S}_o^{Ji} = S_o + (n-1)r_i$.

4. Repeat Steps 1-3 for each of the n plots.

5. Calculate the jackknife estimate of S_t as,

$$\hat{S}_t^J = \frac{1}{n}\sum_{i=1}^{n}\theta_i = S_o + \frac{n-1}{n}\sum_{i=1}^{n} r_i = S_o + \frac{n-1}{n}R , \qquad [6]$$

where $R = \sum_{i=1}^{n} r_i$.

The variance of $\hat{S}_t^J$ is,

$$\mathrm{Var}\left(\hat{S}_t^J\right) = \left(\frac{n-1}{n}\right)^2 \mathrm{Var}\left(\sum_{i=1}^{n} r_i\right) = \left(\frac{n-1}{n}\right)^2 n\mathrm{Var}\left(r_i\right)$$

$$= \frac{(n-1)^2}{n}\left(\frac{1}{n-1}\right)\sum_{i=1}^{n}\left[r_i^2 - \left(\frac{R}{n}\right)^2\right] = \left(\frac{n-1}{n}\right)\sum_{i=1}^{n}\left[r_i^2 - \left(\frac{R}{n}\right)^2\right]. \qquad [7]$$

The above jackknife estimates are characterized as first-order, because the observations from only a single plot are removed. Second-order jackknife estimates based on removing two plots simultaneously may also be calculated, but for this application preliminary analyses indicated they were not substantially better than first-order estimates.

Analyses

Evaluation of the nonlinear models was based on the quality of fit of the model to the data and the standard errors of the estimates of the asymptotes. The non-parametric bootstrap and jackknife approaches were evaluated to ensure that the sample size was adequate to produce defensible estimates of S_t. The issue is whether the estimate of S_t continues to increase as observations for more plots are added to the sample. If so, the sample size is inadequate. To evaluate the adequacy of the sample size, subsets of the total sample of various sizes were randomly drawn 250 times, and the mean of the estimates, $\hat{S}_t$, was calculated for each subset. If the sample size is adequate, the graph of the mean versus the subset sample size should reach and approximately maintain a plateau as the subset sample size increases. Adequacy of the sample size was evaluated for both non-parametric approaches using data for VEG and INT combined and data for each separately.

After evaluating the four approaches, the best one with respect to a basis in probability, adequacy of the sample size, and precision of the estimates was selected. Estimates of S_t were calculated for VEG and INT separately. The difference in the

estimates was considered to be statistically significant if the 2-standard error confidence intervals did not overlap.

Results

The estimates of S_t for VEG and INT combined obtained using the exponential model and jackknife techniques were similar (Table 1). However, the estimate obtained using the bootstrap technique was only slightly greater than the number of species observed, and the estimate obtained using the Michaelis-Menten model was less than the number of species observed. The latter result is due to an extremely poor fit of the model to the data (Figure 2). For the two non-parametric approaches, the sample size was adequate for the jackknife approach but not for the bootstrap approach (Figure 3). Because of the poor fit of the Michaelis-Menten model and the inadequacy of the sample size for the bootstrap approach, these two approaches were not considered further. The much smaller standard errors, $SE\left(\hat{S}_t\right)$, for the model-based approaches was attributed to using the average of S_o over a large number of randomizations of the plot orderings as the dependent variable. This averaging process masks the greater residual variability that would be observed if only a single ordering were used.

Table 1. Estimates of total tree species, $\hat{S}_t$.

Approach	All categories[1]		Intermix[2] (INT)		Vegetated[3] (VEG)	
	$\hat{S}_t$	$SE\left(\hat{S}_t\right)$	$\hat{S}_t$	$SE\left(\hat{S}_t\right)$	$\hat{S}_t$	$SE\left(\hat{S}_t\right)$
Michaelis-Menten model	54.25	0.07	52.48	0.08	38.92	0.07
Exponential model	65.56	0.11	64.85	0.15	41.62	0.10
Bootstrap	57.30	1.30	55.48	1.36	40.68	1.10
Jackknife	63.00	1.68	61.00	1.68	45.00	1.56

[1] 55 species observed
[2] 53 species observed
[3] 39 species observed

Although $SE\left(\hat{S}_t\right)$ was greater for the jackknife approach than for the exponential model approach, the jackknife approach was selected for comparing estimates for VEG and INT because it has a basis in probability while the exponential model is an entirely arbitrary formulation.

Sample sizes for both VEG and INT were adequate for the jackknife approach. For VEG, the number of observed species was $S_o = 39$, and the estimate of total species was $\hat{S}_t^J = 45.00$ with standard error, $SE\left(\hat{S}_t^J\right) = 156$; for INT, $S_o = 53$, and $\hat{S}_t^J = 6100$ with $SE\left(\hat{S}_t^J\right) = 168$. The comparison of the estimates indicated that tree species richness was statistically significantly greater in INT, which included a low level of housing density, than in VEG, which included little or no housing. All species observed in VEG, except one, were also observed in INT. Of the 14 species observed in INT but

not observed in VEG, several were ornamental or exotic species that are often associated with homes.

