NRC Research Press

ARTICLE

Prediction of diameter distributions and tree-lists in southwestern Oregon using LiDAR and stand-level auxiliary information

Francisco Mauro, Bryce Frank, Vicente J. Monleon, Hailemariam Temesgen, and Kevin R. Ford

Abstract: Diameter distributions and tree-lists provide information about forest stocks disaggregated by size and species and are key for informing forest management. Diameter distributions and tree-lists are multivariate responses, which makes the evaluation of methods for their prediction reliant on the use of dissimilarity metrics to summarize differences between observations and predictions. We compared four strategies for selection of *k* nearest neighbors (*k*-NN) methods to predict diameter distributions and tree-lists using LiDAR and stand-level auxiliary data and analyzed the effect of the *k*-NN distance and number of neighbors in the performance of the predictions. Strategies differed by the dissimilarity metric used to search for optimal *k*-NN configurations and the presence or absence of post-stratification. We also analyzed how alternative *k*-NN configurations ranked when tree-lists were aggregated using different DBH classes and species groupings. For all dissimilarity metrics, *k*-NN configurations using random-forest distance and three or more neighbors provided the best results. Rankings of *k*-NN configurations based on different dissimilarity metrics were relatively insensitive to changes on the width of the DBH classes and the definition of the species groups. The selection of the *k*-NN methods was clearly dependent on the choice of the dissimilarity metric. Further research is needed to find suitable ways to define dissimilarity metrics that reflect how forest managers evaluate differences between predicted and observed tree-lists and diameter distributions.

Key words: tree-lists, diameter distributions, *k*-NN, Reynolds index, LiDAR.

Résumé : Les distributions de diamètre et les listes d'arbres fournissent des informations sur les stocks forestiers, ventilées par taille et par espèce, et sont des informations de gestion forestière essentielles. Les distributions de diamètre et les listes d'arbres sont influencées par plusieurs variables, par conséquent l'évaluation des méthodes servant à les prédire dépend de l'utilisation de mesures de dissimilarité pour exprimer les différences entre les observations et leurs prédictions. Nous avons comparé quatre stratégies de sélection des k plus proches voisins (*k*-NN) pour prédire les distributions de diamètre et les listes d'arbres à l'aide de données auxiliaires LiDAR et de peuplement. Nous avons également analysé les effets de la distance *k*-NN et du nombre de voisins sur la qualité des prédictions. Les stratégies différaient par la mesure de dissimilarité utilisée pour obtenir les configurations *k*-NN optimales et par l'utilisation, ou non, de la post-stratification. Nous avons également analysé le classement des différentes configurations de *k*-NN lorsque les listes d'arbres étaient agrégées selon différentes classes de diamètre et groupes d'espèces. Pour toutes les mesures de dissimilarité, les configurations *k*-NN utilisant une distance fondée sur les forêts d'arbres décisionnels et trois voisins ou plus ont donné les meilleurs résultats. Les classements des configurations *k*-NN fondées sur différentes mesures de dissimilarité étaient relativement insensibles aux modifications de : l'étendue des classes de diamètre et la définition des groupes d'espèces. Le choix des méthodes *k*-NN dépendait clairement du choix de la mesure de dissimilarité. Des recherches supplémentaires sont nécessaires pour trouver des moyens appropriés de définir des mesures de dissimilarité qui reflètent l'évaluation que font les aménagistes forestiers des différences entre les listes d'arbres et les distributions de diamètre prédites et observées. [Traduit par la Rédaction]

Mots-clés : liste d'arbres, distribution de diamètre, *k*-NN, indice de Reynolds, LiDAR.

1. Introduction

To make informed decisions, forest managers require information about the current state of the forest. Knowledge of the total stand volume or basal area is not enough for forest managers to accurately characterize forest attributes to project future growth and yield, develop harvest schedules, and prioritize stands for treatments (such as thinning or regeneration harvest). To that end, information about forest stocks disaggregated by size classes and by species is of primary importance. Estimates of the diameter distribution become necessary for those evaluations and are one of the most important outputs that a complete forest inventory seeks to provide.

Diameter distributions provide stocking information for different diameters at breast height (DBH, 1.3 m) by species (or species groups) classes in terms of stand density, basal area, or volume on a per unit area basis. DBH classes are usually defined using widths

Received 31 July 2018. Accepted 16 December 2018.

F. Mauro, B. Frank, and H. Temesgen. Forest Engineering Resources and Management Department, College of Forestry, Oregon State University, Corvallis, Oregon, USA.
V.J. Monleon. US Forest Service, Pacific Northwest Research Station, Corvallis Forestry Sciences Laboratory, Corvallis, Oregon, USA.
K.R. Ford. Bureau of Land Management Oregon, Washington State Office, Portland, Oregon, USA.
Corresponding author: Francisco Mauro (email: francisco.mauro@oregonstate.edu).

that can range from 1 cm to 10–15 cm. Hereafter, we will use the term DBH × species class to refer to the categories or bins that form the diameter distribution. These DBH × species classes are the result of crossing the DBH classes and the species classes or species groups under study. Oftentimes, instead of discrete diameter distributions with pre-determined DBH classes, forest managers prefer outputs from forest inventories that allow them to use different DBH classes and different species groupings to address specific management decisions. Continuous diameter distribution models and tree-lists can be used to satisfy that need. An alternative output is an expected tree-list for an area, which is a table where all trees for a particular area are "enumerated" along with their attributes and their corresponding expansion factors (modified sampling weights that can be used to scale up field measurements to reference area totals) (Bechtold and Patterson 2005; Avery and Burkhart 2015). In addition to being a more flexible output than fixed-width discrete diameter distributions, tree-lists are commonly used as input data for growth or stand development models (Temesgen et al. 2003) such as the U.S. Forest Service Forest Vegetation Simulator (FVS) (Wykoff et al. 1982; Dixon 2002) and ORGANON (Hann and Larsen 1991).

1.1. Prediction of tree-lists and diameter distributions

1.1.1. Prediction methods

Diameter distributions and tree-lists are multivariate. So, to predict them, it is necessary to use techniques that can efficiently handle high-dimensional responses.

Parameter prediction (Gobakken and Næsset 2004; Maltamo et al. 2007), parameter recovery (Gobakken and Næsset 2005; Peuhkurinen et al. 2011; Arias-Rodil et al. 2018), and generalized linear models (Breidenbach et al. 2008) have been used to model continuous diameter distributions. These methods provide good results in even-aged forests and tend to perform worse in uneven-aged forests with bimodal DBH distributions. Variants of these methods using mixture models with two Weibull density functions have been proposed for multi-modal distributions (Thomas et al. 2008). Another commonly used method that can be used with multi-modal distributions consists of predicting the total stocking and a set of 10 or 12 percentiles that are later used to recover a continuous representation of the DBH density function by interpolation (Gobakken and Næsset 2005; Bollandsås and Næsset 2007; Bollandsås et al. 2013). All these strategies require multiple models, and neither adapt very well to the prediction of tree-lists.

Tomppo and Katila (1991) and Moeur and Stage (1995) identified *k*-nearest neighbor (*k*-NN) techniques as very appealing alternatives to parametric regression models, because they allow for estimates of a large number of variables of interest using a single prediction method. In *k*-NN, predictions for an unsampled target unit are generated as weighted averages of the values observed for the variables of interest in the *k* plots that are most similar, in terms of auxiliary information, to the target unit. Similarity between a target and sample plot is measured using a distance metric, which quantifies distance between the target and sample plot in the space defined by the set of auxiliary variables. Maltamo and Kangas (1998) and Temesgen et al. (2003) found *k*-NN methods to be a flexible, accurate, and operative method of predicting tree-lists and diameter distributions. In both cases, stands were the elementary unit used for prediction but later studies used the area-based approach (ABA) (Næsset 2002) in combination with *k*-NN methods finding similar results (Packalén and Maltamo 2007; Peuhkurinen et al. 2008; Bollandsås et al. 2013; Strunk et al. 2017; Lamb et al. 2018). Considering their success in previous experiences, we will focus hereafter on *k*-NN methods to predict diameter distributions and tree-lists.

For a given problem, multiple *k*-NN configurations can be used, depending on factors such as the set of auxiliary variables, the number of nearest neighbors (*k*), the distance metric, or the neighbor weighting scheme (e.g., Ohmann and Gregory (2002); Eskelson et al. (2009); Chirici et al. (2016)). The modeler uses the sample data to search for *k*-NN configurations that minimize certain measure of dissimilarity between observed and predicted values and provide optimal answers with respect to such measure for the problem at hand. Dissimilarity metrics such as the root mean square error (RMSE) in univariate problems are functions that provide a measure of the average error obtained with a given prediction method. For multivariate problems, a variety of dissimilarity metrics are used. However, although all multivariate dissimilarity metrics try to quantify average errors, they all do it in a different manner. Some dissimilarity metrics are based on distance metrics, so to avoid confusion between the distance metrics used to identify sample plots to be imputed into the target unit and dissimilarity metrics used to evaluate the performance of the prediction, we will use the terms *k*-NN distance and dissimilarity, respectively.

In practical applications where a large number of auxiliary variables and responses are available, automatic methods to search for optimal *k*-NN configurations become necessary. These automatic methods require choosing a dissimilarity metric to minimize and an algorithm to search the space of possible *k*-NN configurations (Packalén and Maltamo 2007; Packalén et al. 2012). In this manuscript, we focus on the dissimilarity metrics used in the search for optimal configurations and will not consider the algorithms used in the search. Hence, the term search strategy will be used to refer to searching procedures that optimize certain dissimilarity metric.

