Forestry: An International Journal of Forest Research, 2023, 1–13
https://doi.org/10.1093/forestry/cpad041
Original Article

Downloaded from https://academic.oup.com/forestry/advance-article/doi/10.1093/forestry/cpad041/7242317 by University of Minnesota - Twin Cities user on 04 October 2023

Integrating rapid assessment, variable probability sampling, and machine learning to improve accuracy and consistency in mapping local spatial distribution of plant species richness

Bo-Hao Perng[1], Tzeng Yih Lam[1,*], Sheng-Hsin Su[2], Mohamad Danial Bin Md Sabri[3], David Burslem[4], Dairon Cardenas[5], Álvaro Duque[6], Sisira Ediriweera[7], Nimal Gunatilleke[8], Vojtech Novotny[9,10], Michael J. O'Brien[11] and Glen Reynolds[12]

[1]School of Forestry and Resource Conservation, National Taiwan University, Taipei 10617, Taiwan
[2]Fushan Research Center, Taiwan Forestry Research Institute, Yilan 264013, Taiwan
[3]Forestry and Environment Division, Forest Research Institute Malaysia, Kepong, Selangor 52109, Malaysia
[4]The School of Biological Sciences, University of Aberdeen, Aberdeen AB24 3FX, United Kingdom
[5]Instituto Amazónico de Investigaciones Científicas SINCHI, Leticia, Amazonas, Colombia
[6]Department of Forest Sciences, National University of Colombia, Medellín 050034, Colombia
[7]Faculty of Applied Sciences, Uva Wellassa University, Badulla 90000, Sri Lanka
[8]University of Peradeniya, Peradeniya 20400, Sri Lanka
[9]Biology Centre, Institute of Entomology of the Czech Academy of Sciences, České Budějovice 37005, Czech Republic
[10]Faculty of Sciences, University of South Bohemia, České Budějovice 37011, Czech Republic
[11]Estación Experimental de Zonas Áridas, Consejo Superior de Investigaciones Científicas, Almería 04120, Spain
[12]Danum Valley Field Centre, South East Asia Rainforest Research Partnership (SEARRP), Lahad Datu, Sabah 91112, Malaysia
*Corresponding author. School of Forestry and Resource Conservation, National Taiwan University, Taipei 10617, Taiwan. E-mail: tylam.forest@gmail.com

Abstract

Conserving plant diversity is integral to sustainable forest management. This study aims at diversifying tools to map spatial distribution of species richness. We develop a sampling strategy of using rapid assessments by local communities to gather prior information on species richness distribution to drive census cell selection by sampling with covariate designs. An artificial neural network model is built to predict the spatial patterns. Accuracy and consistency of rapid assessment factors, sample selection methods, and sampling intensity of census cells were tested in a simulation study with seven 25–50-ha census plots in the tropics and subtropics. Results showed that identifying more plant individuals in a rapid assessment improved accuracy and consistency, while transect was comparable to or slightly better than nearest-neighbor assessment, but knowing more species had little effects. Results of sampling with covariate designs depended on covariates. The covariate I_{freq}, inverse of the frequency of the rapidly assessed species richness strata, was the best choice. List sampling and local pivotal method with I_{freq} increased accuracy by 0.7%–1.6% and consistency by 7.6%–12.0% for 5% to 20% sampling intensity. This study recommends a rapid assessment method of selecting 20 individuals at every 20-m interval along a transect. Knowing at least half of the species in a forest that are abundant is sufficient. Local pivotal method is recommended at 5% sampling intensity or less. This study presents a methodology to directly involve local communities in probability-based forest resource assessment to support decision-making in forest management.

Keywords: biodiversity conservation; design-based sampling; forest inventory; rapid biodiversity assessment; species diversity; variable probability sampling

Introduction

Plant diversity is an important forest ecosystem service that benefits human society (Gascon et al. 2015). A common indicator of plant diversity is species richness, which is defined as the total number of plant species in a forest. Species richness is formally adopted by the Montréal Process and the Helsinki Process as an indicator for sustainable forestry (Hall 2001). Mapping the spatial distribution of plant species richness improves understanding of plant community spatial and temporal changes, helps designate nature reserves, and supports forest landscape management decisions (Pearson and Carroll 1998, Devictor et al. 2010, Villero et al. 2017). Producing such a map is challenging because it is impossible to enumerate all plant individuals in a forest. Thus, species richness assessments often rely on sampling to survey plant species richness in some parts of a forest and on statistical models to predict richness on unsampled areas. Thus, building a cost-effective, precise, and probability-based plant diversity inventory system is essential for multipurpose management of forest resources.

Ground plots for a plant diversity survey are usually low in numbers and sparsely distributed (Chong et al. 2001, Haas et al. 2006). Thoughtful placement of ground plots could yield necessary information for management decisions while keeping cost manageable. Basu (1969) stated that a sampling design would be efficient if selection probabilities were conditional on the parameters of interest. In case of surveying species diversity, selection probability of a sampling design would ideally be conditional on

Handling editor: Dr. Rubén Manso
Received: December 1, 2022. **Revised:** May 28, 2023. **Accepted:** July 27, 2023
© The Author(s) 2023. Published by Oxford University Press on behalf of Institute of Chartered Foresters. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com.

true species richness so that an area with high species richness would be more likely sampled, and vice versa. The dilemma is that prior knowledge of true species richness necessary to drive sample selection is usually unavailable. Proxies derived from remote sensing are often used (Pau et al. 2012, Coops et al. 2019). However, there is evidence suggesting that correlation between spectral variability in remotely sensed images and species richness is only subtle (Fassnacht et al. 2022). Alternatively, we propose a rapid assessment strategy that engages local communities familiar with their forests to gather prior information on the spatial distribution of species richness using their knowledge. This prior information is then used to drive sample selection by a probability-based sampling design.

There were very few studies on using knowledge to guide sampling of plant diversity. Goff et al. (1982) introduced the Timed Meandered Search (TMS) method to locate sample points based on field judgment. However, it was not a probability-based design. Lam et al. (2018) developed a procedure for assessing plant diversity by applying Probability Proportional to Prediction (3P) sampling (Grosenbaugh 1964, 1965) to knowledge used in a rapid assessment. According to their procedures, a forest is first tessellated into non-overlapping cells of equal size. Local communities visit a cell and rapidly select and identify a few plant individuals based on their knowledge. Once all cells are rapidly assessed, the number of species in each cell is tallied. The 3P sampling is used to select a subsample of cells for species census with selection probability proportional to the species tally from the rapid assessment. The species census could be carried out by professionals such as a botanist. With both rapid assessment and census information, Lam et al. (2018) assessed accuracy and efficiency in predicting average cell-level species richness. This study applies their sampling methods but expands their work to map spatial distribution of species richness across a forest with additional datasets from the tropical regions and other sampling with covariate designs.

Sampling with covariate designs are a general class of sampling designs that use auxiliary information (covariate) to increase efficiency in forest inventory (Kershaw Jr. et al. 2016). They are also known as sampling design based on auxiliary information in Särndal et al. (1992). The 3P sampling is one of them. Many studies on using sampling with covariate designs focus on estimating primary forest products. For example, Yang et al. (2019) found that sampling with covariate designs with covariates extracted from an airborne Light Detection and Ranging (LiDAR) system combined with estimation models improved estimation of forest volume. Hsu et al. (2020) applied sampling with covariate designs to correct for local bias in LiDAR-assisted estimates of forest volume in small woodlots. However, Iles (2003) noted that a covariate could be any as long as it was highly correlated with the variable of interest. Thus, this supports the use of sampling with covariate designs in any context, especially in our case of mapping distribution of species richness.

Spatial modeling requires modeling trends in observed data, estimating underlying model parameters, and predicting at unobserved locations (Banerjee et al. 2015). Artificial neural network (ANN) is a machine learning framework that adapts its internal structures to external information and is capable of solving a complex nonlinear problem (McRoberts et al. 1991). In the past decades, ANN has been applied in forestry such as predicting occurrence of rare plant species (Williams et al. 2009), modeling volume increment (Bayat et al. 2021), and mapping stand structures (Ingram et al. 2005). In particular, Foody and Cutler (2006) applied ANN to map species richness and composition of a tropical forest in Malaysia with remote sensing. They found that ANN produced strong correlation between predicted and observed values. Furthermore, ANN has been found to outperform other machine learning approaches such as support vector machine and random forest in predicting occurrence of disease (Hung et al. 2017), mapping species richness (Choe et al. 2021), and monitoring of water quality (Hafeez et al. 2019). Taking advantage of its flexibility, this study integrates ANN with the above sampling procedures to predict spatial distribution of species richness across a forest.

With the anthropogenic influence on the world forests, Pimm and Raven (2000) warned that many plant species could disappear before they were known. Hence, there is an urgent need to innovate tools for conserving biodiversity and improving forest planning to meet complex management objectives. Local knowledge is increasingly recognized to benefit biodiversity survey (Walker et al. 1995). Applying local knowledge with sampling with covariate designs could be cost-effective by potentially reducing the cost of biodiversity census. Actively engaging local communities in forest planning could also lead to their social and economic well-being (FSC 2012). The overall goal of this simulation study was to assess a probability-based plant diversity inventory system integrating rapid assessment with local knowledge, sampling with covariate designs, and a machine learning technique for mapping distribution of species richness across a forest. The three specific objectives were (i) to understand how amount of knowledge and rapid assessment methods affect accuracy and consistency in mapping species richness, (ii) to compare accuracy and consistency between sampling with covariate designs and between covariates, and (iii) to evaluate trade-off between number of census cells and accuracy and consistency.

Methods

Data

Seven long-term Forest Dynamics Census Plots from the Forest Global Earth Observatory Network (ForestGEO; forestgeo.si.edu) were used in this study (hereafter as sites; Table 1). Six of the sites were in the tropics with one in the subtropics. All sites were either 25 or 50 ha in area with identical plant census protocols. All woody plants with diameter at breast height ≥1 cm were mapped, measured, and identified to species level. In this study, only the main stem of an individual was retained. Species richness of the seven sites ranged from 110 to 1233.

Rapid assessment method

Following Lam et al. (2018), each site was first tessellated into non-overlapping equally sized 20 × 20 m (0.04 ha) cells (Fig. S1a; Supplementary Materials). The total number of cells (N) of the seven sites was either 625 or 1250 with cell average species richness from 31 to 127 (Table 1). Rapid assessment of plant diversity was carried out in all N cells of a site (hereafter as rapid cells). Three factors were considered when simulating rapid assessment: (i) amount of knowledge (KN), (ii) rapid assessment effort (RE), and (iii) rapid assessment type (RT). KN simulated how much a local community know about the species in a forest, and indirectly, their ability to identify a species. KN had three levels: (i) 50% (KN50), (ii) 75% (KN75), and (iii) 100% (KN100) of the total number of species in a site, i.e. knowing 50%, 75%, or all the species in a site. To simulate KN50 and KN75, species were randomly selected without replacement and with probability proportional to their total abundance in a site. This assumed that a local community was more familiar with a locally abundant species than a rare one.

