

Sample-based Estimation of Regional Forest Tree Species Richness

Steen Magnussen (Corresponding author)

Natural Resources Canada, 506 West Burnside Rd

Victoria BC V8Z 1M5, Canada

Tel: 1-250-298-2423 E-mail: steen.magnussen@nrcan.gc.ca

Received: September 12, 2011

Accepted: September 28, 2011

Published: December 1, 2011

doi:10.5539/enrr.v1n1p2

URL: <http://dx.doi.org/10.5539/enrr.v1n1p2>

Abstract

Sample-based estimates of species richness for large heterogeneous regions may be affected by interactions between the sampling design and sub-regional heterogeneity in species richness and diversity. To investigate this issue a Monte Carlo simulation study was conducted with actual forest inventory tree species incidence data from two large regions, four estimators of species richness, three sampling designs, two sampling units (fixed area plots and tree nearest to plot center), and five sample sizes. The regions represent forested areas in central and eastern Canada and three states in the USA. Design effects in estimates of regional species richness were generally weak, of little practical import, and limited to settings with the smallest sample sizes. The study shows that accurate regional estimates of richness can be obtained by pooling sub-regional species incidence data. Sampling with fixed area plots was, with few estimator-dependent exceptions, considerably more efficient than sampling with one tree per sample location. A proposed ratio-based procedure for combining sub-regional estimates of richness broadens the options for regional estimation.

Keywords: Forest inventory, Species incidence, Sampling design, Stratification, Singleton species, Plot sampling

1. Introduction

Existing regional and national forest inventories can provide reasonable estimates of the number (viz. richness) of distinct forest tree species in a region or a nation (Magnussen, 2011). Large-scale forest inventory sampling with fixed area sample units (plots, quadrants, clusters) are typically characterized by sampling only a small proportion of the population of interest (Magnussen *et al.*, 2007; Winter *et al.*, 2008). Under these circumstances, few species richness estimators have performed well in Monte Carlo simulations with actual forest inventory data (Magnussen, 2011; Magnussen *et al.*, 2010). Heretofore popular estimators of richness like the Jackknife, the bootstrap, and Chao's coverage-based estimators generally disappoint because they were not designed for sampling with multi-tree quadrants (Hellmann & Fowler, 1999; Hwang & Shen, 2010; Lam & Kleinn, 2008; Palmer, 1991; Schreuder *et al.*, 2000; Schreuder *et al.*, 1999).

Species richness in a large heterogeneous region can be decomposed into a within- and among-ecosystem diversity viz. alpha- and beta-diversity (Barnett & Stohlgren, 2003; Engen *et al.*, 2008; Jobe, 2008; O'Dea *et al.*, 2006; Rosenzweig & Clark, 1994; Rosenzweig *et al.*, 2003). Alpha diversity is the biodiversity within a particular area, community or ecosystem, and is usually expressed as the species richness of the area. Beta diversity is also a measure of biodiversity which works by comparing the species diversity between ecosystems or along environmental gradients. Non-parametric and model-based estimators of species richness implicitly assume a spatially homogenous structure of species diversity (Chazdon *et al.*, 1998; Colwell *et al.*, 2004). Hence they target estimation of alpha-diversity (Gimaret-Carpentier *et al.*, 1998; Rosenzweig, *et al.*, 2003). For the estimation of both alpha- and beta-diversity new sampling strategies, plot designs, and estimators that explicitly consider species turnover rates and distance decay of compositional similarity have been proposed (Harte *et al.*, 1999; Krishnamani *et al.*, 2004; Palmer *et al.*, 2002; Ricotta *et al.*, 2002; Tackaberry *et al.*, 1997; Ugland *et al.*, 2003). Others argue for sampling stratified by habitat type (Waldhardt *et al.*, 2004). Practical applications of these new designs (Hortal *et al.*, 2006; Jobe, 2008; Nagaike *et al.*, 2010; O'Dea, *et al.*, 2006) have realized that a diverse origin, structure, and history of species mosaics and patterns prevent definitive conclusions (Whittaker & Field, 2000).

Existing regional and national forest inventories are designed for a multitude of purposes (Köhl *et al.*, 2006, p 3; McRoberts *et al.*, 2010) and are therefore unlikely to adopt a new sampling paradigm for the sake of improved species richness estimation. Most regional and national inventories employ a sampling design with sample locations on a regular or semi-regular grid that allows local variation in sampling intensity. Designs of this kind can only map environmental gradients at the scales of inter-plot distances. In principle, only a design suited for variogram estimation (Di Zio *et al.*, 2004; Lark, 2002; Zhao & Wall, 2004) can quantify ecological gradients in species composition (turnover rates). In areas with managed forests and a significant anthropogenic influence on plant communities, these gradients may be discontinuous and irregular, making attempts to account for beta-diversity difficult.

Simulation studies by Brose *et al.* (2003) and Wagner and Wildi (2003) found no support for a major effect of either ecological gradients or spatial heterogeneity in species habitats and occurrence on the performance of non-parametric species richness estimators. Ecological gradients, habitat patches, and edges did increase Monte-Carlo estimates of sampling variance, but without any obvious effect on bias. Nor did a stratified sampling approach with estimation of richness from samples pooled across strata produce a clear improvement over simple random sampling in a species-rich wet tropical forest in the Western Ghats, India (Gimaret-Carpentier, *et al.*, 1998). Sample size was again the single most influential factor.

Since the spatial distribution of forest tree species within the forested area of a large region is never uniform, the question remains whether stratified sampling (STR) offers an advantage over simple random sampling (SRS), and possibly over sampling with a constant sample size in each stratum (EQ). The answer may depend on the employed estimator of richness and choice of sampling unit (trees inside a fixed area plot, trees intersecting a transect line, a fixed number of trees nearest to a sample location).

To explore this question a Monte Carlo simulation study with *i*) two regional finite area populations of forest tree species, *ii*) four model-based estimators of species richness, *iii*) three sampling designs, *iv*) two sampling units, and *v*) five sample sizes was conducted. Estimators of species richness were limited to four with a demonstrated good performance in sampling with fixed-area plots (Magnussen, 2011; Magnussen, *et al.*, 2010). In addition, to broaden options for regional estimation of species richness, a new procedure for combining sub-regional estimates of species richness into a single regional estimate is proposed and evaluated.

2. Materials and Methods

2.1 Forest species incidence data

Two collections of forest inventory tree species plot incidence data act as two finite area regions (populations) for which an estimate of the number (richness) of distinct forest tree species (S) is desired. Each region was composed of N approximately equal-area primary sampling units (PSU) and stratified to three geographic sub-regions (strata). The raw data for each PSU were the names of all distinct tree species contained in a PSU. The N lists of names of species in each PSU were then converted into a $N \times S$ binary incidence matrix.

The first region is composed of forest inventory species incidence data from five Canadian provinces and is henceforth referred to as PROV. PROV is a spatially balanced collection of $N = 21\,091$ permanent forest inventory plots from eastern Canada (Magnussen & Boudewyn, 2008). The plots represent approximately 108 million ha of forested land in the Canadian provinces of Ontario, Quebec, New Brunswick, Nova Scotia, and Newfoundland and Labrador. All trees with a stem diameter larger than the provincial lower limit (range: 2.5 to 9.0 cm) were identified to species. The median size of a sample plot is 404 m² with an average of 139 trees (sdev. 53) and 8 species (sdev. 5). There were a total of $S = 84$ forest tree species in the data. Three geographic sub-regions (strata) were formed to represent the northern (NO), the central (CE), and the maritime (MA) parts of PROV. The number of plots by sub-region were 2938 (NO), 11 963 (CE), and 6190 (MA). Species richness in the three sub-regions were 33 (NO), 70 (CE), and 71 (MA).

The second region is composed of forest inventory species incidence data from 13 181 US Forest Service (USFS) national forest inventory sample plots (Bechtold *et al.*, 2005) collected between 2005 and 2008 in three states (sub-regions): Georgia (GA, 3213 plots), Michigan (MI, 5851 plots), and Minnesota (MN, 4117 plots). The sampling frame and design has been described by Bechtold and Patterson (2005). All trees in a plot with a stem diameter of 12.7 cm or larger at a reference level of 1.37 m above ground were counted and identified to species. Unknown species were excluded. They were assumed to be non-forest tree species. A sample plot contained, on average, six species (sdev. 3). The region is henceforth called FIA. Species richness in FIA was 187, 143 in GA, 102 in MI, and 66 in MN.

The main ecological differences between PROV and FIA are *i*) the boreal nature of the forests in NO and *ii*) the warm temperate forests in GA dominated by deciduous tree species.