1.1.2. Strategies to select k-NN configurations

Commonly used strategies to search for *k*-NN configurations to predict diameter distributions have typically focused on first identifying a small set of candidate configurations that optimize the joint prediction of a relatively small sets of aggregated variables such as total stand volume and (or) basal area (Packalén and Maltamo 2008; Packalén et al. 2012; Bollandsås et al. 2013; Peuhkurinen et al. 2018) by minimizing a measure of error for these aggregated variables. Then, in a further step, candidates are ranked based on their ability to predict diameter distributions using a dissimilarity metric such as the average Reynolds' error index $\bar{E}$ (Reynolds et al. 1988), which is a dissimilarity metric specially tailored to evaluate the agreement between sets of observed and predicted diameter distributions (e.g., Packalén and Maltamo 2008; Räty et al. 2018). If a given dissimilarity metric for diameter distributions will be used to rank methods, because the diameter distributions are of critical scientific or management interest, and if the identified dissimilarity metric is thought to be a good measure of predictive accuracy, it seems advisable to directly minimize such a metric in the search for optimal *k*-NN configurations. This should render gains in performance under that metric and result in better answers to the science and management questions compared to strategies that optimize the prediction of aggregated variables. To the best of our knowledge, no previous study has directly minimized dissimilarity metrics for diameter distributions when searching for optimal *k*-NN configurations, nor has a previous study compared the resulting predictions to those obtained using common strategies that rely on a first step where only a measure of error for aggregated variables is minimized.

The choice of the dissimilarity metric is an open question. Although $\bar{E}$ is an extensively used dissimilarity metric for diameter distributions, similar metrics such as the recently proposed *H* and *I* indexes (Strunk et al. 2017) can also be used. These metrics ($\bar{E}$, *H*, and *I*) first calculate the differences between observed and predicted values for each DBH × species class. Then, differences for all DBH × species classes in a plot are taken in absolute value in case of $\bar{E}$, and squared in case of I, and summed to obtain a plot level sum of differences. Finally, plot level sums of differences are av-

eraged to obtain a measure of the performance of a given prediction method. These dissimilarity metrics are easy to compute and have a relatively straightforward interpretation, as they provide the average sum of errors across all DBH × species classes. However, the relative simplicity of these metrics prevents them from considering the correlations that might exist among the responses in different DBH × species classes. Considering those correlations might be a desirable feature to account for in a dissimilarity metric.

The generalized root mean square difference (grmsd) is a multivariate dissimilarity metric that has been extensively applied to search for optimal *k*-NN configurations intended to simultaneously predict multiple responses. This metric is the square root of the mean of the squared Mahalanobis distance between observed and imputed values in a space defined by the observed values of the variables of interest (Crookston and Finley 2008). It accounts for the correlations among the variables of interest in the training sample and penalizes the cases in which the predicted values for the multiple variables of interest go against their observed correlations. Even though grmsd is a dissimilarity metric with appealing properties for the evaluation of methods aiming at predicting discrete diameter distributions (i.e., it considers the correlations among classes of the diameter distribution), as far as we know, it has not been used directly for this purpose. Hence, no references exist about how useful this metric is to evaluate dissimilarities of diameter distributions and whether or not using LiDAR auxiliary information improves diameter distribution predictions when this metric is used as the evaluation criteria.

In the remainder of the manuscript, we will consider multiple dissimilarity metrics, including both $\overline{E}$ and grmsd, but we do not attempt to establish any of these metrics as superior to another or as a standard for future applications. Different dissimilarity metrics provide complementary information, and the best dissimilarity metric to use will likely depend on the science and management questions being addressed. Thus, we evaluate multiple dissimilarity metrics with very different strengths and weaknesses.

Finally, a characteristic of dissimilarity metrics with applications to diameter distributions is that they require defining DBH × species classes before they are computed. If a *k*-NN method, identified as optimal for certain DBH × species classes, is then used to generate tree-lists, then the predicted tree-lists will be optimal to compute diameter distributions for the DBH × species classes used in the optimization. If other DBH × species classes are used in application of the predicted tree-lists (e.g., input for growth and yield modeling), then the selected *k*-NN method and the predicted tree-lists may not be optimal for that application. Forest managers could be interested in using different DBH classes or different species groupings for different problems. It seems reasonable to search for candidate *k*-NN configurations using DBH × species classes that are as close as possible to the ones preferred by the forest managers and then consider how different candidates perform when using DBH and species classes that are different from the ones used in the initial search. An analysis of the results obtained with the outlined selection procedure could provide important insights for future efforts to predict tree-lists and diameter distributions.

2. Objectives

This manuscript has two objectives. The main objective is to analyze different strategies to search for optimal *k*-NN configurations for prediction of tree-lists and diameter distributions. We compared the performance of *k*-NN configurations identified as optimal by four different searching strategies. The first two considered dissimilarity metrics directly applied to the diameter distribution (i.e., grmsd applied to diameter distributions and $\overline{E}$), and the last two optimized the prediction of aggregated variables. The factors *k*-NN distance and number of neighbors were considered in the comparisons to obtain insights for future applications.

A secondary objective of this study is to analyze how the ranking of the *k*-NN candidate configurations identified as optimal in the previous step, using a given DBH bin-width and species grouping, are altered if the tree-lists generated with those *k*-NN configurations are aggregated using different DBH bin-widths and (or) species classes.

These two topics are examined with a case study developed in a large-scale forest inventory context, where a combination of LiDAR, topographic variables, and stand-level attributes are available as auxiliary information.

3. Materials

3.1. Study area and field data collection

The study area comprises 98 104 ha of forests administered by the Bureau of Land Management (BLM) within the Southwestern Oregon (SWO) LiDAR dataset acquired by Oregon Department of Geology and Mineral Industries (DOGAMI) in Coos and Curry Counties (Fig. 1). Coastal coniferous forest dominates the landscape. Douglas-fir (*Pseudotsuga menziesii* (Mirb.) Franco) is the most abundant species but other softwood species such as western hemlock (*Tsuga heterophylla* (Raf.) Sarg.), grand fir (*Abies grandis* (Douglas ex D. Don) Lindley), and western redcedar (*Thuja plicata* Donn ex D. Don) are also common. Hardwood species such as red alder (*Alnus rubra* Bong.) and bigleaf maple (*Acer macrophyllum* Pursh) are relatively abundant.

In total, 861 circular, 492.12 m^2 (one-eighth of an acre) field plots were measured in the study area. The sampling design for the study area is described in Shin et al. (2016), and the methods described in Hawbaker et al. (2009) were followed to ensure that plots were uniformly spread over the space of LiDAR auxiliary variables (see Section 3.2). Coordinates of the plot center were obtained by differentially correcting GPS observations taken at the plot center during at least 10 min using a Topcon GRS-1 receiver. Based on previous experiments under closed canopy conditions with similar equipment and recording times, accuracy of coordinates is expected to be 2.5 m or better (Andersen et al. 2009; Valbuena et al. 2010). Species, DBH, height, crown ratio, and a live–dead binary variable were recorded for all trees with DBH larger than 12.7 cm in the plot. Trees with DBH smaller than 12.7 cm were only measured in a concentric, 78.74 m^2 subplot and were assigned a subplot to plot expansion factor of 6.25. For these trees DBH, tree height and species were recorded.

3.2. Auxiliary information

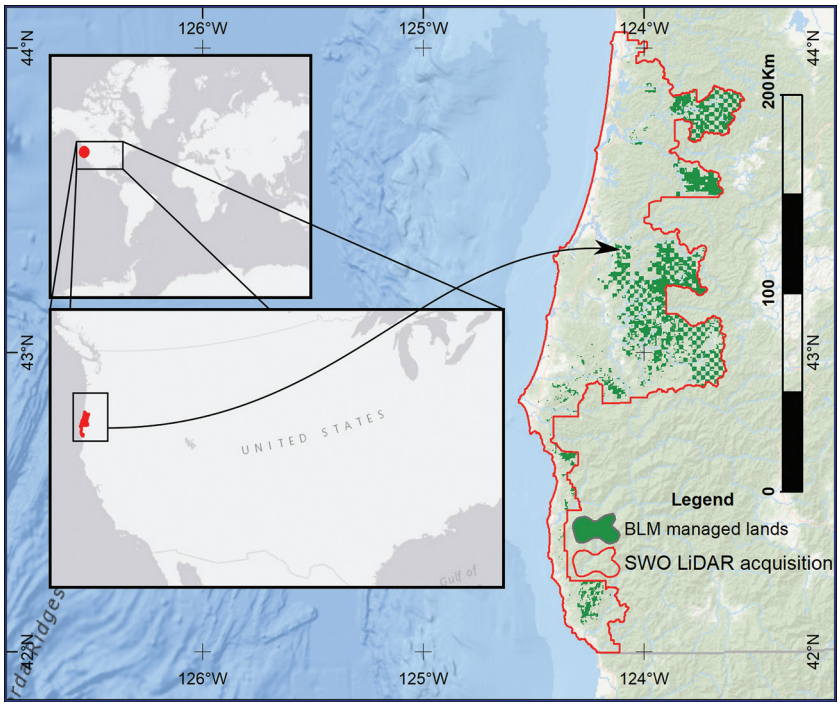
In total, 46 auxiliary variables, grouped in two sets, were available for the study area. A detailed description of these two sets of predictors is available in Supplementary Table S1.[1]

The first set of 21 predictors was obtained using LiDAR data acquired in the study area during the summers of 2008 and 2009. The average number of pulses per square meters was 5.6 pulses·m^{-2}, which resulted in an echo density of 8.1 points·m^{-2}. Nineteen of these predictors were point cloud metrics, including extreme values, percentiles, and fractions of pulses above different thresholds, computed using FUSION v2.80 (McGaughey 2010). In addition, the elevation above the sea level and the slope were also computed for each plot using a LiDAR derived digital terrain model (DTM).

The second set consisted of 24 stand-level predictors. The BLM operates with stands as management units and records stand boundaries, geographic attributes, and descriptors of forest structure in their Forest Operations Inventory (FOI) spatial database (Bureau of Land Management, OR/WA 2015). These stand-level de-

[1]Supplementary data are available with the article through the journal Web site at http://nrcresearchpress.com/doi/suppl/10.1139/cjfr-2018-0332.