Downloaded from https://academic.oup.com/forestry/advance-article/doi/10.1093/forestry/cpad041/7242317 by University of Minnesota - Twin Cities user on 04 October 2023

Table 1. Characteristics of the seven long-term forest dynamics census plots.

Plot	Country	Ecological zone	Plot area (ha)	Plot plant count	Plot species richness	Cell count (*N*)	Cell species richness	Cell plant count	Year	Reference
Amacay-acu	Colombia	Tropical	25	116 570	1233	625	102 (23.4%)	187 (23.9%)	2007	Duque et al. (2017), Zuleta et al. (2020)
Pasoh	Malaysia	Tropical	50	335 343	820	1250	127 (11.9%)	268 (19.3%)	1985	Manokaran and LaFrankie (1990)
Danum	Malaysia	Tropical	50	233 408	693	1250	71 (19.6%)	187 (21.9%)	2011	O'Brien et al. (2022)
Wanang	Papua New Guinea	Tropical	50	253 653	581	1250	84 (18.8%)	192 (28.1%)	2009	Vincent et al. (2015)
BCI	Panama	Tropical	50	221 758	302	1250	52 (17.2%)	177 (19.9%)	2010	(Hubbell et al. 1999; Condit et al. 2019a, 2019b, 2019c)
Sinharaja	Sri Lanka	Tropical	25	207 469	238	625	52 (19.8%)	332 (37.5%)	1996	Anderson--Teixeira et al. (2015)
Fushan	Taiwan	Subtropical	25	114 354	110	625	31 (26.0%)	183 (44.3%)	2004	Su et al. (2007)

Ecological zone is defined by the ForestGEO classification (forestgeo.si.edu). Plot plant count = total number of plant individuals in a plot. Plot species richness = total number of species in a plot. Cell count = total number of non-overlapping 20 × 20 m cells in a plot. Cell species richness = average number of species in a cell with its coefficient of variation in parentheses. Cell plant count = average number of plant individuals in a cell with its coefficient of variation in parentheses. Year = year of data collection.

RE implied the amount of resources invested in a rapid assessment. It was represented by the number of individuals selected for rapid assessment in a cell: (i) 10 (RE10), (ii) 20 (RE20), and (iii) 40 (RE40) plant individuals. RT mimicked the ways of selecting individuals for rapid assessment in a cell. Three types of RT were simulated: (i) random walk (RTrw), (ii) nearest-neighbor (RTnn), and (iii) transect (RTtr). RTrw mimicked selecting individuals by randomly walking about in a cell. It was simulated by randomly selecting individuals in a cell with equal probability (Fig. S1b). RTnn was analogous to the nearest-neighbor sampling (Kershaw Jr. et al. 2016). First, a location in a cell was randomly selected, and then, a number of individuals closest to the location were selected (Fig. S1c). RTtr mimicked selecting individuals when traversing a transect. To simulate RTtr, transects were first randomly placed across rows of cells in a site, i.e. placement of transects was independent between rows (Fig. S1d). The transect in each cell was then subdivided into five equal intervals (Fig. S1d). In each interval, one-fifth of RE (RE/5) individuals closest to the transect in perpendicular distance on either side were selected (Fig. S1d). In summary, there were a total of 27 combinations of rapid assessment methods (3 KN × 3 RE × 3 RT).

Sampling with covariate design

Five sampling designs were used to subsample rapid cells for census of plant species (hereafter as census cells): (i) simple random sampling (SRS), (ii) spatial systematic sampling (SYS), (iii) systematic sampling from an ordered list (SOL), (iv) list sampling (LIST), and (v) local pivotal method (LPM). Subsampling was without replacement at eight sampling intensities: 1%, 5%, 10%, 15%, 20%, 25%, 30%, and 40% of *N* rapid cells in a site. Let *n* be the number of census cells. SRS and SYS were sample selection methods without covariate. For SRS, *n* census cells were randomly selected with equal probability. For SYS, a two-dimensional systematic grid of *n* census cells was established with grid size determined by *n* and the dimensions of a site. A rapid cell was first randomly selected as the starting census cell (Iles 2003), and the remaining $n - 1$ census cells were spread out according to the grid size.

SOL, LIST, and LPM are sampling with covariate designs. SOL was equal probability design, but LIST and LPM were variable probability designs (Yang et al. 2019). For SOL, rapid cells were first ordered by the increasing values of a covariate. If two or more rapid cells shared the same values, they were randomly ordered. A rapid cell was randomly selected as the starting census cell (Iles 2003), and the remaining $n - 1$ census cells were systematically selected at an interval of N/n (Kershaw Jr. et al. 2016). For LIST, rapid cells were first arranged into a list in any order. Cumulative sum of a covariate was sequentially calculated for each rapid cell down the list. Each rapid cell was then assigned two numbers: the cumulative sum of the rapid cell before it in the list and its cumulative sum. A random number was drawn between 0 and the maximum cumulative sum. A rapid cell with the random number falling between its two numbers was selected as the census cell (Yang et al. 2019). This process was repeated *n* times. Grafström et al. (2012) developed the local pivotal method (LPM) so that several covariates could be used to assign selection probability to a sample unit. LPM spread census cells evenly across a multidimensional space formed by multiple covariates, which might include spatial coordinates. LPM assigned either equal or unequal inclusion probabilities to the rapid cells. A measure of distance was calculated between two rapid cells in the multidimensional space. A set of updating rules on inclusion probability was used to select rapid cells as the final set of census cells. A rapid cell was first randomly selected, and its nearest rapid cell was identified. Inclusion probabilities of the two rapid cells were updated according to the rules. This process was repeated until the sampling outcome was decided for one of them (Grafström et al. 2012). In general, two rapid cells that were closed to each other had small joint inclusion probability (Grafström and Ringvall 2013), which lowered the probability of selecting both as census cells. Hence,

Downloaded from https://academic.oup.com/forestry/advance-article/doi/10.1093/forestry/cpad041/7242317 by University of Minnesota - Twin Cities user on 04 October 2023

this negative correlation created a well-spread sample (Grafström et al. 2012).

Two covariates were extracted from each rapid cell: (i) rapid richness (S_{rapid}) and (ii) inverse of the frequency of the rapid richness stratum that it belonged to (I_{freq}). S_{rapid} was defined as the number of species identified from rapid assessment. We assumed that S_{rapid} reflected true richness in a rapid cell. A rapid cell with higher S_{rapid} was more likely to be species rich, and these cells should be chosen for census with a greater probability. I_{freq} was derived from post-stratifying rapid cells by their S_{rapid}. In post-stratification, rapid cells with the same S_{rapid} were grouped into a stratum. The frequency (count) of rapid cells in each stratum was calculated. I_{freq} for a rapid cell was the inverse of the frequency of the stratum that it belonged to. For example, a stratum with S_{rapid} of 2 species had 50 rapid cells in it. Then, I_{freq} for every rapid cell in that stratum was 1/50. The maximum number of strata would be RE. Basically, I_{freq} assigned higher selection probability to rapid cells in a stratum with lower frequency. S_{rapid} was used as a covariate in SOL, LIST, and LPM, while I_{freq} was used in LIST and LPM. Regardless of the covariate, LPM included cell coordinates defined as the row and column indices of rapid cells as an additional set of covariates to achieve spatial balance. For LPM(S_{rapid}), equal inclusion probability was assigned to the rapid cells in the space of S_{rapid}. For LPM(I_{freq}), unequal inclusion probability proportional to I_{freq} was assigned to the rapid cells. In summary, seven sampling designs were simulated: SRS, SYS, SOL, LIST(S_{rapid}), LIST(I_{freq}), LPM(S_{rapid}), and LPM(I_{freq}).

Artificial neural network

A dataset compiled from both rapid assessment and census information was separated into two sets of data: a modeling dataset and a prediction dataset. The modeling dataset consisted of n census cells with both the rapid assessment and census information. The modeling dataset was used to train an ANN model. Information extracted from each census cell in the modeling dataset were S_{rapid}, rapid Shannon index, rapid species list, cell coordinates, and S_{census}. Rapid Shannon index was the Shannon diversity index (Magurran 2004) calculated with species information from the rapid assessment. Rapid species list was a list of species name from the rapid assessment. S_{census} was the total number of species found in a cell from the census. The prediction dataset consisted of $N - n = m$ rapid cells that were not censused, i.e. having only information from the rapid assessment. The prediction dataset was used to assess the performance of the final ANN model.

An ANN model consisted of three main layers arranged in a sequence: an input layer, a network of hidden layers, and an output layer. The input layer organized input data and fed them to the network of hidden layers for model construction. In this study, the input data consisted of S_{rapid}, rapid Shannon index, rapid species list, and cell coordinates from the modeling dataset. The network of hidden layers consisted of five hidden layers with 64 artificial neurons per layer. An artificial neuron was governed by an activation function. An activation function performed pointwise nonlinear transformation of input data into a 'signal'. If the 'signal' was strong enough, the neuron fired its outputs to the neurons in the next hidden layer. This study applied the Mish activation function (Misra 2020). When everything passed through the five hidden layers and an initial ANN model was built, the output layer predicted species richness (S_{pred}) of each n census cells.

At the next step, S_{pred} was compared to S_{census} for the n census cells. The differences were used to compute loss by the Huber loss function (Huber 1964). Gradient of the computed loss was calculated by taking partial derivatives. It was then propagated through the network of hidden layers of the initial ANN model using the backward propagation algorithm from Rumelhart et al. (1986). The Adam gradient descent algorithm from Kingma and Ba (2015) was used to update the parameters in the hidden layers with the backward propagated gradient. The newly updated ANN model was trained again with the same input data. The training-updating process was repeated 50 times to produce the final ANN model. During the training-updating process, two regularizer algorithms (Pereyra et al. 2017) and a dropout algorithm from Srivastava et al. (2014) were applied to avoid model overfitting. Pasupa and Sunhem (2016) and Olson et al. (2018) found that regularization techniques and dropout algorithms were effective in handling model overfitting from small datasets, which would be useful in this study due to the low sampling intensity of census cells. A 4-fold cross-validation was also carried out to assess model performance and for selecting optimal hyperparameters. For the cross-validation, the modeling dataset was randomly split into four groups. Three groups were used for training the model, and one group was used for validation. The cross-validation process was repeated four times.

Lastly, the final ANN model was used to predict species richness of the m rapid cells in the prediction dataset. S_{rapid}, rapid Shannon index, rapid species list, and cell coordinates of the m rapid cells were first extracted from the prediction dataset. These input data were fed into the final ANN model. It then predicted S_{pred} for each m rapid cell for further statistical analyses. The above process of model building and prediction is summarized in Fig. S2b (Supplementary Materials).