2.2 Sampling designs

For the purpose of estimating S in PROV and FIA a simulated (Monte Carlo) random (probability) sample of n fixed-area inventory plots (without replacement, *wor*) of species incidence data was taken from each region. Three designs determined selections of plots: 1) equal sub-regional sample sizes (EQ), 2) sub-regional sample sizes proportional to the number of inventory plots in a sub-region (STR), and 3) simple random selection within entire region (SRS). Five sample sizes, representing sample fractions of approximately 0.5%, 1%, 2%, 3%, and 5%, were used for each design. Data from each sample plot were either *i*) a list of species names of all trees recorded in the selected plot (CLU), *ii*) the species name of the tree located closest to the center of a selected plot (T1). Ties were resolved with a toss of a fair coin. All simulations and computations were carried out in the MATHEMATICA programming language (Wolfram, 1999). The advantage of working with actual inventory data is that actual species associations at the plot-level are captured. However, it also limits the scope of possible conclusions that can be drawn.

The observed species richness (SO) and four model-based estimates of S (\hat{S}_{est} , see next) were computed from each sample. The effects of design and sample size on these estimates were assessed from 400 Monte Carlo replications of the 15 design settings. With this number of replications, the estimated standard error of the mean number of observed species SO in a given design was always less than 0.3 (Koehler *et al.*, 2009).

2.3 Estimators of species richness

Four estimators with a demonstrated good performance (in terms of bias and accuracy) in low-intensity sampling for forest tree species richness with fixed-area plots (CLU) were considered (Magnussen, 2011; Magnussen, *et al.*, 2010). Estimators of species richness may differ in sensitivity to the sampling design due to differential reliance on the number and relative incidence of species observed in a sample (Haas & Stokes, 1998; Mao & Lindsay, 2007; Walther & Morand, 1998); thus the four estimators can only provide an example of how the sample design and choice of sampling unit can affect estimators of richness. On the other hand, their appeal to forest tree species richness estimation will increase if they also demonstrate insensitivity to the design for sample collection (EQ, STR, SRS) and choice of sampling unit (CLU, T1).

The four estimators of species richness used in this study are detailed in the Appendix. A brief outline of each estimator is given next. The first estimator is a Pólya-urn type estimator (\hat{S}_{urn}) which, through replicated chains of iterative urn-drawings, generates predictions of new species discoveries in the $N-n$ non-sampled PSUs, and from that a posterior distribution of predictions of S . The estimate \hat{S}_{urn} is the mean of the M posterior predictions and $\text{var}(\hat{S}_{urn})$ the estimate of variance (Magnussen, *et al.*, 2010). Here $M = 50$.

The second estimator is a Horwitz-Thompson (HT) type (model-based) estimator (\hat{S}_{ht}) that estimates the number of unseen species from sample-based estimates of species inclusion probabilities (Magnussen, 2011). The HT-type formulation of the estimator means that it adapts to any sample design for which the design-based HT-estimator applies (Thompson, 1992, p 49).

The third estimator (\hat{S}_{sh}) was developed by Shen and He for fixed-area plot *wor* sampling from a finite area population (Shen & He, 2008). A zero-truncated binomial distribution is assumed for the number of plots containing a given species, and a modified beta-distribution for the distribution of the probability of the plot-level presence/absence of a species. After fitting the modified beta-distribution to the data, an explicit solution to the probability that a species occurred in x sample quadrats was obtained (EQ 4 p 2053) and used in a multinomial likelihood (Sanathanan, 1972) to estimate the number of unseen species. A maximum likelihood variance estimator is provided (Shen & He, 2008). The fourth estimator (\hat{S}_{hw}) is a generalized jackknife estimator of richness (Hwang & Shen, 2010) expressed as the sum of SO and a prediction of the number of unseen species. The latter was obtained from an exponential regression model that predicts a scaled ratio of $f_i \times f_{i+1}^{-1}$ where f_i is the number of species seen in i sample plots. The regression model was then used to predict f_0 .

2.4 Estimation of regional forest tree species richness

Estimation of regional (FIA, PROV) forest tree species richness under the three sampling designs (EQ, STR, and SRS), five sample fractions, and two types of sampling unit (CLU, T1), proceeded by *i*) a direct estimation of regional species richness from the pooled incidence data (i.e. ignoring the sampling design), and *ii*) combining sub-regional estimates of richness to a regional estimate (\tilde{S}_{comb}) via a novel procedure detailed below. The direct estimation of regional richness followed the procedures detailed in the source document of each estimator. In some situations sub-regional estimates of species richness may already exist, and it could be advantageous to take advantage of these estimates in the pursuit of a regional estimate. Sub-regional estimates are, of course, non-additive when one or more species can occur in more than one sub-region within the region (Guralnick & Van Cleve, 2005). Thus the issue of sub-regional overlap in species composition must be addressed. The simplest approach is to multiply the sum of sub-regional estimates of richness by a “uniqueness” ratio ($0 < R_s \leq 1$) of the regional species richness. Of course, R_s is unknown and must somehow be estimated from available data or knowledge. Expert knowledge, tree atlases, museum data (Beck & Kitching, 2007), or previous studies may provide sufficient information for an estimation of R_s . If sub-regional species incidence data (or equivalent) are available, a possible approximation to R_s can be obtained as

$$(4) \quad \tilde{R}_s \approx SO_{region} \times \left(\sum_{h=1}^H SO_h \right)^{-1} \times \tilde{C}$$

where SO_{region} is the total number of species observed in the pooled sub-regional samples, and SO_h is the number of species observed in sub-region h . The multiplier $\tilde{C} \leq 1$ is a correction factor compensating for a positive bias that arises from the facts that *i*) observed species observed in a relatively small sample tend to be more common than species not captured by the sample (Chao, 1981; Reichert *et al.*, 2010; Starr, 1979) and *ii*) common species are more likely to be shared among sub-regions than less-common species (Cao *et al.*, 2004; Jaccard, 1901; Ugland *et al.*, 2005). After extensive trials it was decided to use Chao's estimator of sample coverage as the correction factor ($C = 1 - \text{relative frequency of species found in just one plot}$) (Bunge & Fitzpatrick, 1993) (Chao & Lee, 1992). Our simulations (not shown) demonstrated that \tilde{R}_s is asymptotically unbiased as $n_h \rightarrow \infty$. The ratio \tilde{R}_s attains a maximum of 1.0 when no species occur in more than one stratum, and a minimum of H^{-1} when species compositions are identical across strata. Accordingly, the estimator for the regional richness becomes $\tilde{S}_{comb}(est) = \tilde{R}_s \sum_h \hat{S}_{h,est}$ where *est* stands for the estimator used to estimate sub-regional richness. Note it is not necessarily that the same estimator is used in every sub-region.

Derivation of a variance estimator for $\tilde{S}_{comb}(est)$ may be possible, it depends on how \tilde{R}_s was obtained and the available data. Here, the variance was estimated by a standard application of the delta-technique to the expression of $\tilde{S}_{comb}(est)$ as a function of SO_{region} , SO_h , $\hat{S}_{h,est}$, and \tilde{C} , $h=1, \dots, H$ (Oehlert, 1992). Variances of SO_{region} and SO_h were obtained via a modified (plot-level) bootstrap procedure (Magnussen & McRoberts, 2011). The variance of $\hat{S}_{h,est}$ was estimated as detailed in the source document for each estimator (c.f. references given in section on estimators). Finally, a variance approximation for \tilde{C} was obtained from a bootstrap estimate of the variance of the number of species that occurred in just one plot (assuming the overall species yield is fixed by design).

2.5 Shannon-Weaver indicator of diversity

The Shannon-Weaver index is a widely used indicator of diversity (Magnussen & Boyle, 1995). To help readers appreciate the data used here, we also report on the Shannon-Weaver indices of diversity for the two regions (FIA and PROV) as well as for their sub-regions (Figures 1 and 2). The Shannon-Weaver index is computed as follow (Krebs, 1989, p 444)

$$(5) \quad H = -\sum_{i=1}^s \phi_i \log(\phi_i)$$

where ϕ_i is the plot-level relative incidence of species i .

