Canadian Journal of Forest Research
Revue canadienne de recherche forestière

Multivariate estimation for accurate and logically-consistent forest-attributes maps at macroscales

Journal:	*Canadian Journal of Forest Research*
Manuscript ID	cjfr-2017-0221.R3
Manuscript Type:	Article
Date Submitted by the Author:	08-Dec-2017

Complete List of Authors:	Lochhead, Kyle; University of British Columbia , Forest Resources Management LeMay, Valerie; University of British Columbia Bull, Gary; University of British Columbia Schwab, Olaf; Natural Resources Canada Halperin, James; University of British Columbia, Forest Resources Management

Keyword:	multivariate imputation, system of models, kriging with external drift, national forest inventory, multi-source inventory
Is the invited manuscript for consideration in a Special Issue? :	N/A

SCHOLARONE™
Manuscripts

Multivariate estimation for accurate and logically-consistent forest-attributes maps at macroscales

Kyle Lochhead, Valerie LeMay, Gary Bull, Olaf Schwab, James Halperin

Kyle Lochhead[a*]

Valerie LeMay[a] Email: valerie.lemay@ubc.ca

Gary Bull[a] Email: gary.bull@ubc.ca

James Halperin[a] Email: j.halperin@alumni.ubc.ca

[a] Department of Forest Resources Management, Faculty of Forestry, University of British Columbia, 2424 Main Mall. Vancouver, BC, V6T 1Z4, Canada

Olaf Swab. Natural Resources Canada, Canadian Forest Service, Ottawa, ON, K1A 0E4, Canada Email: Olaf.Schwab@NRCan-RNCan.gc.ca

[*]Corresponding author: K. Lochhead Email: kyle.lochhead@live.forestry.ubc.ca Tel.: +1 604 822 5689; fax: +1 604 822 9106

Draft

Abstract

Spatially-explicit wall-to-wall forest-attributes information is critically important for designing management strategies resilient to climate-induced uncertainties. Multivariate estimation methods that link forest attributes and auxiliary variables at full-information locations can be used to estimate the forest attributes for locations with auxiliary-variables information only. However, trade-offs between estimation accuracies versus logical consistency among estimated attributes may occur. This is particularly likely for macroscales (i.e., $\geq$ 1 Mha) with large forest-attributes variances and wide spacing between full-information locations. We examined these trade-offs for ~390 Mha of Canada's boreal zone using variable-space nearest neighbours imputation versus two modelling methods (i.e., a system of simultaneous nonlinear models and kriging with external drift). We found logical consistency among estimated forest attributes (i.e., crown closure, average height and age, volume per ha, species percentages) using: 1) $k \leq 2$ nearest neighbours; or 2) careful model selection for the modelling methods. Of these logically-consistent methods, kriging with external drift was the most accurate, but implementing this for a macroscale is computationally more difficult. This extra cost is justified given the importance of assessing strategies under expected climate changes in Canada's boreal forest and in other forest regions.

Keywords: multivariate imputation, system of models, kriging with external drift, national forest inventory, multi-source inventory

Draft

Introduction

Designing resilient landscape strategies for changing environmental conditions has increased the need for forest-attributes information across very large national landscapes or macroscales (Boisvenue et al. 2016a). In the case of the ~552 Mha Canadian boreal zone (Brandt 2013), uncertainties surrounding future climates have raised concerns over possible increases in the frequency and impacts of natural disturbances (Flannigan et al. 2005; Weed et al. 2013). Also, forest management goals increasingly include a broader range of ecosystem services, including a wider variety of forest products, sustaining and providing wildlife habitats, and maintaining water and soil integrity. These changes require policy makers to evaluate the cumulative effects of macroscale economic and ecological changes (Lindner et al. 2002). More comprehensive and complex decision support tools are needed to guide changing forest management and policy; wall-to-wall, spatially-explicit forest-attributes information is needed to support these tools (Bernier et al. 2016; Boisvenue et al. 2016b).

Draft

Multivariate estimation methods can predict forest-attributes across a landscape by using relationships between forest attributes and auxiliary variables at full-information locations to estimate forest attributes at all other locations with only auxiliary variables. However, for scales ≥ 1 Mha (i.e., macroscales), budgetary constraints limit the number of spatial locations with full-information to only a small proportion of the land area. Also, the diversity of ecosystems across this broad spatial scale is often much greater than for smaller spatial scales. As noted by Moeur and Stage (1995), confidence in this macroscale wall-to-wall forest-attributes information is crucial for developing a plausible decision space to assess and design management strategies.

To provide forest-attributes information needed for management, many countries have undertaken a national forest inventory (NFI) that includes ground sampling coupled with remotely sensed imagery

(Vidal et al. 2016). Commonly, a systematic sample of ground plots is repeatedly measured over time, providing a continuous assessment using consistent definitions of many forest attributes (Tomppo 2010). For macroscale NFIs, including Canada, ground plots may be partially or entirely replaced by interpreted large-scale photo-plots as a cost-effective option (Magnussen and Russo 2012). Using standardized protocols and viewing stereo-pairs of photos as 3-D images, professional photo-interpreters can measure the crown closure, the species composition based crown closure of each species, and the average height, but other variables are interpreted based on knowledge of the area, information from ground plots, relationships among variables, and other information (Avery and Burkhart 2002; Kershaw et al. 2017). Remotely-sensed (e.g., Landsat) and other available wall-to-wall map information are then spatially and temporally matched with the NFI plots in a multi-sourced forest inventory (Tomppo et al. 2008a; Nilsson et al. 2016). Overall, this multi-sourced information can be used to obtain wall-to-wall estimates of forest attributes at one point in time; these estimates can also be used as inputs into growth and yield models for forecasting different management scenarios (Bettinger et al. 2005; Boisvenue et al. 2016b).

Draft

Alternative methods have been proposed for obtaining wall-to-wall estimates of forest-attributes using multi-source information. Methods can be univariate, where each forest attribute is separately estimated, or multivariate where a vector or matrix of forest attributes is simultaneously estimated (see overviews by Eskelson et al. 2009 and by Chirici et al. 2016). Further, estimation methods can be model-free, where no model or probability distribution is assumed (i.e., nearest-neighbour imputation methods in real- or in variable-space), or model-based, where a model with an assumed probability distribution (parametric model) or without an assumed probability distribution (nonparametric model) is explicitly described and used in the estimation process (Fehrmann et al. 2008).

In terms of model-free methods, Tomppo (1988) used nearest neighbours imputation methods (i.e., a donor method, termed *k*-NN by Tomppo) based on proximity in variable-space to estimate each forest attribute (i.e., univariate). Since then, many papers have used variations of univariate *k*NN (see Chirici et al. 2016). Alternatively, Moeur and Stage (1995) used a multivariate imputation method they termed most similar neighbour (MSN) to estimate a vector of forest attributes simultaneously based on *k*=1 neighbour. As with variations using *k*NN, many papers have used variations on MSN, termed variable-space nearest neighbour methods (VSNN) in an overview paper by Eskelson et al. (2009). An extension to doubly-multivariate estimation was demonstrated by Temesgen et al. (2003) who investigated the use of the multivariate VSNN for estimating a matrix of species, sizes and stems per ha (i.e., a tree-list) needed to project each forested stand within a forest inventory. In terms of model-based methods, univariate estimation of each forest attribute has a very long history, including a wide range of linear and nonlinear, parametric and non-parametric methods. Multivariate estimation using model-based methods is relatively more recent than univariate model-based methods, but includes using systems of models (e.g., LeMay 1990; Babcock et al. 2013).

Draft

Regardless of the method used, estimates of forest attributes must be accurate and logically consistent to obtain the confidence of forest managers (Moeur and Stage 1995; Ohmann and Gregory 2002). Accuracy indicates the closeness of an estimated attribute value to the real value, often measured by summaries of differences between actual and estimated values for full-information spatial locations (Foody 2002). Logical consistency refers to the preservation of attribute definitions and logical relationships (Morrison 1995), as measured by the degree of adherence to logical rules that test for nonsensical values for each estimated attribute and for impossible combinations among estimated attributes (Kainz 1995).