Simulation

A total of 1512 combinations from the 27 rapid assessment methods, the 7 sampling designs, and the 8 sampling intensities were simulated. Simulation was independently carried out for each site. At first, each of the 27 rapid assessment methods was independently simulated. For the rapid assessment methods with KN50 and KN75, a list of known species was generated as described above. When simulating the rapid assessment with KN50 and KN75, a selected individual whose species did not match those in the list (i.e. 'unknown' species) was dropped. However, the dropped individual was not replaced by selecting a new individual. Thus, the actual number of individuals selected in a cell after the simulation could be lower than the designated RE. We believed this better reflected field operation and the effects of incomplete knowledge. Next, for each of the simulated 27 rapid assessment method, 56 combinations of sampling design and sampling intensity were simulated. In other words, the 56 combinations of sampling design and sampling intensity were nested under a rapid assessment method. This process of first simulating the rapid assessment methods and then the sampling designs and intensities was repeated 100 times. With the 100 iterations, the list of known species was independent between iterations, i.e. the list could be different between iterations. Additionally, the 56 combinations of sampling design and sampling intensity were also independent between iterations. The simulation process is summarized in Fig. S2a. When all simulations were completed, ANN model building and prediction were carried out for each iteration of the 1512 combinations. The simulation was carried out in Python.

Statistical analysis

This study assessed accuracy and consistency of the 1512 combinations in predicting spatial distribution of species richness in

Downloaded from https://academic.oup.com/forestry/advance-article/doi/10.1093/forestry/cpad041/7242317 by University of Minnesota - Twin Cities user on 04 October 2023

m rapid cells in a prediction dataset. Let k be the k-th rapid cell in a prediction dataset, $k = 1, \ldots, m$. Percent absolute bias (PAB) for the k-th rapid cell was calculated as $PAB_k = \left|S_{pred,k} - S_{true,k}\right| / S_{true,k} \cdot 100\%$, where $S_{true,k}$ was the true number of species found in the k-th rapid cell. PAB accounted for both over- and underestimation of S_{true}. It also made comparison between sites comparable due to differences in species richness. Relative accuracy (Q) for the k-th rapid cell was calculated as the ratio of $S_{pred,k}$ to $S_{true,k}$, $Q_k = S_{pred,k}/S_{true,k}$. $Q_k > 1$ implied overestimation with $Q_k < 1$ indicating underestimation.

Accuracy was represented by the mean percent absolute bias (MPAB). For the i-th iteration, MPAB was calculated by averaging PAB of all m rapid cells, $MPAB_i = \sum_{k=1}^{m} PAB_{ik}/m$. A higher MPAB indicated lower accuracy, and vice versa. Consistency was represented by the coefficient of variation in relative accuracy (CVQ). CVQ was the ratio of the standard deviation of relative accuracy to the mean of relative accuracy expressed in percentage, $CVQ = SDQ/MQ \cdot 100\%$. For the i-th iteration, the standard deviation of relative accuracy (SDQ) was calculated as $SDQ_i = \sqrt{\sum_{k=1}^{m} (Q_{ik} - MQ_i)^2/(m-1)}$, and the mean of relative accuracy (MQ) was calculated as $MQ_i = \sum_{k=1}^{m} Q_{ik}/m$. A lower CVQ indicated higher consistency because relative accuracy between m rapid cells was more similar. In other words, a lower CVQ implied that rapid cells with high true species richness were predicted so, and vice versa. CVQ was useful to forest planning if a spatial map of relative species richness was sufficient for management decision-making without the need for unbiased prediction of species richness.

For Objective 1, the effects of the rapid assessment methods were represented by the means of MPAB and CVQ calculated across all sites and iterations for each of the 1512 combinations (27 rapid assessment methods × 7 sampling design × 8 sampling intensity). To supplement the results of the effect sizes (Sullivan and Feinn 2012), a three-way analysis of variance (ANOVA) model was developed. The ANOVA model was fitted with sites treated as blocks to control for underlying variations between sites and was fitted separately to each of the 56 combinations of sampling design and sampling intensity as follows,

$$Y = \mu + site + KN + RE + RT + (KN \times RE) + (KN \times RT) + (RE \times RT) + (KN \times RE \times RT) \quad (1)$$

where Y was either MPAB or CVQ.

For Objective 2, SRS was assigned as the baseline. Differences in MPAB and CVQ between a sampling design and SRS were calculated as follows: within a site, for the j-th rapid assessment method at the i-th iteration (Fig. S2a), the difference in MPAB (dMPAB) and CVQ (dCVQ) was calculated for each of the 56 combinations of sampling design and sampling intensity as

$$dY_{ij,si,sdesign} = Y_{ij,si,sdesign} - Y_{ij,si,SRS} \quad (2)$$

where dY was either dMPAB or dCVQ, $i = 1, \ldots, 100$ iterations, $j = 1, \ldots, 27$ rapid assessment methods, si = sampling intensity (1%, 5%, 10%, 15%, 20%, 25%, 30%, 40%), and sdesign = sampling design (SYS, SOL, LIST(S_{rapid}), LIST(I_{freq}), LPM(S_{rapid}), and LPM(I_{freq})). The dY was calculated in this manner to control for underlying variations between sites and the nesting structure of sampling design and sampling intensity under a rapid assessment method (Fig. S2a). Lastly, all abbreviations above are provided in the Supplementary Materials.

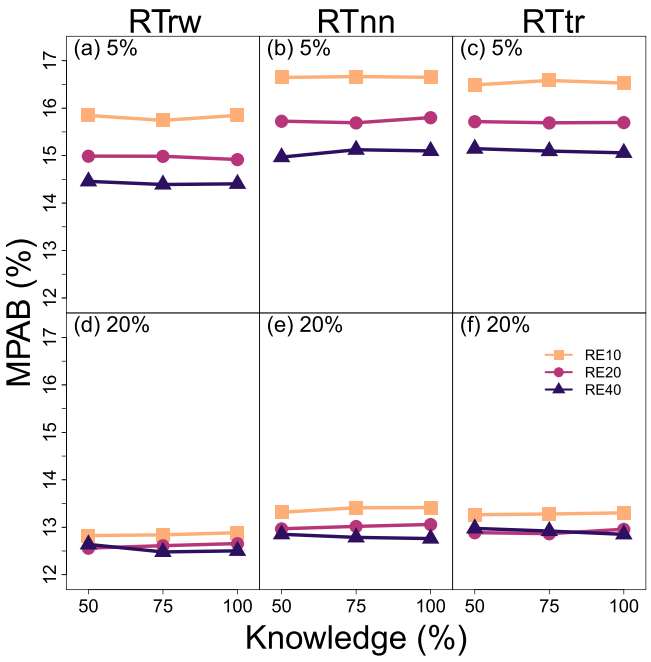

Figure 1. MPAB of the 27 combinations of KN, RE, and RT for SRS at (a–c) 5% and (d–f) 20% sampling intensity. The rapid assessment efforts are 10 (RE10), 20 (RE20), and 40 (RE40) individuals. The rapid assessment types are random walk (RTrw), nearest-neighbor (RTnn), and transect (RTtr). MPAB is averaged across all sites and iterations.

Results

Rapid assessment method

With SRS as the baseline example, there were distinct effects of RE and RT but not KN on the accuracy (MPAB) at both 5% and 20% sampling intensity (Fig. 1). Increasing RE decreased MPAB suggesting an increase in accuracy. At 5% sampling intensity, reduction in MPAB was about 1.5% when RE increased from 10 to 40 individuals, e.g. MPAB decreased from 16.5% to 14.8% on average for RTrw (Fig. 1a–c). At 20% sampling intensity, MPAB was reduced by only about 0.5%, and accuracy of RE40 was either comparable to or slightly higher than RE20 (Fig. 1d–f). RTrw was consistently the most accurate with the lowest MPAB. RTrw reduced MPAB by about 1% and 0.5% on average compared to RTtr and RTnn at 5% and 20% sampling intensity, respectively (Fig. 1). In contrast, RTtr was either comparable to or slightly more accurate than RTnn. KN had no visible effect on accuracy as MPAB was similar across KN levels. The effects of RE, RT, and KN were consistently observed across sampling designs and sampling intensities (Figs. S3–S10; Supplementary Materials). Overall, a rapid assessment method with any KN × RE40 × RTrw was either the most accurate or comparably so. Lastly, the ANOVA models supported the results suggesting significant effects of RE and RT ($P < .001$) and generally insignificant effects of KN ($P > .05$) on accuracy (Fig. S11; Supplementary Materials).

Under SRS, there were noticeable effects of RT on the consistency (CVQ), but the effects of RE and KN were less straightforward (Fig. 2). Increasing RE increased consistency by reducing CVQ, but it depended on sampling intensity. At 5% sampling intensity, reduction in CVQ ranged from 1.3% to 2.6% when RE increased from 10 to 40 individuals, e.g. CVQ decreased from 29.9% to 28.6% on average for RTrw (Fig. 2a–c). On the contrary, at 20% sampling intensity, CVQ was comparable between RE (28.6%–29.5%; Fig. 2d–f) except for RE40 × RTtr, which was slightly higher

Downloaded from https://academic.oup.com/forestry/advance-article/doi/10.1093/forestry/cpad041/7242317 by University of Minnesota - Twin Cities user on 04 October 2023

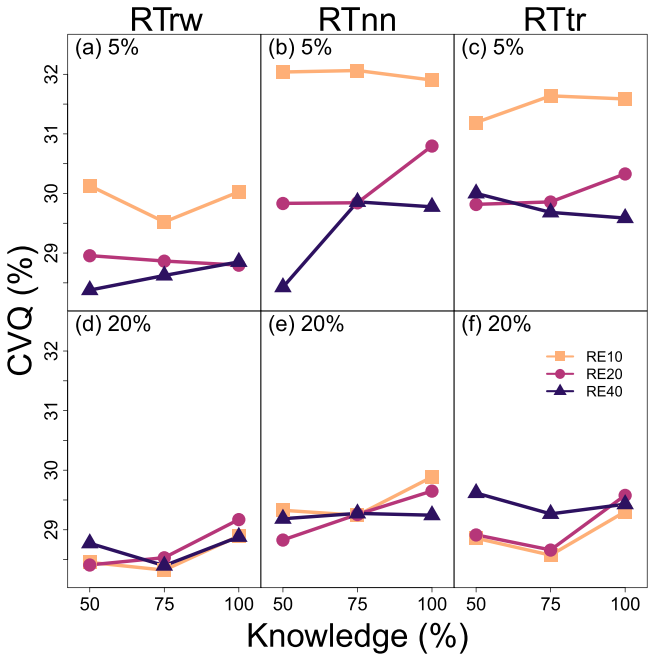

Figure 2. CVQ of the 27 combinations of KN, RE, and RT for SRS at (a–c) 5% and (d–f) 20% sampling intensity. The rapid assessment efforts are 10 (RE10), 20 (RE20), and 40 (RE40) individuals. The rapid assessment types are random walk (RTrw), nearest-neighbor (RTnn), and transect (RTtr). CVQ is averaged across all sites and iterations.