3. Results and Discussion

3.1 Regional and sub-regional species richness

Model-based estimates of species richness depends strongly on the distribution of relative plot-level incidence of species (Chao, 2005; Mao & Colwell, 2005). To obtain a sample of SO species in a community with many rare and a few common species requires a larger sample than in a community with a more even distribution of relative incidence. As expected, the relative plot-level incidence distribution for GA, MI, MN, and combined resembled an inverse J -shaped distribution (Colwell, *et al.*, 2004) characterized by many rare species and fewer common species (Figure 1). Despite large differences in richness (66 in MN to 147 in GA) the sub-regional distributions share several common features: About half the species occur in less than 1% of the plots and only 10% reach a relative incidence of 10% or higher. A large number of species are shared among sub-regions ($\bigcup_{GA,MI} = 62$, $\bigcup_{GA,MN} = 38$, and $\bigcup_{MI,MN} = 56$).

Relative incidence distributions in sub-regions of PROV also resembled an inverse J -shaped distribution but less so than in FIA (Figure 2). Differences among sub-regional richness (33 to 71) were smaller. As for FIA, about half the species occur in less than 1% of the plots; a larger proportion (30-40%) can be regarded as fairly common. A large number of species were shared among sub-regions ($\bigcup_{NO,CE} = 33$, $\bigcup_{NO,MA} = 33$, and $\bigcup_{CE,MA} = 57$).

3.2 Design effects in the number of observed species

The observed species count is, on its own, an estimator of richness and an important component of any estimator of S (Chao, 2005; Mao & Lindsay, 2007). Observed species counts in FIA under EQ, STR and SRS designs were similar (Figure 3 top). In the combination FIA×EQ the samples contained approximately three species more than with STR and SRS designs. Under an EQ design the sample fraction in GA is higher than in MN and MI, and with more species in GA, one should, *a priori*, expect that $SO_{EQ} > SO_{STR} \approx SO_{SRS}$. It is therefore somewhat surprising, that the design effects are minor and do not seem to increase with sample size. Similarities among sub-regions in distributions of relative species plot-incidence are likely a main cause for the modest design effects. Design effects for CLU and T1 are similar once the sample fraction is 1% or higher.

In PROV the species count with an EQ design was lower than with STR and SRS designs (Figure 3 bottom). In an EQ design the sample fraction is highest in species-poor NO and lowest in CE. Hence, an increase in the NO sample fraction at the expense of CE and ME triggers a decline in the number of species encountered per sampling unit. As expected, the decline is more pronounced for T1 than for CLU sampling.

Sampling errors of observed richness in FIA and PROV were, for a given sample size and sampling unit, similar for EQ, STR, and SRS designs (Figure 3). With fixed-area plot (CLU) sampling units in FIA the sampling error declined with an increase in sample fraction from 5.3 (0.5% sample) to 3.8 (5% sample). Yet it increased (slightly) from 2.9 to 3.8 when only the centermost tree in each plot is selected (T1) due to larger increases in SO with increasing sample sizes.

3.3 Design effects in the number of observed singleton species

Species observed in just one sample unit (singletons, SI) exert a disproportionate effect on estimates of richness (Mao & Lindsay, 2007; Reichert, *et al.*, 2010). In both FIA and PROV the number of observed SI -species was similar across the EQ, STR, and SRS designs. With fixed-area plot sampling (CLU) in FIA the number of singletons drops from approximately 25 with the lowest sample fraction to 20 with the largest. With selection of the centermost tree per plot the number of singletons at first increased and then decreased as the sample fraction is increased. In PROV one observed a steady decline in the number of SI -species with increasing sample size when only one tree is selected per plot (T1).

The impact of singletons on estimates of richness is relative to the number of observed species (SO) and hence sample size (Chao, 2005; Thrush *et al.*, 2006). However, when examining the ratio of the number of observed singleton species to the number of observed species no important trend emerged across designs, sample size, and choice of sampling unit.

3.4 Direct estimation of regional tree species richness

A comprehensive comparison of the estimators under SRS and CLU sampling has been done elsewhere (Magnussen, 2011). In terms of root mean squared errors this study confirmed that the HT-type estimator ranks first for the three largest sample fractions while the urn estimator ranks first at the two lowest sample fractions (not shown).

In FIA the four studied estimators of richness were apparently insensitive to design effects (Figure 4). The only indication of an important design effect was in the combination of the Shen & He estimator, the two lowest sample fractions, and the T1 sampling unit (for which the estimator was not designed). Here, estimates under SRS designs were lower (15–30) than those under EQ and STR design. In PROV the \hat{S}_{sh} and \hat{S}_{hw} estimators indicated a sensitivity to the sampling design, whereas estimates with \hat{S}_{urn} and \hat{S}_{ht} were indifferent (Figure 5). With \hat{S}_{sh} and an EQ design higher (less biased) estimates of S were obtained than with STR and SRS designs. With the \hat{S}_{hw} there appears to be complex design \times sample size interactions. Sampling with fixed-area plots (CLU) produced, as expected, less biased estimates of S than sampling with selection of a single central tree per plot (T1). The \hat{S}_{urn} estimator was the most sensitive to the choice of sampling unit and \hat{S}_{hw} the least sensitive.

Bias in regional and sub-regional estimates of species richness were strongly correlated (> 0.8) and followed single linear trend across sample fractions and choice of sampling unit. The slope of the trend line for the \hat{S}_{ht} estimator was close to 1, otherwise the slope was in the range from 0.5 to 0.7.

An absence of design effects suggest that the influence of sub-regional heterogeneity in species diversity and incidence (beta-diversity, Jobe, 2008) is outweighed by random variation introduced by low-intensity probability sampling. Interactions between estimators, design, and sampling unit were all minor and of no practical importance. Consequently, the performance of the four richness estimators is controlled almost exclusively by sample size and choice of sampling unit (Brose, *et al.*, 2003; Chazdon, *et al.*, 1998). Similar results, but limited to considerably higher sample fractions, were reached in simulations with artificial populations and non-parametric richness estimators (e.g. Jackknife, bootstrap, and Chao's estimators) (Brose, *et al.*, 2003; Wagner & Wildi, 2002). Gimaret-Carpentier (1998) reported a slim advantage of stratifying samples by topography when estimating species richness in the Western Ghats of India.

The results also suggest that fairly accurate estimates of species richness in a large and heterogeneous region can be obtained by pooling sample data from sub-regional surveys. As well, the relative bias of a regional estimate may be lower than the average relative bias of sub-regional estimates.

Results from this study cannot be extended indiscriminately to regions with steep ecological gradients, and distance-dependent species' turn-over rates. When it is possible to verify, for example, a power law species–area relationship (Arrhenius, 1921), it should be exploited in pursuit of less biased and more efficient estimates of species richness than possible with estimators of the kind used here, since they ignore information about fundamental ecosystem processes (Chazdon, *et al.*, 1998; Engen, *et al.*, 2008; Harte, *et al.*, 1999; Jobe, 2008; Krishnamani, *et al.*, 2004; O'Dea, *et al.*, 2006; Thrush, *et al.*, 2006; Ugland & Gray, 2004; Ugland, *et al.*, 2005). Nevertheless, the potential advantages of a ecological stratification may not materialize unless the difficult task of identifying an effective stratification criterion (metric) can be completed with confidence (Barnett & Stohlgren, 2003; Stohlgren *et al.*, 1997).

3.5 Combining sub-regional estimates to a regional estimate of richness

The primary intent behind the proposed procedure for combining sub-regional estimates of species richness to a regional estimate is to give the analyst more options for combining and using existing information and data. Application in this study was limited to the case where the analyst has full access to all data gathered at the sub-regional level. As such, the example merely demonstrated that the direct and combined approach in this case generated very similar results.

As for the direct estimates of regional species richness, the design effects in the combined estimate of richness were unimportant (Figures 6 and 7). The performance of the combined estimate depends on the properties of the component estimates and the multiplier \tilde{R}_s . With the \hat{S}_{sh} and \hat{S}_{urn} estimators and CLU, the combination of sub-regional estimates of richness into a regional estimate produced a reduction in bias (1%–13%) in all designs with the three lowest sample fractions. A reduction linked to a positive bias in \tilde{R}_s that, as expected, declined rapidly with an increase in sample fraction. Otherwise, with \hat{S}_{ht} , \hat{S}_{sh} , and \hat{S}_{urn} the performance of \hat{S}_{comb} was similar to that of a direct estimator (Figures 6 and 7). However, combining sub-regional CLU-based \hat{S}_{hw} estimates reversed a negative bias of richness to a positive bias in both FIA and PROV. A noticeable among

sub-regional heterogeneity in the estimates of the generalized jackknife regression function for the prediction of the number of unseen species is the most likely explanation for this anomaly.

4. Conclusions

Fairly accurate regional sample-based estimates of species richness can be obtained with any of four studied richness estimators by pooling species incidence data from low-intensity, sub-regional sampling with fixed-area plots. A new algorithm for combining sub-regional estimates of species richness affords added flexibility and more options for the task of regional estimation.