Using univariate *kNN*, optimal accuracy for an estimated forest attribute can be achieved via choosing an optimal combination of the auxiliary variables, the weights associated with each auxiliary variable, the distance metric, and the number of neighbours (McRoberts 2009). Logical consistency for each estimated forest attribute is assured using *k*NN, since *k* neighbours are selected from full-information locations and the measured values for the forest attribute are averaged to obtain the estimate for each location with auxiliary variables only. Using univariate model-based methods, careful selection of the model can also ensure logical consistency for each estimated forest attribute. However, logical inconsistencies among attributes may occur using model-free or model-based univariate methods since each forest attribute is separately estimated. Using VSNN with k=1 neighbour selected from full-information locations, logical consistency for each estimated forest attribute as well as across the vector (or matrix) of attributes is obtained (Moeur and Stage 1995; Mauro et al. 2015). This may not be the case using VSNN with k>1 neighbour, since the vector of averages calculated using *k* donor locations may not be a logically consistent combination of forest attributes (e.g., species compositions than do not occur in nature). Also, estimation accuracy for each forest attribute may be smaller using VSNN with k≥1 than univariate *k*NN, since optimal selection of: 1) auxiliary variables, weights for each auxiliary variable, the distance metric and the number of neighbours may not be possible given the dimensionality of the multivariate problem (Indyk and Mowati 1998); and 2) accuracy compromises must be made among the vector (or matrix) of estimated forest attributes. Using a multivariate model-based method may provide greater accuracy than VSNN by: 1) developing a simultaneous system of recursive models that allows forest attributes estimated using a model earlier in the system to be used in estimating forest attributes later in the system (Pindyck and Rubinfeld 1981; LeMay 1990); and 2) carefully selecting the auxiliary variable(s) and the model form for each model of the system. While both optimal accuracy and

Draft

logical consistency are desirable, providing both may be cost-prohibitive for very large spatial scales or macroscales (e.g., Tomppo and Czaplewski 2002; McRoberts 2008; Tomppo et al. 2008b).

In this research paper, we addressed the following main question: Which multivariate estimation method provides the greatest accuracy for a macroscale problem, while maintaining logical consistency among forest attributes? To investigate this, we compared three multivariate estimation methods using a ~390 Mha sub-area of Canada's boreal forest. For this area, high-resolution multivariate maps of forest attributes needed for macroscale strategic analysis are currently lacking or are outdated (Beaudoin et al 2014). Specifically, we compared two model-based approaches, a system of simultaneous nonlinear models (SNLM) and kriging with external drift (KED) with the model-free VSNN method to estimate: crown closure percent (CC), average height of dominant trees (Ht), average age of dominant trees (Age), volume per ha for all trees (Vol), and tree species percentages. These attributes describe the current forest and are often the input variables used in stand-level growth models (Bokalo et al. 2010) that underlie many decision-support tools. Given prior research results for smaller spatial scales, macroscale mapping issues raised by Beaudoin et al. (2014), and basic principles underlying these three methods, we hypothesized that: 1) VSNN would be more accurate, since it is model-free; 2) using VSNN with k>1 would increase accuracy, but may adversely affect logical consistency; 3) carefully designing an SNLM would ensure logical consistency of forest attributes, while obtaining accurate estimates of each attribute; and 4) greater accuracy could be achieved by allowing the parameters of the SNLM to vary spatially (KED method). Based on our results, we selected one method and produced multivariate maps (90 m) required for macroscale strategic analysis of Canada's boreal forest management areas within which forest companies operate (to view these maps see doi: 10.14288/1.0354319).

Draft

Materials and methods

Study area

The boreal zone of Canada (hereafter, referred to as "boreal") has a total area of ~ 552 M ha, including ~270 Mha of forest (Brandt et al. 2013). Large areas of pure or mixed coniferous tree species occur, including white spruce (*Picea glauca* (Moench) Voss), black spruce (*Picea mariana* (Mill.) BSP), tamarack (*Larix laricina* (Du Roi) K. Koch), balsam fir (*Abies balsamea* (L.) Mill.), jack pine (*Pinus banksiana* Lamb.), and lodgepole pine (*Pinus contorta* Dougl. var *latifolia* Engelm.). Deciduous species, particularly aspen (*Populus tremuloides* Michx.), balsam poplar (*Populus balsamifera L.*), and paper birch (*Betula papyrifera* Marsh.), occur in either pure stands or in mixtures with conifers (Brandt 2013). The boreal is bounded in the north by tundra within the arctic zone, in the south by grasslands or temperate forests, in the west by the Rocky Mountains, and in the east by the maritime forests near the Atlantic Ocean. For this study, we confined our study area to south of 60° N, since tree density becomes sparse as the forest transitions to tundra north of this limit (Fig. 1). Further, phenological differences between satellite images are more pronounced at these higher latitudes complicating image acquisition and processing (Banskota et al. 2014). Using this northern boundary and excluding major lakes, ~390 Mha remained in the study area.

Draft

Imagery and other auxiliary data

Multivariate estimation methods rely on a suite of X- (aka, predictor or auxiliary) variables assembled from multiple data sources. For this study, 38 possible X-variables were derived from surface reflectance imagery, climate, topographic and other data assembled for the study area (Table 1).

Surface reflectance imagery for the boreal forest was retrieved from the Landsat Climate Data Record (USGS Earth Explorer 2013), a Landsat-5 (for scenes selected before 2000 and after 2003) or 7 (between 2000 and 2003) level 2-A product generated by the Landsat Ecosystem Disturbance

Adaptive Processing System (Masek et al. 2006). These images provided wall-to-wall, orthorectified, maximally cloud-free coverage at a 30 m resolution. Included with this product were masks for clouds, cloud shadows, water and ice (Zhu and Woodcock 2012). A total of 1 004 images between 1987 and 2010 were acquired to temporally match the varying acquisition years of the NFI photo-plot data (described later). Scene selection was set to the peak growing season (mid-June to August) to reduce phenological differences while recognizing that small differences would be unavoidable over a national geographic extent (Tipton et al. 2010). Images were then masked to remove clouds, shadows and waterbodies. The resulting processed surface reflectance images provided the reflectance measures and vegetation indices described in Table 1.

Climate, topographic, and other variables were also considered as possible X-variables (Table 1). Topographic variables were computed using hydrology tools in ArcGIS v 10.2, including elevation (Elv), slope (Slp), and aspect (Asp) along with 11 interaction terms recommended by Stage and Salas (2007), and CTI (compound topographic index; a variable describing topographic position). The final X-variable was a raster layer of the presence or absence of saturated soils, based on a land cover layer of wetlands and poorly drained soils extracted from the Natural Resource Canada CanVec+ dataset (Geogratis 2013). All layers were resampled to the 30 m pixels to match the surface reflectance imagery using cubic convolution.

Canada's National Forest Inventory photo-plots

The aerial photo-plots of Canada's NFI (see https://nfi.nfis.org/) provided the common forest attributes information (i.e., the Y-variables) used in this study. Although ground plots are available, they were measured on only a subset (1 in 10) of photo-plot locations (Gillis et al. 2005). Using 20 km by 20 km grid spacing across the boreal, a stereo-photo pair (color, 1:10, 000 or 1:20 000) was acquired at each grid intersection. Professional photo-interpreters then used 3D viewing to stratify

Draft

the 2 km by 2 km photo-plot into many irregularly-shaped polygons according to the harmonized definitions of Canada's NFI Land-Cover Classification System (Gillis et al. 2005). They classified each polygon as vegetated or non-vegetated (i.e., waterbodies, snow, rock, etc.) land-cover classes. Non-vegetated polygons were not further considered in this study. Since these areas have been mapped across Canada in the CanVec+ dataset, they can be masked out of estimated forest attributes maps. Within the study area, 3 298 photo-plots were classified as vegetated and had cloud-free Landsat-TM/ETM imagery matching the photo-plot acquisition time (Fig. 1). Vegetated polygons had been further classified by crown closure percent as treed (≥ 10 %) or non-treed (<10%) based on the FAO (2015) definition, and a series of forest attributes were photo-interpreted for each treed polygon. A subset of these forest attributes was used as the Y-variables in this study (Table 2). To reduce the number of Y-variables, species percentages were aggregated into species groups corresponding with those commonly used in stand-level growth and yield models.

