Forestry An International Journal of Forest Research

Institute of Chartered Foresters

Forestry 2023; **96**, 411–424, https://doi.org/10.1093/forestry/cpac055
Advance Access publication 6 February 2023

sgsR: a structurally guided sampling toolbox for LiDAR-based forest inventories

Tristan R. H. Goodbody[1,*], Nicholas C. Coops[1], Martin Queinnec[1], Joanne C. White[2], Piotr Tompalski[2], Andrew T. Hudak[3], David Auty[4], Ruben Valbuena[5], Antoine LeBoeuf[6], Ian Sinclair[7], Grant McCartney[8], Jean-Francois Prieur[9] and Murray E. Woods[7]

[1]*Faculty of Forestry, Department of Forest Resources Management, University of British Columbia, 2424 Main Mall, Vancouver, BC V6T 1Z4, Canada*
[2]*Canadian Forest Service (Pacific Forestry Centre), Natural Resources Canada, 506 West Burnside Road, Victoria, BC V8Z 1M5, Canada*
[3]*Rocky Mountain Research Station, USDA Forest Service, Moscow, ID 83843, USA*
[4]*School of Forestry, Northern Arizona University, 200 East Pine Knoll Drive, Flagstaff, AZ 86011-5018, USA*
[5]*Division of Forest Remote Sensing, Department of Forest Resource Management, Swedish University of Agricultural Sciences, Bangor University, School of Natural Sciences, Umeå, Thoday building, LL57 2UW Bangor, Sweden, UK*
[6]*Ministère des Forêts, de la Faune et des Parcs, 5700, 4ième Avenue Ouest, local A 108 Québec, QC G1H 6R1, Canada*
[7]*Ontario Ministry of Natural Resources and Forestry, 1235 Queen St E, Sault Ste Marie, ON P6A 2E5, Canada*
[8]*Forsite Consultants Ltd, 330-42 St SW, Salmon Arm, BC V1E 4R1, Canada*
[9]*Département de Géomatique Appliquée Faculté de Lettres et de Sciences humaines, Centre d'Applications et de Recherches en Télédétection (CARTEL), 2500 Bd de l'Université, Sherbrooke, Quebec J1K 2R1, Canada*

*Corresponding author E-mail: goodbody.t@gmail.com

Received 20 May 2022

Establishing field inventories can be labor intensive, logistically challenging and expensive. Optimizing a sample to derive accurate forest attribute predictions is a key management-level inventory objective. Traditional sampling designs involving pre-defined, interpreted strata could result in poor selection of within-strata sampling intensities, leading to inaccurate estimates of forest structural variables. The use of airborne laser scanning (ALS) data as an applied forest inventory tool continues to improve understanding of the composition and spatial distribution of vegetation structure across forested landscapes. The increased availability of wall-to-wall ALS data is promoting the concept of structurally guided sampling (SGS), where ALS metrics are used as an auxiliary data source driving stratification and sampling within management-level forest inventories. In this manuscript, we present an open-source R package named *sgsR* that provides a robust toolbox for implementing various SGS approaches. The goal of this package is to provide a toolkit to facilitate better optimized allocation of sample units and sample size, as well as to assess and augment existing plot networks by accounting for current forest structural conditions. Here, we first provide justification for SGS approaches and the creation of the *sgsR* toolbox. We then briefly describe key functions and workflows the package offers and provide two reproducible examples. Avenues to implement SGS protocols according to auxiliary data needs are presented.

Introduction

Mensuration is a cornerstone of forest management. Quantitative and qualitative measurements acquired for trees, plots and stands are essential for forest planning, policy and research. Although constrained by cost, labour availability and logistics, innovation in field measurement methods via increased technological integration and augmentation is promising (Katila and Tomppo, 2001; Melville *et al.*, 2015; Puliti *et al.*, 2018). Methods for leveraging field data through space and time to support forest inventories are also well understood (Lefsky *et al.*, 1999; Næsset, 2002; Bechtold and Patterson, 2005; White *et al.*, 2013; Dash *et al.*, 2015; Tompalski *et al.*, 2019; van Ewijk *et al.*, 2020).

Information needs dictate the scope and scale of forest inventories, and by association, the data and methods used to derive that information (McRoberts and Tomppo, 2007; Tomppo *et al.*, 2010). Fundamentally, the purpose of any forest inventory and associated sampling framework is to obtain knowledge about the population under investigation and estimate target population parameters. Tomppo (2010) comprehensively outlined

Handling Editor: Dr. Fabian Fassnacht

how a variety of national forest inventories (NFIs) around the world use long-standing probability-based approaches where permanent and non-permanent field samples provide an effective means of acquiring data on a variety of forest attributes for multi-purpose resource management. The re-measurement of permanent sample units (plots) in NFIs, rather than routine plot re-establishment, effectively manages costs and provides time-series information facilitating estimates of forest attribute change, a critical component of strategic inventories. Operational and tactical inventories, unlike NFIs, are smaller in scale and are generally focused on production and management of timber and non-timber resources. The scope of these inventories is therefore less focused on long-term changes and more on providing managers with the best possible predictions of forest attributes (e.g. timber volume and stem density) to support product demand and effective environmental stewardship.

Given the highly dynamic nature of forest environments and the socio-economic factors dictating their management, routine sampling is important for ensuring up-to-date operational and tactical inventory attribute estimates. The costly process of establishing new sample networks, and augmenting existing ones, stands to benefit from objective methods outlining where and how to optimally locate plots to enable effective management and planning decisions (Melville *et al*., 2015). Research continues to suggest that sampling algorithms leveraging auxiliary remotely sensed datasets can be used to improve estimates of forest population parameters and inform on key management objectives (Hawbaker *et al*., 2009; McRoberts, 2012; Junttila *et al*., 2013; Papa *et al*., 2020).

Forest attribute models yield greater error rates when calibration datasets fail to capture the full range of predictor variability (e.g. Demaerschalk and Kozak, 1974; Hawbaker *et al*., 2009; Maltamo *et al*., 2011). Models that operate within the bounds of the original calibration data, i.e. the smallest convex set containing all the design points (Cook, 1975), will often perform poorly when extrapolating beyond this region (Montgomery *et al*., 2006). Outside the calibration region, parameter estimation methods are in 'extrapolation mode' where predictions are less reliable (Demaerschalk and Kozak, 1974; Montgomery *et al*., 2006). The same applies to (oft-used) non-parametric modeling approaches (such as k-nearest neighbours) that depend on the close proximity of nearest neighbors in a reference set to impute plausible predictions. In the scope of management-level, production-based forestry, such approaches are likely to produce extrapolation errors when imputing less abundant forest structural types that are rare in the population, and consequent sample, when not treated differently in the sampling design. With sustainable forest management objectives becoming increasingly pertinent, regions with rare spatial and structural complexity must be sampled given their potential to contain disproportionately large quantities of biomass or habitat value (Davies and Asner, 2014). Hence, regardless of whether parametric or non-parametric modelling approaches are applied, model-based inference is well suited for small area estimation (McRoberts, 2012; Melville *et al*., 2015). That being said, sampling the full range of structural variability is more important for non-parametric methods, which are also becoming more common for the implementation of area-based inventory approaches (White *et al*., 2017).

In the effort to improve spatial and compositional variability within model calibration data, *a priori* integration of spatially explicit remotely sensed data and geospatial layers has become increasingly common (Maltamo *et al*., 2011; Tomppo *et al*., 2014; Dash *et al*., 2016; Papa *et al*., 2020). The use of aerial and/or satellite imagery to provide estimates of condition and composition, and land cover classifications to differentiate forested and non-forested ecosystems has aided in the design of field surveys (Corona, 2010; McRoberts *et al*., 2014). Remotely sensed data that differentiate, stratify or characterize areas of interest prior to field surveys have helped to improve sampling strategies while maintaining the statistical precision required for decision-making (Gobakken *et al*., 2013; Melville *et al*., 2015; Papa *et al*., 2020). Photointerpretation of aerial imagery to delineate stand boundaries and estimate species composition has provided a basis for delineating forest strata (Leckie, 2003; Maltamo *et al*. 2021). Sampling designs that leverage these data have been found to directly integrate landscape heterogeneity, increasing the empirical range of predictors in an acquired sample and improving accuracy and precision of model predictions (McRoberts *et al*., 2012).

Forest structure is both a product and a driver of ecosystem processes and diversity that provides insight into forest yield, condition, disturbance history, habitat characteristics and operational development (Spies, 1998; Thom and Keeton, 2019). Active remote sensing technologies such as airborne laser scanning (ALS) can be analyzed to derive measures of vegetation structure and underlying terrain (Holopainen *et al*., 2014; Tompalski *et al*., 2019; Queinnec *et al*., 2021a). The ability to structurally characterize forest vegetation using ALS metrics has led to the development and implementation of enhanced forest inventories (EFIs) (White *et al*., 2017). EFIs leverage data, such as ALS, to integrate previously inaccessible or cost-prohibitive information (e.g. forest structural descriptions, digital terrain models) to develop forest information layers and predictions that are wall-to-wall, spatially explicit and at a finer spatial resolution than traditional inventories. The operational and economic benefits provided by EFIs for forest management and planning are well documented (Eid *et al*., 2004; Borders *et al*., 2008; White *et al*., 2013; Ayrey *et al*., 2019), and as the acquisition of ALS continues, so too does the creation and augmentation of EFI frameworks. Empirical characterizations of forest structure within EFIs have been demonstrated to correlate well with desirable field measured forest attributes like height and timber volume (Tompalski *et al*., 2021). The linkage between field measurements and ALS data facilitate area-based inventory approaches (Næsset, 2002), wherein forest structure may be characterized using a parsimonious set of ALS point cloud metrics in conjunction with field plot data (e.g. tree height, canopy cover; Lefsky *et al*., 1999; Bouvier *et al*., 2015) to predict a wide range of inventory attributes (e.g. above ground biomass, stem volume, Lorey's height; Wästlund *et al*., 2018; Yu *et al*., 2020).