Acknowledgements

The PROV data were kindly provided by the Nova Scotia Department of Natural Resources, the Prince Edward Island Department of Agriculture and Forestry, the New Brunswick Department of Natural Resources and Energy, the Québec Ministère des Ressources naturelles, de la Faune et des Parcs, and the Ontario Ministry of Natural Resources. FIA data came from the US Forest Service Forest Inventory Programme and were kindly made available through Dr. Brad Smith, Wash. D.C.

References

- Arrhenius, O. (1921). Species and man. *Journal of Ecology*, 9, 95-99. <http://dx.doi.org/10.2307/2255763>
- Barnett, D. T., & Stohlgren, T. J. (2003). A nested-intensity design for surveying plant diversity. *Biodiversity and Conservation*, 12(2), 255-278. <http://dx.doi.org/10.1023/A:1021939010065>
- Bechtold, W. A., & Patterson, P. L. (2005). *The enhanced forest inventory and analysis program - National sampling design and estimation procedures* (U. F. Service, Trans.) (pp. 85). Asheville, NC
- Beck, J., & Kitching, I. J. (2007). Estimating regional species richness of tropical insects from museum data: a comparison of a geography-based and sample-based methods. *Journal of Applied Ecology*, 44(3), 672-681. <http://dx.doi.org/10.1111/j.1365-2664.2007.01291.x>
- Brose, U., Martinez, N. D., & Williams, R. J. (2003). Estimating species richness: Sensitivity to sample coverage and insensitivity to spatial patterns. *Ecology*, 84(9), 2364-2377. <http://dx.doi.org/10.1890/02-0558>
- Bunge, J., & Fitzpatrick, M. (1993). Estimating the number of species: A review. *Journal of the American Statistical Association*, 88(421), 364-373. <http://dx.doi.org/10.2307/2290733>
- Cao, Y., Larsen, D. P., & White, D. (2004). Estimating regional species richness using a limited number of survey units. *Ecoscience*, 11(1), 23-35.
- Chao, A. (1981). On estimating the probability of discovering a new species. *Annals of Statistics*, 9(6), 1339-1343. <http://dx.doi.org/10.1214/aos/1176345651>
- Chao, A. (2005). *Species richness estimation*. In S. Kotz, N. Balakrishnan, C. B. Read & B. Vidakovic (Eds.), *Encyclopedia of Statistical Sciences* (2nd ed., Vol. 2, pp. 7909-7916). New York: Wiley.
- Chao, A., & Lee, S. M. (1992). Estimating the number of classes via sample coverage. *Journal of the American Statistical Association*, 87(417), 210-217. <http://dx.doi.org/10.2307/2290471>
- Chazdon, R. L., Colwell, R. K., Denslow, J. S., Guariguata, M. R., Dallmeier, F., & Comiskey, J. (1998). *Statistical methods for estimating species richness of woody regeneration in primary and secondary rain forests of NE Costa Rica Forest biodiversity research, monitoring and modeling: Conceptual background and old world case studies* (pp. 285-309). Paris: Parthenon.
- Colwell, R. K., Mao, C. X., & Chang, J. (2004). Interpolating, extrapolating, and comparing incidence-based species accumulation curves. *Ecology*, 85(10), 2717-2727. <http://dx.doi.org/10.1890/03-0557>
- Di Zio, S., Fontanella, L., & Ippoliti, L. (2004). Optimal spatial sampling schemes for environmental surveys. *Environmental and Ecological Statistics*, 11(4), 397-414. <http://dx.doi.org/10.1007/s10651-004-4186-9>
- Engen, S., Saether, B. E., Sverdrup-Thygeson, A., Grotan, V., & Odegaard, F. (2008). Assessment of species diversity from species abundance distributions at different localities. [Article]. *Oikos*, 117(5), 738-748. <http://dx.doi.org/10.1111/j.0030-1299.2008.16466.x>
- Gimaret-Carpentier, C., Pelissier, R., Pascal, J. P., & Houllier, F. (1998). Sampling strategies for the assessment of tree species diversity. *Journal of Vegetation Science*, 9(2), 161-172. <http://dx.doi.org/10.2307/3237115>

- Guralnick, R., & Van Cleve, J. (2005). Strengths and weaknesses of museum and national survey data sets for predicting regional species richness: comparative and combined approaches. *Diversity and Distributions*, 11(4), 349-359. <http://dx.doi.org/10.1111/j.1366-9516.2005.00164.x>
- Haas, P., & Stokes, L. (1998). Estimating the number of classes in a finite population. *Journal of the American Statistical Association*, 93(444), 1475-1487. <http://dx.doi.org/10.2307/2670061>
- Harte, J., McCarthy, S., Taylor, K., Kinzig, A., & Fischer, M. L. (1999). Estimating species-area relationships from plot to landscape scale using species spatial-turnover data. *Oikos*, 86(1), 45-54. <http://dx.doi.org/10.2307/3546568>
- Hellmann, J. J., & Fowler, G. W. (1999). Bias, precision, and accuracy of four measures of species richness. *Ecological Applications*, 9(3), 824-834. [http://dx.doi.org/10.1890/1051-0761\(1999\)009\[0824:BPAAOF\]2.0.CO;2](http://dx.doi.org/10.1890/1051-0761(1999)009[0824:BPAAOF]2.0.CO;2)
- Hortal, J., Borges, P. A. V., & Gaspar, C. (2006). Evaluating the performance of species richness estimators: sensitivity to sample grain size. *Journal of Animal Ecology*, 75(1), 274-287. <http://dx.doi.org/10.1111/j.1365-2656.2006.01048.x>
- Hwang, W.-H., & Shen, T.-J. (2010). Small-Sample Estimation of Species Richness Applied to Forest Communities. *Biometrics*, 66(4), 1052-1060. <http://dx.doi.org/10.1111/j.1541-0420.2009.01371.x>
- Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37, 547-579.
- Jobe, R. T. (2008). Estimating landscape-scale species richness: Reconciling frequency- and turnover-based approaches. *Ecology*, 89(1), 174-182. <http://dx.doi.org/10.1890/06-1722.1>
- Koehler, E., Brown, E., & Haneuse, J.-P. A. (2009). On the assessment of Monte Carlo error in simulation-based statistical analyses. *The American Statistician*, 63(2), 155-162. <http://dx.doi.org/10.1198/tast.2009.0030>
- Köhl, M., Magnussen, S., & Marchetti, M. (2006). *Sampling methods, remote sensing and GIS multiresource forest inventory*. Berlin: Springer
- Krebs, C. (1989). *Ecological Methodology*. New York: Harper Collins
- Krishnamani, R., Kumar, A., & Harte, J. (2004). Estimating species richness at large spatial scales using data from small discrete plots. *Ecography*, 27(5), 637-642. <http://dx.doi.org/10.1111/j.0906-7590.2004.03790.x>
- Lam, T. Y., & Kleinn, C. (2008). Estimation of tree species richness from large area forest inventory data: Evaluation and comparison of jackknife estimators. *Forest Ecology and Management*, 255, 1002-10010. <http://dx.doi.org/10.1016/j.foreco.2007.10.007>
- Lark, R. M. (2002). Optimized spatial sampling of soil for estimation of the variogram by maximum likelihood. *Geoderma*, 105(1-2), 49-80. [http://dx.doi.org/10.1016/S0016-7061\(01\)00092-1](http://dx.doi.org/10.1016/S0016-7061(01)00092-1)
- Magnussen, S. (2011). A Horvitz-Thompson type estimator of species richness. *Envirometrics*. 22(7), 901-910. doi:10.1002/env.1117
- Magnussen, S., & Boudewyn, P. (2008). Estimators of tree species richness: An assessment for Central and Eastern Canada. *Journal of Sustainable Forestry*, 26(1), 77-96. http://dx.doi.org/10.1300/J091v26n01_04
- Magnussen, S., & Boyle, T. J. B. (1995). Estimating sample size for inference about the Shannon-Weaver and the Simpson indices of species diversity. *Forest Ecology and Management*, 78(1-3), 71-84. [http://dx.doi.org/10.1016/0378-1127\(95\)03596-1](http://dx.doi.org/10.1016/0378-1127(95)03596-1)
- Magnussen, S., & McRoberts, R. (2011). A modified bootstrap procedure for cluster sampling variance estimation of species richness. *Journal of Applied Statistics*, 38(5-6), 1223-1238. <http://dx.doi.org/10.1080/02664763.2010.491861>
- Magnussen, S., Smith, B., Kleinn, C., & Sun, I. F. (2010). An urn model for species richness estimation in quadrat sampling from fixed-area populations. *Forestry*, 83(3), 293-306. <http://dx.doi.org/10.1093/forestry/cpq012>
- Magnussen, S., Smith, B., & Uribe, A. S. (2007). National Forest Inventories in North America for monitoring forest species diversity. *Plant Biosystems*, 141(1), 113-122. <http://dx.doi.org/10.1080/11263500601153925>
- Mao, C. X., & Colwell, R. K. (2005). Estimation of species richness: Mixture models, the role of rare species, and inferential challenges. *Ecology*, 86(5), 1143-1153. <http://dx.doi.org/10.1890/04-1078>