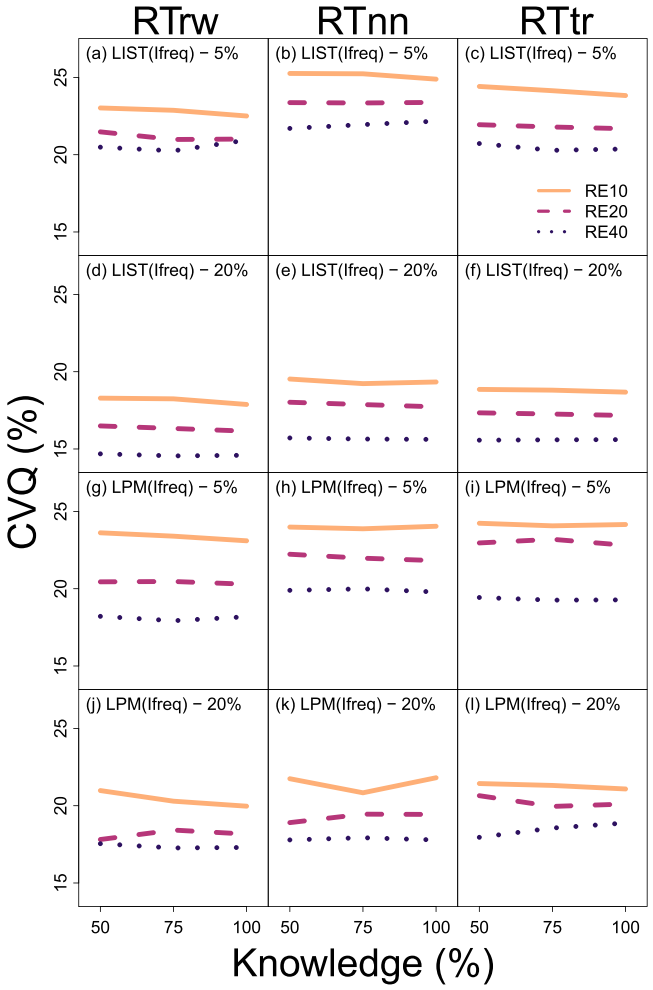

Figure 3. CVQ of the 27 combinations of KN, RE, and RT for (a–f) LIST(I_{freq}) and (g–l) LPM(I_{freq}) at 5% and 20% sampling intensity. The rapid assessment efforts are 10 (RE10), 20 (RE20), and 40 (RE40) individuals. The rapid assessment types are random walk (RTrw), nearest-neighbor (RTnn), and transect (RTtr). CVQ is averaged across all sites and iterations.

at KN50 and KN75 (Fig. 2f). Similar to MPAB, RTrw was the most consistent with the lowest CVQ. At 5% sampling intensity, RTrw reduced CVQ by 0.8% to 2.1% on average compared to RTtr and RTnn (Fig. 2a–c), but the reduction was less than 1% at 20% sampling intensity (Fig. 2d–f). Also, RTtr was either comparable to or slightly more consistent than RTnn. KN did not have an apparent effect on consistency mainly due to large variation in CVQ. For example, at 5% sampling intensity, the range in CVQ was 28.4%–32.0%, 28.6%–32.1%, and 28.8%–31.9% for KN50, KN75, and KN100, respectively (Fig. 2a–c). As a result, averages of CVQ were similar across KN. In contrast, results of LIST(I_{freq}) and LPM(I_{freq}) were different to SRS (Fig. 3), but they resembled the trends in MPAB (Fig. 1). At 5% and 20% sampling intensity, there were distinct effects of RE and RT on consistency but not KN (Fig. 3). The results were consistent across sampling intensities (Figs. S12–S19; Supplementary Materials). Overall, a rapid assessment method with any KN × RE40 × RTrw was either the most consistent or comparably so. Lastly, the ANOVA models supported the results suggesting significant effects of RE for most sampling designs and sampling intensities, significant effects of RT ($P < .001$), and generally insignificant effects of KN ($P > .05$) on consistency (Fig. S20; Supplementary Materials).

In summary, results showed evident effects of RE and RT on the accuracy and consistency but not KN. This was broadly observed across different sampling designs and a range of sampling intensities. In general, increasing RE and/or applying RTrw during rapid assessment improved the accuracy and consistency in mapping spatial distribution of species richness.

Sampling design

To explore potential underlying variations between sites, MPAB and CVQ were averaged across all rapid assessment methods and iterations by site and sampling intensity for SRS. Under SRS, both Danum and Amacayacu had the largest MPAB (17.3%–18.6% and 16.0%–19.3%, respectively), while Pasoh had the lowest MPAB (7.2%–9.3%) at 5% and 20% sampling intensity (Table 2). For CVQ, Danum consistently had the highest CVQ (74.8%–85.8%), while Pasoh had the lowest value (9.8%–12.1%) at 5% and 20% sampling intensity (Table 2). The results also were consistently observed across sampling intensity (Table S1; Supplementary Materials). This suggested that sites were inherently variable in achieving a level of accuracy and consistency.

To examine accuracy and consistency of a sampling design with respect to SRS, dMPAB and dCVQ were averaged across all sites, rapid assessment methods, and iterations by sampling design and sampling intensity. Among the sampling designs, LIST(I_{freq}) and LPM(I_{freq}) outperformed SRS, and both achieved the largest reduction in MPAB and CVQ for all sampling intensities apart from 1% sampling intensity for MPAB (Tables 3 and S2). At 5% sampling intensity, LIST(I_{freq}) and LPM(I_{freq}) reduced MPAB by 0.78% and 0.70%, respectively, while LIST(I_{freq}) and LPM(I_{freq}) reduced CVQ by 7.64% and 8.44%, respectively, compared to SRS (Table 3). As expected, higher sampling intensity increased the reduction by both sampling designs. At 20% sampling intensity, reduction in MPAB was almost double, while it was 1.1 to 1.6 times for CVQ (Table 3). Across sampling intensity, MPAB and CVQ were

Downloaded from https://academic.oup.com/forestry/advance-article/doi/10.1093/forestry/cpad041/7242317 by University of Minnesota - Twin Cities user on 04 October 2023

Table 2. MPAB and CVQ by site at 5% and 20% sampling intensity for SRS

Site	MPAB (%)		CVQ (%)	
	5%	20%	5%	20%
Amacayacu	19.33 (2.09)	15.98 (1.83)	35.81 (4.62)	32.45 (4.84)
BCI	12.97 (0.81)	10.30 (0.65)	16.58 (0.98)	13.64 (0.79)
Danum	18.55 (2.21)	17.28 (2.05)	74.79 (24.80)	85.75 (19.41)
Fushan	18.37 (1.94)	15.65 (1.64)	25.09 (2.27)	22.73 (2.27)
Pasoh	9.28 (0.46)	7.23 (0.43)	12.05 (0.51)	9.78 (0.49)
Sinharaja	14.55 (1.04)	11.27 (0.82)	18.68 (1.07)	15.20 (1.01)
Wanang	15.83 (0.96)	12.74 (0.94)	27.10 (2.00)	23.69 (2.59)

MPAB and CVQ are averages across all rapid assessment methods and iterations with standard deviations in parentheses.

Table 3. Difference in mean percent absolute bias (dMPAB) and difference in coefficient of variation in relative accuracy (dCVQ) by sampling design at 5% and 20% sampling intensity with SRS as the baseline

Sampling design	dMPAB (%)		dCVQ (%)	
	5%	20%	5%	20%
SYS	−0.07 (1.67)	−0.06 (1.59)	−0.24 (13.50)	−0.27 (9.94)
SOL	−0.01 (1.65)	0.02 (1.58)	0.16 (12.46)	0.44 (9.52)
LIST(S_{rapid})	0.16 (1.86)	0.42 (1.74)	−0.62 (14.49)	0.23 (10.72)
LIST(I_{freq})	−0.78 (1.87)	−1.64 (2.51)	−7.64 (19.45)	−11.97 (24.74)
LPM(S_{rapid})	0.01 (1.63)	0.06 (1.55)	0.33 (12.44)	0.52 (9.50)
LPM(I_{freq})	−0.70 (1.83)	−1.57 (2.31)	−8.44 (20.34)	−9.65 (20.89)

The dMPAB and dCVQ are averages across all sites, rapid assessment methods, and iterations with standard deviations in parentheses.

reduced up to 1.9% and 13.1%, respectively (Table S2). On the other hand, SOL was not different from SRS with reduction in MPAB and CVQ less than 0.1% and 0.8%, respectively, across sampling intensity (Table S2). In summary, LIST(I_{freq}) performed slightly better than LPM(I_{freq}) in accuracy, but it slightly underperformed in consistency when sampling intensity was less than 10% (Table S2).

The standard deviations in dMPAB and dCVQ indicated that variations in dMPAB and dCVQ for LIST(I_{freq}) and LPM(I_{freq}) were higher than the other sampling designs across sampling intensity (Tables 3 and S2). For dMPAB at 5% sampling intensity, the standard deviation in LIST(I_{freq}) and LPM(I_{freq}) were 1.9% and 1.8%, respectively (Table 3). For dCVQ, they were 19.5% and 20.3%, respectively (Table 3). At 20% sampling intensity, standard deviations of LIST(I_{freq}) and LPM(I_{freq}) were up to 1.6 and 2.6 times higher than those of the other sampling designs (Table 3). For comparison, SYS, SOL, and LPM(S_{rapid}) had the lowest standard deviation in dMPAB and dCVQ across sampling intensity (Table S2). In summary, this suggested that although LIST(I_{freq}) and LPM(I_{freq}) increased accuracy and consistency compared to SRS, the effect was more variable between sites and rapid assessment methods.

Looking more closely, the variations in dMPAB and dCVQ of LIST(I_{freq}) and LPM(I_{freq}) between the 27 rapid assessment methods at 5% sampling intensity was largely due to RE with smaller influence from RT and KN (Fig. 4). In general, larger RE led to greater reduction in dMPAB and dCVQ for a combination of KN × RT. For example, dMPAB of the combination KN50 × RTrw × RE10 for LPM(I_{freq}) was −0.48% but increased to −0.89% at RE40 (Fig. 4a). For a combination of KN × RE, RTtr had larger dMPAB and dCVQ than RTnn under LIST(I_{freq}), but RTtr was either comparable to or had smaller dMPAB and dCVQ than RTnn under LPM(I_{freq}) (Fig. 4b, c, e, f). In summary, even though LIST(I_{freq}) and LPM(I_{freq}) achieved overall greater accuracy and consistency, their performances depended on the choice of a rapid assessment method with larger RE offering better results in general.