Fig. 1. Location of the study area. The study area included the areas managed by the BLM (dark green areas) within the SWO LiDAR acquisition (red boundary). [Colour online.]



scriptors are available for all stands in the study area and included the age and species composition of up to three layers of vegetation (i.e., upper, middle, and bottom layers). The FOI database covers the entire study area and its information proceeds from multiple sources including photo-interpretation, field visits, previous inventories, and vegetation surveys. The first 19 stand-level auxiliary variables were obtained summarizing the information about the vegetation structure available in the FOI database (Supplementary Table S1[1]). The remaining five predictors consisted of the average slope in the stand and the percentage of stand area facing north, east, south, and west. These five auxiliary variables were computed using the LiDAR-derived DTM as an input, but their calculation requires aggregation of pixels at the stand level. All stand-level auxiliary variables were attributed to the field plots by intersecting the stand database and the field plots layer.

4. Methods

4.1. Objective 1. Comparison of search strategies

Strategies considered for the first objective were defined based on the dissimilarity metric used in the algorithm that searched for optimal *k*-NN configurations. The dissimilarity metrics were grouped into two categories. The first category (Group 1) contained the two multivariate metrics directly applied to diameter distributions. These metrics were computed in the algorithm that searched for optimal configurations using predefined DBH × species classes. The second category (Group 2) contained two metrics that are applied to aggregated variables (i.e., total volume without separating by species or DBH classes). These strategies were considered as the reference, as they are similar to commonly used search strategies employed in previous studies.

4.1.1. Strategies with dissimilarity metrics applied to diameter distributions (Group 1)

For the two strategies in Group 1, the search for optimal configuration was based on comparisons of observed and predicted diameter distributions providing volume for DBH × species classes. We used volume as the criterion because it correlates with other variables such as biomass or monetary value better than stand density or basal area. The dissimilarity metrics used for this purpose were grmsd applied to diameter distributions, $\text{grmsd}_{\text{dist}}$ hereafter, and the average Reynolds' index, $\bar{E}$.

The formula for $\text{grmsd}_{\text{dist}}$ is

$$\text{grmsd}_{\text{dist}} = \sqrt{\sum_{i=1}^{n} \boldsymbol{\Delta}_i^t \hat{\boldsymbol{\Sigma}}_{\text{dist}}^{-1} \boldsymbol{\Delta}_i} \quad (1)$$

where the superscript t is the transpose operator, and $\boldsymbol{\Delta}_i = (y_{i,1} - \hat{y}_{i,1}, \ldots, y_{i,m} - \hat{y}_{i,m})^t$ is the vector of differences, for plot i, between observed ($y_{i,j}$), and predicted volumes, ($\hat{y}_{i,j}$), for the $j = 1 \ldots m$ DBH × species classes. The number of field plots is n and $\hat{\boldsymbol{\Sigma}}_{\text{dist}}^{-1}$ is the inverse of the estimated variance–covariance matrix of the volumes by DBH × species classes observed in the sample.

$\bar{E}$ provides the average sum of errors across all DBH × species classes at a plot and has the same units as the tree characteristic calculated for each DBH × species class (in this case, $m^3 \cdot ha^{-1}$):

$$\bar{E} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} \left| y_{i,j} - \hat{y}_{i,j} \right| \quad (2)$$

Both $\text{grmsd}_{\text{dist}}$ and $\bar{E}$ require defining DBH × species classes before they are computed. DBH classes used in the search were 12.7 cm wide. A class width of 12.7 cm (5 inches) was found as an appropriate compromise that provided a reasonably narrow resolution for the diameter distribution but kept the number of DBH × species classes reasonably low for computational efficiency. In total, 11 species groups were considered. The first 9 groups corresponded to those species that accounted for at least 5% of the volume in at least five sampled plots. The remaining species were classified as "other broadleaves" and "other conifers". In total, 171 DBH × species classes were present in the study area. Finally, DBH × species group classes with very little variation

in the study area (i.e., with a coefficient of variation smaller than 0.01) were not considered when searching for *k*-NN configuration but were included in all other analyses. After the removal of size classes with negligible variation, the number of DBH × species group classes considered in the search for *k*-NN configurations was 123. The effect of removing these rare classes when searching for optimal *k*-NN configurations is expected to be negligible. If the majority of the plots have very similar volumes in a given DBH × class, the effect of imputing one plot or another will not result in a meaningful difference in $\text{grmsd}_{\text{dist}}$ and $\overline{E}$.

4.1.2. *Strategies with dissimilarity metrics applied to aggregated variables (Group 2)*

The first strategy in this group was similar to the ones used by Packalén and Maltamo (2008); Bollandsås et al. (2013). It involved searching for optimal *k*-NN configurations for the joint prediction of total volume, basal area, and stand density (trees per hectare). The dissimilarity metric used for these searches was grmsd applied to aggregated variables (*V*, BA, and *N*), with no differentiation by DBH or species, and we will denote it as $\text{grmsd}_{\text{V,BA,N}}$:

$$\text{(3)}\qquad \text{grmsd}_{\text{V,BA,N}} = \sqrt{\sum_{i=1}^{n} (V_i - \hat{V}_i, \text{BA}_i - \widehat{\text{BA}_i}, N_i - \hat{N}_i)^{\text{t}} \hat{\Sigma}^{-1}_{\text{V,BA,N}} (V_i - \hat{V}_i, \text{BA}_i - \widehat{\text{BA}_i}, N_i - \hat{N}_i)}$$

where V_i and $\hat{V}_i$, BA_i $\widehat{BA_i}$, and N_i and $\hat{N}_i$ are, respectively, the observed and predicted values of volume, basal area, and stand density for the *i*th plot and $\hat{\Sigma}^{-1}_{V,BA,N}$ is the inverse of the estimated variance covariance matrix for these three variables.

An additional strategy in Group 2 was used as a reference. It consisted of searching for optimal *k*-NN configurations for the prediction of *V*. This strategy was included in the analysis to assess the effect of not considering any degree of disaggregation by DBH or by species. The dissimilarity metric used for these searches was the RMSE_{V} for volume ($\text{m}^3 \cdot \text{ha}^{-1}$):

$$\text{(4)}\qquad \text{RMSE}_{\text{V}} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (V_i - \hat{V}_i)^2}$$

4.1.3. *Search algorithm*

For each strategy, a semi-automated procedure was used to search for optimal *k*-NN configurations. Hereafter, we will use the term *k*-NN configuration to refer to a particular *k*-NN variant defined by the following.

1. A number of neighbors, which can range from 1 to 5.
2. One of the following five distance metrics: (*i*) the Euclidean distance for normalized predictors, (*ii*) the Mahalanobis distance, (*iii*) the most similar neighbor distance (msn), (*iv*) the generalized nearest neighbor distance (gnn) distance, and (*v*) the random forest distance. All of these distance metrics and details about their computation can be found in the package yaImpute (Crookston and Finley 2008) for R 3.4.4 (R Core Team 2018).
3. A subset of the 45 predictors described in Section 3.2.

We did not consider different functions to weight donors in our analysis, because this factor has been found to have minor influence in predictions. Thus, we used the inverse distance weighting function of the yaImpute package for all *k*-NN configurations. The weight w_{ir} assigned by this function to the *r*th donor plot used when making predictions for target unit *i* is $w_{ir} = \frac{1}{1 + d_{ir}} \Big/ \sum_{r=1}^{k} \frac{1}{1 + d_{ir}}$, where d_{ir} is the distance between target unit *i* and its *r*th donor in the auxiliary variable space.

For a given distance metric and number of neighbors, the number of possible subsets of predictors is 2^{46}–1, which is approximately 7.04×10^{13}. This problem is very common in any model selection exercise involving LiDAR auxiliary information, as LiDAR provides a large number of auxiliary variables that are often strongly correlated with each other. This high-dimensionality problem is exacerbated in this study by the inclusion of stand-level auxiliary variables. Due to the large number of potential *k*-NN configurations, it is impossible to test all of them. Thus, the sequential removal algorithm implemented in the function varSelection in combination with the function bestVars of the package yaImpute were used with each of the combinations of distance metric and number of neighbors. By default, the function varSelection calls the function grmsd that computes this metric every time variables are added or removed. For the strategy that used $\overline{E}$ as the dissimilarity metric, we coded an R function to calculate its values. The function varSelection was modified accordingly, so instead of calling the function grmsd in every step of the optimization, it called the new function to compute $\overline{E}$.

4.1.4. *Analysis for objective 1. Comparisons of search strategies based on the DBH × species group classes used in the search for optimal configurations*

For each of the four compared strategies, the outcome of the automated search procedure was a set of 25 optimal *k*-NN configurations providing the best subset of predictors for each combination of number of neighbors and distance metric. The resulting 100 *k*-NN configurations (4 strategies and 25 configurations per strategy) were compared with each other.