- Mao, C. X., & Lindsay, B. G. (2007). Estimating the Number of Classes. *Annals of Statistics*, 35, 917-930. <http://dx.doi.org/10.1214/009053606000001280>
- McRoberts, R. E., Tomppo, E. O., & Næsset, E. (2010). Advances and emerging issues in national forest inventories. *Scandinavian Journal of Forest Research*, 25(4), 368-381. <http://dx.doi.org/10.1080/02827581.2010.496739>
- Nagaike, T., Hayashi, A., & Kubo, M. (2010). Is it necessary to change the number of samples for different forest types when evaluating plant species richness? A case study in a forested landscape in central Japan. *Journal of Forest Research*, 15(3), 169-175. <http://dx.doi.org/10.1007/s10310-009-0169-x>
- O'Dea, N., Whittaker, R. J., & Ugland, K. I. (2006). Using spatial heterogeneity to extrapolate species richness: a new method tested on Ecuadorian cloud forest birds. *Journal of Applied Ecology*, 43(1), 189-198. <http://dx.doi.org/10.1111/j.1365-2664.2006.01143.x>
- Oehlert, G. W. (1992). A Note on the Delta Method. *The American Statistician*, 46(1), 27-29. <http://dx.doi.org/10.2307/2684406>
- Palmer, M. W. (1991). Estimating species richness: The second-order jackknife reconsidered. *Ecology*, 72(4), 1512-1513. <http://dx.doi.org/10.2307/1941127>
- Palmer, M. W., Earls, P. G., Hoagland, B. W., White, P. S., & Wohlgemuth, T. (2002). Quantitative tools for perfecting species lists. *Environmetrics*, 13(2), 121-137. <http://dx.doi.org/10.1002/env.516>
- Patil, G. P., & Taillie, C. (2001). Estimation of species richness based on species range. *Community Ecology*, 2(2), 209-211. <http://dx.doi.org/10.1556/ComEc.2.2001.2.8>
- Pla, L. (2004). Bootstrap confidence intervals for the Shannon biodiversity index: A simulation study. *Journal of Agricultural Biological and Environmental Statistics*, 9(1), 42-56. <http://dx.doi.org/10.1198/1085711043136>
- Reichert, K., Ugland, K. I., Bartsch, I., Hortal, J., Bremner, J., & Kraberg, A. (2010). Species richness estimation: Estimator performance and the influence of rare species. *Limnology, Oceanography. Methods*, 8, 294-303. <http://dx.doi.org/10.4319/lom.2010.8.294>
- Ricotta, C., Carranza, M. L., & Avena, G. (2002). Computing beta-diversity from species-area curves. *Basic and Applied Ecology*, 3(1), 15-18. <http://dx.doi.org/10.1078/1439-1791-00082>
- Rosenzweig, M. L., & Clark, C. W. (1994). Island Extinction Rates from Regular Censuses. *Conservation Biology*, 8(2), 491-494. <http://dx.doi.org/10.1046/j.1523-1739.1994.08020491.x>
- Rosenzweig, M. L., Turner, W. R., Cox, J. G., & Ricketts, T. H. (2003). Estimating Diversity in Unsourced Habitats of a Biogeographical Province. *Conservation Biology*, 17(3), 864-874. <http://dx.doi.org/10.1046/j.1523-1739.2003.01281.x>
- Sanathanan, L. (1972). Estimating the size of a multinomial population. *Annals of Mathematical Statistics*, 43(1), 142-152. <http://dx.doi.org/10.1214/aoms/1177692709>
- Schreuder, H. T., Lin, J. M. S., & Teply, J. (2000). *Estimating the number of tree species in forest populations using current vegetation survey and forest inventory and analysis approximation plots and grid intensities* (pp. 7). Ft Collins CO: USDA Forest Service
- Schreuder, H. T., Williams, M. S., & Reich, R. M. (1999). Estimating the number of tree species in a forest community using survey data. *Environmental Monitoring and Assessment*, 56(2), 293-303. <http://dx.doi.org/10.1023/A:1005988402245>
- Shen, T. J., Chao, A., & Lin, C. F. (2003). Predicting the number of new species in further taxonomic sampling. *Ecology*, 84(3), 798-804. [http://dx.doi.org/10.1890/0012-9658\(2003\)084\[0798:PTNONS\]2.0.CO;2](http://dx.doi.org/10.1890/0012-9658(2003)084[0798:PTNONS]2.0.CO;2)
- Shen, T. J., & He, F. L. (2008). An incidence-based richness estimator for quadrats sampled without replacement. *Ecology*, 89(7), 2052-2060. <http://dx.doi.org/10.1890/07-1526.1>
- Starr, N. (1979). Linear estimation of the probability of discovering a new species. *Annals of Statistics*, 7(3), 644-652. <http://dx.doi.org/10.1214/aos/1176344684>
- Stohlgren, T. J., Chong, G. W., Kalkhan, M. A., & Schell, L. D. (1997). Rapid assessment of plant diversity patterns: A methodology for landscapes. *Environmental Monitoring and Assessment*, 48, 25-43. <http://dx.doi.org/10.1023/A:1005796618823>

- Tackaberry, R., Brokaw, N., Kellman, M., & Mallory, E. (1997). Estimating Species Richness in Tropical Forest: The Missing Species Extrapolation Technique. *Journal of Tropical Ecology*, 13(3), 449-458. <http://dx.doi.org/10.1017/S0266467400010610>
- Thompson, S. K. (1992). *Sampling*. New York: Wiley
- Thrush, S. F., Gray, J. S., Hewitt, J. E., & Ugland, K. I. (2006). Predicting the effects of habitat homogenization on marine biology. *Ecological Applications*, 16(5), 1636-1642. [http://dx.doi.org/10.1890/1051-0761\(2006\)016\[1636:PTEOHH\]2.0.CO;2](http://dx.doi.org/10.1890/1051-0761(2006)016[1636:PTEOHH]2.0.CO;2)
- Ugland, K. I., & Gray, J. S. (2004). Estimation of species richness: analysis of the methods developed by Chao and Karakassis. *Marine Ecology-Progress Series*, 284, 1-8. <http://dx.doi.org/10.3354/meps284001>
- Ugland, K. I., Gray, J. S., & Ellingsen, K. E. (2003). The species-accumulation curve and estimation of species richness. *Journal of Animal Ecology*, 72(5), 888-897. <http://dx.doi.org/10.1046/j.1365-2656.2003.00748.x>
- Ugland, K. I., Gray, J. S., & Lamshead, P. J. D. (2005). Species accumulation curves analysed by a class of null models discovered by Arrhenius. *Oikos*, 108(2), 263-274. <http://dx.doi.org/10.1111/j.0030-1299.2005.13532.x>
- Wagner, H. H. (2003). Spatial Covariance in Plant Communities: Integrating Ordination, Geostatistics, and Variance Testing. *Ecology*, 84(4), 1045-1057. [http://dx.doi.org/10.1890/0012-9658\(2003\)084\[1045:SCIPCI\]2.0.CO;2](http://dx.doi.org/10.1890/0012-9658(2003)084[1045:SCIPCI]2.0.CO;2)
- Wagner, H. H., & Wildi, O. (2002). Realistic simulation of the effects of abundance distribution and spatial heterogeneity on non-parametric estimators of species richness. *Ecoscience*, 9(2), 241-250.
- Waldhardt, R., Simmering, D., & Otte, A. (2004). Estimation and prediction of plant species richness in a mosaic landscape. *Landscape Ecology*, 19(2), 211-226. <http://dx.doi.org/10.1023/B:LAND.0000021722.08588.58>
- Walther, B. A., & Morand, S. (1998). Comparative performance of species richness estimation methods. *Parasitology*, 116, 395-405. <http://dx.doi.org/10.1017/S0031182097002230>
- Whittaker, R. J., & Field, R. (2000). Tree Species Richness Modelling: An Approach of Global Applicability? *Oikos*, 89(2), 399-402. <http://dx.doi.org/10.1034/j.1600-0706.2000.890222.x>
- Winter, S., Chirici, G., McRoberts, R. E., Hauk, E., & Tomppo, E. (2008). Possibilities for harmonizing national forest inventory data for use in forest biodiversity assessments. *Forestry*, 81(1), 33-44. <http://dx.doi.org/10.1093/forestry/cpm042>
- Wolfram, S. (1999). *The Mathematica Book*. Champaign, IL: Wolfram Media / Cambridge University Press
- Zhao, Y. L., & Wall, M. M. (2004). Investigating the use of the variogram for lattice data. *Journal of Computational and Graphical Statistics*, 13(3), 719-738. <http://dx.doi.org/10.1198/106186004X2444>