A major benefit of using remotely sensed data to help guide sampling frameworks is that they provide information about the population that can be leveraged in sampling designs (Benedetti *et al*., 2015). Statistical approaches that minimize bias and variance of parameter estimates can be referred to as optimal designs (Smith, 1918; O'Brien and Funk, 2003; Silvey, 2013). A growing number of studies have demonstrated that

the use of structural metrics from ALS as auxiliary information for allocating sample units offers opportunities to optimally design experiments (Hawbaker *et al*., 2009; Maltamo *et al*., 2011; Melville *et al*., 2015; Queinnec *et al*., 2021a, 2021b). These approaches can provide transparent, repeatable, tuneable and spatially explicit methods to accurately represent the full structural variation of a forest resource and/or augment sample networks to include less frequently occurring, yet managerially or ecologically important forest structural conditions. Practical approaches to the operational implementation of optimal designs using ALS data for forest management, which we refer to as structurally guided sampling (SGS) approaches, have also been outlined (White *et al*., 2013; White *et al*., 2017).

Leiterer *et al*. (2015) and Kane *et al*. (2010) suggested that information contained in ALS data can be condensed to a few parsimonious metrics and through dimensionality reduction approaches such as principal component analysis, which has been prevalent in the literature (see Table 1). Maltamo *et al*. (2011) emphasized that an appropriate method to decide among ALS metrics is to determine their relative importance to attributes of interest (e.g. if accurate estimates of stand height or volume are desired, then metrics such as the 90th percentile of height are logical to include). Methods of stratifying ALS metric populations to guide sampling strategies have varied. Standardized breaks (Hawbaker *et al*., 2009; Gobakken *et al*., 2013) and statistical summaries (e.g. principal components) (Fedrigo *et al*., 2018; Papa *et al*., 2020; Queinnec *et al*., 2021a, 2021b) are common, while sampling algorithms that incorporate spatial variation without the need for stratification are growing in application (Grafström and Ringvall, 2013; Grafström *et al*., 2014; Melville *et al*., 2015). Outcomes from these studies have suggested that model performance can be improved by ensuring representativeness both spatially over the forest estate and within the empirical predictor space. This increases the potential to achieve a greater degree of efficiency in sampling while optimizing sample size, leading to potential reductions in overall inventory costs (Junttila *et al*., 2013; Queinnec *et al*., 2021a). One caveat to this approach is that optimizing stratification and sampling efforts to estimate a particular population parameter (e.g. mean biomass) may lead to a greater degree of uncertainty in estimates of other attributes (e.g. mean stem density). This emphasizes the importance of establishing inventory objectives and associated attributes of interest early in the planning stage.

SGS approaches using ALS data have been shown to provide empirical rigor for determining the quantity and spatial allocation of sample units (Table 1). Herein, we present *sgsR*, a free, open-source, customizable and efficient software toolbox implemented in the R statistical computing environment (R Core Team, 2022). *sgsR* is primarily focused on management-level forest inventory applications where model-based inference will be used to estimate population attributes (Gregoire, 1998; Chen *et al*., 2016). This manuscript is structured to provide an overview of *sgsR* functionality and outline theoretical processing strategies with applied and reproducible examples. We then comment on the managerial value of a consolidated toolbox capable of implementing SGS approaches and its ability to support operational EFI planning.

Methods

The fundamental processing objectives of *sgsR* are to facilitate incorporation of forest structural metrics, such as those derived from the *lidR* package (Roussel *et al*., 2020), in the design of forest management inventories. ALS metrics are the intended inputs to *sgsR* because they provide information on three-dimensional forest structure. Other remotely sensed or auxiliary data can however be used (e.g. optical imagery) if the spatial resolution of inputs is consistent. Here, we focus on the use of ALS metrics in the context of management-level inventory applications. *sgsR* is an open-source and robust toolbox to enable objective methods for creating, augmenting and analyzing forest inventory samples. The package is currently hosted on the comprehensive R archive network (https://cran.r-project.org/package=sgsR) and Github (https://github.com/tgoodbody/sgsR).

Processing strategies using sgsR

sgsR provides a collection of sampling algorithms that use auxiliary information for allocating sample units over an areal sampling frame. Probability-based sampling methods (e.g. simple random sampling, systematic sampling), where inclusion probabilities of sample units are known or can be derived (see Gregoire, 1998) are included, though these methods are not the principal focus of the package.

Functions within *sgsR* can be separated into two main processing streams: *stratification* and *sampling* (Table 2). By separating the two processing components into individual steps, users can easily design and customize both the initial stratification and subsequent sampling to best fit auxiliary data and inventory objectives. Supplementary processing steps include *calculating* functions, which perform various intermediary tasks such as calculating covariate values, as well as *extracting* functions, which extract co-located covariate values for each sample unit. The processing workflow therefore starts with a *stratification*, followed by *sampling* and ends with *extracting*. A summary of potential processing workflows is presented in Figure 1.

A stratified sampling example entails a user first providing a population (wall-to-wall coverage of ALS data where each raster cell is a potential sample unit), which is then stratified (e.g. using the *strat_breaks* function) to produce a stratified raster output. The generated stratification is then used as an input, along with the desired sample size, to sample proportionally among strata using the *sample_strat* algorithm. The output sample is then paired with the input ALS metrics using *extract_*metrics, attributing co-located metric values to each sample unit to enable downstream forest attribute modelling.

Stratification algorithms

Auxiliary rasters

Functions denoted with *strat_* are stratification algorithms within *sgsR*. These algorithms use auxiliary raster data (e.g. ALS metric populations) as inputs and provide stratified areas of interest as outputs. Algorithms are either supervised (e.g. *strat_breaks*), where the user provides empirical values that drive stratifications, or unsupervised (e.g. *strat_quantiles*), where the user specifies

Table 1 Literature where ALS and/or forest inventory information were used as auxiliary data for sampling.

Study	General topic	Data/metric(s) used	Attribute of interest
van Aardt *et al.* (2006)	Stratification	Canopy height model (CHM)	Volume and biomass
Junttila *et al.* (2008)	Sample selection	Height and intensity metrics	Volume, stem count, basal area
Hawbaker *et al.* (2009)	Stratification	Mean and standard deviation (SD)	Vegetation structure and biomass
Maltamo *et al.* (2011)	Plot selection	Height, density and ground *vs* canopy echoes	Volume and stem count
Junttila *et al.* (2013)	Plot selection	Principal components height	Biomass
Gobakken *et al.* (2013)	Plot selection	70^{th} percentile of height, density	Volume
Leiterer *et al.* (2015)	Stratification	Relative frequency distribution of echoes (full waveform)	Canopy structure
Melville *et al.* (2015)	Stratification and plot selection	Height and stocking density. Canopy cover and occupied volume	Volume
Melville and Stone (2016)	Plot selection	Monte Carlo random selection of height and density	Volume
Niemi and Vauhkonen (2016)	Stratification	Textural CHM metrics	Volume, basal area and mean diameter
Fedrigo *et al.* (2018)	Stratification	Height, plant area volume density, plant area index	Forest stand types
Papa *et al.* (2020)	Stratification and plot selection	Canopy height, leaf area density and index, voxel-based vegetation density	Structural variation
Queinnec *et al.* (2021b)	Stratification	Height, canopy cover and height variability	Canopy height, cover and canopy height variability
Queinnec *et al.* (2021a)	Stratification and plot selection	Principal components of height, cover and variability	Lorey's height, basal area, diameter, stem density, volume and biomass

Table 2 Stratification *(strat_) and sampling (sample_)* algorithms implemented in the R package sgsR v1.3.4

	Function name	Description
Stratification	*strat_breaks*	Stratify using user-defined breaks.
	strat_quantiles	Stratify using quantile breaks.
	strat_kmeans	Stratify using kmeans[1].
	strat_poly	Stratify based on a polygon coverage.
	strat_map	Combine (map) two stratifications.
Sampling	*sample_srs*	Simple random sampling.
	sample_systematic	Systematic sampling within a regular or hexagonal tessellation.
	sample_strat	Stratified random sampling. Requires stratified raster as input (Queinnec *et al.*, 2021a).
	sample_clhs	Conditioned Latin hypercube sampling using *clhs*[2] functionality. Requires metrics as input (Roudier, 2011).
	sample_balanced	Balanced raster sampling using *lcube*[3] and *lpm2_kdtree*[4] methods. Requires metrics as input (Grafström and Lisic, 2018).
	sample_ahels	Adapted hypercube evaluation of a legacy sample (ahels). Requires metrics as input (Malone *et al.*, 2019).
	sample_nc	Nearest centroid algorithm (Melville and Stone, 2016). Requires metrics as input.
	sample_existing	Sub-sample an existing sample using *clhs*[2] functionality.

