Estimating Tree Species Richness From Forest Inventory Plot Data

Ronald E. McRoberts[1] **and Dacia M. Meneguzzo**[2]

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. A crucial difficulty in estimating species richness from sample data obtained from sources such as inventory plots is that no assurance exists that all species occurring in a geographic area of interest are observed in the sample. Several model-based and nonparametric techniques have been developed to estimate tree species richness from sample data. Three such approaches were compared using data obtained from forest inventory plots in Minnesota, United States of America. The results indicate that an exponential model method and a nonparametric jackknife method were superior to the nonparametric bootstrap method.

Introduction

Of the international forest sustainability initiatives, the Montréal Process (1998) is geographically the largest, involving 12 countries on 5 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 that, 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.

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. Complete tree censuses, however, 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. Unfortunately, although tree species richness is an intuitive measure, it is difficult to estimate using sample data because no assurance exists that all species in a geographic area of interest have been observed in the sample, particularly rare or highly clustered species.

The objective of the study was to compare one model-based and two nonparametric approaches for estimating tree species richness from forest inventory plot data.

Data

Forest inventory data are 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 is conducted by the Forest Inventory and Analysis (FIA) program of the U.S. Department of Agriculture Forest Service. The program collects and analyzes inventory data and reports on the status and trends of

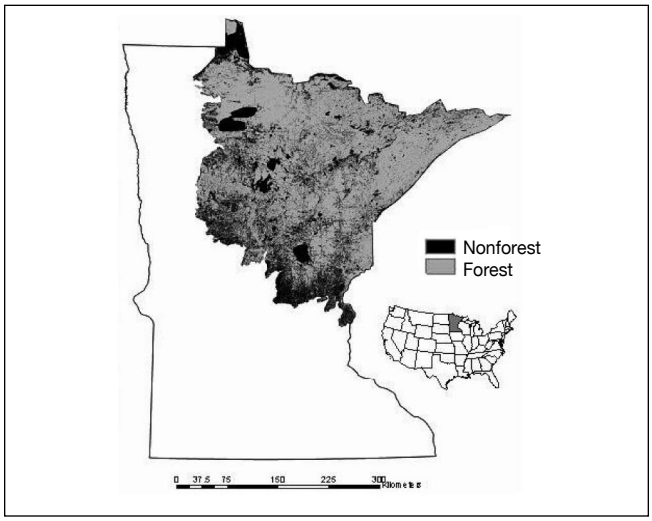
[1] Mathematical Statistician, U.S. Department of Agriculture (USDA), Forest Service, North Central Research Station, 1992 Folwell Avenue, St. Paul, MN 55108. E-mail: rmcroberts@fs.fed.us.

[2] Forester, USDA Forest Service, North Central Research Station, 1992 Folwell Avenue, St. Paul, MN 55108. E-mail: dmeneguzzo@fs.fed.us.

the Nation's forests. 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 (Bechtold and Patterson 2005, McRoberts *et al.* 2005). The sample was systematically divided into five interpenetrating, nonoverlapping 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 national FIA plot consists of four 7.32-m (24-ft) radius circular subplots that 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 study area was in Minnesota, United States of America, 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) (fig. 1). Forests in the study area are generally naturally regenerated, uneven aged, and mixtures of conifer and deciduous species. Data for 3,300 plots with centers in the study area and observed between 1999 and 2003 were available.

Figure 1.—*Study area.*



Methods

Several model-based and nonparametric 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 .

Exponential Model

Model-based approaches are generally based on empirical species accumulation curves (Soberón and Llorente 1993) depicting the relationship between the total number of species observed and the cumulative area of the sample. Nonlinear statistical models with horizontal asymptotes are fit to the empirical curves, and the estimates of the asymptotes are considered estimates of S_t. The exponential model

$$E\left(S_o\right) = b_1\left[1 - \exp\left(b_2 A^{b_3}\right)\right] \quad (1)$$

where E(.) is statistical expectation, S_o is the number of species observed, A is the cumulative area of the sample, and the βs are parameters, is a flexible curve, although admittedly it has no biological basis for describing a species accumulation curve. With this model, β_1 corresponds to the asymptote, and its estimate provides an estimate of S_t. The covariance matrix of the model parameter estimates is estimated as

$$\hat{V} = \hat{\sigma}_\varepsilon^2 \left(Z'Z\right)^{-1} \quad (2)$$

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

Bootstrap

For a sample of size n, Smith and van Belle (1984) describe the bootstrap procedure (Efron 1979) 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 i^{th} 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 j^{th} 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

$$\hat{S}_t^B = \frac{1}{N}\sum_{i=1}^{N}\hat{S}_t^{Bi} \tag{3}$$

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

$$Var\left(\hat{S}_t^B\right) = \sum_{j=1}^{S_o}\left(1-\frac{Y_j}{n}\right)^n\left[1-\left(1-\frac{Y_j}{n}\right)^n\right] + \sum_{j\neq k}\sum\left[\left(\frac{Z_{jk}}{n}\right)^n - \left(1-\frac{Y_j}{n}\right)^n\left(1-\frac{Y_k}{n}\right)^n\right] \tag{4}$$

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

Jackknife

For a sample of size n, Smith and van Belle (1984) describe how the Jackknife estimate of S_t may be obtained in five steps:

1. Remove the observations corresponding to the i^{th} plot, and let r_i be the number of species that were observed only on the i^{th} plot.