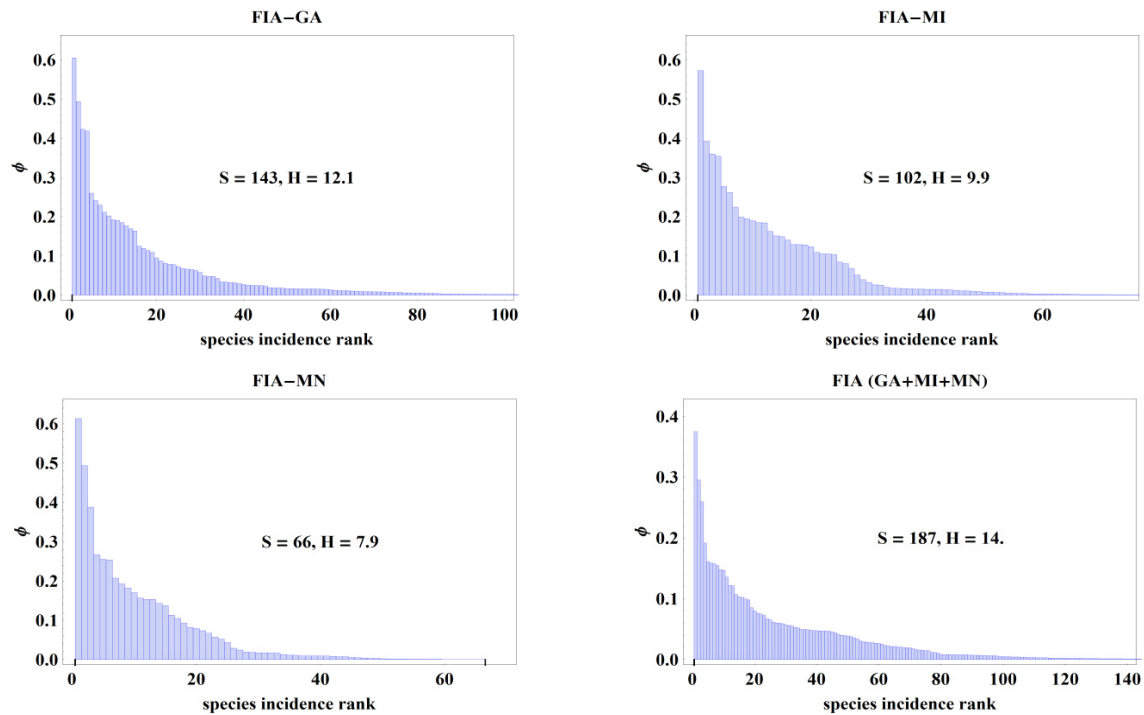


Figure 1. Relative plot incidence (ϕ) of species in FIA sub-regions GA, MI, and MN. Species are ranked (1, ..., S) in descending order of plot-level incidence. H = Shannon-Weaver index of diversity (Pla, 2004)

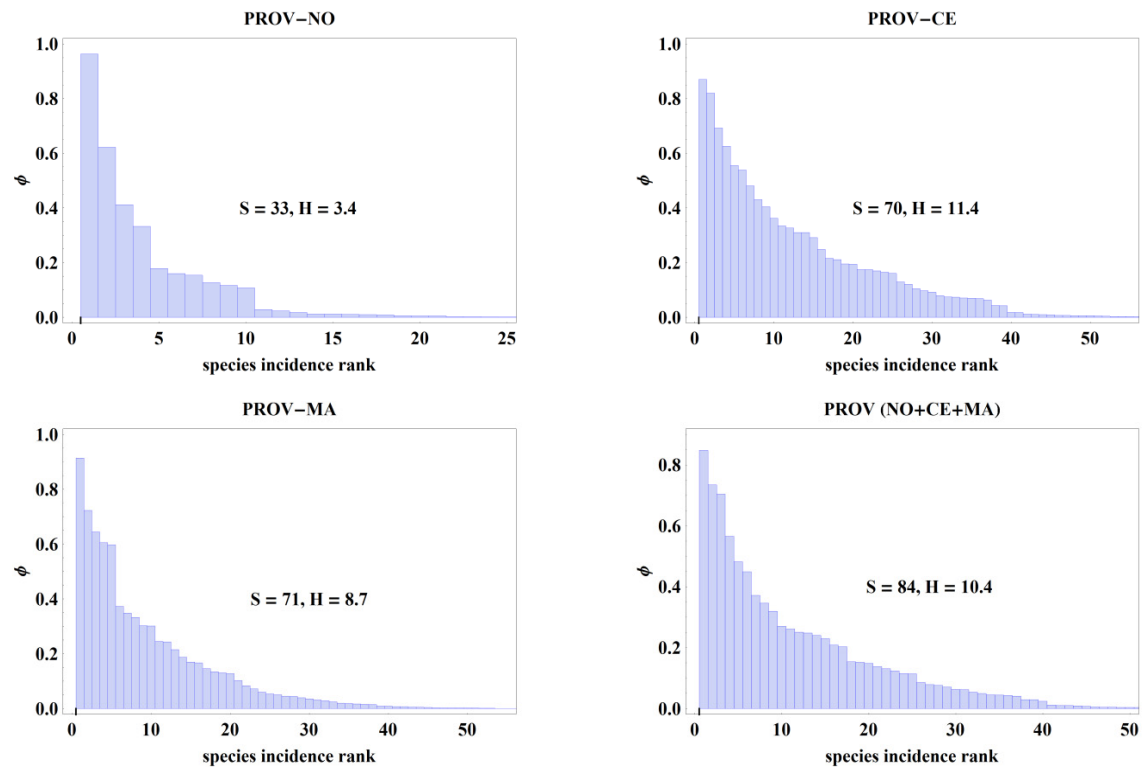


Figure 2. Relative plot incidence (ϕ) of species in PROV sub-regions NO, CE, and MA. Species are ranked (1, ..., S) in descending order of plot-level incidence. H = Shannon-Weaver index of diversity (Pla, 2004)

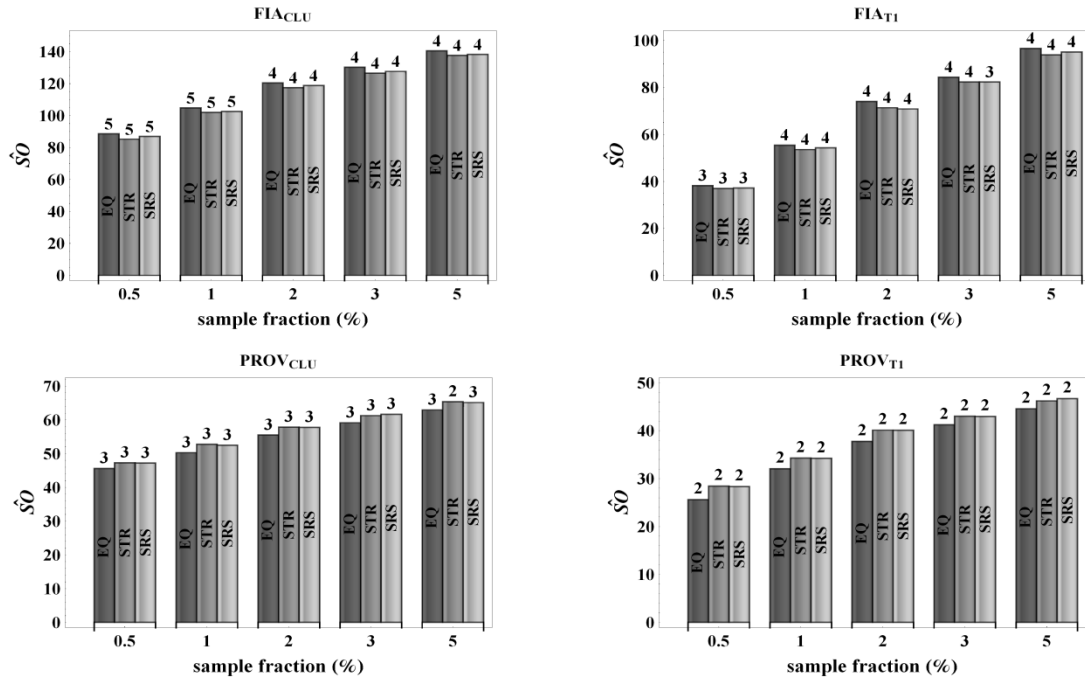


Figure 3. Number of distinct species (\hat{S}_O) observed in regions FIA and PROV samples under three designs (EQ: equal sample allocation to sub-regions, STR: allocation proportional to sub-region size (area), SRS: simple random sampling), and five sample fractions (%). Subscripts: CLU refers to fixed-area plot sampling (multiple trees per plot), and T1 to sampling of one (central) tree per plot. Monte Carlo estimates of sampling error (rounded) are indicated on top of each bar