[1]stats package – kmeans – https://rdrr.io/r/stats/kmeans.html [2]clhs package – clhs – https://rdrr.io/cran/clhs/man/clhs.html [3]BalancedSampling package – lcube – https://rdrr.io/cran/BalancedSampling/man/lcube.html [4]SamplingBigData – lpm2_kdtree – https://rdrr.io/cran/SamplingBigData/man/lpm2_kdtree.html

the desired number of output strata (*nStrata*) and stratification is handled by the algorithm.

All stratification algorithms allow the user to supply individual or multiple input metric rasters. For example, as of *sgsR* v1.3.4 the *strat_breaks* algorithm allows for single and dual metric stratifications. A single metric (e.g. 90^{th} percentile of height) can be provided where desired breakpoints are supplied by the user. In these cases, raster cells situated between breaks along

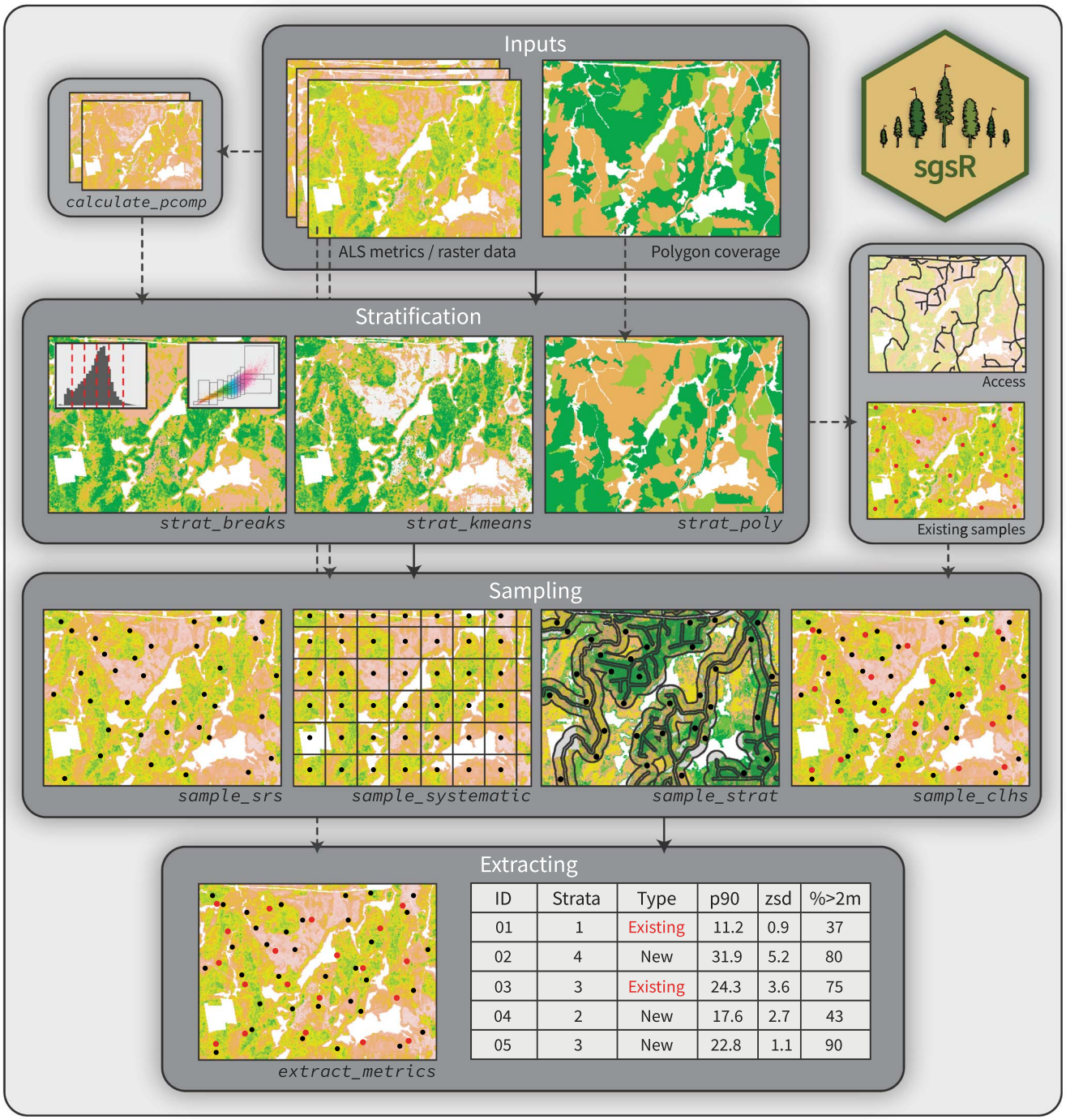

Figure 1 *sgsR* processing workflow detailing potential inputs, a selection of stratification and sampling algorithms, and supplementary processing strategies.

the metric distribution are allocated to an individual stratum (Figure 2A,B). The same can also be done in cases where two metrics are provided (e.g. 90th percentile of height and standard deviation of height), with corresponding breaks specified for each metric. In this case, strata are defined by breakpoints in both metrics, where raster cells situated within the dual bounds of a stratum define their delineation (Figure 2C,D; Hawbaker *et al.*, 2009; Gobakken *et al.*, 2013). It is also possible for the user to supply metric summaries such as principal components (calculated using *calculate_pcomp*) as metric rasters (see Queinnec *et al.*, 2021a), with the same stratification rules applying.

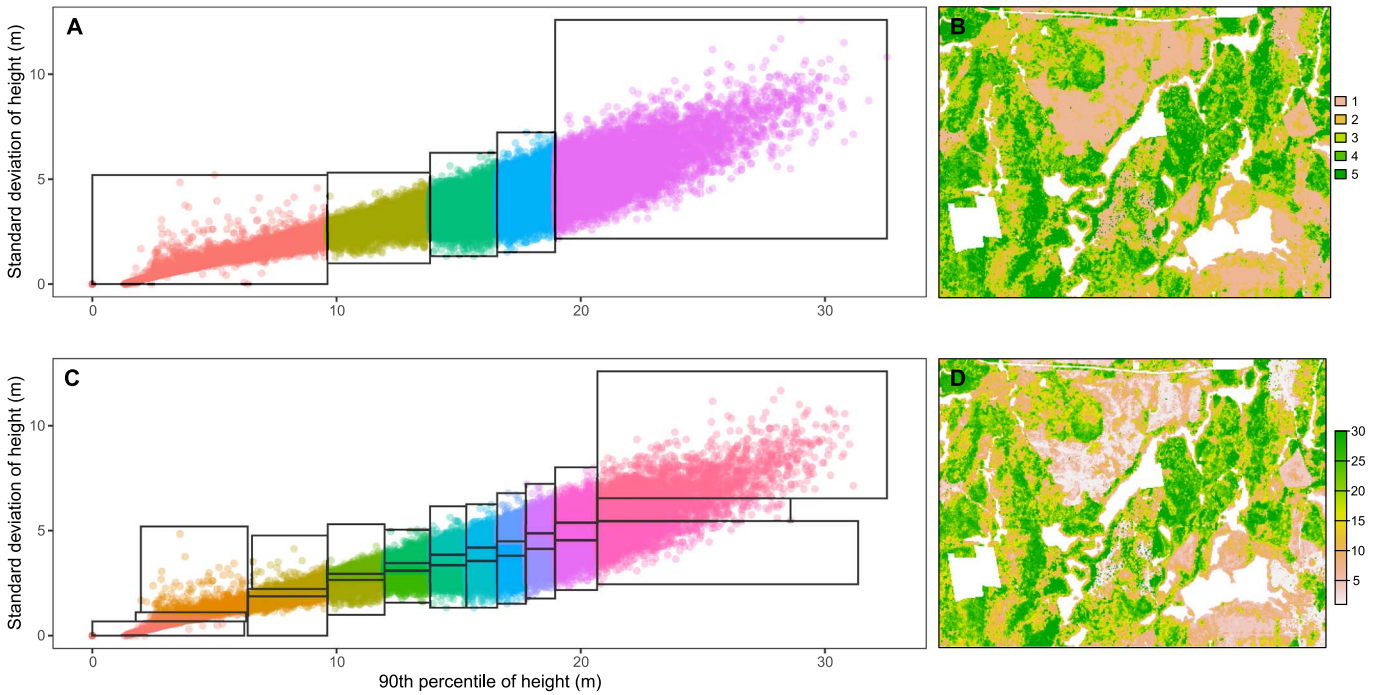

Figure 2 Single (90th percentile of height) and dual (90th percentile of height and standard deviation of height) stratification scatter plots (A and C, respectively) and rasters (B and D, respectively) colored to distinguish classes and delineated with black boundaries. Stratifications were generated using the *strat_quantiles* function in *sgsR*.

Polygon coverages

Forest inventories with polygon coverages summarizing forest attributes such as species, management type or photo-interpreted estimates of volume can be stratified using the *strat_poly* algorithm. The algorithm requires that the user defines the inventory of interest, as well as an auxiliary raster to define the output extent and spatial resolution. Users can then specify the population of interest (e.g. species) for stratification and provide vectors for how categorical or numeric values should be stratified (e.g. strata representing pine/spruce/fir). Grouping of values is also possible (e.g. strata representing pine/spruce and fir) to allow for a greater degree of control in stratum composition and outputs.