For each of the 100 *k*-NN candidate configurations, we computed the four dissimilarity metrics used in the searches for optimal configurations. We compared (graphically and numerically) the values of the four dissimilarity metrics to identify the most important factors controlling the performance of alternative *k*-NN configurations. Comparisons were based on both the absolute and relative values obtained for these metrics. The relative loss in performance (eq. 5) of a given *k*-NN configuration (A) for a given dissimilarity metric (Diss) was measured as

$$\text{(5)}\qquad \text{Relative loss} = 100 \frac{\text{Diss}_{\text{A}} - \text{Diss}_{\text{Optim}}}{\text{Diss}_{\text{Optim}}} (\%)$$

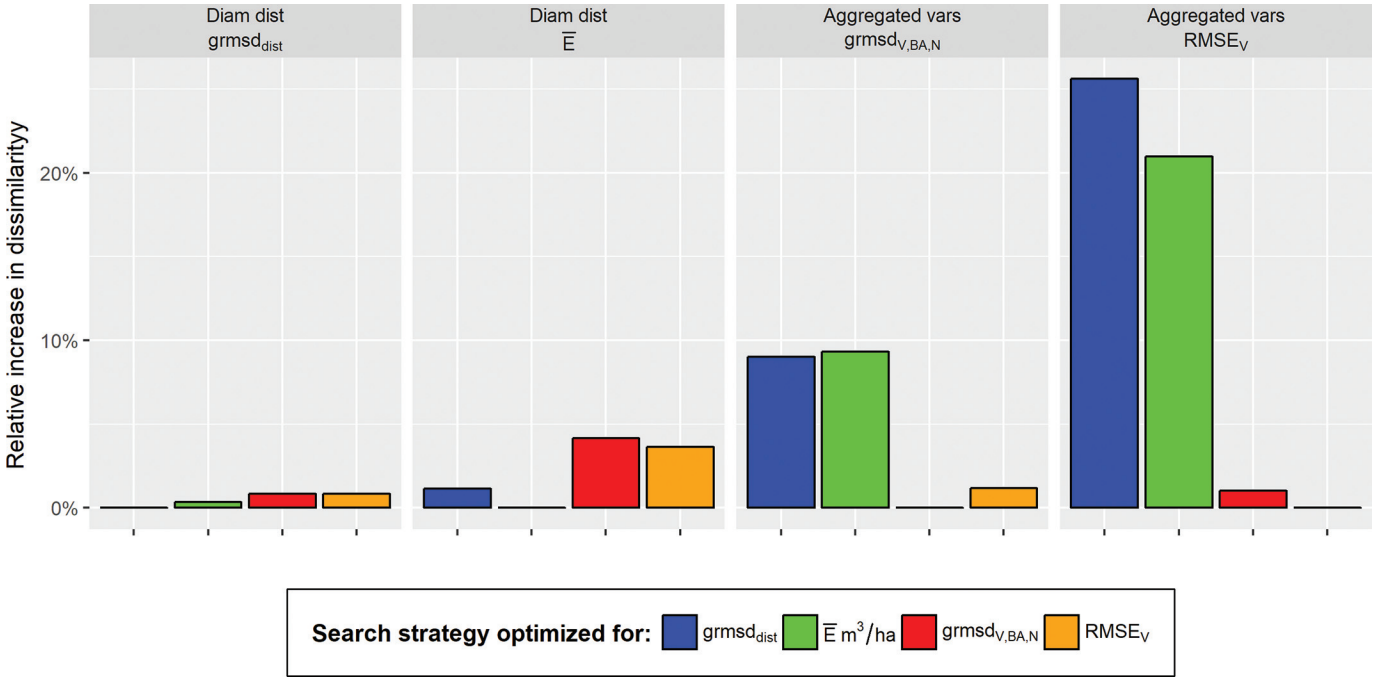
where Diss_{A} is the dissimilarity observed for the *k*-NN configuration (A), and $\text{Diss}_{\text{Optim}}$ is the minimum value of dissimilarity among the 100 *k*-NN candidates. Finally, we also used the values of the four dissimilarity metrics computed for the 100 candidate *k*-NN configurations for the entire study area to calculate the Kendall's tau correlation coefficients (Kendall 1938) between the four dissimilarity metrics. These correlations were used to assess how the rankings of *k*-NN configurations based on different dissimilarity metrics were related and to what extent different dissimilarity metrics were complementary.

To obtain a measure of the improvement with respect to a situation where no auxiliary information was available, the sample mean was taken as a baseline (i.e., all plots were imputed the sample mean). Then, the values of the four dissimilarity metrics were computed using the DBH × species class means as predicted values. The resulting values of $\text{grmsd}_{\text{dist}}$, $\overline{E}$, $\text{grmsd}_{\text{V,BA,N}}$, and RMSE_{V} were considered as references and compared with the ones obtained for those dissimilarity metrics using 100 *k*-NN configurations.

4.2. Objective 2. Performance of *k*-NN candidates when using different DBH × species classes

For objective 2, we focused on how the alternative *k*-NN configurations, identified as candidates when using the DBH × species described in Section 4.1.1, performed when using different DBH ×

Fig. 2. Relative differences in performance for $grmsd_{dist}$, $\overline{E}$, $grmsd_{V,BA,N}$, and $RMSE_V$. Different panels show the results for different dissimilarity metrics. Differences are computed between the best *k*-NN configuration obtained with a given search strategy and the overall minimum for that dissimilarity metric. [Colour online.]



species groupings. For this analysis, it is necessary to predict tree-lists for all plots using the 100 *k*-NN candidate configurations and, then, bin all the predicted tree-lists using different DBH × species classes. We first describe the process used to generate imputed tree-lists and then describe the comparisons made for objective 2.

4.2.1. Generation of imputed lists

The generation or prediction of plot-level tree-lists follows the same principles used when predicting other quantitative variables. *k*-NN predictions are obtained as weighted sums of the values of the variables of interest measured in the donors. Weights are typically standardized to sum to one, so predicted values can be seen as the result of averaging all donors, with each of them represented in a proportion specified by their *k*-NN weight in the target unit. This interpretation of the *k*-NN weights readily connects with the idea of expansion factors that allows, as explained below, the generation of predicted tree-lists even when *k* is larger than 1.

Predicted tree-lists for target units can be obtained by first multi-plying the expansion factor of each tree in each donor plot by the *k*-NN weight of the plot (see e.g., Packalén and Maltamo (2008) and Strunk et al. (2017)). Then, the updated tree-lists of all donor plots for a target unit are appended one after another. The updated expansion factors ensure that the portion of the resulting tree-list that originated from a specific donor scales up to a fraction of the target unit's predicted tree-list as specified by the donor plot's *k*-NN weight. Therefore, predicted tree-lists will be fully compatible with the logic followed in *k*-NN imputation to generate predictions for any variable of interest that is a sum of tree-level attributes in the target area (e.g., basal areas, volumes, trees per acre). These predicted tree-lists can also be used as if they were measured tree-lists to predict variables of interest that are functions of sums of tree-level attributes (e.g., stand density index). However, tree-lists generated with the outlined procedure will fail at estimating variables such as species richness (Magurran 2013) that are not the result of an aggregation of tree-level attributes. In addition, combining tree-lists from multiple donor plots may lead to unrealistic combinations of species and size classes.

4.2.2. Analysis of performance of k-NN alternatives for different DBH × species classes

For each of the 100 *k*-NN candidates, we generated the predicted tree-list for every plot using the method described in Section 4.2.1. Then, predicted and observed tree-lists were used to compute diameter distributions using 15 alternative sets of DBH × species bins. The width of the DBH classes took the values 2.54, 5.08, 12.70, and 25.40 cm (1, 2, 5, and 10 inches, respectively) plus no differentiation by diameter classes (i.e., no differentiation by size, only total volume by species group). The species groups that we considered were as follows: (*i*) the species groups described in Section 4.1.1, (*ii*) conifers vs hardwood, and (*iii*) all species together (no differentiation by species groups). The classes with coefficients of variation smaller than 0.01 were only removed in the search for *k*-NN candidates, but they were included in this analysis.

For each set of DBH × species bins, the 100 *k*-NN configurations were ranked according to $grmsd_{dist}$ and $\overline{E}$. We compared the rankings obtained for the 100 *k*-NN configurations when using the DBH × species bins of the search for optimal configurations (optimization rankings) with the rankings obtained for the *k*-NN configurations using each of the 14 alternative sets of DBH × species bins (alternative rankings). Kendal's tau correlation coefficient was used as an indicator to assess how *k*-NN configurations, identified as optimal for the DBH × species bins employed in the search for *k*-NN candidates, ranked when using a different DBH × species bins. Graphical comparisons of the rankings were also used to identify and compare the effects of changing the DBH class width and the effect of changing the species groupings.

5. Results

5.1. Objective 1

5.1.1. Dissimilarity metric

In general, the differences among strategies that minimized metrics specially tailored to analyze diameter distributions (i.e., strategies in Group 1) were of small magnitude (Fig. 2). However, strategies that relied on the joint optimization of *V*, BA, and *N*, or the optimization of *V* alone, resulted in worse values for $\overline{E}$ and

$grmsd_{dist}$. For the strategies optimizing $\overline{E}$ and $grmsd_{dist}$, differences between *k*-NN configurations were of smaller magnitude for $\overline{E}$ than for $grmsd_{dist}$. When optimizing for the prediction of *V*, BA, and *N*, the losses in performance were 1.36% for $grmsd_{dist}$ and 6.37% for $\overline{E}$. When optimizing for volume alone, losses in performance were 1.23% for $grmsd_{dist}$ and 3.87% for $\overline{E}$. On another hand, optimizing for error metrics applied to the diameter distribution results in losses in performance when predicting volume or when jointly predicting *V*, BA, and *N* (Group 2 metrics). Strategies that minimized $grmsd_{dist}$ or $\overline{E}$ provided values of $grmsd_{V,BA,N}$ that were from 8.09% to 10.06% larger than the optimum. Similarly strategies that minimized $grmsd_{dist}$ or Reynolds index yielded values of $RMSE_V$ that were from 16.01% to 22.17% larger than the minimum attained for this metric (Fig. 2).

Kendall's tau correlation coefficient for $\overline{E}$ and $grmsd_{dist}$ was 0.52, which indicates that rankings based on these two variables are moderately correlated but not completely equivalent (Supplementary Fig. S2[1]). Kendall's tau correlation coefficient between $grmsd_{V,BA,N}$ and $RMSE_V$ was 0.72, which indicates that these two dissimilarity metrics for aggregated variables are closely related. Both $\overline{E}$ and $grmsd_{dist}$ had significantly lower rank correlations with $grmsd_{V,BA,N}$ and $RMSE_V$. The values of the rank correlations between $\overline{E}$ and $grmsd_{V,BA,N}$ and between $\overline{E}$ and $RMSE_V$ were 0.34 and 0.41, respectively. Kendall's tau correlation coefficient for $grmsd_{dist}$ and $grmsd_{V,BA,N}$ and for $grmsd_{dist}$ and $RMSE_V$ were 0.66 and 0.59, respectively. These results show that each dissimilarity metric represents a different aspect of the differences between predicted and observed diameter distributions. Optimizing for any of them, will necessary imply sub-optimal predictions under other metrics.