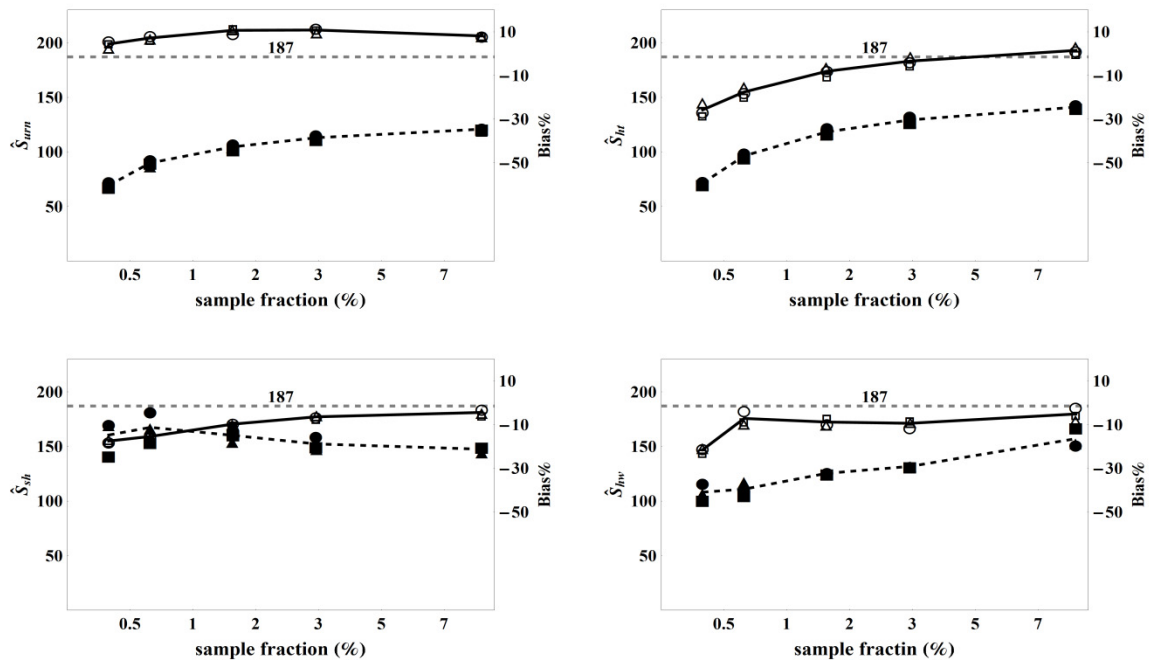


Figure 4. Joint estimation of S in FIA with four estimators of tree species richness, three sampling designs, five sample fractions (%), and two sampling units. Legend: \circ : EQ-CLU, \square : STR-CLU, Δ : SRS-CLU, \bullet : EQ-T1, \blacksquare : STR-T1, \blacktriangle : SRS-T1. Trend lines for CLU (full) and T1 (dashed) are indicated

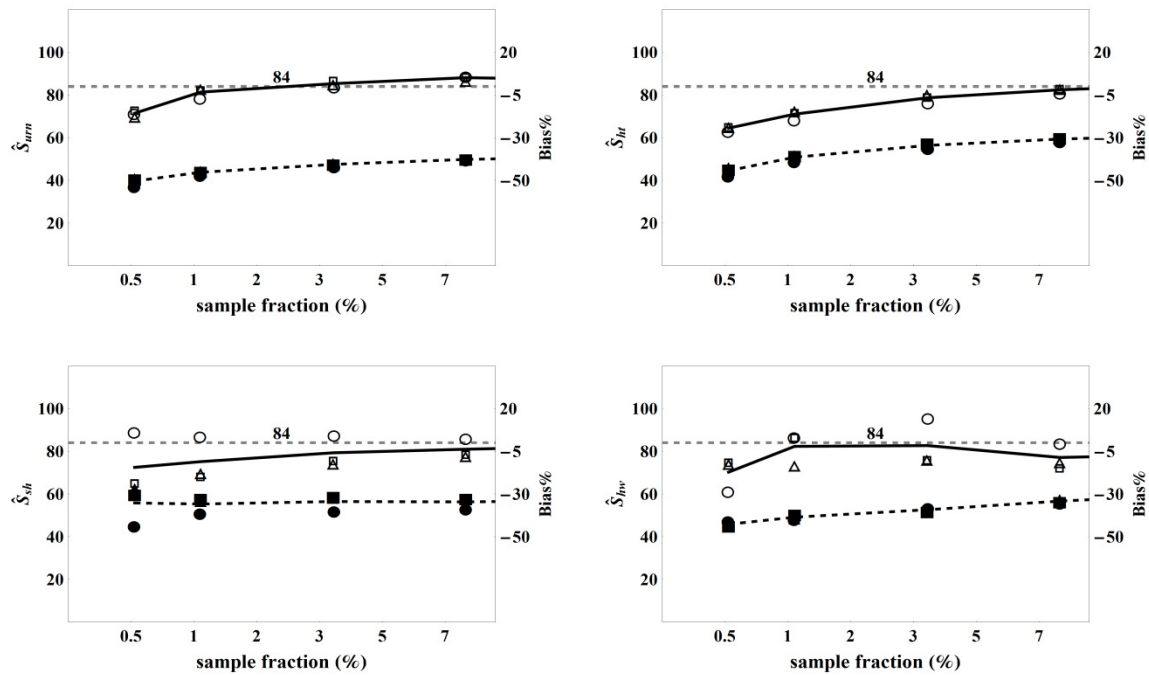


Figure 5. Direct estimation of S in PROV with four estimators, three designs, five sample sizes, and two sampling units. Legend: \circ : EQ-CLU, \square : STR-CLU, Δ : SRS-CLU, \bullet : EQ-T1, \blacksquare : STR-T1, \blacktriangle : SRS-T1 Trend lines for CLU (full) and T1 (dashed) are indicated