With the recognition that managers may wish to pair polygon-based stratifications with those derived using ALS metrics, the combination of two stratifications is possible using the *strat_map* algorithm. Two stratifications of matching extent and resolution are provided as input (Figure 3A,B), which are then mapped against one another to generate unique strata based on stratum pairings (Figure 3C). This facilitates the user to generate stratifications detailing quantitative and qualitative measures, such as structure by species, or multiple qualitative measures, such as species by management type.

Sampling algorithms

Functions denoted with *sample_* are sampling algorithms in *sgsR*. Depending on the sampling algorithm, users are able to provide either auxiliary metrics or stratifications derived from *strat_* functions as inputs. A number of customizable parameters can be set including the sample size (*nSamp*), a minimum distance threshold (*mindist*) between allocated sample units and the ability for the user to define an access network (*access*) and assign minimum (*buff_inner*) and maximum (*buff_outer*) buffer distances to constrain sampling extents.

Probability-based sampling

Commonly used probability-based sampling algorithms, including simple random sampling (*sample_srs*) and systematic sampling (*sample_systematic*), are implemented in *sgsR*. The *sample_systematic* function provides the user with the ability to choose a tessellation shape (regular or hexagonal grid) and uses a random start point and rotation. Users define a sampling interval distance and sample location within the tessellation (tessellation center, corners or random). The use of non-default values for select parameters (e.g. *mindist* and *access*) will result in changes to sample unit inclusion probabilities and a resulting shift to a model-based inference approach (Gregoire, 1998; Gregoire and Valentine, 2007).

sample_strat

The stratified random sampling algorithm (*sample_strat*) has two implemented approaches. First, traditional stratified random sampling, where inclusion probabilities for all sample units within strata are equal given default parameterization (*method* = '*random*'), and the method described in Queinnec et al. (2021a) (*method* = '*Queinnec*') is described in the following text.

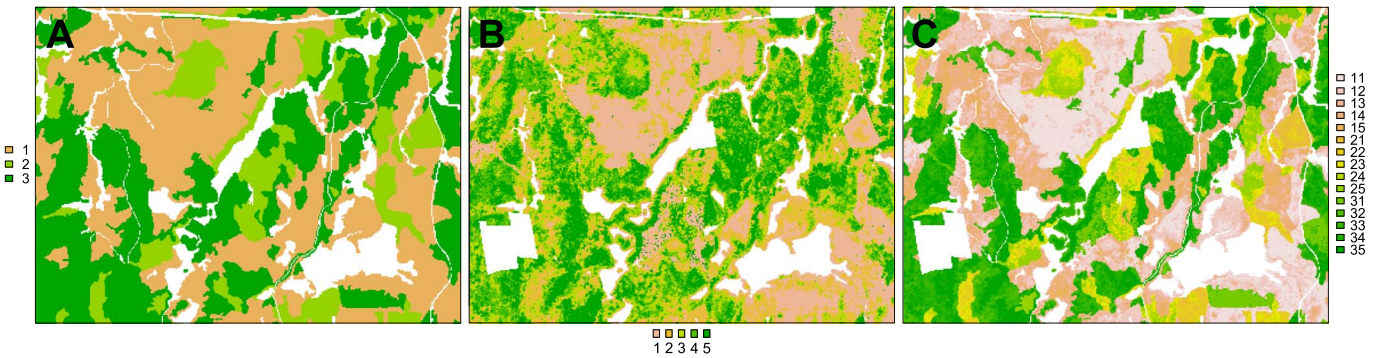

Figure 3 Stratification into three classes using *strat_poly* (A); 5 class p90 stratification derived using *strat_quantiles* (B); mapped combination of strata from A and B derived using *strat_map* (C). Mapped raster stratum indices combine to form the final stratum values (e.g. stratum A1 and B1 combine to make C11).

sample_strat is the only sampling algorithm where a stratification is a mandatory input. By default, the algorithm uses the *Queinnec* method, where inclusion probabilities are proportional to stratum size (spatial coverage), and sample units are allocated without replacement, using the following two-step hierarchical rule. First, grouped stratum pixels are identified using a moving window (e.g. a 3×3 set of pixels of the same strata), and random sampling within grouped pixel subsets is performed to meet the desired overall sample size. Second, if the desired overall sample size cannot be met due to a lack of available pixel groups, stratum pixels not meeting the group definition are randomly sampled until the desired overall sample size is obtained. This two-step approach is designed to prioritize sampling in areas of homogenous strata coverage, instead of isolated strata pixels.

The method in which sample allocation is performed can be changed using the *allocation* parameter (calculated internally using the *calculate_allocation* function). By default, samples are allocated proportionally to the size of each stratum. Additional allocation options include optimal allocation with equal sampling cost (*allocation = 'optim'*) as outlined in Gregoire and Valentine (2007), equal allocation (*allocation = 'equal'*), where the same number of samples are allocated to each stratum, and manual allocation (*allocation = 'manual'*), where users can assign relative weights to each stratum.

Latin hypercube and balanced sampling

Functions developed to simplify and consolidate prominent sampling approaches have been implemented via *sample_balanced* and *sample_clhs*. Both of these algorithms take auxiliary metrics as inputs and provide an efficient means of sampling from multivariate distributions. *sample_balanced* leverages functionality from the *BalancedSampling* (Grafström and Lisic, 2018) and *SamplingBigData* (Lisic and Grafström, 2018) R packages, providing users the ability to choose between the local pivotal method (*lpm2_kdtree*), *lcube,* and *lcubestratified* sampling algorithms described in Grafström and Lisic (2018). These algorithms were included within *sgsR* given their fast implementation and ability to balance sample unit selection in physical and predictor space (Grafström *et al.*, 2014; Melville and Stone, 2016). Balanced sampling approaches are especially valuable when there are spatial trends in input variables (Grafström *et al.*, 2012).

Conditioned Latin hypercube sampling (*clhs*; (Minasny and McBratney, 2006) has been implemented via *sample_clhs*, leveraging functionality from the *clhs: Conditional Latin Hypercube Sampling* R package (Minasny and McBratney, 2006; Roudier, 2011). *clhs* is a stratified random sampling approach that has been used in soil science and environmental research as an efficient method to representatively sample within multivariate distributions and assess uncertainty in model predictions (Minasny and McBratney, 2002). Proposed benefits of this sampling approach are numerous, though the principal reason for its inclusion in *sgsR* is that the distribution and multivariate correlation of input metrics can be preserved to ensure a representative sample (Yang *et al.*, 2020). Unique to the *sample_clhs* algorithm is the *cost* parameter. The user can provide either an index or name of an input metric to be used to constrain *clhs* sampling. Common examples of these constraints could be distance from roads, inaccessible areas, land tenure or terrain elevation.

Augmenting an existing sample

Select functions (e.g. *sample_strat* and *sample_clhs*) allow the user to augment an existing sample. This functionality is implemented to acknowledge that the acquisition of an entirely new sample is likely to be uncommon in areas with existing management histories, or in locations where historical data exist (e.g. permanent sample networks).

The *sample_ahels* algorithm within *sgsR* has been adapted from that described in Malone *et al.* (2019) to perform the adapted hypercube evaluation of a legacy sample (ahels). The algorithm takes an existing sample and auxiliary metrics to allocate new sample units where the ratio between metric density and existing sampling intensity indicates under-representation. Simply put, *sample_ahels* divides auxiliary metrics into a user-determined number of quantiles and generates a density matrix of auxiliary metric data falling into each. Quantile and density matrices can be supplied to the *sample_ahels* algorithm to improve sampling processing speeds using the *calculate_pop* function. Quantiles that do not cover at least 1 per cent of the study area are omitted from the analysis by default, making inclusion probabilities for respective raster cells zero. Existing sample units are then co-located with metric values and allotted into corresponding metric quantiles to evaluate sampling

intensity. Over- or under-representation within quantiles is determined according to the ratio of metric and sample densities (<1 being under-represented and >1 being over-represented).

Once preliminary sampling ratios have been computed, the user can elect to add a discrete number of sample units (*nSamp*) or specify a sampling ratio threshold (*threshold*) where sample units will not be added beyond a specified level. Once a method is defined, an iterative sampling approach is applied where sample units are randomly allocated to prioritize the quantile with the smallest sampling ratio, which is recomputed each time a sample unit is added. Sample units are allocated until the given *nSamp* or *threshold* is reached.

Examples of stratification and sampling using *sgsR*

We present two fully reproducible simulations to illustrate the stratification and sampling functionality of *sgsR*. R code used for all examples can be found in Appendix A. Using a test site in Ontario, Canada, we provide example workflows where:

1. The population of the 90th percentile of height metric is stratified using defined break points to generate a stratified raster, which is then sampled (*sample_strat*) to allocate a total of 100 sample units. A road access layer is provided to constrain the allocation of samples to be a maximum of 400 m and a minimum of 50 m away from access.
2. An existing sample network of 50 sample units is provided, with the goal of allocating an additional 100 (*sample_*ahels) sample units based on the distribution of the 90th percentile of height.

Inputs

ALS data were acquired in June 2018 over the Romeo Malette Forest near Kapuskasing, Ontario. A subset of ~4000 ha was chosen and the 90th percentile of height (*zq90*) and standard deviation of height (*zsd*) were calculated at a 20-m spatial resolution using the *lidR* package (Roussel *et al.*, 2020; Roussel and Auty, 2022). Metrics were then masked to remove non-forested areas and parks. A road access layer with 167 road segments totaling 66 km was used during stratified sampling in Example 1 to constrain the location of sample units to be a maximum and minimum distance from the road centerline. These data are internal to the *sgsR* package.