5.1.2. Global trends in performance

General patterns in performance for the four dissimilarity metrics (Fig. 3) indicate that

1. Random forest distance tends to provide the best results for all dissimilarity metrics.
2. All dissimilarity metrics improved when the number of neighbors increased.
3. For $\overline{E}$, which is based on the absolute value of the errors, improvement due to increasing the number of neighbors was steady. For $grmsd_{dist}$, which is based on the squared value of the errors, the improvement when increasing *k* was very steep for values of *k* smaller than 3 and then tended to stabilize. Averages optimize square error losses, which seems to explain the larger impact of increasing the number of neighbors, and therefore the averaging, in this dissimilarity metric.

The importance of the *k*-NN distance compared with the effect of *k* showed different patterns depending on the dissimilarity metric. For $grmsd_{dist}$, differences between the best and worst *k*-NN configurations with the same *k*-NN distance but different number of neighbors were always larger than 0.2 and exceeded 0.3 multiple times (Fig. 4), whereas differences between the best and worst *k*-NN configurations with the same number of neighbors but different *k*-NN distance never reached 0.08 (Fig. 5). For $grmsd_{V,BA,N}$ and $RMSE_V$ increasing the number of neighbors had a larger influence on precision than changing the *k*-NN distance metric (Figs. 4 and 5). Finally, for $\overline{E}$, increasing the number of neighbors or changing the *k*-NN distance metric had similar effect. Differences between the best and worst configuration with the same number of neighbors and differences between best and worst configuration with a given *k*-NN distance were typically in the range of 50–150 $m^3 \cdot ha^{-1}$.

5.2. Objective 2. Performance when using DBH × species classes different than the ones used in the optimization

The second objective examines whether *k*-NN configurations identified as best for a specific DBH by species group classification still perform best when different groupings are considered. As can be expected, the more different the DBH × species classes to the ones used in the search, the weaker the correlation between rankings of *k*-NN configurations. However, rankings of *k*-NN alternatives are stable for a relatively large number of sets of DBH × species bins (Fig. 6). Increasing to 25.4 cm or decreasing to 2.5 cm the width of the DBH classes had a minor effect in the rankings of *k*-NN candidates. Using different species groupings seemed to affect rank correlations less, but this can be explained by the dominance of Douglas-fir in the study area, which causes that most of the volume will, at least, belong to the correct species.

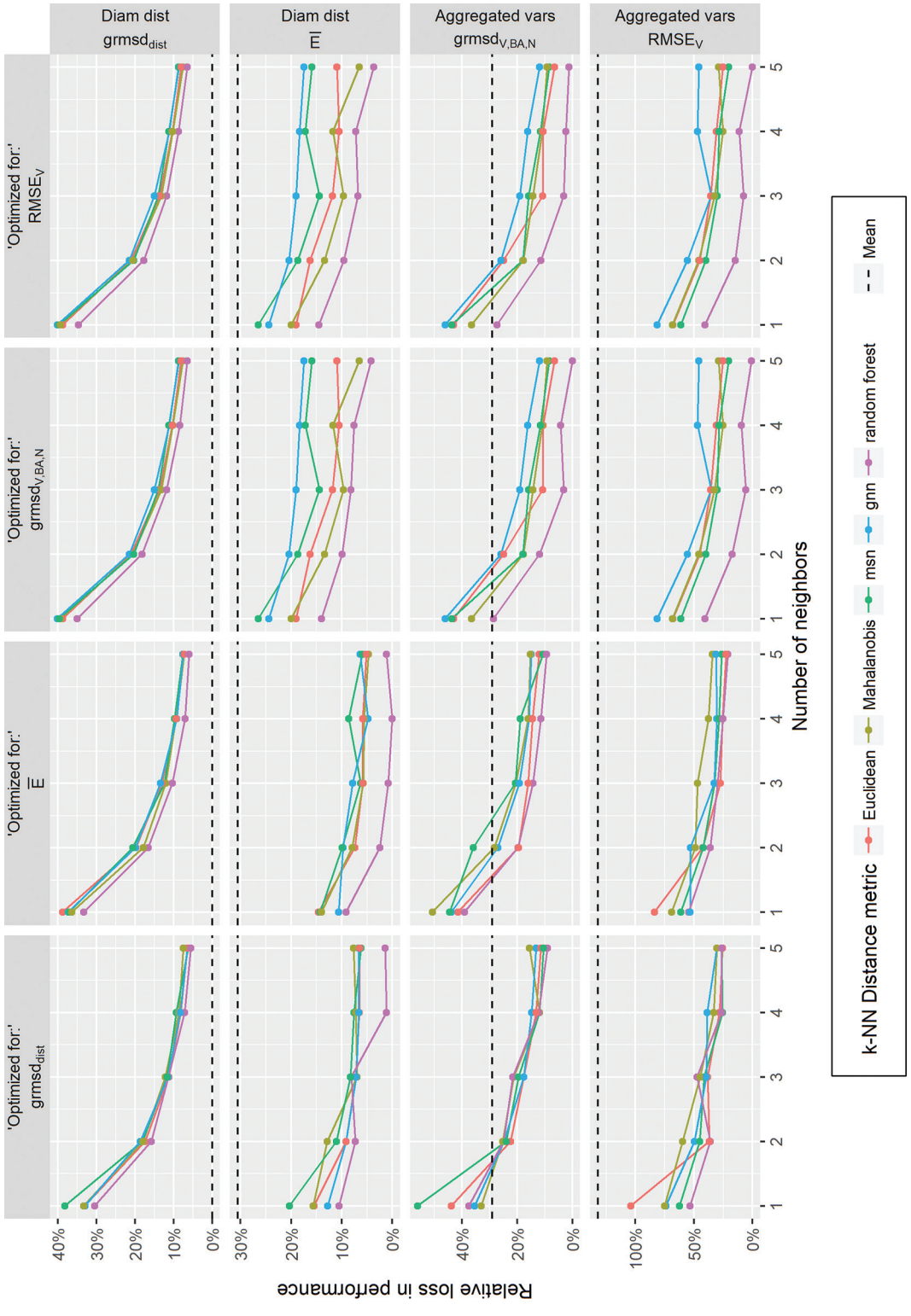
6. Discussion

An inherent quality of diameter distributions is their highly multivariate nature, which makes it necessary to rely on dissimilarity metrics to summarize discrepancies between observed and predicted values. The dissimilarity metrics used as ranking criteria imply different degrees of aggregation and inherent losses of information that may or may not result in realistic representations of the preferences of the forest managers. This makes it very difficult to translate the values of these dissimilarity metrics into a representation of the lack of resemblance between observed and predicted distributions that is meaningful for forest managers. This problem clearly appears when alternative *k*-NN configurations are compared (Fig. 2). For example, when $\overline{E}$ is optimized, $RMSE_V$ is a 20.97% larger than the value obtained when $RMSE_V$ is optimized. However, optimizing for $RMSE_V$ results in values of $\overline{E}$ that are 3.65% larger than the value obtained when $\overline{E}$ is optimized. To evaluate which one of those two *k*-NN configuration is better for a given problem, the forest manager and the modeler would need to evaluate the trade-off of an increase of 20.97% in $RMSE_V$ against an increase of 3.65% in $\overline{E}$. Although a 20.97% increase in $RMSE_V$ can be translated into something meaningful, evaluating the trade-off is likely difficult, especially if one considers situations like the one under study where 171 DBH × species classes are present in the study area.

Similarly, the best *k*-NN alternative with respect to $grmsd_{dist}$ produces values for this metric 5.18% larger than the values obtained with the sample mean, which implies that the incorporation of the auxiliary information does not produce gains with respect to this metric (Fig. 2). This result should be interpreted carefully. On one hand, grmsd or the average Mahalanobis distance has been used to account for correlations in many studies involving prediction of multivariate responses; however, to the best of our knowledge, it has never been applied to measure the agreement between observed and predicted diameter distributions. Further research is needed to confirm whether similar results are obtained for this metric in other study areas when predicting diameter distributions using similar auxiliary data. Although choosing the mean as a predictor for the whole study area would be the best option if only $grmsd_{dist}$ was considered, such a choice seems to be highly unsatisfactory under all other criteria. Using the mean as a predictor results in values for $\overline{E}$, $grmsd_{V,BA,N}$, and $RMSE_V$ that are, respectively, 30.55%, 28.97%, and 131.99% worse than the optimum value attained for each of these dissimilarity metrics and 29.65%, 18.32%, and 84.70% worse than values obtained with the *k*-NN configuration that provides the best values for $grmsd_{dist}$.

The choice of a dissimilarity metric for a given problem will condition all further stages for the selection of a *k*-NN configuration and, therefore, will affect all the final products and maps. The type of metric used for a particular problem strongly depends on the science and management questions that will be addressed using the predictions provided by the selected *k*-NN configuration. For example, if a manager primarily seeks to identify stands with high volume per acre values that lead to cost-effective regeneration harvests, use of a dissimilarity metric like $RMSE_V$ would be