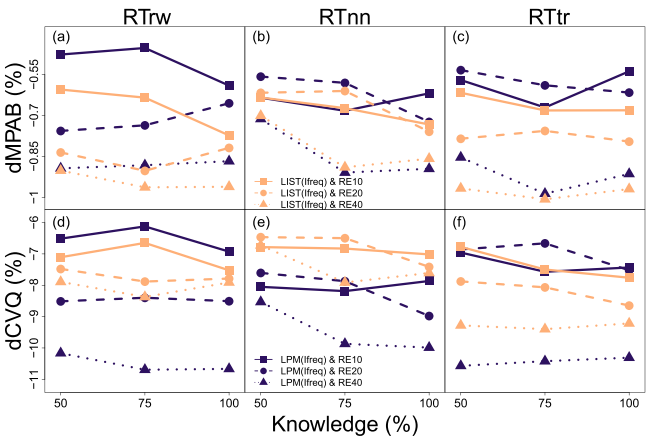

Figure 4. Difference in mean percent absolute bias (dMPAB; a–c) and difference in coefficient of variation in relative accuracy (dCVQ; d–f) by the 27 combinations of KN, RE, and RT for LIST(I_{freq}) and LPM(I_{freq}) at 5% sampling intensity. The rapid assessment efforts are 10 (RE10), 20 (RE20), and 40 (RE40) individuals. The rapid assessment types are random walk (RTrw), nearest-neighbor (RTnn), and transect (RTtr). The dMPAB and dCVQ are averages across all sites and iterations.

Sampling intensity

Increasing sampling intensity of census cells reduced MPAB and CVQ nonlinearly with the largest rate of reduction between 1% and 5% sampling intensity, and the decrease gradually leveled off at higher sampling intensity (Figs. 5, S21, and S22; Supplementary Materials). LIST(I_{freq}) and LPM(I_{freq}) generally outperformed the other sampling designs, but the trends were different between MPAB and CVQ. For MPAB, LIST(I_{freq}) and LPM(I_{freq}) performed equally well, but both only outperformed the other sampling designs when sampling intensity was greater than 5% (Figs. 5a, b and S21). For example, there was about 1.7% difference in MPAB between LIST(I_{freq}) and the other sampling designs at

Downloaded from https://academic.oup.com/forestry/advance-article/doi/10.1093/forestry/cpad041/7242317 by University of Minnesota - Twin Cities user on 04 October 2023

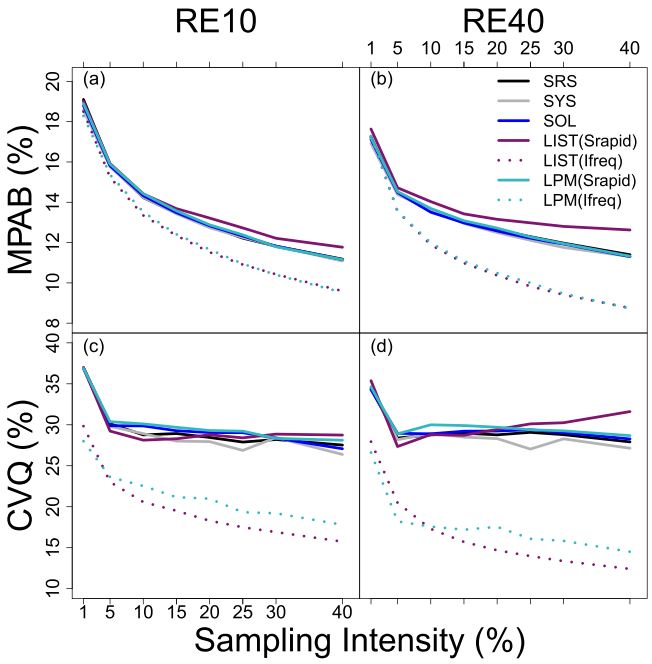

Figure 5. MPAB and CVQ by sampling intensity for the seven sampling designs. The relationship is depicted for the combination of knowledge level of 50% (KN50), random walk rapid assessment type (RTrw), and rapid assessment effort of (a, c) 10 (RE10) and (b, d) 40 (RE40) individuals. MPAB and CVQ are averaged across all sites.

10% sampling intensity for the KN50 × RTrw × RE40 combination (Fig. 5b).

On the other hand, LIST(I_{freq}) and LPM(I_{freq}) consistently produced lower CVQ across all sampling intensities (Figs. 5c, d and S22). For example, there was about 6.5% difference in CVQ between LIST(I_{freq}) and the other sampling designs at 1% sampling intensity for the KN50 × RTrw × RE40 combination (Fig. 5d). When sampling intensity was below 5%, LPM(I_{freq}) performed comparable to or slightly better than LIST(I_{freq}), but LIST(I_{freq}) was distinctly better than LPM(I_{freq}) when sampling intensity was at and above 10% sampling intensity (Figs. 5c, d and S22). Interestingly, increasing sampling intensity under the rest of the sampling designs had no effect on reducing CVQ. Lastly, LIST(S_{rapid}) performed the poorest overall. In short, LIST(I_{freq}) and LPM(I_{freq}) were preferred given that they performed well in terms of both accuracy and consistency when sampling intensity was lower than 5%.

Discussion

Our study proposes an alternative approach to mapping distribution of woody plant species richness across a local forest by integrating rapid assessment by local communities, variable probability sampling theory, and machine learning techniques. Because assessing species diversity is resource intensive, Oliver and Beattie (1996) proposed a 'restricted sampling' strategy by reducing sampling units, time, or methods while trying to capture relative differences in diversity between sites. Our approach could be regarded as an informed restricted sampling because prior information is collected from a rapid assessment that supposedly reflects underlying spatial differences in diversity. The information is used to better allocate limited resources to the resource-intensive census by taking advantage of variable probability sampling theory. Studies applying sampling with covariate designs to species diversity assessment are few. Hence, it is advantageous to explore their potential benefits at the very least from the perspective of managing resources for biodiversity assessment.

While it is expected that there would be variations between sites, the large ranges in the accuracy and consistency between sites come as a surprise. For example, the consistency of Danum was about six to eight times lower than that of Pasoh depending on sampling intensity. A possible explanation could be that sites such as Danum and Amacayacu consist of a few small pockets of cells that are exceptionally high or low in species richness (Fig. S23a, c; Supplementary Materials). When some of these cells are not selected for census, the final ANN model would not be able to adequately predict richness in these clusters. In contrast, Pasoh does not show such localization of very high or very low richness cells (Fig. S23b). Instead, cell species richness is quite similar across Pasoh (Fig. S23b). We suspect that the localization is due to the small cell size (20 × 20 m) used in our study. Under the well-recognized species–area relationship (Condit et al. 1996), delineating larger cell size might mitigate this effect because a larger cell size will have greater number of species per cell, which in turn reduces heterogeneity in spatial distribution of species richness. Future study could conduct an in-depth analysis on the effect of changing cell size and its trade-off. Having comparable accuracy and consistency between sites will help our proposed sampling strategy to be generalizable to other forests in the tropics and subtropics.

Knowledge of plant species is integral to a rapid assessment because it determines the ability to identify a plant, and in turn, the amount of information gained. Results suggest that knowledge has minimal effects on the accuracy and consistency in mapping distribution of species richness. Lam et al. (2018) found similar results when predicting cell-level species richness with 3P sampling. Moreover, results suggest that knowing at least half of the species that are abundant is sufficient for a rapid assessment. This supports engaging local communities with their knowledge in a rapid assessment. Local knowledge concerns naming plants by a local community in an area (Khasbagan and Soyolt 2008). There are challenges using local knowledge. Folk names assigned to plants should ideally be a one-to-one correspondence to scientific nomenclature, but one-to-multitude or multitude-to-one correspondence often happens (Khasbagan and Soyolt 2008, Lam and Kleinn 2008). Any conflict in nomenclature should be resolved prior to a rapid assessment such as with methods listed in Khasbagan and Soyolt (2008). Notwithstanding this issue, traditional knowledge has been recognized for its contribution to sustainable forest management (Parrotta et al. 2016). For example, Cummings and Read (2016) applied local knowledge of Makushi and Wapishiana Amerindians to assess impact of land-use change on tree species in southern Guyana. Hernández-Stefanoni et al. (2006) found that indigenous Mayan knowledge was comparable to satellite imaging in assessing plant species composition and vegetation structures in a Mexican tropical forest. It should be cautioned that in practical applications, knowledge of species in a forest is highly dependent on a local community such as their ties to the surrounding forests. Thus, it is likely that knowledge level (KN) could well be below 50%. Prior to a rapid assessment, knowledge of a community could be assessed by an expert and appropriate training could be provided to build their capacity. A future simulation study could explore lower KN to reflect practical applications. Nonetheless, these studies and ours support creatively exploring ways to integrate local knowledge in forest management decision-making.

Downloaded from https://academic.oup.com/forestry/advance-article/doi/10.1093/forestry/cpad041/7242317 by University of Minnesota - Twin Cities user on 04 October 2023

Rapid assessment effort and type influence the amount of information gained. It appears that RE, the number of rapidly assessed individuals, has the strongest effect on the accuracy and consistency regardless of sampling design and intensity. This agrees with the general species–individual relationship in that the number of observed species increases with the number of sampled individuals (Condit et al. 1996). However, higher RE increases resources required for a rapid assessment. Thus, the choice of RE depends on resource constraints and capacity of local communities. Building capacity of local communities benefits a rapid assessment. RE could be increased without a huge increase in required resources because local communities could identify species faster at a set time or simply by having more people involved in the activity. This study recommends RE of 20 individuals regardless of knowledge and RT. This is because increasing RE from 10 to 20 individuals improves the accuracy and consistency at a greater rate than increasing RE from 20 to 40 individuals.

Among the RTs, the random walk is the most accurate and consistent when compared to the transect and the nearest-neighbor approach. Heterogeneity in local environmental (Fayolle et al. 2012) and dispersal limitation (Seidler and Plotkin 2006) could lead to local aggregation of individuals of a plant species. The three methods address this spatial autocorrelation differently. With random selection, RTrw should reduce the chance of selecting individuals of the same species due to local aggregation, which in turn should observe a more variety of species. Analogous to RTrw is the TMS method by Goff et al. (1982). Huebner (2007) found that TMS was better at detecting invasive plants than systematic plot and transect. We expect RTtr to produce higher accuracy and consistency than RTnn because RTnn is more likely to select individuals of the same species. However, results suggest that RTtr is comparable to or only slightly better than RTnn. A possible explanation is the small cell size (20 × 20 m). We speculate that RTtr would perform better than RTnn if the cell size is larger, which causes selected individuals to spread over a wider area, thus capturing a more variety of species. Analogously, Quon et al. (2020) found that subplots of a cluster plot captured more diverse plant species when they were spread further apart. This could be explored in a future study along with the influence of cell size on the site effect. Nevertheless, RTtr and RTnn are more field operational than RTrw. Strictly speaking, implementing RTrw in the field requires a full list of trees in each cell prior to random selection, which is impractical. Hence, field application of RTrw would be more haphazard than actual randomness. In this study, RTrw serves more as the baseline comparison to RTnn and RTtr. On the other hand, RTnn and RTtr are familiar methods in sampling tree diversity (Gordon and Newton 2006, Wohlgemuth et al. 2008). While both are comparable, this study recommends RTtr because it is easier to select a random starting point and lay out a transect following a specific azimuth across a row of cells. However, ground training is still needed to implement RTtr in the field to reduce haphazardness in selecting plants for rapid assessment.