Example 1: stratified sampling

In the first example, a stratified sampling approach was used to generate a network of 100 sample units no greater than 400 m and no less than 50 m from road access.

ALS metrics are first loaded into R using the *rast* function from the terra package (Hijmans, 2022). Using the *strat_quantiles* function, the *zq90* metric is set as the *mraster* parameter. *sgsR* parameters for raster inputs are *mrasters* (metrics rasters) and *srasters* (stratified raster outputs from *strat_* functions). The desired number of strata (*nStrata*) is set to 5. The output *sraster*, which will be used as the input in the subsequent sampling algorithm, is a five-class stratification based on the division of the *zq90* metric population into five equally sized strata.

Prior to sampling, the road access network of the study site is loaded into R using the *st_read* function from the *sf* package (Pebesma, 2018) and is set as the *access* parameter within *sample_strat*. In order to confine sample allocation to be within a particular distance from the road network, the maximum distance (*buff_outer)* and minimum distance (*buff_inner*) are specified as 400 and 50 m for example data, respectively. The unit of measure for buffering distances will always be the same as the unit of measure for access data (e.g. meters, feet, kilometers, degrees etc.). *access* buffering is handled internally within *sample_strat*. We compared cumulative frequency distributions for *zq90* and *zsd* to ensure that the access constrained sampling frame had minimal differences to the ALS metric population (Figure 4). Differences between both cumulative frequency distributions were assessed visually and were found to be negligible.

The output of *sample_strat* is a simple features collection (Pebesma, 2018) of 100 point objects (sample units) with *strata, type, rule* and *geometry* attributes. *strata* refers to the strata each

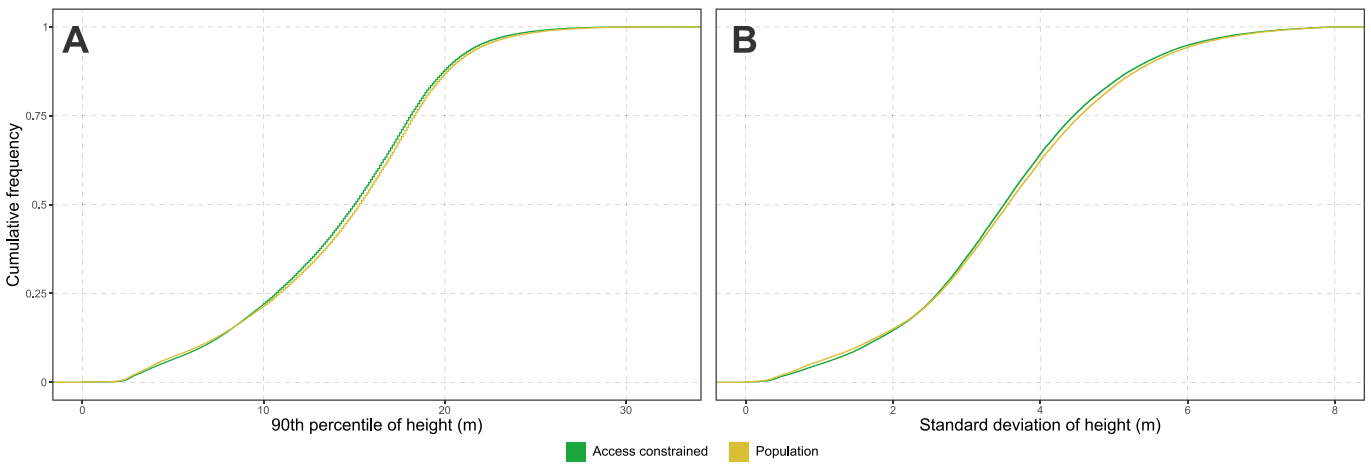

Figure 4 Cumulative frequency distributions for *zq90* (A) and *zsd* (B) that compare the access constrained region (green) with populations (yellow).

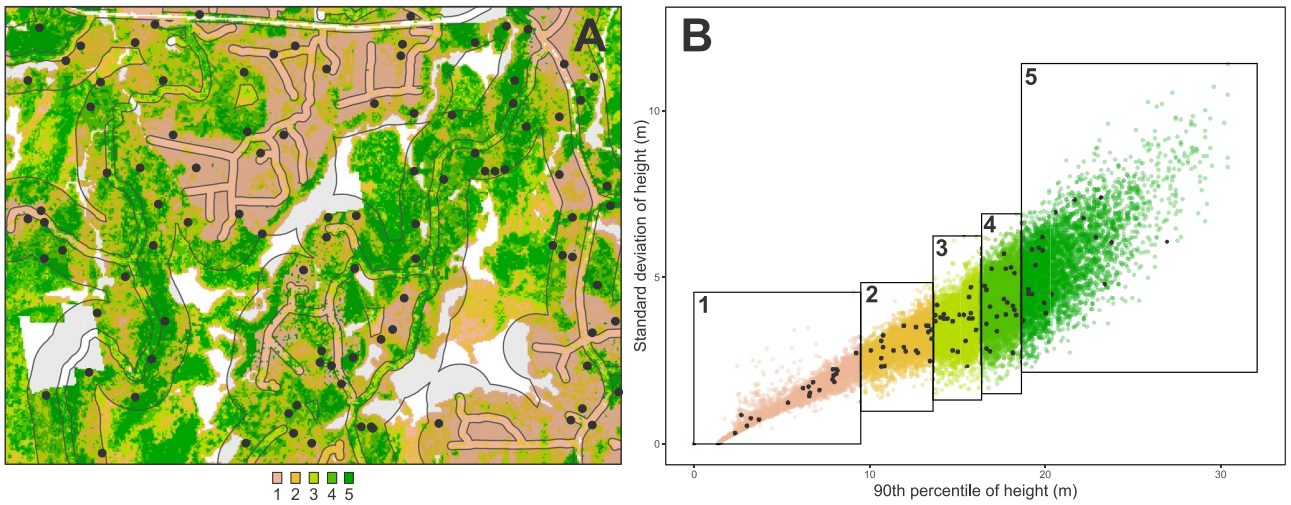

Figure 5 Graphical result of sampling from Example 1 with mapped sample units and road access buffers (A), and scatter plot showing distribution of sample units (black) in relation to the overall population (colored by strata; B). Stratification was performed on *zq90* only. The scatter plot with *zsd* on the y-axis is purely to enable visualization.

sample unit belongs to, *type* refers to whether the samples are newly created (*new*) or provided during sampling (*existing*), *rule* refers to sampling rules outlined in the *sample_strat* section of this manuscript and *geometry* contains spatial information for each sample. Following allocation, co-located ALS metrics are then extracted for each sample unit using the *extract_metrics* function, appending co-located ALS metrics with each sample unit. The output from *sample_strat* and the visualization of the allocated sample based on their distribution within strata are shown in Figure 5.

Example 2: augmenting an existing sample network

In the second example, 100 sample units were added to an existing sample to improve the structural representation of sample units within the 90th percentile of height. For this example, we first generated a set of 50 sample units to simulate an *existing* sample. For the purposes of this example, we assumed these sample units were already being used for forest management. To do this, we performed simple random sampling using the *sample_srs* function, setting the *zq90* metric as the *raster* parameter. *sample_srs* allows both *mrasters* and *srasters* as inputs, given that the algorithm does not use metric content, only extent, into account during sampling.

Once the *existing* sample network was created, we augmented it to improve representation based on *zq90*. For this, we use the *sample_ahels* algorithm. The *zq90* metric was set as *mraster* in the *sample_ahels* algorithm and we set the random sample generated above as *existing*. In this example, we specified that we wanted to add 100 new sample units (*nSamp*). We used the default *nQuant* parameter. The optional parameter to display details related to the algorithm output is specified (*details = TRUE*).

sample_ahels outputs a list where *samples* is a simple features collection (Pebesma, 2018) of 150 point objects (50 *existing* and 100 *new* sample units) with co-located ALS metrics, and *details* is a sub-list with three sampling ratio matrices. The *existingRatio* matrix contains the sampling ratios of the *existing* random sample, the *sampledRatio* matrix contains the sampling ratios of the combination of *existing* and *new* sample units added by *sample_ahels* and the *diffRatio* is the difference between *sampledRatio* and *existingRatio*. With the understanding that an optimal ratio for each quantile is 1, we compared *existingRatio* and *sampledRatio* matrices and noted that sampling ratios using *sample_ahels* became closer to the ideal of 1 within metric quantiles (Figure 6).

The proportional representation of sample units within quantiles is presented in Figure 7. By default, *sample_ahels* divides the *zq90* metric distribution into 10 quantiles (*nQuant*), dropping quantiles (inclusion probabilities for included sample units become zero) containing less than 1 per cent of the population from sampling. In this example, quantile 10 was dropped, leaving 9 remaining quantiles for sampling to occur (Figure 7). The proportional frequency of stratum coverage and the sampling frequency within quantiles are shown to be unequal prior to sampling using *sample_ahels* (Figure 7A). Notably, quantiles 1 and 9 have no sampling representation, while quantile 7 shows oversampling. This indicates that sampling intensity within quantiles is not representative of the proportional coverage of the 90th percentile of height metric. We found that after the addition of 100 samples using *sample_ahels*, the proportional frequency of sample plots and quantile coverage became close to equal for all quantiles (Figure 7A). This is corroborated by the sampling ratio values from the *sampledRatio* matrix presented in Figure 6B. This indicates that additional sampling using *sample_ahels* resulted in sample units that were more representative of the 90th percentile of height across the study area.