Draft

Spatial matching of multi-source data

All layers representing the X- and Y-variables were spatially and temporally registered (i.e., matched). A 90 m by 90 m pixel window was extracted from the centroid of each irregularly-shaped polygon (Fig. 1). Extracting one pixel window avoided within-polygon dependencies and a larger pixel size mitigated spatial registration issues. Using the centroid avoided polygon edge effects; any 90 m by 90 m pixel window not entirely contained within a polygon boundary was excluded. Then, values for each X- and Y-variable were extracted for each pixel. A total of 78 453 full-information locations with both X- and Y- variables was obtained.

Data splitting

The full-information locations data were split into a reference (aka, donor for VSNN or model-fitting for SNLM and KED) versus a target (aka, test or validation) dataset as used in other studies (e.g., LeMay and Temesgen 2005; Hall et al. 2006). Although Snee (1977) recommended using *n*-way validation, Roecker (1991) found marginal improvement over random splitting in the variable-selection setting. Data were split at the photo-plot level to better mimic the multivariate estimation applied to the entire land area (i.e., in application, the reference dataset would contain all full-information locations in a photo-plot, but the target dataset would include only spatial locations outside of photo-plots). The resulting validation dataset had ~20% of the full-information locations (n_{targ} = 15 025) and the reference dataset contained the remaining ~80% (n_{ref} = 63 428).

Multivariate estimation methods

System of simultaneous nonlinear models

For the first method, a system of simultaneous nonlinear models (SNLM; aka simultaneous system of nonlinear equations; see Judge et al. 1985) representing the relationships between the X- and Y-variables was fitted using the reference dataset (i.e., full-information locations) and this was applied to the target dataset (i.e., locations with auxiliary variables only) following Ver Hoef and Temesgen (2013). Using nonlinear model forms can better reflect known biological relationships (Littell et al. 2006) and careful choices of these forms can ensure logical consistency for each Y-variable of the system. Then, using a system of models allows for different X-variables in each model of the system. Variables that do not impact the estimated conditional mean of a particular Y-variable (i.e., the estimated Y-variable given the particular set of X-variable values) can be dropped from the model. This allows for more accurate, parsimonious models relative to using a fixed set of X-variables for all Y-variables. Further, a system of models allows for across-model constraints to ensure logical consistency across Y-variables (Babcock et al. 2013). Finally, allowing for a Y-variable of one model to appear as a predictor variable in another model of a simultaneous equations system can improve both accuracy and logical consistency (LeMay 1990; Gujarati et al. 2011). We specifically used a

Draft

recursive system of simultaneous nonlinear models, thus enabling an estimated Y-variable to be used as a predictor variable for models later in the ordered system of models (i.e., an instrumental variables (IV) method; see Pindyck and Rubinfeld 1981; Judge et al. 1985; Gujarati et al. 2011)

The SNLM was carefully developed as a logically ordered, recursive system of models to reflect logical, biological relationships for each Y-variable and across Y-variables. Specifically, the SNLM preserved the [0,100] limits of CC and species percentages, the additivity of species percentages (must sum to 100%), and accounted for the interdependencies of CC, Age, Ht, and Vol. The first model was the CC model using a logistic model form that constrains estimates within [0,100]. The estimated CC value ($\widehat{CC}$) was then used to indicate if a location can be considered as treed ($\widehat{CC} \geq 10$ %) or non-treed, since all non-treed locations have logical zero values for estimated species percentages, Age, Ht, and Vol. $\widehat{CC}$ was also available as a possible predictor variable (i.e., using an IV approach) for later models. The species percentages model was next, again using a logistic model form to constrain all estimated percentages to [0,100]. Age, Ht, and finally Vol models followed using nonlinear models to ensure all estimated values were >0, and allowing Y-variables earlier in the system to be possible predictor variables. Details for each model of the system are presented next.

As noted earlier, the CC model was the first model of the SNLM. For this, we used a logistic model:

$$[1] \qquad \frac{Y_i}{100} = \pi_i = \frac{\exp(\eta_i)}{1+\exp(\eta_i)} \qquad \text{with} \quad \eta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \boldsymbol{\beta}'\boldsymbol{x_i} + \varepsilon_i$$

where Y_i is the CC for the i^{th} observation expressed as a percent; π_i is CC expressed as a proportion; η_i is the log odds ratio (i.e., logit); β_0 is the constant parameter (i.e., intercept of the logit model); $\boldsymbol{\beta} = (\beta_1, \dots \beta_l)'$ is a vector of l parameters associated with the $\boldsymbol{x_i}$ vector of predictor variables; and ε_i is the error term. This model was fit using all full-information locations of the reference dataset and the SAS (v9.4) LOGISTIC procedure. The selection of X-variables was performed using the branch-and-bound algorithm of Furnival and Wilson (1974) to find the model with the smallest AIC, but giving preference to variables that exploit known biological relationships with the Y-variables. The preference variables were AlbY and SS to ensure that crown closure changes with latitude (i.e., should decrease with increasing latitude; Sirois 1992) and changes for saturated versus not saturated soils (i.e., should be lower for saturated soils; Glebov and Korzukhin 1992).

For the vector of species percentages model of the SNLM, a generalized version of Eq. 1 to a multinomial logistic model was used, where species percentages were considered proportions (e.g., Thompson 1987).

[2] $$\frac{Y_{ij}}{100} = \pi_{ij} = \frac{\exp(\eta_{ij})}{\sum_{j=1}^{J}\exp(\eta_{ij})} \text{ with } \eta_{ij} = \log\left(\frac{\pi_{ij}}{\pi_{iJ}}\right) = \beta_{0j} + \boldsymbol{\beta_j}'\boldsymbol{x_i} + \varepsilon_{ij}$$

where Y_{ij} is the percentage of the j^{th} species group (j=1…5) for the i^{th} observation and $\sum_{j=1}^{J=5} Y_{ij} = 100$; π_{ij} is species percentage expressed as a proportion; η_{ij} is the log odds ratio for each species group relative to the baseline species group (J=5), meaning $\eta_{i5} = \log(1) = 0$; and ε_{ij} is the error term. This model ensured that all estimated species percentages were in the [0,100] interval and the sum of all the species groups were equal to 100 for each observation. This model was fitted using the subset of the reference dataset considered treed based on $\widehat{CC}$ using Eq. [1], and using the SAS (v9.4) LOGISTIC procedure with the Newton-Raphson maximum likelihood algorithm. X-variables were selected following the same method as for the CC model. Specifically, SS, MAP, and Slp were the preference variables, since SS would be expected to relate to the presence of black spruce (Sb) which is commonly associated with wetland areas (i.e., saturated soils, Brandt 2009), and MAP and Slp are important abiotic drivers of species composition (Soja et al. 2007).

The remaining models for Age, Ht and Vol were fit as a system of simultaneous nonlinear models using the subset of locations in the reference dataset considered treed based on $\widehat{CC}$. For Age and Ht,

https://mc06.manuscriptcentral.com/cjfr-pubs

we chose an asymptotic nonlinear model to limit the maximum values to logical biological limits, while ensuring non-negative values.

$$[3] \qquad Y_i = \frac{\alpha}{1+\exp(\theta_i)} + \varepsilon_i \quad \text{with} \quad \theta_i = \beta_0 + \boldsymbol{\beta}'\boldsymbol{x_i} + \boldsymbol{\delta}'\boldsymbol{\hat{y}_i}$$

where Y_i is the Ht or Age for the i^{th} observation in the reference dataset; α is the maximum possible estimated value (i.e., asymptote); $\boldsymbol{\delta}'$is a vector of parameters associated with the $\boldsymbol{\hat{y}_i}$, estimated Y-variables from models earlier in the recursive system; and ε_i is the error.