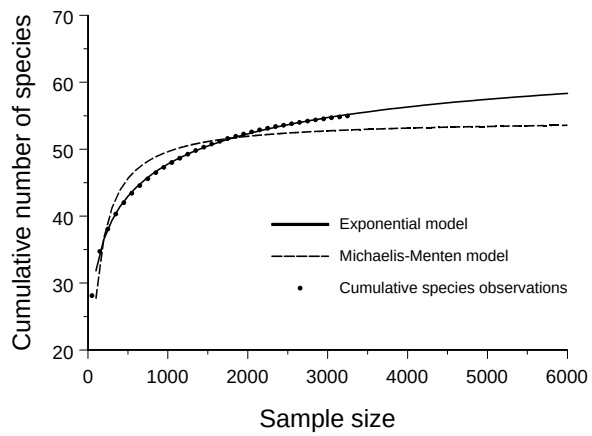

Figure 2. Quality of fit for model-based approaches.

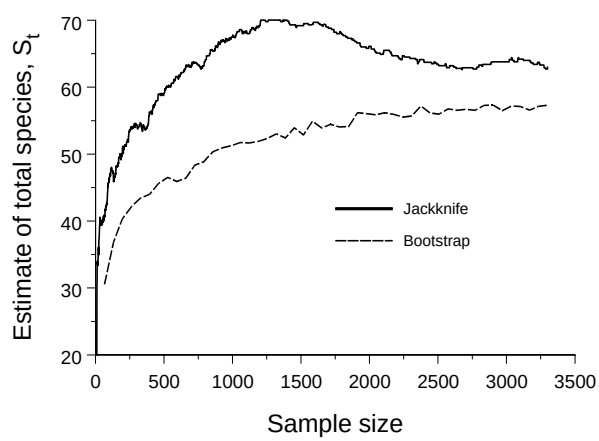

Figure 3. Adequacy of sample size for non-parametric approaches.

Conclusions

Four conclusions may be drawn from the study. First, forest inventory plot data may be used to estimate tree species richness, although caution must be exercised to ensure the adequacy of the sample size. Second, for this data set, the exponential model exhibited considerable more flexibility in fitting the data than did the Michaelis-Menten model. In fact, the quality of the fit of the Michaelis-Menten model was so poor that the estimate of total species was less than the number of species observed. Third, for this data set the sample size was adequate for the non-parametric jackknife

approach but not for the non-parametric bootstrap approach. However, this conclusion should not be construed to suggest that the jackknife approach will always be superior to the bootstrap approach. Fourth, the estimate of tree species richness for the Intermix category (INT) which included low levels of housing density was statistically significantly greater than the estimate for the vegetated category (VEG) which included very few houses. This result is partially attributed to the ornamental and exotic species associated with homes in INT.

Reference

Bailey, R.G. (1995) *Description of the ecoregions of the United States. Ed. 2. Revised and expanded (1st ed. 1980)*. USDA Forest Service Miscellaneous Publication No. 1391 (revised). 108 p. with separate map.

Clench, H. (1979) *How to make regional lists of butterflies: Some thoughts*. Journal of Lepidopterists' Society 33: 216-231.

Efron, B. (1979) *Bootstrap methods: Another look at the jackknife*. Annals of statistics 7: 1-26.

Loveland, T.L., and Shaw, D.M. (2001) Multiresolution land characterization: building collaborative partnerships. In: Scott, J.M., Tear, T., and Davis, F. (Eds). Gap analysis: a l;andscape approach to biodiversity planning. *Proceedings of the ASPRS/GAP Symposium*, Charlotte, North Carolina, USA 1996. pp. 83-89.

McRoberts, R.E., McWilliams, W.H., Reams, G.A., Schmidt, T.L., Jenkins, J.C., O'Neill, K.P., Miles, P.D., and Brand, G.J. (2004) *Assessing sustainability using data from the Forest Inventory and Analysis program of the United States Forest Serv*ice. Journal of Sustainable Forestry 18: 23-46.

Raaijmakers, J.G.W. (1987) *Statistical analysis of the Michaelis-Menten equation*. Biometrics 43: 792-803.

Radeloff, V.C., Hammer, R.B., Stewart, S.I, Fried, J.S., Holcomb, S.S., and McKeefry, J.F. (2005) *The wildland–urban interface in the United States*. Ecological Applications 15:799–805.

Smith, E.P., and van Belle, G. (1984) *Nonparametric estimation of species richness.* Biometrics 40: 19-129.

Soberón, J., and Llorente, J. (1993) *The use of species accumulation functions for the prediction of species richness*. Conservation Biology 7: 480-488.

Vogelmann, J.E., Howard, S.M., Yang, L., Larson, C.R., Wylie, B.K., and Van Driel, N. (2001) *Completion of the 1990s National Land Cover Data Set for the conterminous United States from Landsat Thematic Mapper data and ancillary data sources.* Photogrammetric Engineering and Remote Sensing 67: 650-662.