Discussion

sgsR has been designed to provide users with a robust, free and open-source toolbox that enables implementation, testing and comparison of varying sampling approaches for management-level inventories. The processing workflows presented highlight

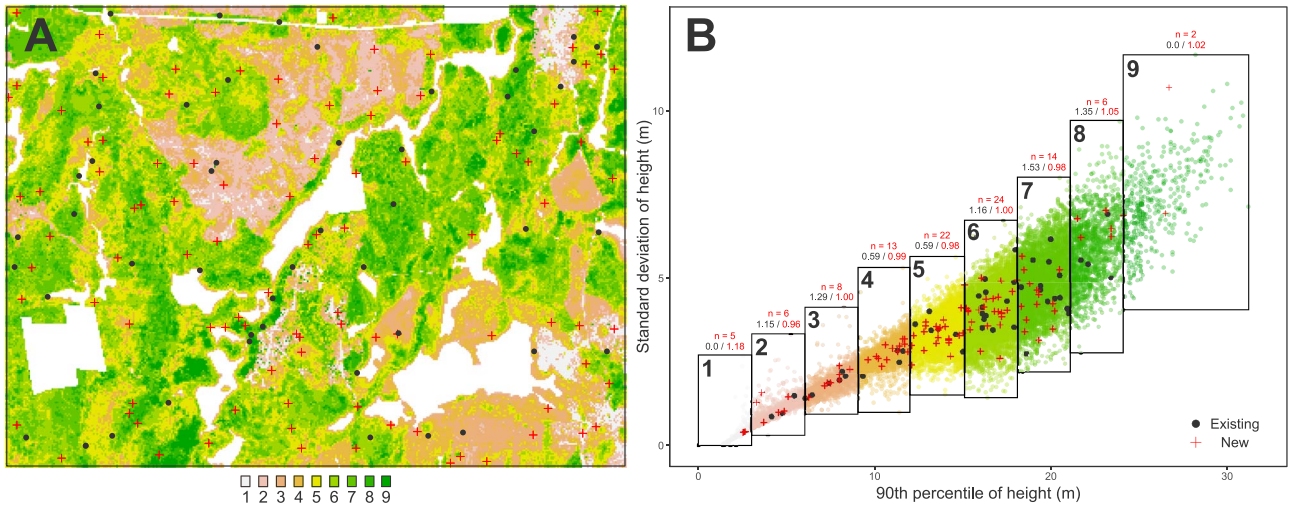

Figure 6 Example 2 processing outputs. Map of quantile strata (A) with existing (black) and new (red) sample units. Scatter plot (B) with quantile delineations by color and black boundaries. Existing and new sample units are presented to indicate their location within the distribution of the 90th percentile of height and standard deviation of height. Text above quantile boundaries denotes the initial *existingRatio* of sample units (black) and *sampledRatio* following addition of new sample units (red). The number of sample units added to each quantile is listed for each quantile (e.g. $n = 2$).

the customizable nature of the software. The goal of this package is to provide a collection of stratification and sampling algorithms to help users effectively represent structural variation across a forest management area (Maltamo *et al.*, 2011; White *et al.*, 2013). Algorithms have been developed to improve sample representation in rare or under-represented forest types. Sampling is a time-, labour- and cost-intensive task, and methods that help managers balance sample size and intended precision of model predictions to enable accurate forest inventories are therefore critical for effective management (Papa *et al.*, 2020; Queinnec *et al.*, 2021a).

The objective of the examples provided in this manuscript was to outline some of the capabilities of *sgsR*. The toolbox provides users with a variety of options and flexibility in parameterization to allocate sample units. The operational feasibility and/or optimization of these sample units in an applied field measurement context however will always require objective scrutiny from end users. Particular stratification and/or sampling approaches may not be suited to certain sampling schemes or consequent modelling strategies (Queinnec *et al.*, 2021a). Critical thinking and scientific rigor for justifying methods used are fundamental to the SGS process. Additionally, extraneous factors related to logistics, safety and field optimization must also be considered. Within reason, a potential strategy of including redundancy (i.e. allocating more sample units than necessary) in a generated sample has no additional cost to the user; however, it provides flexibility to make on-the-ground decisions about which sample units to visit while in the field, given conditions encountered. This strategy is likely to help maintain representation within forest structures, while providing field crews with the flexibility to implement efficient sampling programs. It is important to note, however, that such decisions could affect inference given that sample units may not be chosen at random.

sgsR is intended to provide methods for optimizing the allocation of sample units to derive comprehensive data about the state of forest structure. Many stratification and sampling approaches presented in *sgsR* have been created to implement methodologies found in published literature (e.g. Gobakken *et al.*, 2013; Melville and Stone, 2016; Malone *et al.*, 2019; Queinnec *et al.*, 2021a). A challenge to fully evaluating the degree to which alternative sampling schema improve model estimates is the need to conduct field measurements for validation purposes; hence there is often a need to rely on simulations to assess the relative merits of different sampling strategies. Not having field measured attributes (e.g. timber volume or biomass) in these allocated sample units makes it challenging to immediately demonstrate their value for improving model-based inference. It is therefore fundamental that ALS metrics being used to drive SGS approaches are correlated to the attributes of modelling interest (Maltamo *et al.*, 2011; Gobakken *et al.*, 2013; Grafström *et al.*, 2014; Melville *et al.*, 2015).

The primary intended use case for *sgsR* is to enable SGS approaches using ALS data; however, the inclusion of alternate remotely sensed covariates is entirely possible and encouraged. Spectral data, climate variables, or species information could be essential to a given sampling strategy or desired inventory outcome. Incorporation of these auxiliary data into sampling strategies can also be implemented in *sgsR* if overall requirements for the data such as matching of the spatial extent and spatial resolution are met. The ability to combine additional covariates with structural data, or use them in isolation, provides users with an avenue to test, analyze and improve their justification for sampling approaches. Research into the effectiveness of using additional remote sensing covariates for improving forest inventory or other resources management sampling frameworks are welcome and encouraged.

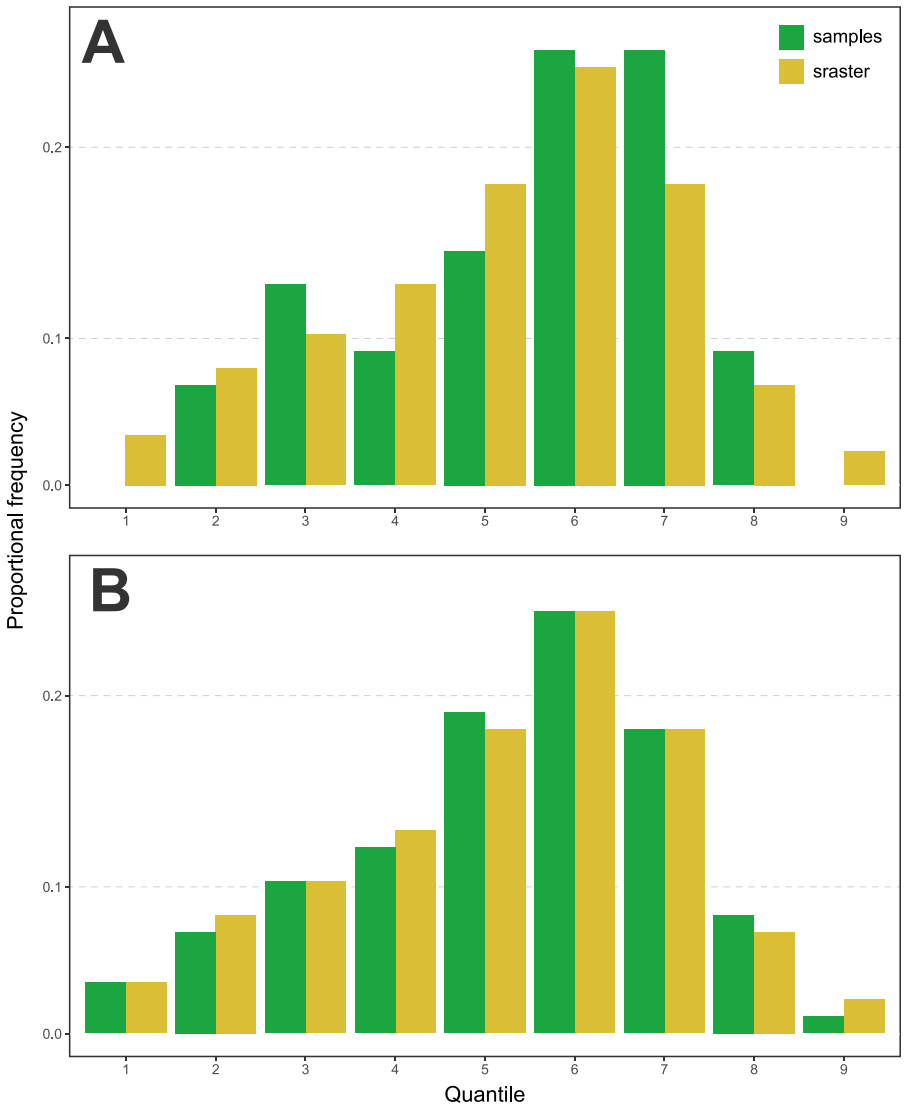

Figure 7 The sampling frequency (green) and stratum coverage frequency (yellow) for *existing* sample units only (A) and existing and new sample units (B) for the 90th percentile of height. Bars with an unequal value for a given quantile indicate that the number of sample units and the coverage of the strata are not proportional.