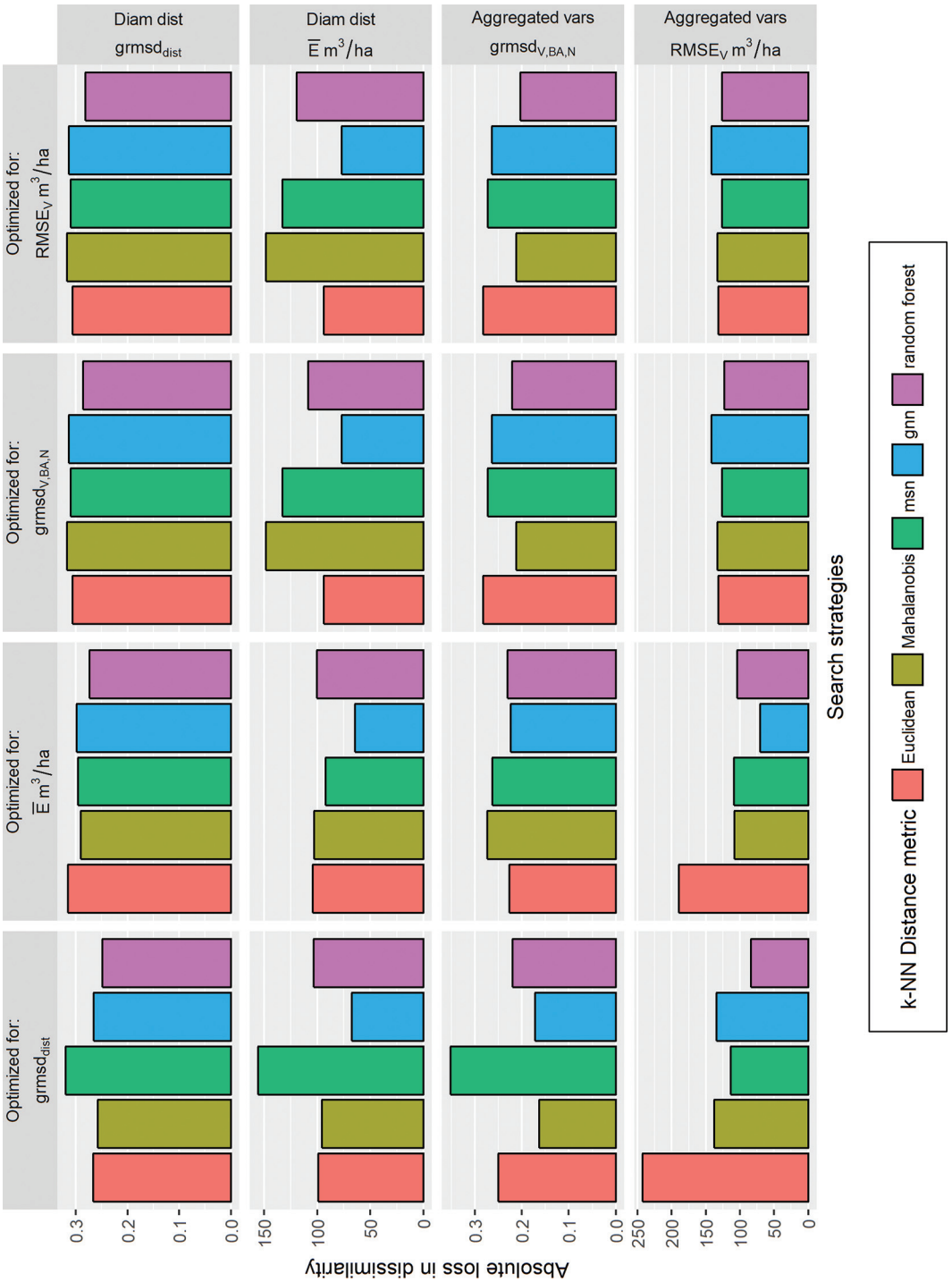
Fig. 3. Relative losses in performance as a percentage of the overall minimum attained for each dissimilarity metric. The dissimilarity metric used in the search is indicated in the column headings. Row headings indicate the dissimilarity metric that is evaluated. The reference value obtained using the sample mean as a prediction is indicated as a horizontal dashed line for each dissimilarity metric. [Colour online.]



best because it minimizes uncertainty in volume. Although the greater uncertainty in the distribution of that volume among DBH and species classes would be unfortunate because that distribution affects harvest value, it may be less important than total volume in certain landscapes.

If a manager seeks to thin stands for multiple objectives, such as improving stand-level drought resilience while also implementing an economically viable timber sale, a dissimilarity metric like grmsd$_{V,BA,N}$ might be best. This dissimilarity metric would balance the need for identifying dense stands most in need of treat-

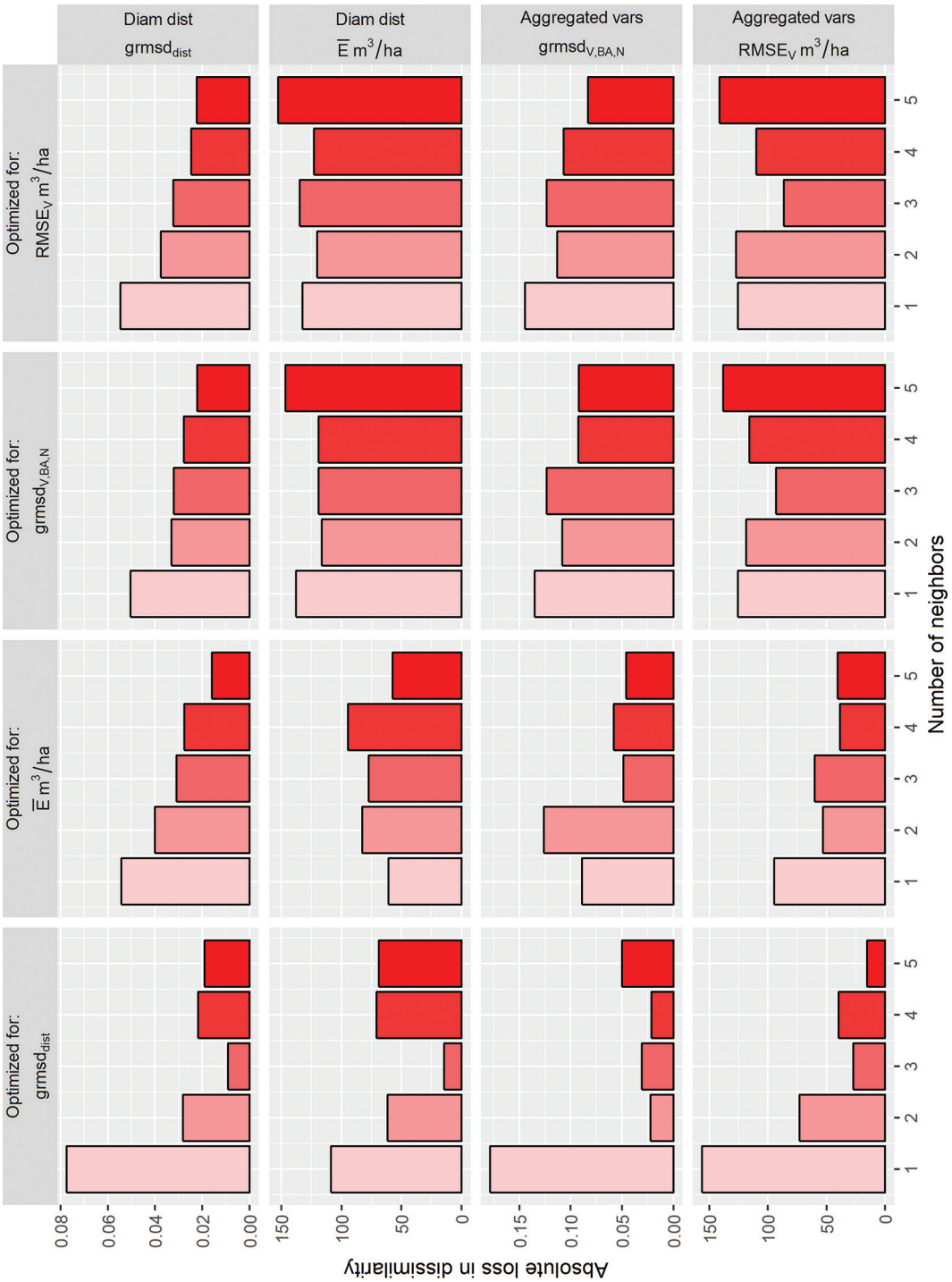
Fig. 4. Absolute loss in dissimilarity for $grmsd_{dist}$, $\overline{E}$, $grmsd_{V,BA,N}$, and $RMSE_V$ by search strategy, when comparing *k*-NN configurations obtained with the same distance metric. The dissimilarity metric used in the search is indicated in the column headings. [Colour online.]



ment (potentially quantifying this with Reineke's stand density index, which itself is based on basal area and the number of trees per unit area) that also have enough volume per acre to support an economically viable timber sale, which might be the only feasible way to conduct a treatment that promotes forest health.

If a manager wants to develop a harvest schedule or other type of long-term plan, then it will be important to know the current state of the forest and project how it will develop in the future using a growth and yield model. In this case, the manager might prefer dissimilarity metrics applied to DBH and species distributions (such as $\overline{E}$ or $grmsd_{dist}$) as opposed to dissimilarity metrics applied to aggregated variables (such as $RMSE_V$ or $grmsd_{V,BA,N}$). That is because stands with different DBH distributions and species compositions but similar initial stand-level attributes can develop into stands with very different stand-level attributes over time. Dissimilarity metrics that account for correlations among DBH × species classes (e.g., $grmsd_{dist}$) might be preferable to those that do not (e.g., $\overline{E}$) when interactions among trees within a stand are complex such as when competition is strongly asymmetric or species differ strongly in shade tolerance. In these cases, stand