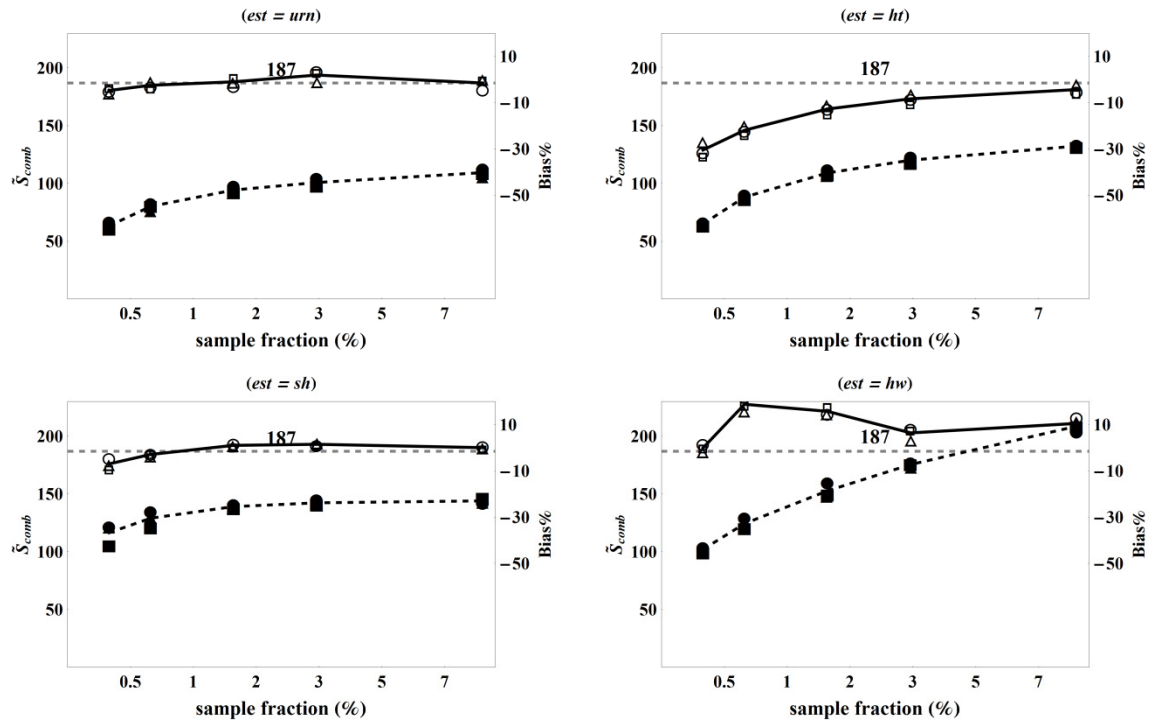


Figure 6. Combined estimation of S in FIA with four estimators of tree species richness, three sampling designs, five sample fractions (%), and two sampling units. Legend: \circ : EQ-CLU, \square : STR-CLU, Δ : SRS-CLU, \bullet : EQ-T1, \blacksquare : STR-T1, \blacktriangle : SRS-T1. Trend lines for CLU (full) and T1 (dashed) are indicated

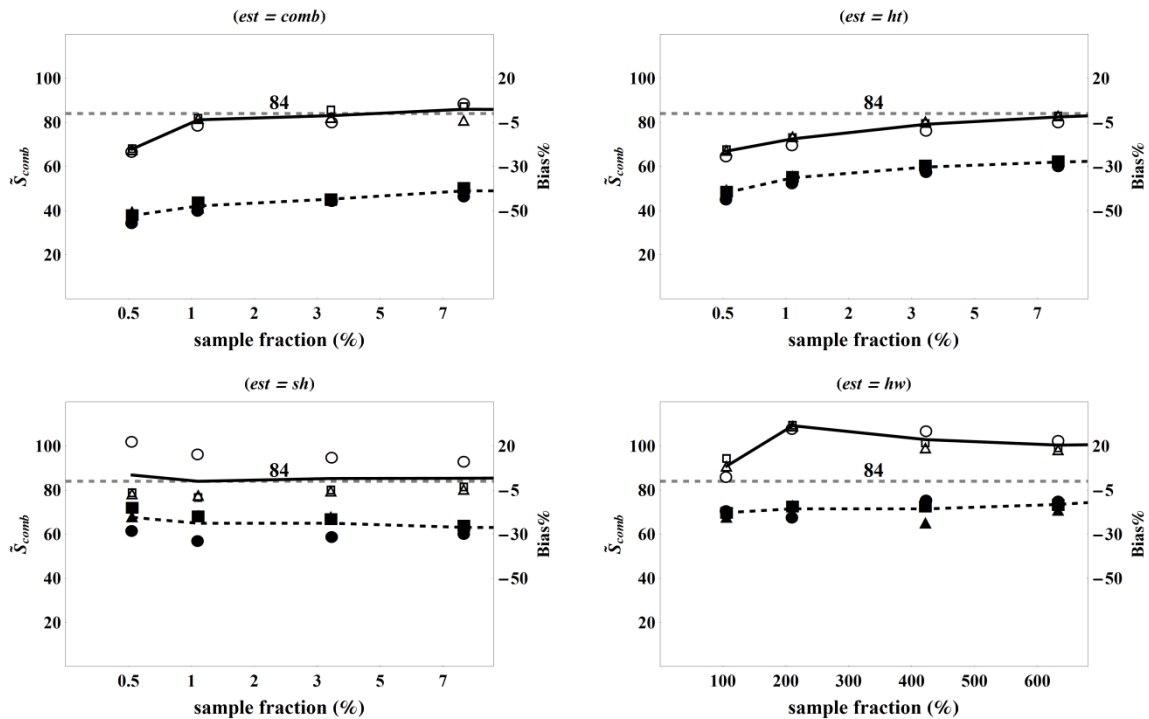


Figure 7. Combined estimation of S in PROV with four estimators of tree species richness, three sampling designs, five sample fractions (%), and two sampling units. Legend: \circ : EQ-CLU, \square : STR-CLU, Δ : SRS-CLU, \bullet : EQ-T1, \blacksquare : STR-T1, \blacktriangle : SRS-T1. Trend lines for CLU (full) and T1 (dashed) are indicated

Appendix

The algorithm for the *urn* estimator by (Magnussen, *et al.*, 2010) is as follows:

Step	Description
0	Place n balls - each representing a survey plot - in an urn. On each ball is written the number of species that was seen in only this plot (singleton species, $SI_i, i = 1, \dots, n$); set $n_{urn} = n$.
1	Set urn species counter equal to the observed species richness in the sample, $S_{urn} = SO$.
2	Draw at random a ball from the urn with index, say, $i^*, i^* = 1, \dots, n_{urn}$. Create a new ball, representing a new plot, with singleton value δSI_{i^*} where δ is a realization of a random Bernoulli variable (0,1). Where: $\begin{cases} \delta = 1 \text{ with prob. } p_{new} \text{ if } SI_{i^*} > 0 \\ \delta = 0 \text{ with prob. } p_{new} \text{ if } SI_{i^*} = 0 \end{cases}$ and $p_{new} = n \times \left\{ n + 1 + \left(1.25 - \text{erf} \left[0.2105 \times \bar{s}^2 \right] \right) (n_{urn} - n) \right\}^{-1}$ with \bar{s} = observed mean number of species per plot. The outcome $\delta = 0 \wedge SI_{i^*} > 0$ constitutes a rediscovery of a singleton species, which therefore loses status as such. Note: $\text{erf}[z]$ is the Gaussian error function, i.e. the integral of a standard normal variable (z) from zero to z . Return two copies of this ball to the urn ($n_{urn} = n + 1$).
3	Update species counter: $S_{urn} = S_{urn} + \delta$.
4	If $n_{urn} < N \wedge \sum_{i=1}^{n_{urn}} SI_i > 0$ GOTO 2.
5	S_{urn} is a single urn-replication prediction. Steps 0-5 are repeated M times to generate a resampling distribution of S_{urn} .

The HT-type estimator of species richness (Magnussen, 2011) is computed from the number of observed species (SO), sample-based estimates of species inclusion probabilities $\hat{\pi}_i, i = 1, \dots, SO$ (Patil & Taillie, 2001), population size (N) and estimated size (number of plots) of species $i = 1, \dots, SO$. In short,

$$(1) \quad \hat{S}_{ht} = \sum_{i=1}^{SO} \left(1 + \sum_{m=1}^{N-n} m \binom{N - \hat{N}_i}{\min[n, N - \hat{N}_i]}^m \binom{N}{n}^{-m} \right) \hat{\pi}_i^{-1}$$

The estimator by Shen and He (Shen *et al.*, 2003)

$$(2) \quad \hat{S}_{sh} = SO \left[\frac{1 - \frac{\Gamma(\hat{\alpha} + \hat{\beta})}{\Gamma(\hat{\beta})} \frac{\Gamma(N + \hat{\beta})}{\Gamma(N + \hat{\alpha} + \hat{\beta})}}{1 - \frac{\Gamma(\hat{\alpha} + \hat{\beta})}{\Gamma(\hat{\beta})} \frac{\Gamma(n + \hat{\beta})}{\Gamma(n + \hat{\alpha} + \hat{\beta})}} \right]$$

where $\hat{\alpha}$ and $\hat{\beta}$ are maximum likelihood estimates of the parameters in a modified beta-distribution of the relative plot-incidence of observed species (see equations 3 and 4 on page 2053).

Hwang and Shen (2010) proposed a generalized jackknife estimator of richness

$$(3) \quad \hat{S}_{hw} = SO + \hat{f}_0$$

where \hat{f}_0 is an estimate of the ‘unseen’ species, which, in turn, is estimated from $\hat{g}(0) f_1 n^{-1}$ with f_1 equal to the number of species occurring in 1 sample plot (singletons). the function $g(k) = (n-k) E[f_k] \{ (k+1) E[f_{k+1}] \}^{-1}$ is approximated by an exponential regression model and estimated from the data via constrained nonlinear least squares regression (see Hwang and Shen for details). $\hat{g}(0)$ is obtained by extrapolation.