A Chapman-Richards (C-R) model (Richards 1959) was selected for Vol, because this model form has been widely applied in forestry due to its flexibility, accuracy and biologically meaningful properties (e.g., Zhao-gang and Feng-ri 2003). However, we used $\widehat{\boldsymbol{Ht}}$ instead of $\widehat{\boldsymbol{Age}}$, since it more directly relates to Vol (e.g., Garcia 2003).

Draft

$$[4] \qquad Y_i = \alpha_i\left(1 - Be^{k\,\widehat{\boldsymbol{Ht}}_i}\right)^{\frac{1}{1-m}} + \varepsilon_i \quad \text{with} \quad \alpha_i = (A_0 + \boldsymbol{A}'\boldsymbol{x_i} + \boldsymbol{\delta}'\boldsymbol{\hat{y}_i})$$

where Y_i is the Vol for the i^{th} observation in the reference dataset; α_i is the asymptote or maximum Vol; B is a shape parameter; k is a parameter associated with $\widehat{\boldsymbol{Ht}}_{\boldsymbol{i}}$; m is also a shape parameter; $\boldsymbol{A} = (A_1, \dots A_l)'$ is a vector of l parameters associated with the $\boldsymbol{x_i}$ vector of predictor variables; $\boldsymbol{\delta}'$is a vector of parameters associated with the $\boldsymbol{\hat{y}_i}$ estimated Y-variables from models earlier in the recursive system (e.g., CC, species percentage, and Age); and ε_i is the error. The asymptote parameter was allowed to vary with X-variables and previously estimated Y-variables, since this represents the maximum potential volume at a location and varies with site factors and genetics (Stage and Salas 2007).

Although the models for Age, Ht, and Vol were fitted as a system, the selection of X-variables was first performed for each model of the system separately. Linearized versions of Age and Ht were obtained by setting the asymptotes (the α parameters) to the maximal values in the reference dataset

(Table 2). A maximum R^2 improvement algorithm (e.g., MAXR in REG procedure) was then used with preference for variables that exploit known biological relationships with the Y-variables. Given that B_4, B_5 and B_7 spectral wavelengths are associated with forest disturbance (Key and Benson 2006) and shadowing effects indicative of older stand structures (Kuusinen et al. 2014), these X-variables were used as preference variables for Age. For the Ht model, B_4 was given preference given the sensitivity of vegetation structure to this spectral wavelength (Hall et al. 2006). For the Vol model, a linearized version of the C-R model was obtained by fixing the B and m parameters to 1 and 0, respectively. Following the selection of X-variables, the system was fit using FIML implemented using PROC MODEL of SAS (v9.4). No error variance models were added since there was no evidence of heteroscedastic error variances in diagnostic graphs.

Kriging with external drift

For kriging with external drift (KED), the fitted SNLM was localized by estimating random effects using the reference data (i.e., representing full-information locations) and then spatially interpolating these random effects for target locations (i.e., simulating locations with auxiliary variables only; Schabenberger and Gotway 2005). For KED using a linear model, often the error term is allowed to spatially vary (i.e., residual kriging), which is equivalent to adding in a spatially-varying intercept (Littell et al. 2006; Lloyd 2007). However, each model of the SNLM system has a zero y-intercept (i.e., no y-intercept). Allowing a spatially-varying error term could result in estimated percentages outside of the [0,100] interval for Eq. 1 and 2, and in negative estimates for Y-variables of Eq. 3 and 4, thereby affecting logical consistency. Instead, we modified one or more parameters of the model of the SNLM to be random parameters (Schabenberger and Gotway 2005; Littell et al. 2006), as used by Merz and Bloschl (2003).

After preliminary investigations, we included a spatially-varying β_0 in the CC model (Eq. 1):

Draft

[5] $$\frac{Y_i}{100} = \pi_i = \frac{\exp(\eta_i)}{1+\exp(\eta_i)} \quad \text{with} \quad \eta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + z_i + \boldsymbol{\beta}'\boldsymbol{x_i} + \varepsilon_i$$

where $z_i \sim N(0, \Sigma)$ is the spatially-varying parameter estimated for each photo-plot; and all other parameters and variables were previously defined for Eq. 1. This was repeated for the species percentages model (Eq. 2). For Ht and Age (Eq. 3) and for Vol (Eq. 4), we introduced random effects as follows:

[6] $$Y_i = \frac{\alpha}{1+\exp(\theta_i + z_i)} + \varepsilon_i$$

[7] $$Y_i = \alpha_i\left(1 - Be^{k\hat{y}_i}\right)^{\frac{1}{1-m}} + \varepsilon_i \quad \text{with} \quad \alpha_i = (A_0 + \boldsymbol{A}'\boldsymbol{x_i} + \boldsymbol{\delta}'\boldsymbol{\hat{y}_i} + z_i)$$

To estimate z_i, we used the SAS (v9.4) NLMIXED procedure for each model separately, where all other parameters were retained from the previous fit of the SNLM. To apply the spatially-explicit models to the target dataset, the simple kriging (SK) predictor was used to spatially predict z_S for each spatial location (s) based on the a spatial neighbourhood (Schabenberger and Gotway 2005):

Draft

[8] $$\hat{z}_s = \sum_{i=1}^{n_{\text{ref}}} \lambda_s z_i$$

The λ_s were estimated from a model of the semi-variogram of z_i. Several semi-variogram models were fit using ArcGIS v10.2; these were visually compared to the empirical semi-variogram and one model was selected. This was repeated for each of eight cardinal directions (i.e., N, NE, E, SE, S, SW, W, and NW) to check if the assumed stationarity was met with regards to direction (Schabenberger and Gotway 2002).

Variable space nearest neighbour estimation

Unlike the other two methods, VSNN is a model-free method (see Eskelson et al. 2009). The X-variables are used to determine the variable-space distances between reference locations and a target location; the closest neighbours (i.e., k≥1) among the reference locations are used as donors of the Y-variable information for the target location. As k increases, the variability of the estimated Y-variables across the target dataset (i.e., the spatial extent if used in mapping) decreases as the estimated Y-variables for each spatial location approach the vector of means using the entire reference dataset. Two steps were again used, following the use of VSNN by others (e.g., Moisen and Frescino 2002; Halperin et al. 2016). First, univariate *k*NN was used to estimate CC for locations in the target dataset. Then, $\widehat{CC}$ (estimated) was used to divide the target dataset into treed versus non-treed locations as with SNLM and KED. For non-treed locations, all other Y-variables were estimated as zero for the target dataset. For treed locations, VSNN was used to estimate the remaining Y-variables for each location in the target dataset using only treed locations of the reference dataset. The X-variables used for the SNLM and KED methods were also used for the VSNN method. All X-variables were standardized (i.e., subtracted the mean and divided by standard deviation) to remove the effects of different measurement scales as in other studies (e.g., LeMay and Temesgen 2005). Although other distance metrics could be used to select neighbours in multivariate variable-space (see Eskelson et al. 2009), we used the distance metric proposed by Moeur and Stage (1995) based on canonical correlation analysis (CCA) between Y- and X-variables, as used by Beaudoin et al. (2014). Unlike Moeur and Stage, we varied *k* from 1 to 15 and used weighted averages (i.e., inverse-distance in variable-space) of Y-variables from selected neighbours. The R package YaImpute (Crookston and Finley 2008) was used to implement the VSNN methods.

Draft

Comparisons

The accuracy and logical consistency of the three multivariate estimation methods were compared using the target dataset. For accuracy, reality was defined by the actual Y-variables of the target dataset. For logical consistency, reality was defined using other non-data driven information to create a rule-based set of criteria, as recommended by Kainz (1995). For CC, Ht, Age, and Vol, we tested the null hypothesis $\boldsymbol{H_0}: \boldsymbol{\mu_Y} = \boldsymbol{0}$ (i.e., vector of mean differences between actual and estimated forest attributes is a zero vector) against the alternative hypothesis that at least one mean difference is not equal to zero. For this, we used Hotelling's paired T^2 statistic (Hotelling 1951), a multivariate generalization of Student's paired t-statistic.