Conclusion

Increasing investments in ALS acquisitions by public and private stakeholders in the forest sector is influencing forest inventory approaches and has created a need for comprehensive tools for generating representative sample networks for the improvement of forest inventory attribute models. *sgsR* was created to provide a toolbox the forest management community could use that integrates multiple stratification and sampling approaches, existing sample networks and forest management data to improve the representative allocation of sample units. Building on the existing scientific literature as described herein, we have established the case for objective and transparent methods to improve and implement SGS routines that take full advantage of the synoptic characterization of forest structure that is afforded by wall-to-wall ALS acquisitions. We have described key algorithms and workflows to enable SGS approaches and presented reproducible examples of how *sgsR* can be implemented and adapted to develop SGS protocols according to the end user's specific information needs.

Supplementary data

Supplementary data are available at *Forestry* online.

Data Availability

Data internal to the sgsR package used in this article are included within the CRAN package release (https://cran.r-project.org/package=sgsR) *and Github repository–* (https://github.com/tgoodbody/sgsR), *and can be downloaded without the need for an access code.*

Conflict of interest statement

None declared.

Funding

We are very thankful to the Canadian Wood Fibre Centre's Forest Innovation Program (Project#: AWD-016477) for funding this work.

Acknowledgements

We are thankful to the editorial staff of *Forestry* and those who anonymously reviewed this manuscript.

References

van Aardt, J.A.N., Wynne, R.H. and Oderwald, R.G. 2006 Lidar-distributional parameters on a per-segment basis. *For. Sci.* **52**, 636–649.

Ayrey, E., Hayes, D.J., Fraver, S., Kershaw, J.A. Jr. and Weiskittel, A.R. 2019 Ecologically-based metrics for assessing structure in developing area-based, enhanced forest inventories from LiDAR. *Can. J. Remote. Sens.* **45**, 88–112. https://doi.org/10.1080/07038992.2019.1612738.

Bechtold, W.A. and Patterson, P.L. 2005 The enhanced forest inventory and analysis program – national sampling design and estimation procedures. *Gen. Tech. Rep.* SRS-80. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station. 85.

Benedetti, R., Piersimoni, F. and Postiglione, P. 2015 *Sampling Spatial Units for Agricultural Surveys*. Springer, https://doi.org/10.1007/978-3-662-46008-5.

Borders, B.E., Harrison, W.M., Clutter, M.L., Shiver, B.D. and Souter, R.A. 2008 The value of timber inventory information for management planning. *Can. J. For. Res.* **38**, 2287–2294. https://doi.org/10.1139/X08-075.

Bouvier, M., Durrieu, S., Fournier, R.A. and Renaud, J.P. 2015 Generalizing predictive models of forest inventory attributes using an area-based approach with airborne LiDAR data. *Remote Sens. Environ.* **156**, 322–334. https://doi.org/10.1016/j.rse.2014.10.004.

Chen, Q., McRoberts, R.E., Wang, C. and Radtke, P.J. 2016 Forest aboveground biomass mapping and estimation across multiple spatial scales using model-based inference. *Remote Sens. Environ.* **184**, 350–360. https://doi.org/10.1016/j.rse.2016.07.023.

Cook, R.D. 1975 Influential observations in linear regression. *J. Am. Stat. Assoc.* **74**, 169–174.

Corona, P. 2010 Integration of forest mapping and inventory to support forest management. *iForest* **3**, 59–64. https://doi.org/10.3832/ifor0531-003.

Dash, J. *et al.* 2016 Remote sensing for precision forestry. *N. Z. J. For.* **60**, 15–24.

Dash, J.P., Marshall, H.M. and Rawley, B. 2015 Methods for estimating multivariate stand yields and errors using k-NN and aerial laser scanning. *Forestry* **88**, 237–247. https://doi.org/10.1093/forestry/cpu054.

Davies, A.B. and Asner, G.P. 2014 Advances in animal ecology from 3D-LiDAR ecosystem mapping. *Trends Ecol. Evol.* **29**, 681–691. https://doi.org/10.1016/j.tree.2014.10.005.

Demaerschalk, J.P.P. and Kozak, A. 1974 Suggestions and criteria for more effective regression sampling. *Can. J. For. Res.* **4**, 341–348. https://doi.org/10.1139/x74-051.

Eid, T., Gobakken, T. and Næsset, E. 2004 Comparing stand inventories for large areas based on photo-interpretation and laser scanning by means of cost-plus-loss analyses. *Scand. J. For. Res.* **19**, 512–523. https://doi.org/10.1080/02827580410019463.

van Ewijk, K., Tompalski, P., Treitz, P., Coops, N.C., Woods (ret.), M. and Pitt (ret.), D. 2020 Transferability of ALS-derived forest resource inventory attributes between an eastern and western Canadian boreal forest mixedwood site. *Can. J. Remote. Sens.* **46**, 214–236. https://doi.org/10.1080/07038992.2020.1769470.

Fedrigo, M., Newnham, G.J., Coops, N.C., Culvenor, D.S., Bolton, D.K. and Nitschke, C.R. 2018 Predicting temperate forest stand types using only structural profiles from discrete return airborne lidar. *ISPRS J. Photogramm. Remote Sens.* **136**, 106–119. https://doi.org/10.1016/j.isprsjprs.2017.11.018.

Gobakken, T., Korhonen, L. and Næsset, E. 2013 Laser-assisted selection of field plots for an area-based forest inventory. *Silva Fennica* **47**, 1–20. https://doi.org/10.14214/sf.943.

Grafström, A. and Lisic, J. 2019 *BalancedSampling: Balanced and Spatially Balanced Sampling*. Available at: https://CRAN.R-project.org/package=BalancedSampling.Hijmans.

Grafström, A., Lundström, N.L.P. and Schelin, L. 2012 Spatially balanced sampling through the pivotal method. *Biometrics* **68**, 514–520. https://doi.org/10.1111/j.1541-0420.2011.01699.x.

Grafström, A. and Ringvall, A.H. 2013 Improving forest field inventories by using remote sensing data in novel sampling designs. *Can. J. For. Res.* **43**, 1015–1022. https://doi.org/10.1139/cjfr-2013-0123.

Grafström, A., Saarela, S. and Ene, L.T. 2014 Efficient sampling strategies for forest inventories by spreading the sample in auxiliary space. *Can. J. For. Res.* **44**, 1156–1164. https://doi.org/10.1139/cjfr-2014-0202.

Gregoire, T.G. 1998 Design-based and model-based inference in survey sampling: appreciating the difference. *Can. J. For. Res.* **28**, 1429–1447. https://doi.org/10.1139/x98-166.

Gregoire, T.G. and Valentine, H.T. 2007 *Sampling Strategies for Natural Resources and the Environment, Paper Knowledge. Toward a Media History of Documents*. Chapman and Hall/CRC Press. Available at: https://ca1lib.org/book/2639890/aac78b, https://doi.org/10.1201/9780203498880.

Hawbaker, T.J., Keuler, N.S., Lesak, A.A., Gobakken, T., Contrucci, K. and Radeloff, V.C. 2009 Improved estimates of forest vegetation structure and biomass with a LiDAR-optimized sampling design. *J Geophys Res Biogeosci* **114**, n/a–n/a. https://doi.org/10.1029/2008JG000870.

Hijmans, R. 2022 *terra: Spatial Data Analysis*. Available at: https://rspatial.org/terra/.

Holopainen, M., Vastaranta, M. and Hyyppä, J. 2014 Outlook for the next generation's precision forestry in Finland. *Forests* **5**, 1682–1694. https://doi.org/10.3390/f5071682.

Junttila, V., Finley, A.O., Bradford, J.B. and Kauranne, T. 2013 Strategies for minimizing sample size for use in airborne LiDAR-based forest inventory. *For. Ecol. Manag.* **292**, 75–85. https://doi.org/10.1016/j.foreco.2012.12.019.

Junttila, V., Maltamo, M. and Kauranne, T. 2008 Sparse Bayesian estimation of forest stand characteristics from airborne laser scanning. *For. Sci.* **54**, 543–552. https://doi.org/10.1093/forestscience/54.5.543.

Kane, V.R., McGaughey, R.J., Bakker, J.D., Gersonde, R.F., Lutz, J.A. and Franklin, J.F. 2010 Comparisons between field- and LiDAR-based measures of stand structural complexity. *Can. J. For. Res.* **40**, 761–773. https://doi.org/10.1139/X10-024.

Katila, M. and Tomppo, E. 2001 Selecting estimation parameters for the Finnish multisource National Forest Inventory. *Remote Sens. Environ.* **76**, 16–32.

Leckie, D. 2003 Stand delineation and composition estimation using semi-automated individual tree crown analysis. *Remote Sens. Environ.* **85**, 355–369. https://doi.org/10.1016/S0034-4257(03)00013-0.

Lefsky, M.A., Cohen, W.B., Acker, S.A., Parker, G.G., Spies, T.A. and Harding, D. 1999 Lidar remote sensing of the canopy structure and biophysical properties of Douglas-fir western hemlock forests. *Remote Sens. Environ.* **70**, 339–361. https://doi.org/10.1016/S0034-4257(99)00052-8.

Leiterer, R., Furrer, R., Schaepman, M.E. and Morsdorf, F. 2015 Forest canopy-structure characterization: a data-driven approach. *For. Ecol. Manag.* **358**, 48–61. https://doi.org/10.1016/j.foreco.2015.09.003.