This study highlights that constructing a covariate for mapping distribution of species richness is not trivial. The choice of a covariate affects the accuracy and consistency of the sampling with covariate designs. When estimating forest volume, sampling with covariate designs generally produce unbiased estimates (Magnussen 2015) and are more efficient than SRS (Yang et al. 2019), but Hsu et al. (2021) showed that 3P sampling produced different levels of precision when using different covariates. In our study, there is no advantage in using S_{rapid} as the covariate because sampling with covariate designs with S_{rapid} are either comparable to or worse than SRS. This contradicts the common perception that sampling with covariate designs would be effective if a covariate is highly correlated with the variable of interest. On the other hand, the covariate I_{freq} works as expected.

We speculate this is the issue of regression analysis under complex survey designs (Lohr 2010; p. 429–444) as an analysis is sensitive to its predictor space (Draper and Smith 1998; p. 721). Lohr (2010) showed that a sampling design influenced a predictor space, especially when inclusion probability was different between samples or was related to the response variable. As evident in Fig. 6, the predictor spaces formed by sampling with covariate designs with S_{rapid} are different to those formed by the sampling designs with I_{freq}. SOL, LIST(S_{rapid}), and LPM(S_{rapid}) tend to select census cells from S_{rapid} strata with high frequency of rapid cells similar to the way SRS select census cells (Fig. 6). On the other hand, LIST(I_{freq}) and LPM(I_{freq}) tend to select census cells more evenly distributed across S_{rapid} strata, which is drastically different from SRS—especially Danum (Fig. 6a, b). Yang et al. (2019) found similar issues in estimating forest volume by variable probability sampling with random forest imputation and nonlinear models. They used selection probabilities as weights during model fitting to reduce bias. Thus, this issue of covariate requires consideration because there is a growing interest to predict spatial distribution of primary forest products across a large forest area with remote sensing (Lisańczuk et al. 2020, D'Amico et al. 2022). Furthermore, tools have also been developed to use wall-to-wall airborne laser scanning data to assist allocation of samples such as sgsR (Goodbody et al. 2023), which could benefit from carefully designed covariates. Nevertheless, different covariates related to diversity such as the Shannon diversity index should be tested in future studies. If one or more covariates could be found to reduce the strong site effects, that would help in generalizing our study.

Both LPM(I_{freq}) and LIST(I_{freq}) are viable choices of sampling designs to map distribution of species richness. We recommend LPM(I_{freq}) for several reasons. For one, sampling intensity is usually low. Chiarucci et al. (2003) found that many plant diversity inventories had sampling intensity less than 1%. Sampling intensity of some national forest inventories in Europe could be only 0.2% (Tomppo et al. 2010). Results show that LPM(I_{freq}) is as good as or slightly better than LIST(I_{freq}) in terms of accuracy and consistency when sampling intensity is below 5%. Thus, implementing LPM(I_{freq}) benefits forest planning regardless of whether decision-making is based on accuracy or consistency. LPM also has the advantage of selecting samples from a multidimensional sampling frame (Grafström et al. 2012). For example, Pitkänen et al. (2022) found that LPM increased efficiency up to three times in selecting sample locations for pre-harvest assessment when using multiple covariates derived from ground inventory and various sources of remote sensing. Because species diversity is a multifaceted concept (Lam and Maguire 2012), it is possible to consider this when selecting census cells with LPM. However, LIST is easier to be understood and implemented by practitioners not trained in sampling theory compared to LPM. If this is of concern and also if sampling intensity is greater than 10%, we would recommend LIST(I_{freq}).

Increasing the number of individuals for rapid assessment or sampling intensity will directly increase the cost of our methods. According to Iles (2003, p. 437), sampling with covariate designs reward judgment. Better judgment reduces sample size for detailed measurements, which in turn reduces inventory cost (Iles 2003). In our context, increase in the experience of local communities increases speed in the field, reduces identification

Downloaded from https://academic.oup.com/forestry/advance-article/doi/10.1093/forestry/cpad041/7242317 by University of Minnesota - Twin Cities user on 04 October 2023

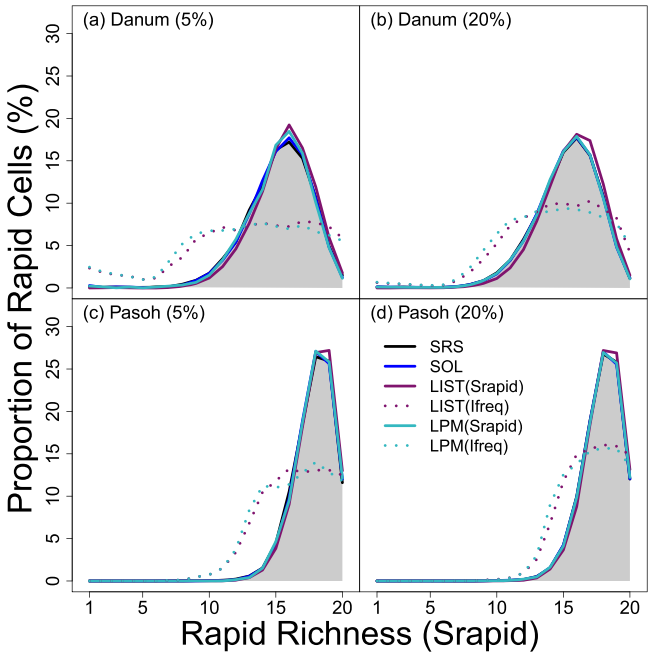

Figure 6. Proportion of rapid cells selected for census at each level of rapid richness (S_{rapid}) for the five sampling designs: SRS, SOL, LIST(S_{rapid}), LIST(I_{freq}), LPM(S_{rapid}), and LPM(I_{freq}). The proportions are depicted for (a, b) Danum and (c, d) Pasoh at 5% and 20% sampling intensity. The gray shaded region depicts proportion of all rapid cells with respect to S_{rapid} in the two sites. The rapid assessment combination under which census cells are selected is KN100 × RE20 × RTrw.

error, or even improves the ability to identify rare species. This could lead to increasing RE within the same set of time, which has the upside of maintaining costs and improving accuracy and consistency. Hence, there is an incentive to invest in training and engaging local communities. Census is resource intensive, which makes deciding a sampling intensity critical. The nonlinear decreasing trends of accuracy and consistency over sampling intensity mirror the trends of precision in volume estimates over sample size from Yang et al. (2019). This implies that approaches to decide sample size for estimating volume could be considered for diversity assessment such as sample size determination formulas from Kershaw Jr. et al. (2016) or nonlinear models relating precision to sample size in Yang et al. (2019). Looking at the decay rate, our study suggests 5% sampling intensity because the rate of improvement in accuracy and consistency is the largest when increasing sampling intensity from 1% to 5%. However, this suggestion is provisional because a detailed cost-plus-loss analysis (Lynch 2017, Yang et al. 2017), which examines the trade-off between cost components, accuracy, and consistency, is needed. This aspect of biodiversity assessment has not been well studied and is required if one is to design an efficient multipurpose inventory system that adequately samples timber products and species diversity.

A limitation to our approach is spatial coverage. A rapid assessment by local communities is no match for geographical coverage by remote sensing. It is possible to integrate remote sensing and our sampling strategy into a hierarchical variable probability sampling design. At the landscape level, wall-to-wall metrics such as normalized difference vegetation index (Pau et al. 2012), fraction of photosynthetically active radiation (Coops et al. 2019), or topoclimatic variables and phenology-related information (Fassnacht et al. 2016) could be derived from remote sensing. They serve as the covariates for sampling with covariate designs to select local sites for rapid assessment and census. The information from the local sites are then fed back to the remote sensing metrics to predict spatial distribution of species richness at the landscape level. The number of required levels depends on geographical coverage and sources of remote sensing. Nevertheless, covariates developed at multiple levels are intrinsically linked through the hierarchical variable probability sampling design.

Conclusion

Decision making in sustainable forestry requires adequate information and appropriate tools (Baskerville 1986). This study develops a sampling strategy of assessing spatial distribution of species richness by integrating rapid assessment by local communities, sampling with covariate designs, and machine learning techniques. The aim is to diversify tools to generate information for sustainable management of plant diversity. Using knowledge to guide sample selection has been around for decades (Grosenbaugh 1964) but has seldom been applied. This study has demonstrated the feasibility of using local knowledge to construct covariates useful for diversity assessment. It also highlights the impact of complex survey designs on the choice of covariate to achieve the desired results. This study uses ANN to build prediction models for its non-parametric nature and flexibility. However, parametric models such as the geostatistical methods of kriging used in Ferrier and Guisan (2006) could be explored in future study. As forest management objectives are increasingly complex and multifaceted to meet diverse societal demands, a forest inventory system needs to be innovative and multipurpose. It ideally generates information on timber products and forest ecosystem services on the same level of precision. In conclusion, our study seeks to empower local communities by finding means for them to directly engage in a forest management process.

Acknowledgements

We would like to thank three anonymous reviewers and the editors for their valuable comments that have significantly improved the manuscript. We also like to thank Dr George Weiblen from the University of Minnesota for assisting with the Wanang dataset. The 25-ha Long-Term Ecological Research Project of Amacayacu is a collaborative project of the Instituto Amazónico de Investigaciones Científicas Sinchi and the Universidad Nacional de Colombia Sede Medellín, in partnership with the Unidad de Manejo Especial de Parques Naturales Nacionales and the Forest Global Earth Observatory of the Smithsonian Tropical Research Institute (ForestGEO). The Amacayacu Forest Dynamics Plot is part of ForestGEO a global network of large-scale demographic tree plots. We acknowledge the Director and staff of the Amacayacu National Park for supporting and maintaining the project in this National Park. BCI: The BCI forest dynamics research project was made possible by National Science Foundation grants to Stephen P. Hubbell: DEB-0640386, DEB-0425651, DEB-0346488, DEB-0129874, DEB-00753102, DEB-9909347, DEB-9615226, DEB-9615226, DEB-9405933, DEB-9221033, DEB-9100058, DEB-8906869, DEB-8605042, DEB-8206992, DEB-7922197, support from the Forest Global Earth Observatory, the Smithsonian Tropical Research Institute, the John D. and Catherine T. MacArthur Foundation, the Mellon Foundation, the Small World Institute Fund, and numerous private individuals, and through the hard work of over 100 people from 10 countries over the past three decades. The plot project is part the Forest

Downloaded from https://academic.oup.com/forestry/advance-article/doi/10.1093/forestry/cpad041/7242317 by University of Minnesota - Twin Cities user on 04 October 2023