[9] $$paired\ T^2 = n_{\text{targ}}\bar{\boldsymbol{y}}'{\boldsymbol{S_Y}}^{-1}\bar{\boldsymbol{y}}$$

where n_{targ} is the number of full-information locations in the target dataset; $\bar{\boldsymbol{y}}$ is the mean vector of differences between actual and estimated values for the Y-variables; and $\boldsymbol{S_Y}$ is the estimated variance-covariance matrix of these differences. Other accuracy metrics separately calculated for CC, Ht, Age and Vol using the actual (Y_s) versus estimated values $(\widehat{Y_s})$ for the target dataset were:

1. Root Mean Squared Prediction Error (RMSPE) defined as:

[10] $$\text{RMSPE} = \sqrt{\sum_{s=1}^{n_{\text{targ}}} \frac{(Y_S - \widehat{Y_S})^2}{n_{\text{targ}}}}$$

2. Percent RMSPE defined as:

[11] $$\%\ \text{RMSPE} = 100\ (\frac{\text{RMSPE}}{\bar{Y}})$$

where $\bar{Y}$ is the mean of actual values for forest attribute Y in the target dataset.

3. Mean difference (MD) between actual and estimated Y-variable values, defined as:

[12] $$\text{MD} = \frac{1}{n_{\text{targ}}} \sum_{s=1}^{n_{\text{targ}}} \left(Y_s - \widehat{Y_s}\right).$$

4. Pearson's correlation coefficient between Y_s and $\widehat{Y_s}$.

To indicate accuracy for extremes of each Y-variable, RMSPE and MD were also calculated using data representing the 0 to 10th and then the 90 to 100th percentiles of the range of actual values for each Y-variable in the target dataset. Accuracy of species percentages was assessed by a confusion matrix of broad species classes.

The rules applied to assess adherence to logical consistency were: 1) estimated percent crown closure and all species percentages must be within the [0,100] interval; 2) estimated species percentages must

Draft

sum to 100; 3) estimated Ht, Age, and Vol values must be non-negative; 4) across-variables ratios must be within bounds of biological reality; and 5) species percentages must be possible based on ecological information. Rules 1, 2 and 3 were met given the steps described in the methods section. Adherence to Rule 4 was not assured using any of the three methods. Although a variety of relationships across Y-variables could be evaluated for Rule 4, we used the ratio of $\widehat{Vol}$ to $\widehat{Age} = \widehat{MAI}$ to look for across-variable inconsistencies, since this growth measure is often used in forest management planning. The cumulative distributions of the actual and estimated *MAI* values in the target dataset were compared.

To evaluate Rule 5, we used ternary diagrams of actual and estimated species percentages to visually check for illogical combinations. Ternary diagrams map the frequency of percent variables in a two-dimensional space on an equilateral triangle (van den Boogaart and Tolosana-Delgado 2013). Points closer to a vertex of the equilateral triangle represent a larger percentage of the species attributed to that vertex. These species percentages ternary diagrams were obtained for two ecological communities (photo-interpreted land areas), namely: 1) lowlands or areas saturated with water long enough to promote hydrophilic vegetation; and 2) uplands, defined as non-wetland ecosystems.

Results

SNLM and KED models

Many combinations of spectral, climate, topographical and other X-variables were evaluated. Based on the AIC values, three models for CC and species percentages and two systems of Age, Ht, and Vol models were initially selected (see Table S1 in Supplementary Materials). The CC model with B_5, NSI, NDMI, PPT_{sm}, AlbY and SS resulted in the smallest AIC. The species percentage model using SS, NDMI, B_5, MT_{sm}, CMD, MAP, Elv and Slp was selected. For the system of Age, Ht and Vol

Draft

models, the previously estimated $\widehat{\mathrm{CC}}$ (i.e., log($\widehat{\mathrm{CC}}$)) was selected for Age and Ht models, $\widehat{\mathrm{Pj}}$ and $\widehat{\mathrm{Aw}}$ were selected for the Ht model, and $\widehat{\mathrm{Ht}}$ was selected for the Vol model. Also, allowing the asymptote (α_i) of the Vol model to change with SS resulted in a smaller AIC for the system.

For KED, a random parameter was added to each model of the SNLM as described earlier (Eqs. 5-7). The estimated random parameter variances for CC, Aw , Pj, Sb, Sw, Age, Ht and Vol were 0.11, 1.14, 7.38, 2.75, 3.53, 0.31, 0.36, and 79 292.93, respectively. Empirical semi-variograms were constructed using the estimates of random effects for each random parameter by location (EBLUP; Schabenberger and Gotway 2005). We found no evidence of directional dependence. The Gaussian semi-variogram model for species percentages and the exponential semi-variogram model for the remaining Y-variables fit the data (Supplementary Fig. S1); spatial correlation was found up to 71 km for the species percentages and up to 300 km and beyond for the other Y-variables. Overall, the evidence indicated that KED should improve the accuracy relative to the SNLM, particularly for CC, Ht, Age and Vol.

Comparisons

Accuracy

Applying the selected SNLM to the target dataset (i.e., validation data) resulted in an average %RMSPE across CC, Ht, Age and Vol of 72% and produced a mean vector of differences nearly equal to a zero vector (paired T^2=8.09, p=0.08, Fig. 2). RMSPE values for CC, Ht, Age and Vol were 26.4%, 7.4 m, 43.1 years and 76.8 m^3 ha^{-1}, and MD values were -2.0%, 1.2 m, -3.6 years, and -0.1 m^3 ha^{-1}, respectively (Table 3). Estimated values were more accurate for the 0 to 10th percentiles of actual CC, Ht, Age and Vol than for the 90th to 100th percentiles. Correlations between actual and estimated Y-variables ranged from 0.49 for Age to 0.56 for Ht (Fig. 3). The overall classification accuracy for the nine species groups was 48% (Table 4), with smaller accuracies for more mixed

Draft

species groups (C-Ot, CD and DC) compared to the more homogenous species groups (C-Sb, D-Aw and NT). In general, coniferous species groups were primarily confused with other coniferous species groups, while the mixed species groups (CD and DC) and D-Aw were primarily confused with each other. The NT group was confused predominantly with the C-Sb group.

The KED method improved the accuracy relative to SNLM for CC, Ht, Age and Vol (Table 3) and for species percentages (Table 4). The vector of mean differences was also not different from a zero vector (paired T^2=7.72, p =0.10, Fig. 2). For CC, Ht, Age and Vol, the KED method resulted in the smallest RMSPE among the three methods tested; however, the MD was slightly larger for CC and Vol relative to the SNLM method (Table 3). KED also resulted in greater accuracies at the extremes of the 0 to 10th and the 90th to 100th percentiles, with the exception of Age where accuracies were greater using SNLM at the smaller percentile range and for VSNN at the upper percentile range. Correlations between actual and estimated values were largest using KED (Fig. 3). Some species percentages accuracies were also slightly greater using the KED method, notably for C-Pj, C-Sb, C-Sw, and D-Aw species groups (Table 4). However, classification confusion among species groups was similar to that using SNLM.

With VSNN, *k* greatly affected the average %RMSPE across CC, Ht, Age and Vol, ranging from 64% for *k* =15 to 83% for *k* =1 (Fig. 2). Using *k* = 1 or 2 resulted in vectors of mean differences close to a zero vector (paired T^2 = 0.22 with p=0.99 and 7.0 with and p=0.13, respectively). However, for *k* > 2, there was at least one mean difference that was significantly different from zero (T^2 > 15.01, p <0.0001, Fig. 2). Based on these results, *k* = 2 was selected for most comparisons to the other two methods. Using *k* = 2, RMSPE values for CC, Ht, Age and Vol were larger than the other methods (Table 3). However, the MD for CC was smaller than either SNLM or KED. Also, estimates at extremes were more accurate using VSNN for Age, but not for other Y-variables.

Draft

Correlations between actual and estimated CC, Ht, Age and Vol were the smallest among the three methods, ranging from 0.39 for CC to 0.52 for Ht (Fig. 3). Similarly, the overall classification accuracy for species groups was the smallest (Table 4). This was particularly true for NT, resulting from smaller accuracies at the 0 to 10^{th} percentile of CC, which did not improve with increasing *k* (Table 3). However, VSNN with *k* =2 produced the greatest accuracy for mixed species, namely, CD and DC.