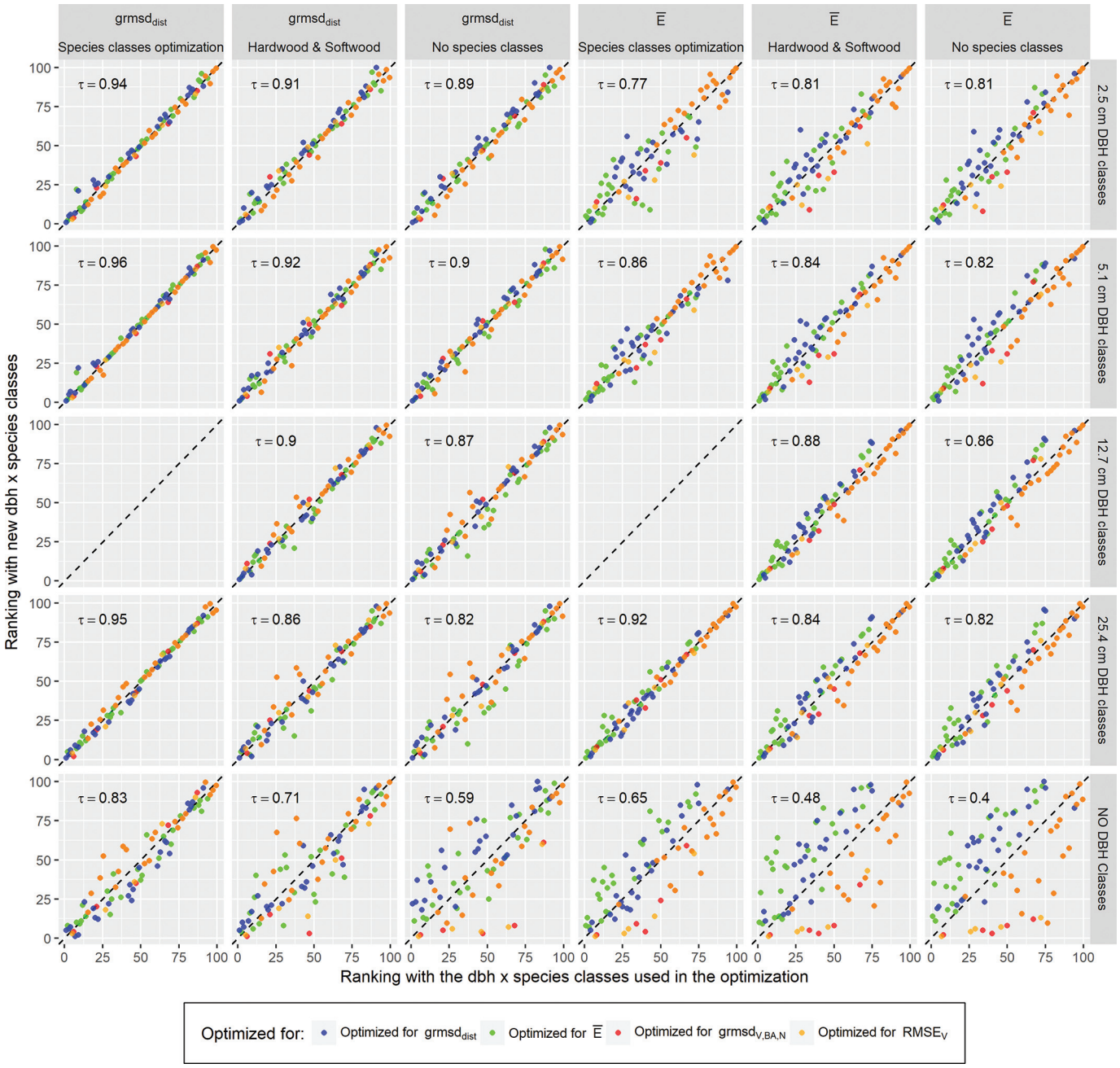
Fig. 5. Absolute loss in dissimilarity for $grmsd_{dist}$, $\overline{E}$, $grmsd_{V,BA,N}$, and $RMSE_V$ by search strategy, when comparing *k*-NN configurations with the number of neighbors. The dissimilarity metric used in the search is indicated in the column headings. [Colour online.]



dynamics are more likely to be an emergent property of the interactions among trees of different sizes and species and not simply the sum of dynamics within each DBH × species class, making it important to impute tree-lists with realistic combinations of DBH and species classes.

In any case, the difficulty of evaluating trade-offs between dissimilarity metrics makes comparisons cumbersome, and rankings of alternative methods based on a single dissimilarity metric must be considered only as a first approximation. Framing the selection of a *k*-NN configuration for prediction of diameter distributions as a decision with multiple criteria seems to be a better option (Tomppo and Halme 2004). In that case, expert systems such as the one described by Martínez-Falero et al. (2017) could be applied to (*i*) characterize the preferences of forest managers, or account for the practical consequences of over-predicting or under-predicting in a given DBH × species class and (*ii*) design ad hoc dissimilarity metrics that adapt better to a particular problem and context. Then, the obtained dissimilarity metric could be used in the selection of optimal *k*-NN configurations ensuring that the metric used to rank *k*-NN alternatives will satisfy the forest

Fig. 6. Rankings for the 100 *k*-NN candidates when using $grmsd_{dist}$ and Reynolds index using the original DBH × species groupings against rankings for the 100 *k*-NN candidates using different DBH × species classes. The *x* axis represents the ranking for the DBH × species classes used in the optimization. The dashed line represents the 1:1 line. [Colour online.]



manager's preferences as much as possible. To the best of our knowledge, this approach has not been explored and clearly deserves attention in forestry research.

Another important topic that merits further research is how to extrapolate results obtained with a plot-level validation to stand-level measures of performance. In most remote sensing applications, validations are performed at the level of the plot or pixel, areas typically around 400–1000 m^2. However, management decisions are typically made at larger scales such as stand, landscape, or regional scales. Unfortunately, errors or dissimilarities at the plot or pixel scale do not directly translate to errors or dissimilarities at larger scales such as a stand that is composed of many pixels. In addition, most of the time, it is unaffordable to have large enough samples to develop stand specific models and validations, and stands have to be regarded as small areas (Rao and Molina 2015). Advances in this field like the multivariate hierarchical models used by Finley et al. (2014) are strongly needed to provide forest managers with unbiased stand-level predictions of diameter distributions and reliable measures of their accuracy or, at least, with some quantitative information about the risks of using synthetic predictions in a particular scenario.

The distance metric and number of neighbors used had an impact on both $grmsd_{dist}$ and $\bar{E}$ similar to those found in previous studies analyzing dissimilarity metrics for aggregated variables (e.g., Hudak et al. (2008) and Zhu et al. (2017)). Best results tended to be obtained using the random forest distance and three or more neighbors. In comparative terms, the effect of *k* was larger than the effect of the distance metric for $grmsd_{dist}$, $grmsd_{V,BA,N}$,

and $RMSE_V$, whereas for $\bar{E}$, distance metric and number of neighbors had a similar effect in performance.

Changing the DBH × species classes had a small effect on the ranking of alternative *k*-NN configurations based on $grmsd_{dist}$ and $\bar{E}$. Rankings of *k*-NN configurations when using DBH classes as narrow as 2.5 cm or as wide as 25.4 cm were very similar to those obtained with the 12.7 cm DBH classes used in the search for optimal configurations. This result has important practical applications. Although it is not possible to ensure that a *k*-NN configuration selected based on certain DBH × species classes is going to be optimal for other DBH × species groupings, the fact that rankings only showed minor changes when changing the DBH × species classes (Fig. 6) supports the idea that it is possible to use coarser DBH classes in the search for effective *k*-NN configurations. Thus, candidates identified using coarser DBH × species groups are likely going to be close to optimal when using finer DBH × species classes. This allows for significant reductions in the computations required to calculate the dissimilarity metrics used in the search for optimal configurations, which is especially important for $grmsd_{dist}$, as this index requires inverting a square matrix of dimensions equal to the number of DBH × species classes considered in the search.

7. Conclusions

1. The multivariate nature of diameter distributions and tree-lists makes it difficult to interpret differences between alternative *k*-NN configurations and highlights the need to find suitable methods to incorporate the criteria and preferences of the forest managers in the evaluation of differences between predicted and observed tree-lists and diameter distributions.
2. For the four dissimilarity metrics considered in this study, the performance of the *k*-NN imputation methods improved when increasing the number of neighbors and when the *k*-NN distance metric was the random forest distance.
3. Rankings of *k*-NN alternatives are stable for different sets of DBH × species bins, which suggests that the selection of optimal *k*-NN configurations can be done using moderately larger DBH classes, which results in faster computations.
4. The selection of the *k*-NN methods was dependent on the choice of the dissimilarity metric used to rank alternatives.

Acknowledgements

We are grateful to two anonymous reviewers and to the Associate Editor for their positive and constructive comments; their suggestions were extremely helpful and improved the quality of this manuscript. This study was supported by the following grants: Using Airborne LiDAR Data to Predict Selected Stand Metrics and Generate Tree-Lists to Improve the Accuracy of Forest Inventory Variables (USDA Forest Service, PNW Research Station) and Predicting Thinning Eligibility with Airborne LiDAR Data. (Bureau of Land Management).

References

Andersen, H.-E., Clarkin, T., Winterberger, K., and Strunk, J. 2009. An accuracy assessment of positions obtained using survey- and recreational-grade Global Positioning System receivers across a range of forest conditions within the Tanana Valley of Interior Alaska. West. J. Appl. For. **24**(9): 128–136.

Arias-Rodil, M., Diéguez-Aranda, U., Álvarez-González, J.G., Pérez-Cruzado, C., Castedo-Dorado, F., and González-Ferreiro, E. 2018. Modeling diameter distributions in radiata pine plantations in Spain with existing countrywide LiDAR data. Ann. For. Sci. **75**(2): 36. doi:10.1007/s13595-018-0712-z.

Avery, T.E., and Burkhart, H.E. 2015. Forest measurements. 5th ed. Waveland Press, Long Grove, Ill.

Bechtold, W.A., and Patterson, P.L. (*Editors*). 2005. The enhanced forest inventory and analysis program-national sampling design and estimation procedures. USDA For. Serv. Gen. Tech. Rep. SRS-80. U.S. Department of Agriculture, Forest Service, Southern Research Station, Asheville, N.C. doi:10.2737/SRS-GTR-80.

Bollandsås, O.M., and Næsset, E. 2007. Estimating percentile-based diameter distributions in uneven-sized Norway spruce stands using airborne laser scanner data. Scand. J. For. Res. **22**(1): 33–47. doi:10.1080/02827580601138264.

Bollandsås, O.M., Maltamo, M., Gobakken, T., and Næsset, E. 2013. Comparing parametric and non-parametric modelling of diameter distributions on independent data using airborne laser scanning in a boreal conifer forest. Forestry, **86**(4): 493–501. doi:10.1093/forestry/cpt020.

Breidenbach, J., Gläser, C., and Schmidt, M. 2008. Estimation of diameter distributions by means of airborne laser scanner data. Can. J. For. Res. **38**(6): 1611–1620. doi:10.1139/x07-237.