Global Earth Observatory (ForestGEO), a global network of large-scale demographic tree plots. Danum: The Danum plot is a core project of the Southeast Asia Rain Forest Research Partnership (SEARRP). We thank SEARRP partners, especially Yayasan Sabah for their support, and HSBC Malaysia and the University of Zurich for funding. We are grateful to the research assistants who are conducting the census, in particular the team leader Alex Karolus, and to Mike Bernados and Bill McDonald for species identifications. We thank Stuart Davies and Shameema Esufali for advice and training. Fushan: The establishment and first census of the Fushan 25-ha plot is a collaborative project by the Taiwan Forestry Research Institute (data provider), the Taiwan Forestry Bureau, the National Taiwan University (Institute of Ecology and Evolutionary Biology) and the ForestGEO (formerly the CTFS). We thank the agencies for providing the datasets and all field staff. S.H.S. also thanks Dr. I-Fang Sun for his long-term support for the Fushan project. Pasoh: Data from the Pasoh Research Forest was provided by the Forest Research Institute Malaysia-Forest Global Earth Observatory, Smithsonian Tropical Research Institute collaborative research project. Negeri Sembilan Forestry Department is the custodian of Pasoh Research Forest and I/we acknowledge the department for preserving the research forest. Sinharaja: The 25-ha Long-Term Ecological Research Project at Sinharaja World Heritage Site is a collaborative project of the Uva Wellassa University, University of Peradeniya, the Forest Global Earth Observatory (ForestGEO) of the Smithsonian Tropical Research Institute, with supplementary funding received from the John D. and Catherine T. Macarthur Foundation, the National Institute for Environmental Science, Japan, and the Helmholtz Centre for Environmental Research-UFZ, Germany, for past censuses. The PIs gratefully acknowledge the Forest Department, Uva Wellassa University, and the Post-Graduate Institute of Science at the University of Peradeniya, Sri Lanka for supporting this project, and the local field and laboratory staff who tirelessly contributed in the repeated censuses of this plot. Wanang: The 50-ha Wanang Forest Dynamics Plot is a collaborative project of the New Guinea Binatang Research Center, the Forest Global Earth Observatory (ForestGEO) of the Smithsonian Tropical Research Institute, the Forest Research Institute of Papua New Guinea, the Czech Academy of Sciences (19-28126X), and the University of Minnesota supported by NSF DEB-1027297 and NIH ICBG 5UO1TW006671. We acknowledge the government of Papua New Guinea and the customary landowners of Wanang for supporting and maintaining the plot.

Author contributions

Bo-Hao Perng (Formal analysis, Investigation, Methodology, Software, Visualization, Writing—original draft), Tzeng Yih Lam (Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Validation, Visualization, Writing—review and editing), Sheng-Hsin Su (Data curation, Project administration, Writing—review and editing), Mohamad Danial Md Sabri (Data curation, Writing—review and editing), David Burslem (Data curation, Writing—review and editing), Dairon Cardenas (Data curation, Writing—review and editing), Álvaro Duque (Data curation, Writing—review and editing), Sisira Ediriweera (Data curation, Writing—review and editing), Nimal Gunatilleke (Data curation, Writing—review and editing), Vojtech Novotny (Data curation, Writing—review and editing), Michael O'Brien (Data curation, Writing—review and editing), and Glen Reynolds (Data curation, Writing—review and editing)

Conflict of interest: None declared.

Funding

The funding of this project is provided by the National Science and Technology Council Taiwan (Grant no. MOST 111-2628-B-002-042 and MOST 111-2326-B-002-005-MY3).

Supplementary data

Supplementary data are available at *Forestry* online.

Data availability

The datasets analyzed for this study are available upon request from The Forest Global Earth Observatory (ForestGEO) database (forestgeo.si.edu).

References

Anderson-Teixeira KJ, Davies SJ, Bennett AC *et al.* CTFS-ForestGEO: a worldwide network monitoring forests in an era of global change. *Glob Ecol Biol* 2015;**21**:528–49. https://doi.org/10.1111/gcb.12712.

Banerjee S, Carlin BP and Gelfand AE 2015 *Hierarchical Modeling and Analysis for Spatial Data*. 2 edn. Boca Raton, FL, USA: Chapman and Hall/CRC, p.558, https://doi.org/10.1201/b17115.

Baskerville GL. Understanding forest management. *For Chron* 1986;**62**:339–47. https://doi.org/10.5558/tfc62339-4.

Basu D. Role of the sufficiency and likelihood principles in sample survey theory. *Sankhyā Ser A* 1969;**31**:441–54.

Bayat M, Bettinger P, Hassani M *et al.* Ten-year estimation of oriental beech (*Fagus orientalis* Lipsky) volume increment in natural forests: a comparison of an artificial neural networks model, multiple linear regression and actual increment. *Forestry* 2021;**94**:598–609. https://doi.org/10.1093/forestry/cpab001.

Chiarucci A, Enright NJ, Perry GLW *et al.* Performance of nonparametric species richness estimators in a high diversity plant community. *Divers Distrib* 2003;**9**:283–95. https://doi.org/10.1046/j.1472-4642.2003.00027.x.

Choe H, Chi J and Thorne JH. Mapping potential plant species richness over large areas with deep learning, MODIS, and species distribution models. *Remote Sens (Basel)* 2021;**13**:2490. https://doi.org/10.3390/rs13132490.

Chong GW, Reich RM, Kalkhan MA *et al.* New approaches for sampling and modeling native and exotic plant species richness. *West N Am Nat* 2001;**61**:328–35.

Condit R, Hubbell SP, Lafrankie JV *et al.* Species-area and species-individual relationships for tropical trees: a comparison of three 50-ha plots. *J Ecol* 1996;**84**:549–62. https://doi.org/10.2307/2261477.

Condit R, Pérez R, Aguilar S *et al. Census Data from 65 Tree Plots in Panama, 1994–2015*. Santa Barbara, CA, USA: DataONE, Dataset, 2019a.

Condit R, Pérez R, Aguilar S *et al. BCI 50-ha Plot Taxonomy, 2019 Version*. Davis, CA, USA: Dataset, Dryad, 2019b.

Condit R, Pérez R, Aguilar S *et al. Complete Data from the Barro Colorado 50-ha Plot: 423617 Trees, 35 Years, 2019 Version*. Davis, CA, USA: Dryad, Dataset, 2019c.

Coops NC, Bolton DK, Hobi ML *et al.* Untangling multiple species richness hypothesis globally using remote sensing habitat indices. *Ecol Indic* 2019;**107**:105567. https://doi.org/10.1016/j.ecolind.2019.105567.

Downloaded from https://academic.oup.com/forestry/advance-article/doi/10.1093/forestry/cpad041/7242317 by University of Minnesota - Twin Cities user on 04 October 2023

Cummings AR and Read JM. Drawing on traditional knowledge to identify and describe ecosystem services associated with northern Amazon's multiple-use plants. *Int J Biodivers Sci Ecosyst Serv Manag* 2016;**12**:39–56. https://doi.org/10.1080/21513732.2015.1136841.

D'Amico G, McRoberts RE, Giannetti F *et al.* Effects of lidar coverage and field plot data numerosity on forest growing stock volume estimation. *Eur J Remote Sens* 2022;**55**:199–212. https://doi.org/10.1080/22797254.2022.2042397.

Devictor V, Mouillot D, Meynard C *et al.* Spatial mismatch and congruence between taxonomic, phylogenetic and functional diversity: the need for integrative conservation strategies in a changing world. *Ecol Lett* 2010;**13**:1030–40. https://doi.org/10.1111/j.1461-0248.2010.01493.x.

Draper NR and Smith H. 1998 *Applied Regression Analysis*. 3rd edn. New York, NY, USA: John Wiley & Sons Inc, 706, https://doi.org/10.1002/9781118625590.

Duque A, Muller-Landau HC, Valencia R *et al.* Insights into regional patterns of Amazonian forest structure, diversity, and dominance from three large terra-firme forest dynamics plots. *Biodivers Conserv* 2017;**26**:669–86. https://doi.org/10.1007/s10531-016-1265-9.

Fassnacht FE, Latifi H, Stereńczak K *et al.* Review of studies on tree species classification from remotely sensed data. *Remote Sens Environ* 2016;**186**:64–87. https://doi.org/10.1016/j.rse.2016.08.013.

Fassnacht FE, Müllerová J, Conti L *et al.* About the link between biodiversity and spectral variation. *Appl Veg Sci* 2022;**25**:e12643.

Fayolle A, Engelbrecht B, Freycon V *et al.* Geological substrates shape tree species and trait distributions in African moist forests. *PloS One* 2012;**7**:e42381. https://doi.org/10.1371/journal.pone.0042381.

Ferrier S and Guisan A. Spatial modelling of biodiversity at the community level. *J Appl Ecol* 2006;**43**:393–404. https://doi.org/10.1111/j.1365-2664.2006.01149.x.

Foody GM and Cutler MEJ. Mapping the species richness and composition of tropical forests from remotely sensed data with neural networks. *Ecol Model* 2006;**195**:37–42. https://doi.org/10.1016/j.ecolmodel.2005.11.007.

FSC. *FSC Principles and Criteria for Forest Stewardship. FSC-STD-01-001 V5-0 EN*. Bonn, Germany: Forest Stewardship Council, 2012.

Gascon C, Brooks TM, Contreras-MacBeath T *et al.* The importance and benefits of species. *Curr Biol* 2015;**25**:R431–38. https://doi.org/10.1016/j.cub.2015.03.041.

Goff FG, Dawson GA and Rochow JJ. Site examination for threatened and endangered plant species. *Environ Manag* 1982;**6**:307–16. https://doi.org/10.1007/BF01875062.

Goodbody TRH, Coops NC, Queinnec M *et al.* sgsR: a structurally guided sampling toolbox for LiDAR-based forest inventories. *Forestry* 2023;**cpac055**:1–14.

Gordon JE and Newton AC. Efficient floristic inventory for the assessment of tropical tree diversity: a comparative test of four alternative approaches. *For Ecol Manage* 2006;**237**:564–73. https://doi.org/10.1016/j.foreco.2006.10.002.

Grafström A, Lundström NLP and Schelin L. Spatially balanced sampling through the pivotal method. *Biometrics* 2012;**68**:514–20. https://doi.org/10.1111/j.1541-0420.2011.01699.x.

Grafström A and Ringvall AH. Improving forest field inventories by using remote sensing data in novel sampling designs. *Can J For Res* 2013;**43**:1015–22. https://doi.org/10.1139/cjfr-2013-0123.

Grosenbaugh LR. 1964 Some suggestions for better sample-tree measurement. In: *Proceedings of the Society of American Foresters Meeting*. Boston, MA, USA: U.S. Department of Agriculture, Forest Service, pp. 36–42.

Grosenbaugh LR. *Three-Pee Sampling Theory and Program: "THRP" for Computer Generation of Selection Criteria*. Berkeley, CA, USA: Research Paper PSW-RP-21, U.S. Department of Agriculture, Forest Service, Pacific Southwest Research & Range Experiment Station, 1965, 53.

Haas PJ, Liu Y and Stokes L. An estimator of number of species from quadrat sampling. *Biometrics* 2006;**62**:135–41. https://doi.org/10.1111/j.1541-0420.2005.00390.x.

Hafeez S, Wong MS, Ho HC *et al.* Comparison of machine learning algorithms for retrieval of water quality indicators in case-II waters: a case study of Hong Kong. *Remote Sens (Basel)* 2019;**11**:617. https://doi.org/10.3390/rs11060617.

Hall JP. Criteria and indicators of sustainable forest management. *Environ Monit Assess* 2001;**67**:109–19. https://doi.org/10.1023/A:1006433132539.