Overall, KED was the most accurate of the methods tested, for estimating CC, Ht, Age and Vol and species percentages. The resulting multivariate maps using KED (Fig. 4; see also doi: 10.14288/1.0354319) can be used in decision support analyses and also illustrate the logical consistencies among the estimated forest attributes. For example, the inset maps show taller heights but younger ages nearer the southern boundary, indicating higher site productivities given the more favorable climate for tree growth. Across the Y-variables, Age was one of the most challenging to estimate using SNLM or KED. For the VSNN methods, the most challenging Y variable was CC, especially for the extremes of the 0 to 10^{th} and 90^{th} to 100^{th} percentiles.

Logical consistency

All three methods were designed to meet the criteria described in Rules 1 to 3. To assess Rule 4, we examined $\widehat{MAI}$ (Supplementary Fig. S2). All three methods were able to estimate cumulative MAI distributions that were similar in shape to the target dataset and values were within biological expectations for the boreal (range = nearly 0.0 to 5.0 m^3 ha^{-1} yr^{-1} as represented in the reference dataset). However, the VSNN methods using large *k* values resulted in fewer estimated small and large MAI values representing a loss in variability relative to actual distributions. For Rule 5, under both lowland and upland communities, SNLM, KED and VSNN with *k* =2 resulted in species assemblages similar to those actual in the target dataset (Fig 5). In wetland communities, the target

Draft

dataset had a large frequency of Sb groupings, which was estimated in all of the methods; however, for the VSNN methods the frequency of mixed species grouping was larger than the SNLM or KED (Table 4 and Fig. 5). This effect was greater for VSNN with k =15 than k =2.

Discussion

Interest in designing resilient landscape strategies for changing environmental conditions have driven policy makers to use decision support tools that are based on wall-to-wall forest-attributes information. In this study, we compared two model-based (SNLM and KED) and one model-free (VSNN) multivariate estimation methods to examine possible trade-offs between accuracy and logical consistency for forest attributes across a macroscale. A cautionary note is that we did not compare all possible multivariate estimation methods, nor all variations of methods we did test.

Draft

Using the model-free VSNN with k > 2 did provide more accurate results than using SNLM for CC, Ht, Age and Vol. Then, increasing k to 15 (i.e., greater smoothing) provided more accurate results for Ht, Age and Vol relative to KED. These results indicated that using the model-free approach can provide more accurate results as anticipated. Also, increasing k in univariate kNN has been shown to decrease the RMSPE until an optimum k is reached (McRoberts 2009). However, as k increased, the range of estimated CC values shrunk, resulting in less accurate estimates for the 0 to 10^{th} percentile of CC in particular. This is particularly important since these small CC percentiles are used to define non-treed areas. Also, using VSNN with k > 2 compressed the range of MAI values (Supplementary Fig. S2), underestimating areas of both very low and very high productivity forests, greatly impacting macroscale decision support analyses. Further, using VSNN with larger k-values resulted in estimated species compositions that included more species. This would lead to an overestimate of forest area with large species diversity, affecting estimates of ecological services from forests. At the extreme using very large k-values, all areas would be estimated to have all species which could be biologically impossible. Further, the ability to estimate rare species and to assess forest fire risks that change with species composition (Bernier et al. 2016) would be greatly curtailed. Overall, VSNN with $k \leq 2$ was needed to meet logical consistency rules, but this adversely affected the accuracies.

We found that an SNLM can be carefully designed to meet logical consistency rules, while remaining competitive with VSNN with regards to accuracy. Knowledge of the system being modeled is required, since careful selection of model forms and predictor variables is needed to obtain logically consistent predictions. Haara and Kangas (2012) showed that model-based methods result in greater accuracies relative to VSNN when the specified model was correct. Further, more accurate estimates were obtained for the lower and upper limits of some forest attributes using SNLM versus VSNN. This is particularly important for estimated CC, given its use in delineating treed versus non-treed areas in forest monitoring frameworks (Halperin et al. 2016). Similar results were obtained by Bollandsås et al. (2013) who showed that using a system of models method to estimate diameter percentiles led to greater accuracies for smaller percentiles relative to a VSNN method. As in this study, Hall et al. (2006) demonstrated the use of a recursive system of models to estimate above ground biomass and volume using estimates of CC and Ht from earlier models in the system. They found that nonlinear models were accurate given that forest attributes tend to have a nonlinear spectral reflectance pattern which can be explained by the influence of canopy development, amount of shadow within the canopy, and forest understory effects on spectral response. However, they cautioned the use of locally-fitted models for larger spatial scales. Räty and Kangas (2008) further emphasized the need to allow parameters to vary for local conditions.

To allow for locally varying conditions, we used KED and allowed some parameters of the SNLM to spatially vary, resulting in more accurate estimates relative to SNLM. Other researchers showed

Draft

results similar to our study with accuracy improvements via spatial localization using kriging without (i.e., no predictor variables) and with external drift (Räty and Kangas 2012; Babcock et al. 2013). For our study, we calculated the random effects for each spatial varying parameter at a 20 km spatial scale reflecting distances between NFI photo-plots. This removed abrupt changes at smaller spatial lags noted by Tuominen et al (2003). We found spatial correlations up to 300 km for some forest attributes in our study area (Supplementary Fig. S1). In their study, Liang et al. (2016) mapped global forest productivity and found spatial correlation over thousands of kilometers using residual errors. We found more limited spatial correlation ranges for species percentages (Supplementary Fig. S1) and, correspondingly, the accuracy of KED was similar to SNLM for these attributes. Of the logically consistent methods we tested, KED gave the best results. Overall, accuracies for these estimated forest attributes were similar to other studies using multi-sourced inventories (e.g., Ohmann and Gregory 2002; Hall et al. 2006; Beaudoin et al. 2014) and for macroscale studies looking to develop global-scale maps of forest attributes (Simard et al. 2011).

Draft

Conclusions

Both accuracy and logical consistency of estimated forest attributes are critical for reliable strategic forest analyses. Given the extensive land area of a macroscale, and the extremely limited accessibility of much of Canada's forests, the photo-plots used in this study provided a viable option for providing the information needed in decision support systems. Of the methods we tested, KED provided both accuracy and logical consistency if based on a carefully designed SNLM. While VSNN methods with larger k values can be more accurate, we found logical inconsistencies for $k >$ 2 that would affect strategic analyses using this information. Overall, KED is our recommended method for providing the forest attribute information needed for decision support systems.

Acknowledgements

Funding for this work was provided by the SMartTForests research program (see www.smartforests.ca) and BioFuelNet (see www.biofuelnet.ca). The authors would like to thank Dr. Tongli Wang (University of British Columbia, UBC) for helping gain access to climate data and Drs. Ron Hall (formerly of Natural Resources Canada), Nicholas Coops (UBC) and Ryan Frazier (formerly of UBC, now Arizona State University) for advice on processing Landsat imagery. We would also like to thank the helpful staff at the Canadian Forest Service within Natural Resources Canada, in particular, Graham Stinson, Frank Eichel and Glenda Russo for providing assistance with access to Canada's NFI.

References

Avery, T. E., and Burkhart, H.E. 2002. Forest measurements, 2nd ed. McGraw-Hill, Toronto. pp. 280-289.

Babcock, C., Matney, J., Finley, A. O., Weiskittel, A., and Cook, B. D. 2013. Multivariate spatial regression models for predicting individual tree structure variables using LiDAR data. J. Select. Top. Appl. Earth Obs. Remote Sens. **6**(1):6-14. doi: 10.1109/JSTARS.2012.2215582.

Banskota, A., Kayastha, N., Falkowski, M., Wulder, M.A., Froese, R.E. and White, J.C. 2014. Forest monitoring using landsat time series data: a review. Can. J. Remote. Sens. **40**(5):362–384. doi:10.1080/07038992.2014.987376.