Lisic, J. and Grafström, A. 2018 *Sampling Big Data: Sampling Methods for Big Data*. Available at: https://CRAN.R-project.org/package=SamplingBigData.

Malone, B.P., Minansy, B. and Brungard, C. 2019 Some methods to improve the utility of conditioned Latin hypercube sampling. *Peer J.* **7**, e6451. https://doi.org/10.7717/peerj.6451.

Maltamo, M., Bollandsas, O.M., Naesset, E., Gobakken, T. and Packalen, P. 2011 Different plot selection strategies for field training data in ALS-assisted forest inventory. *Forestry* **84**, 23–31. https://doi.org/10.1093/forestry/cpq039.

Maltamo, M., Packalen, P. and Kangas, A. 2021 From comprehensive field inventories to remotely sensed wall-to-wall stand attribute data-a brief history of management inventories in the Nordic countries. *Can. J. For. Res.* **51**, 257–266. https://doi.org/10.1139/cjfr-2020-0322.

McRoberts, R.E. 2012 Estimating forest attribute parameters for small areas using nearest neighbors techniques. *For. Ecol. Manag.* **272**, 3–12. https://doi.org/10.1016/j.foreco.2011.06.039.

McRoberts, R.E., Gobakken, T. and Næsset, E. 2012 Post-stratified estimation of forest area and growing stock volume using lidar-based stratifications. *Remote Sens. Environ.* **125**, 157–166. https://doi.org/10.1016/j.rse.2012.07.002.

McRoberts, R.E. and Tomppo, E.O. 2007 Remote sensing support for national forest inventories. *Remote Sens. Environ.* **110**, 412–419. https://doi.org/10.1016/j.rse.2006.09.034.

McRoberts, R.E., Tomppo, E.O. and Czaplewski, R.L. 2014 Sampling designs for national forest assessments: knowledge reference for national forest assessments. *FAO* 23–40. https://scholar.google.ca/scholar?hl=en&as_sdt=0,5&q=Sampling+designs+for+national+forest+assessments:+knowledge+reference+for+national+forest+assessments&btnG=

Melville, G. and Stone, C. 2016 Optimising nearest neighbour information – a simple, efficient sampling strategy for forestry plot imputation using remotely sensed data. *Aust. For.* **79**, 217–228. https://doi.org/10.1080/00049158.2016.1218265.

Melville, G., Stone, C. and Turner, R. 2015 Application of LiDAR data to maximise the efficiency of inventory plots in softwood plantations. *N. Z. J. For. Sci.* **45**, 9. https://doi.org/10.1186/s40490-015-0038-7.

Minasny, B. and McBratney, A.B. 2002 Uncertainty analysis for pedotransfer functions. *Eur. J. Soil Sci.* **53**, 417–429. https://doi.org/10.1046/j.1365-2389.2002.00452.x.

Minasny, B. and McBratney, A.B. 2006 A conditioned Latin hypercube method for sampling in the presence of ancillary information. *Comput. Geosci.* **32**, 1378–1388. https://doi.org/10.1016/j.cageo.2005.12.009.

Montgomery, D.C., Peck, E.A. and Vinning, G.G. 2006 *Introduction to Linear Regression Analysis*. Wiley.

Næsset, E. 2002 Predicting forest stand characteristics with airborne scanning laser using a practical two-stage procedure and field data. *Remote Sens. Environ.* **80**, 88–99. https://doi.org/10.1016/S0034-4257(01)00290-5.

Niemi, M. and Vauhkonen, J. 2016 Extracting canopy surface texture from airborne laser scanning data for the supervised and unsupervised prediction of area-based forest characteristics. *Remote Sens.* **8**, 582. https://doi.org/10.3390/rs8070582.

O'Brien, T.E. and Funk, G.M. 2003 A gentle introduction to optimal design for regression models. *Am. Stat.* **57**, 265–267.

de Almeida Papa, D., Almeida, D.R.A., Silva, C.A., Figueiredo, E.O., Stark, S.C., Valbuena, R. *et al.* 2020 Evaluating tropical forest classification and field sampling stratification from lidar to reduce effort and enable landscape monitoring. *For. Ecol. Manag.* **457**, 117634. https://doi.org/10.1016/j.foreco.2019.117634.

Pebesma, E. 2018 Simple features for R: standardized support for spatial vector data. *R J.* **10**, 439. https://doi.org/10.32614/RJ-2018-009.

Puliti, S., Saarela, S., Gobakken, T., Ståhl, G. and Næsset, E. 2018 Combining UAV and Sentinel-2 auxiliary data for forest growing stock volume estimation through hierarchical model-based inference. *Remote Sens. Environ.* **204**, 485–497. https://doi.org/10.1016/j.rse.2017.10.007.

Queinnec, M., Coops, N.C., White, J.C., McCartney, G. and Sinclair, I. 2021a Developing a forest inventory approach using airborne single photon lidar data: from ground plot selection to forest attribute prediction. *For. Int. J. For. Res.* **95**, 347–362. https://doi.org/10.1093/forestry/cpab051.

Queinnec, M., White, J.C. and Coops, N.C. 2021b Comparing airborne and spaceborne photon-counting LiDAR canopy structural estimates across different boreal forest types. *Remote Sens. Environ.* **262**, 112510. https://doi.org/10.1016/j.rse.2021.112510.

R Core Team 2022 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Available at: http://www.r-project.org/.

Roudier, P. 2011 *clhs: A R Package for Conditioned Latin Hypercube Sampling*. Oxford University Press.

Roussel, J.-R. and Auty, D. 2022 *lidR: Airborne LiDAR Data Manipulation and Visualization for Forestry Applications*. Available at: https://cran.r-project.org/package=lidR.

Roussel, J.-R.J.R., Auty, D., Coops, N.C., Tompalski, P., Goodbody, T.R.H., Meador, A.S. *et al.* 2020 lidR: an R package for analysis of Airborne Laser Scanning (ALS) data. *Remote Sens. Environ.* **251**, 112061. https://doi.org/10.1016/j.rse.2020.112061.

Silvey, S. 2013 *Optimal Design: An Introduction to the Theory for Parameter Estimation*. Springer Science & Business Media.

Smith, K. 1918 On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations. *Biometrika* **12**, 1–85. https://doi.org/10.2307/2331929.

Spies, T.A. 1998 Forest structure: a key to the ecosystem. *Northwest Sci.* **72**, 34–36.

Thom, D. and Keeton, W.S. 2019 Stand structure drives disparities in carbon storage in northern hardwood-conifer forests. *For. Ecol. Manag.* **442**, 10–20. https://doi.org/10.1016/j.foreco.2019.03.053.

Tompalski, P., White, J.C., Coops, N.C. and Wulder, M.A. 2019 Demonstrating the transferability of forest inventory attribute models derived using airborne laser scanning data. *Remote Sens. Environ.* **227**, 110–124. https://doi.org/10.1016/j.rse.2019.04.006.

Tompalski, P., White, J.C., Coops, N.C., Wulder, M.A., Leboeuf, A., Sinclair, I. *et al.* 2021 Quantifying the precision of forest stand height and canopy cover estimates derived from air photo interpretation. *For. Int. J. For. Res.* **94**, 611–629. https://doi.org/10.1093/forestry/cpab022.

Tomppo, E., Gschwantner, T., Lawrence, M., Mcroberts, R.E., Godinho-Ferreira, P. *et al.* 2010 *National Forest Inventories Pathways for Common Reporting*. Springer. doi:https://doi.org/10.1007/978-90-481-3233-1.

Tomppo, E., Malimbwi, R., Katila, M., Mäkisara, K., Henttonen, H.M., Chamuya, N. *et al.* 2014 A sampling design for a large area

forest inventory: case Tanzania. *Can. J. For. Res.* **44**, 931–948. https://doi.org/10.1139/cjfr-2013-0490.

Wästlund, A., Holmgren, J., Lindberg, E., Olsson, H. *et al.* 2018 Forest variable estimation using a high altitude single photon lidar system. *Remote Sens.* **10**, 1442. https://doi.org/10.3390/rs10091442.

White, J.C., Wulder, M.A., Varhola, A., Vastaranta, M., Coops, N.C., Cook, B.D. *et al.* 2013 A best practices guide for generating forest inventory attributes from airborne laser scanning data using an area-based approach. *For. Chron.* **89**, 722–723. https://doi.org/10.5558/tfc2013-132.

White, J.C., White, J.C., Tompalski, P., Vastaranta, M., Wulder, M., Saarinen, N. *et al.* 2017 *A Model Development and Application Guide for Generating an Enhanced Forest Inventory Using Airborne Laser Scanning Data and an Area-based Approach*. Canadian Forest Service, Canadian Wood Fibre Centre, Pacific Forestry Centre, pp. 1–48.

Yang, L., Li, X., Shi, J., Shen, F., Qi, F., Gao, B. *et al.* 2020 Evaluation of conditioned Latin hypercube sampling for soil mapping based on a machine learning method. *Geoderma* **369**, 114337. https://doi.org/10.1016/j.geoderma.2020.114337.

Yu, X., Kukko, A., Kaartinen, H., Wang, Y., Liang, X., Matikainen, L. *et al.* 2020 Comparing features of single and multi-photon lidar in boreal forests. *ISPRS J. Photogramm. Remote Sens.* **168**, 268–276. https://doi.org/10.1016/j.isprsjprs.2020.08.013.