Hernández-Stefanoni JL, Pineda JB and Valdes-Valadez G. Comparing the use of indigenous knowledge with classification and ordination techniques for assessing the species composition and structure of vegetation in a tropical forest. *Environ Manag* 2006;**37**:686–702. https://doi.org/10.1007/s00267-004-0371-8.

Hsu Y-H, Chen Y, Yang T-R *et al.* Sample strategies for bias correction of regional LiDAR-assisted forest inventory estimates on small woodlots. *Ann For Sci* 2020;**77**:1–12.

Hsu Y-H, Kershaw JA, Ducey MJ *et al.* Sampling with probability proportional to prediction (3P sampling) using covariates derived from spherical images. *Can J For Res* 2021;**51**:1140–47. https://doi.org/10.1139/cjfr-2020-0498.

Hubbell SP, Foster RB, O'Brien ST *et al.* Light-gap disturbances, recruitment limitation, and tree diversity in a neotropical forest. *Science* 1999;**283**:554–57. https://doi.org/10.1126/science.283.5401.554.

Huber PJ. Robust estimation of a location parameter. *Ann Math Statist* 1964;**35**:73–101. https://doi.org/10.1214/aoms/1177703732.

Huebner CD. Detection and monitoring of invasive exotic plants: a comparison of four sampling methods. *Northeast Nat* 2007;**14**:183–206. https://doi.org/10.1656/1092-6194(2007)14[183:DAMOIE]2.0.CO;2.

Hung C.-Y, Chen W.-C, Lai P.-T, Lin C.-H. and Lee C.-C. 2017 Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database. In: *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society 2017*. McFarland, WI, USA: Jeju, pp. 3110–13.

Iles K. *A Sampler of Inventory Topics*. 2nd edn. Nanaimo, BC, Canada: Kim Iles and Associates, 2003.

Ingram JC, Dawson TP and Whittaker RJ. Mapping tropical forest structure in southeastern Madagascar using remote sensing and artificial neural networks. *Remote Sens Environ* 2005;**94**:491–507. https://doi.org/10.1016/j.rse.2004.12.001.

Kershaw Jr., J.A, Ducey M.J, Beers TW and Husch B. 2016 *Forest Mensuration*. 5th edn. West Sussex, UK: John Wiley & Sons Ltd, 632, https://doi.org/10.1002/9781118902028.

Khasbagan and Soyolt. Indigenous knowledge for plant species diversity: a case study of wild plants' folk names used by the Mongolians in Ejina desert area, Inner Mongolia, P. R. China. *J Ethnobiol Ethnomedicine* 2008;**4**:2. https://doi.org/10.1186/1746-4269-4-2.

Kingma DP and Ba J. 2015 Adam: a method for stochastic optimization. In: *3rd International Conference on Learning Representations ICLR 2015*, 7–9 May 2015. Y. Bengio and Y. LeCun (eds). San Diego, CA, USA, pp. 1–15.

Lam TY, Hsu Y-H, Yang T-R *et al.* Sampling with probability proportional to prediction: rethinking rapid plant diversity assessment. *Forestry* 2018;**91**:17–26. https://doi.org/10.1093/forestry/cpx044.

Lam TY and Kleinn C. Estimation of tree species richness from large area forest inventory data: evaluation and comparison of

Downloaded from https://academic.oup.com/forestry/advance-article/doi/10.1093/forestry/cpad041/7242317 by University of Minnesota - Twin Cities user on 04 October 2023

jackknife estimators. *For Ecol Manage* 2008;**255**:1002–10. https://doi.org/10.1016/j.foreco.2007.10.007.

Lam TY and Maguire DA. Structural equation modeling: theory and applications in forest management. *Int J Forest Res* 2012;**2012**:263953, 1–263916. https://doi.org/10.1155/2012/263953.

Lisańczuk M, Mitelsztedt K, Parkitna K *et al.* Influence of sampling intensity on performance of two-phase forest inventory using airborne laser scanning. *For Ecosyst* 2020;**7**:65. https://doi.org/10.1186/s40663-020-00277-6.

Lohr SL. *Sampling: Design and Analysis*. 2nd edn. Boston, MA, USA: Cengage Learning, 2010, 608.

Lynch TB. Optimal sample size and plot size or point sampling factor based on cost-plus-loss using the Fairfield Smith relationship for plot size. *Forestry* 2017;**90**:697–709. https://doi.org/10.1093/forestry/cpx024.

Magnussen S. Arguments for a model-dependent inference? *Forestry* 2015;**88**:317–25. https://doi.org/10.1093/forestry/cpv002.

Magurran AE. *Measuring Biological Diversity*. 1st edn. Oxford, UK: Blackwell Science Ltd., 2004, 264.

Manokaran N and LaFrankie JV. Stand structure of Pasoh Forest reserve, a lowland rain forest in peninsular Malaysia. *J Trop For Sci* 1990;**3**:14–24.

McRoberts RE, Schmoldt DL and Rauscher HM. Enhancing the scientific process with artificial intelligence: forest science applications. *AI Appl* 1991;**5**:5–26.

Misra D. Mish: a self regularized non-monotonic neural activation function. *BMVC* 2020, 2020 (Virtual);1–14.

O'Brien MJ, Hector A, Kellenberger RT *et al.* Demographic consequences of heterogeneity in conspecific density dependence among mast-fruiting tropical trees. *Proc R Soc B* 2022;**289**:20220739.

Oliver I and Beattie AJ. Designing a cost-effective invertebrate survey: a test of methods for rapid assessment of biodiversity. *Ecol Appl* 1996;**6**:594–607. https://doi.org/10.2307/2269394.

Olson M, Wyner AJ and Berk R. 2018 Modern neural networks generalize on small data sets. In: Bengio, S, Wallach, HM, Larochelle, H, Grauman, K, Cesa-Bianchi, N. (eds). *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., pp. 3623–32.

Parrotta J, Yeo-Chang Y and Camacho LD. Traditional knowledge for sustainable forest management and provision of ecosystem services. *Int J Biodivers Sci Ecosyst Serv Manag* 2016;**12**:1–4. https://doi.org/10.1080/21513732.2016.1169580.

Pasupa K. and Sunhem W. 2016 A comparison between shallow and deep architecture classifiers on small dataset. In: *2016 8th International Conference on Information Technology and Electrical Engineering*. Red Hook, NY, USA: Curran Associates Inc., pp. 1–6.

Pau S, Gillespie W and Wolkovich EM. Dissecting NDVI–species richness relationships in Hawaiian dry forests. *J Biogeogr* 2012;**39**:1678–86. https://doi.org/10.1111/j.1365-2699.2012.02731.x.

Pearson DL and Carroll SS. Global patterns of species richness: spatial models for conservation planning using bioindicator and precipitation data. *Conserv Biol* 1998;**12**:809–21. https://doi.org/10.1046/j.1523-1739.1998.96460.x.

Pereyra G, Tucker G, Chorowski J *et al.* Regularizing neural networks by penalizing confident output distributions. *arXiv* 2017;1701.06548. https://doi.org/10.48550/arXiv.1701.06548.

Pimm SL and Raven P. Extinction by numbers. *Nature* 2000;**403**:843–45. https://doi.org/10.1038/35002708.

Pitkänen TP, Räty M, Hyvönen P *et al.* Using auxiliary data to rationalize smartphone-based pre-harvest forest mensuration. *Forestry* 2022;**95**:247–60. https://doi.org/10.1093/forestry/cpab039.

Quon C, Lam TY and Lin H-T. Designing cluster plots for sampling local plant species composition for biodiversity management. *Forest Syst* 2020;**29**:e002. https://doi.org/10.5424/fs/2020291-15894.

Rumelhart DE, Hinton GE and Williams RJ. Learning representations by back-propagating errors. *Nature* 1986;**323**:533–36. https://doi.org/10.1038/323533a0.

Särndal C.-E, Swensson B. and Wretman J. 1992 *Model Assisted Survey Sampling*. 1st edn. Springer New York, 695, https://doi.org/10.1007/978-1-4612-4378-6.

Seidler TG and Plotkin JB. Seed dispersal and spatial pattern in tropical trees. *PLoS Biol* 2006;**4**:2132–37.

Srivastava N, Hinton G, Krizhevsky A *et al.* Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res* 2014;**15**:1929–58.

Su S-H, Chang-Yang C-H, Lu C-L *et al. Fushan Subtropical Forest Dynamics Plot: Tree Species Characteristics and Distribution Patterns*. 1st edn. Taipei, Taiwan: Taiwan Forestry Research Institute, 2007, 272.

Sullivan GM and Feinn R. Using effect size or why the P value is not enough. *J Grad Med Educ* 2012;**4**:279–82. https://doi.org/10.4300/JGME-D-12-00156.1.

Tomppo E, Gschwantner T, Lawrence M. and McRoberts RE 2010 *National Forest Inventories: Pathways for Common Reporting*. 1st edn. Springer Netherlands, p. 612, https://doi.org/10.1007/978-90-481-3233-1.

Villero D, Pla M, Camps D *et al.* Integrating species distribution modelling into decision-making to inform conservation actions. *Biodivers Conserv* 2017;**26**:251–71. https://doi.org/10.1007/s10531-016-1243-2.

Vincent JB, Henning B, Saulei S *et al.* Forest carbon in lowland Papua New Guinea: local variation and the importance of small trees. *Austral Ecol* 2015;**40**:151–59. https://doi.org/10.1111/aec.12187.

Walker DH, Sinclair FL and Thapa B. Incorporation of indigenous knowledge and perspectives in agroforestry development. *Agroforest Sys* 1995;**30**:235–48. https://doi.org/10.1007/BF00708923.

Williams JN, Seo C, Thorne J *et al.* Using species distribution models to predict new occurrences for rare plants. *Divers Distrib* 2009;**15**:565–76. https://doi.org/10.1111/j.1472-4642.2009.00567.x.

Wohlgemuth T, Nobis MP, Kienast F *et al.* Modelling vascular plant diversity at the landscape scale using systematic samples. *J Biogeogr* 2008;**35**:1226–40. https://doi.org/10.1111/j.1365-2699.2008.01884.x.

Yang T-R, Hsu Y-H, Kershaw JA *et al.* Big BAF sampling in mixed species forest structures of northeastern North America: influence of count and measure BAF under cost constraints. *Forestry* 2017;**90**:649–60. https://doi.org/10.1093/forestry/cpx020.

Yang T-R, Kershaw JA, Weiskittel AR *et al.* Influence of sample selection method and estimation technique on sample size requirements for wall-to-wall estimation of volume using airborne LiDAR. *Forestry* 2019;**92**:311–23. https://doi.org/10.1093/forestry/cpz014.

Zuleta D, Russo SE, Barona A *et al.* Importance of topography for tree species habitat distributions in a terra firme forest in the Colombian Amazon. *Plant and Soil* 2020;**450**:133–49. https://doi.org/10.1007/s11104-018-3878-0.

Downloaded from https://academic.oup.com/forestry/advance-article/doi/10.1093/forestry/cpad041/7242317 by University of Minnesota - Twin Cities user on 04 October 2023