Beaudoin , A., Bernier, P.Y., Guindon, L., Villemaire, P., Guo, X.J., Stinson, G., Bergeron, T., Magnussen, S., and Hall, R.J. 2014. Mapping attributes of Canada's forests at moderate resolution through kNN and MODIS imagery. Can. J. For. Res. **44**(5):521-532. doi: 10.1139/cjfr-2013-0401.

Draft

Bernier, P.Y., Gauthier, S., Jean, P.O., Manka, F., Boulanger, Y., Beaudoin, A., and Guindon, L. 2016. Mapping local effects of forest properties on fire risk across Canada. For. **7**(8):157. doi: 10.3390/f7080157.

Bettinger, P. Lennette, M., Johnson, K.N., and Spies, T. 2005. A hierarchical spatial framework for forest landscape planning. Ecol. Model. **182**(1):25-48. doi: 10.1016/j.ecolmodel.2004.07.009.

Boisvenue, C., Smiley, B.P., White, J.C., Kurz, W.A., and Wulder, M.A. 2016a. Improving carbon monitoring and reporting in forests using spatially-explicit information. Carbon Balance Manag. **11**(1):23. doi: 10.1186/s13021-016-0065-6.

Boisvenue, C., Smiley, B.P., White, J.C., Kurz, W.A., and Wulder, M.A. 2016b. Integration of Landsat time series and field plots for forest productivity estimates in decision support models. For. Ecol. Manage. **376**(1): 284-297. doi: 10.1016/j.foreco.2016.06.022.

Bokalo, M., Stadt, K.J., Comeau, P.G., and Titus, S.J. 2010. Mixedwood Growth Model (MGM) (version MGM2010.xls). University of Alberta, Edmonton, Alberta, Canada. Available from http://www.rr.ualberta.ca/Research/MixedwoodGrowthModel.aspx [accessed 2 January 2015].

Bollandsås, OM., 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. For. **86**(4):493–501. doi:10.1093/forestry/cpt020.

Brandt, J.P. 2009. The extent of the North American boreal zone. Environ. Rev. **17**(1): 101-161.doi: 10.1139/A09-004.

Brandt, J.P., Flannigan, M.D., Maynard, D.G., Thompson, I.D., and Volney, W.J.A. 2013. An introduction to Canada's boreal zone: ecosystem processes, health, sustainability, and environmental issues. Environ. Rev. **21**(4): 207-226. doi: 10.1139/er-2013-0040.

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**(1): 282-294. doi: 10.1016/j.rse.2016.02.001.

Crookston, N.L., and Finley, A.O. 2008. YaImpute: An R package for *k*NN imputation. J. Stat. Softw. **23**(10): 1-16. doi: 10.18637/jss.v023.i10.

Eskelson, B.N.I., 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.

FAO (Food and Agriculture Organization). 2015. Global forest resources assessment 2015: Terms and definitions. United Nations, FAO, Rome, Italy. FAO For. Pap. 36.

Fehrmann, L., Lehtonen, A., Kleinn, C., and Tomppo, E. 2008. Comparison of linear and mixed-effect regression models and a k-nearest neighbour approach for estimation of single-tree biomass. Can. J. For. Res. **38**(1):1-9. doi: 10.1139/X07-119.

Flannigan, M.D., Logan, K.A., Amiro, B.D., Skinner, W.R., and Stocks, B.J. 2005. Future area burned in Canada. Clim. Change. **72**(1–2): 1-16. doi: 10.1007/s10584-005-5935-y.

Foody, G.M. 2002. Status of land cover classification accuracy assessment. Remote. Sens. Environ. **80**(1):185-201. doi: 10.1016/S0034-4257(01)00295-4.

Furnival, G.M., and Wilson, R.W. 1974. Regressions by leaps and bounds. Technometrics. **16**(4):499-511. doi: 10.2307/1267601.

Garcia, O. 2003. Dimensionality reduction in growth models: an example. For. Biom. Mod. Infor. Sci. **1**(1):1-15. Available from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.458.2915 [accessed 10 May 2017]

Draft

Geogratis 2013. Data catalogue [online]. Available from http://geogratis.gc.ca [accessed 4 December 2013].

Gillis, M. D., Omule, A. Y., and Brierley, T. 2005. Monitoring Canada's forests: the national forest inventory. For. Chron. **81**(2):214-221. doi: 10.5558/tfc81214-2.

Glebov, F.Z, and Korzukhin, M.D. 1992.Chapter 9: Transitions between boreal forest and wetland. *In* A systems analysis of the global boreal forest. *Edited by* H.H. Shugart, R. Leemans, and G.B. Bonan. Cambridge University Press, Cambridge. pp. 241–266.

Gujarati, D.N., Porter, D.C., and Gunasekar, S. 2011. Simultaneous equation methods. In Basic econometrics. 5th ed. McGraw Hill Book Co, New York. pp. 717 – 785.

Haara, A., and Kangas, A. 2012. Comparing k nearest neighbours methods and linear regression–is there reason to Select one over the other? Math. Comput. For. Nat. Res. Sci. **4**(1):50-65. Available from http://mcfns.com/index.php/Journal/article/view/MCFNS.4%3A50/MCFNS.4%3A50 [accessed 10 May 2017].

Hall, R. J., Skakun, R. S., Arsenault, E. J., and Case, B. S. 2006. Modeling forest stand structure attributes using Landsat ETM+ data: Application to mapping of aboveground biomass and stand volume. For. Ecol. Manage. **225**(1):378-390. doi: 10.1016/j.foreco.2006.01.014.

Halperin, J., LeMay, V., Coops, N., Verchot, L., Marshall, P., and Lochhead, K. 2016. Canopy cover estimation in miombo woodlands of Zambia: comparison of Landsat 8 OLI versus RapidEye imagery using parametric, nonparametric, and semiparametric methods. Remote Sens. Environ. **179**(1):170-182. doi: 10.1016/j.rse.2016.03.028.

Hotelling, H. 1951. A generalized T-test and measure of multivariate dispersion. *In* Proceedings of the second Berkeley symposium on mathematical statistics and probability. Berkeley: Univ.

Draft

Calif. Press. pp. 23–42. Available from https://projecteuclid.org/euclid.bsmsp/1200500217 [accessed 10 May 2017].

Indyk, P., and Motwani, R. 1998. Approximate Nearest Neighbor: Towards Removing the Curse of Dimensionality. *In* Proceedings of the 30th Symposium on Theory of Computing, Dallas, Texas, USA May 24 – 26, 1998. pp. 604-613.doi: 10.1145/276698.276876.

Jin, S., and Sader, S.A. 2005. Comparison of time series tasseled cap wetness and the normalized difference moisture index in detecting forest disturbances. Remote Sens. Environ. **94**(3):364-372. doi: 10.1016/j.rse.2004.10.012.

Judge, G.G., Griffiths, W.E., Hill, R.C., Lütkepohl, H., and Lee T-C. 1995. Inference in simultaneous equation models. *In* The theory and practice of econometrics. John Wiley and Sons, Toronto, Canada. pp. 622-631.

Kainz, W. 1995. Chapter 6: Logical consistency. *In* Elements of spatial data quality. *Edited by* S.T. Guptil, and J.L. Morrison. Elsevier Science, Oxford, UK. pp. 109–137.

Kershaw, J.A., Ducey, M.J., Beers, T.W., and B. Husch. 2017. Forest mensuration, 5th ed. John Wiley & Sons, Hoboken,NJ. pp. 446-448.

Key, C.H., and Benson, N.C. 2006. Landscape assessment: Ground measure of severity, the Composite Burn Index; and remote sensing of severity, the normalized burn ratio. *In* FIREMON: Fire Effects Monitoring and Inventory System. *Edited by* D.C. Lutes, R.E. Keane, J.F Caratti, C.H, Key, N.C. Benson, S. Sutherland, and L.J. Gangi, USDA For. Serv. R.M. Res.,, Ogden, UT. Gen. Tech. Rep. RMRS-GTR-164-CD: LA1-51.

Kuusinen, N., Tomppo, E., Shuai, Y., and Berninger, F. 2014. Effects of forest age on albedo in boreal forests estimated from MODIS and Landsat albedo retrievals. Remote Sens. Environ.**145**(5):145-153. doi: 10.1016/j.rse.2014.02.005.