2. Using only observations from the remaining plots, calculate the i^{th} 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

$$\hat{S}_t^J = \frac{1}{n}\sum_{i=1}^{n}\theta_i = S_o + \frac{n-1}{n}\sum_{i=1}^{n} r_i = S_o + \frac{n-1}{n}R \tag{5}$$

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

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

$$Var\left(\hat{S}_t^J\right) = \left(\frac{n-1}{n}\right)^2 Var\left(\sum_{i=1}^{n} r_i\right) = \left(\frac{n-1}{n}\right)^2 nVar\left(r_i\right) \tag{6}$$

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

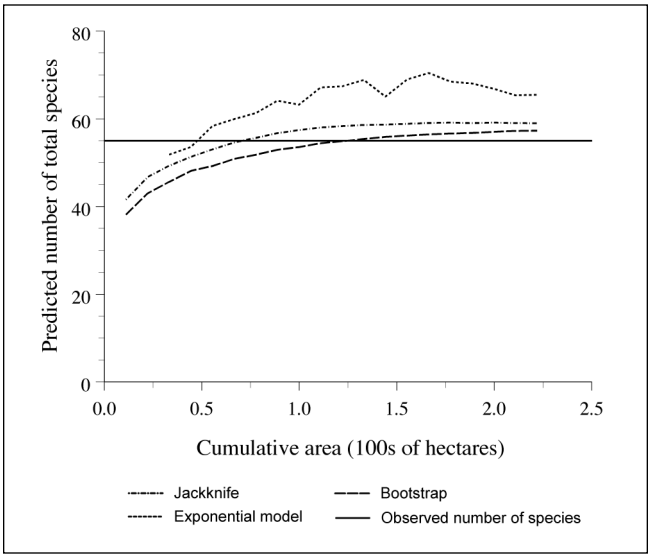
The above jackknife estimates are characterized as first order, because the observations from only a single plot are removed. Second-order jackknife estimates based on removing two plots simultaneously may also be calculated, but for this application preliminary analyses indicated they were not substantially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

Analyses

All three approaches were evaluated to determine sample sizes necessary to produce defensible estimates of S_t. The issue is whether $\hat{S}_t$ continues to increase as the sample size increases.

If so, the sample size is inadequate. If the sample size is adequate, the graph of $\hat{S}_t$ versus the cumulative sample area should reach and maintain an approximately constant value. For all three methods, samples sizes (i.e., number of plots) from 165 to 3,300 in steps of 165 were considered, and 250 samples of each size were randomly drawn. For the exponential model, the order in which plots of a particular sample are considered affects the species accumulation curve and, as a result, $\hat{S}_t$ and Var($\hat{S}_t$). To compensate, the plots in each sample were randomly reordered 1,000 times, the mean species accumulation curve was determined, and the exponential model was fit to the mean curve. For the nonparametric bootstrap and jackknife methods, the order of the plots in the sample is not an issue. For all three methods, the means of $\hat{S}_t$ and $\sqrt{Var\left(\hat{S}_t\right)}$ over the 250 samples were calculated for each sample size.

Figure 2.—*Estimates of total number of species.*



Results and Discussion

For all three methods, 250 samples were sufficient to stabilize the means of the estimates of $\hat{S}_t$ and $\sqrt{Var\left(\hat{S}_t\right)}$. In addition, on the basis of comparisons of residual error estimates, 1,000 random reorderings of the samples were sufficient to eliminate individual sample deviations in the mean species accumulation curves. Graphs of $\hat{S}_t$ versus the cumulative sample area for the three methods indicate that the sample size of 3,300 plots, representing 221.92 ha of sample area, is adequate for the jackknife and exponential model methods but possibly not for the bootstrap method (fig. 2). For the total sample size of 3,300 plots, $\hat{S}_t \pm \sqrt{\hat{Var}\left(\hat{S}_t\right)}$ was 65.46 ± 0.10 for the exponential model, 57.30 ± 1.68 for the bootstrap approach, and 59.00 ± 1.68 for the jackknife approach. All three methods produced estimates of S_t that were greater than $S_o = 55$, as should be expected.

In general, for a given sample size, the precision of the exponential model estimate of S_t was greater than the bootstrap and jackknife estimates which were comparable. For all three methods, the precision increased with greater sample sizes.

The exponential model and jackknife methods appear preferable to the bootstrap method because estimates using the former two methods stabilize with increasing sample size, while estimates using the latter method appear to be continuing to increase. Because the true number of species in the study area is unknown, it is uncertain as to which, if any, of these three methods produces superior estimates. This preliminary study suggests that additional model forms should be considered, that additional samples may be required to obtain a more definitive comparison of the methods, and that the comparisons should be made for other geographical areas.

Literature Cited

Bailey, R.G. 1995. Description of the ecoregions of the United States. 2nd ed. Misc. Pub. 1391 (revised). Washington, DC: U.S. Department of Agriculture, Forest Service. 108 p.

Bechtold, W.A.; Patterson, P.L., eds. 2005. The enhanced Forest Inventory and Analysis program—national sampling design and estimation procedures. Gen. Tech. Rep. SRS-80. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station. 85 p.

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.; Shaw, D.M. 2001. Multiresolution land characterization: building collaborative partnerships. In: Scott, J.M.; Tear, T.; Davis, F., eds. Gap analysis: a landscape approach to biodiversity planning. Proceedings, ASPRS/GAP symposium. Moscow, ID: National Biological Service: 83-89.

McRoberts, R.E.; Bechtold, W.A.; Patterson, P.L.; Scott, C.T.; Reams, G.A. 2005. The enhanced Forest Inventory and Analysis program of the USDA Forest Service: historical perspective and announcement of statistical documentation. Journal of Forestry. 103(6): 304-308.

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

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