Bureau of Land Management, OR/WA. 2015. Forest Operations Inventory vegetation: spatial data standard [online]. Available from https://www.blm.gov/sites/blm.gov/files/FOI_Vegetation_Spatial_Data_Standard_1.pdf [accessed 24 October 2018].

Chirici, G., Mura, M., McInerney, D., Py, N., Tomppo, E.O., Waser, L.T., Travaglini, D., and McRoberts, R.E. 2016. A meta-analysis and review of the literature on the *k*-nearest neighbors technique for forestry applications that use remotely sensed data. Remote Sens. Environ. **176**: 282–294. doi:10.1016/j.rse.2016.02.001.

Crookston, N.L., and Finley, A.O. 2008. yaImpute: an R package for kNN imputation. J. Stat. Softw. **23**(10): 1–16. doi:10.18637/jss.v023.i10.

Dixon, G.E. 2002. Essential FVS: a user's guide to the Forest Vegetation Simulator. USDA For. Serv. Intern. Rep. U.S. Department of Agriculture, Forest Service, Forest Management Service Center, Fort Collins, Colo.

Eskelson, B.N., Temesgen, H., Lemay, V., Barrett, T.M., Crookston, N.L., and Hudak, A.T. 2009. The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases. Scand. J. For. Res. **24**(3): 235–246. doi:10.1080/02827580902870490.

Finley, A.O., Banerjee, S., Weiskittel, A.R., Babcock, C., and Cook, B.D. 2014. Dynamic spatial regression models for space-varying forest stand tables. Environmetrics, **25**(8): 596–609. doi:10.1002/env.2322.

Gobakken, T., and Næsset, E. 2004. Estimation of diameter and basal area distributions in coniferous forest by means of airborne laser scanner data. Scand. J. For. Res. **19**(6): 529–542. doi:10.1080/02827580410019454.

Gobakken, T., and Næsset, E. 2005. Weibull and percentile models for lidar-based estimation of basal area distribution. Scand. J. For. Res. **20**(6): 490–502. doi:10.1080/02827580500373186.

Hann, D.W., and Larsen, D.R. 1991. Diameter growth equations for fourteen tree species in Southwest Oregon. Res. Bull. Series 69. Forest Research Lab, College of Forestry, Oregon State University, Corvallis, Ore.

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**(G2): G00E04. doi:10.1029/2008JG000870.

Hudak, A.T., Crookston, N.L., Evans, J.S., Hall, D.E., and Falkowski, M.J. 2008. Nearest neighbor imputation of species-level, plot-scale forest structure attributes from LiDAR data. Remote Sens. Environ. **112**(5): 2232–2245. doi:10.1016/j.rse.2007.10.009.

Kendall, M.G. 1938. A new measure of rank correlation. Biometrika, **30**(1–2): 81–93. doi:10.1093/biomet/30.1-2.81.

Lamb, S.M., MacLean, D.A., Hennigar, C.R., and Pitt, D.G. 2018. Forecasting forest inventory using imputed tree lists for LiDAR grid cells and a tree-list growth model. Forests, **9**(4): 167. doi:10.3390/f9040167.

Magurran, A.E. 2013. Measuring biological diversity. Wiley-Blackwell.

Maltamo, M., and Kangas, A. 1998. Methods based on *k*-nearest neighbor regression in the prediction of basal area diameter distribution. Can. J. For. Res. **28**(8): 1107–1115. doi:10.1139/x98-085.

Maltamo, M., Suvanto, A., and Packalén, P. 2007. Comparison of basal area and stem frequency diameter distribution modelling using airborne laser scanner data and calibration estimation. For. Ecol. Manage. **247**(1–3): 26–34. doi:10.1016/j.foreco.2007.04.031.

Martínez-Falero, J.E., Ayuga-Tellez, E., Gonzalez-Garcia, C., Grande-Ortiz, M.A., and Garrido, A.S.M. 2017. Experts' analysis of the quality and usability of SILVANET software for informing sustainable forest management. Sustainability, **9**(7): 1200. doi:10.3390/su9071200.

McGaughey, R.J. 2010. FUSION\LDV: software for LIDAR data analysis and visualization. U.S. Department of Agriculture, Forest Service, Pacific Northwest Research Station, Portland, Ore.

Moeur, M., and Stage, A.R. 1995. Most similar neighbor: an improved sampling inference procedure for natural resource planning. For. Sci. **41**(2): 337–359.

Næsset, E. 2002. Predicting forest stand characteristics with airborne scanning laser using a practical two-stage procedure and field data. Remote Sens. Environ. **80**(1): 88–99. doi:10.1016/S0034-4257(01)00290-5.

Ohmann, J.L., and Gregory, M.J. 2002. Predictive mapping of forest composition and structure with direct gradient analysis and nearest-neighbor imputation in coastal Oregon, USA. Can. J. For. Res. **32**(4): 725–741. doi:10.1139/x02-011.

Packalén, P., and Maltamo, M. 2007. The k-MSN method for the prediction of species-specific stand attributes using airborne laser scanning and aerial photographs. Remote Sens. Environ. **109**(3): 328–341. doi:10.1016/j.rse.2007.01.005.

Packalén, P., and Maltamo, M. 2008. Estimation of species-specific diameter distributions using airborne laser scanning and aerial photographs. Can. J. For. Res. **38**(7): 1750–1760. doi:10.1139/X08-037.

Packalén, P., Temesgen, H., and Maltamo, M. 2012. Variable selection strategies

for nearest neighbor imputation methods used in remote sensing based forest inventory. Can. J. Remote Sens. **38**(5): 557–569. doi:10.5589/m12-046.

Peuhkurinen, J., Maltamo, M., and Malinen, J. 2008. Estimating species-specific diameter distributions and saw log recoveries of boreal forests from airborne laser scanning data and aerial photographs: a distribution-based approach. Silva Fenn. **42**(4): 625–641. doi:10.14214/sf.237.

Peuhkurinen, J., Mehtätalo, L., and Maltamo, M. 2011. Comparing individual tree detection and the area-based statistical approach for the retrieval of forest stand characteristics using airborne laser scanning in Scots pine stands. Can. J. For. Res. **41**(3): 583–598. doi:10.1139/X10-223.

Peuhkurinen, J., Tokola, T., Plevak, K., Sirparanta, S., Kedrov, A., and Pyankov, S. 2018. Predicting tree diameter distributions from airborne laser scanning, SPOT 5 satellite, and field sample data in the Perm Region, Russia. Forests, **9**(10): 639. doi:10.3390/f9100639.

R Core Team. 2018. R: a language and environment for statistical computing [online]. R Foundation for Statistical Computing, Vienna, Austria. Available from https://www.R-project.org/.

Rao, J.N.K., and Molina, I. 2015. Introduction. *In* Small area estimation. John Wiley & Sons, Hoboken, N.J. pp. 1–8. doi:10.1002/9781118735855.ch1.

Räty, J., Packalen, P., and Maltamo, M. 2018. Comparing nearest neighbor configurations in the prediction of species-specific diameter distributions. Annals of Forest Science, **75**(1): 26. doi:10.1007/s13595-018-0711-0.

Reynolds, M.R., Burk, T.E., and Huang, W.-C. 1988. Goodness-of-fit tests and model selection procedures for diameter distribution models. For. Sci. **34**(2): 373–399.

Shin, J., Temesgen, H., Strunk, J.L., and Hilker, T. 2016. Comparing modeling methods for predicting forest attributes using LiDAR metrics and ground measurements. Can. J. Remote Sens. **42**(6): 739–765. doi:10.1080/07038992.2016.1252908.

Strunk, J.L., Gould, P.J., Packalen, P., Poudel, K.P., Andersen, H.-E., and Temesgen, H. 2017. An examination of diameter density prediction with *k*-NN and airborne lidar. Forests, **8**(11): 444. doi:10.3390/f8110444.

Temesgen, B.H., LeMay, V.M., Froese, K.L., and Marshall, P.L. 2003. Imputing tree-lists from aerial attributes for complex stands of south-eastern British Columbia. For. Ecol. Manage. **177**(1–3): 277–285. doi:10.1016/S0378-1127(02)00321-3.

Thomas, V., Oliver, R.D., Lim, K., and Woods, M. 2008. LiDAR and Weibull modeling of diameter and basal area. For. Chron. **84**(6): 866–875. doi:10.5558/tfc84866-6.

Tomppo, E., and Halme, M. 2004. Using coarse scale forest variables as ancillary information and weighting of variables in *k*-NN estimation: a genetic algorithm approach. Remote Sens. Environ. **92**(1): 1–20. doi:10.1016/j.rse.2004.04.003.

Tomppo, E., and Katila, M. 1991. Satellite image-based National Forest Inventory of Finland for publication in the IGARSS'91 Digest. *In* Proceedings of the 1991 International Geoscience and Remote Sensing Symposium — IGARSS'91 Remote Sensing: Global Monitoring for Earth Management, Espoo, Finland, 3–6 June 1991. *Edited by* J. Putkonen. IEEE. pp. 1141–1144. doi:10.1109/IGARSS.1991.579272.

Valbuena, R., Mauro, F., Rodriguez-Solano, R., and Manzanera, J.A. 2010. Accuracy and precision of GPS receivers under forest canopies in a mountainous environment. Span. J. Agric. Res. **8**(4): 1047–1057. doi:10.5424/sjar/2010084-1242.

Wykoff, W.R., Crookston, N.L., and Stage, A.R. 1982. User's guide to the Stand Prognosis Model. USDA For. Serv. Gen. Tech. Rep. INT-133. U.S. Department of Agriculture, Forest Service, Intermountain Forest and Range Experiment Station, Ogden, Utah. doi:10.2737/INT-GTR-133.

Zhu, J., Huang, Z., Sun, H., and Wang, G. 2017. Mapping forest ecosystem biomass density for Xiangjiang River Basin by combining plot and remote sensing data and comparing spatial extrapolation methods. Remote Sens. **9**(3): 241. doi:10.3390/rs9030241.