Draft

LeMay, V. 1990. MSLS: A linear least squares technique for fitting a simultaneous system of equations with a generalized error structure. Can. J. For. Res. **20**(12):1830-1839. doi:10.1139/x90-246.

LeMay, V., and Temesgen, H. 2005. Comparison of nearest neighbor methods for estimating basal area and stems per hectare using aerial auxiliary variables. For. Sci. **51**(2):109-119.

Liang, J., Crowther, TW., Picard, N., Wiser, S., Zhou, M., Alberti, G., Schulze, E-D., McGuire, AD., Bozzato, F., Pretzsch, H., et al. 2016. Positive biodiversity-productivity relationship predominant in global forests [online]. Sci. **354** (6309). doi:10.1126/science.aaf8957.

Lindner, M., Sohngen, B., Joyce, L. A., Price, D. T., Bernier, P. Y., and Karjalainen, T. 2002. Integrated forestry assessments for climate change impacts. For. Ecol. Manage. **162**(1):117-136. doi: 10.1016/S0378-1127(02)00054-3.

Littell, R.C., Milliken, G.A., Stroup, W.W., Wolfinger, R.D., and Schabenberger, O. 2006. Spatial variability. In SAS for mixed models. 2nd ed. SAS institute, Cary, NC. pp. 437-478.

Lloyd, C.D. 2007.Spatial Relations. In Local models for spatial analysis. London: CRC/Taylor and Francis. pp. 109-150.

Magnussen, S., and Russo, G. 2012. Uncertainty in photo-interpreted forest inventory variables and effects on estimates of error in Canada's National Forest Inventory. For. Chron. **88(**4): 439-447. doi: 10.5558/tfc2012-080.

Masek, J.G., Vermote, E.F., Saleous, N., Wolfe, R., Hall, F.G., Huemmrich, F., Gao, F., Kutler, J., and Lim, T.K. 2006. A Landsat surface reflectance data set for North America 1990-2000. Geosci. Remote Sens. Lett. **3**(1):68-72. doi: 10.1109/LGRS.2005.857030.

Mauro, F., Monleon, V.J., and Temesgen, H. 2015. Using small area estimation and Lidar-derived variables for multivariate prediction of forest attributes. *In* Forest Inventory and Analysis (FIA) Symposium pushing boundaries: new directions in inventory techniques and

Draft

applications Portland, OR. December 8–10. USDA For. Serv, P. N. Res. Gen. Tech. Rep. PNW-GTR-931. pp. 73-77.

McRoberts, R.E. 2008. Using satellite imagery and the k-nearest neighbors technique as a bridge between strategic and management forest inventories. Remote Sens. Environ. **112**(5): 2212-2221. doi: 10.1016/j.rse.2007.07.025.

McRoberts, R.E. 2009. Diagnostic tools for nearest neighbors techniques when used with satellite imagery. Remote Sens. Environ. **113**(3):489-499. doi: 10.1016/j.rse.2008.06.015.

McRoberts, R.E., Chen, Q., and Walters, B.F. 2017. Multivariate inference for forest inventories using auxiliary airborne laser scanning data. For. Ecol. Manage. **401**(1):295-303. doi 10.1016/j.foreco.2017.07.017.

Merz, R., and Blöschl, G. 2004. Regionalisation of catchment model parameters. J. Hydro. **287**(1):95-123. doi: 10.1016/j.jhydrol.2003.09.028.

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

Moisen, G.G., and Frescino, T.S. 2002. Comparing five modelling techniques for predicting forest characteristics. Ecol. model. **157**(2):209-225. doi: 10.1016/j.biocon.2011.11.013.

Morrison, J. L., 1995. Chapter 1: Spatial data quality. *In* Elements of spatial data quality. *Edited by* S.T Guptil, and J.L Morrison. Elsevier Science, Oxford, UK. pp. 1–12.

Nilsson, M., Nordkvist, K., Jonzén, J., Lindgren, N., Axensten, P., Wallerman, J., Egberth, M., Larsson, S., Nilsson, L., Eriksson, J., and Olsson, H. 2016. A nationwide forest attribute map of Sweden predicted using airborne laser scanning data and field data from the National Forest Inventory. Remote Sens. Environ. **194**(1): 447-454. doi: 10.1016/j.rse.2016.10.022.

Draft

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.

Pindyck, R.S., and Rubinfeld, D.L. 1981. Econometric models and economic forecasts. 2nd ed. McGraw-Hill Book Co., New York.

Räty, M., and Kangas, A. 2008. Localizing general models with classification and regression trees. Scand. J. For. Res. **23**(5):419-430. doi: 10.1080/02827580802378826.

Räty, M., and Kangas, A. 2012. Reprint of: Comparison of k-MSN and kriging in local prediction. For. Ecol. Manage. **272**(1): 51-60. doi: 10.1016/j.foreco.2011.12.046.

Richards, F. 1959. A Flexible Growth Function for Empirical Use. J. Exp. Bot. **10**(2):290-301. doi: 10.1093/jxb/10.2.290.

Roecker, E. 1991. Prediction error and its estimation for subset-selected models. Technometrics. **33**(4): 459–468. doi: 10.2307/1269417.

Rouse, J.W. Jr., Haas, R.H., Deering, D.W., Schell, J.A., and Harlan, J.C. 1974. Monitoring the vernal advancement and retrogradation (green wave effect) of natural vegetation. NASA/GSFC Type III Final Report, Greenbelt, MD. pp. 371.

Roy, P.S., Sharma, K.P., and Jain, A. 1996. Stratification of density in dry deciduous forest using satellite remote sensing digital data-an approach based on spectral indices. J. Biosci. **21**(5):723–734. doi: 10.1007/BF02703148.

Simard, M., Pinto, N., Fisher, J.B., and Baccini, A. 2011. Mapping forest canopy height globally with spaceborne lidar. J. Geophys.Res: Biogeosci. **116** (G4). doi: 10.1029/2011JG001708.

Sirois, L. 1992. Chapter 7: The transition between boreal forest and tundra. In A systems anaylsis of the global boreal forest. *Edited by* H.H. Shugart, R. Leemans, and G.B. Bonan. Cambridge University Press, Cambridge. pp. 196–215.

Draft

Snee, R. 1977. Validation of regression models: methods and examples. Technometrics. **19**(4):415–428. doi: 10.1080/00401706.1977.10489581.

Soja, A.J., Tchebakova, N.M., French, N.H., Flannigan, M.D., Shugart, H.H., Stocks, B.J., Sukhinin, A.I., Parfenova, E.I., Chapin, F.S., and Stackhouse, P.W. 2007. Climate-induced boreal forest change: Predictions versus current observations. Glob. Planet. Change. **56**(3): 274-296. doi: 10.1016/j.gloplacha.2006.07.028.

Stage, A. R., and Salas, C. 2007. Interactions of elevation, aspect, and slope in models of forest species composition and productivity. For. Sci. **53**(4):486-492.

Schabenberger, O., and Gotway, C. 2005. Spatial prediction and kriging. *In* Statistical methods for spatial analysis. Chapman and Hall. Boca Raton, FL. pp. 215-295.

Tarboton, D.G. 1997. A new method for the determination of flow directions and contributing areas in grid digital elevation models. Water Resour. Res. **33**(2):309–319. doi: 10.1029/96WR03137.

Temesgen, 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):277-285. doi: 10.1016/S0378-1127(02)00321-3.

Thompson, S.K. 1987. Sample size for estimating multinomial proportions. Am. Stat. **41**(1): 42-46. doi: 10.1080/00031305.1987.10475440.

Tipton, J., Moisen, G., Patterson, P., Jackson, T.A., and Coulston, J. 2010. Sampling intensity and normalizations: Exploring cost-driving factors in nationwide mapping of tree canopy cover. *In:* Monitoring across borders: 2010 Joint Meeting of the forest inventory and analysis symposium and the southern mensurationists. *Edited by* W. McWilliams, and F. A. Roesch. USDA For. Serv., S. Res. e-Gen. Tech. Rep. SRS-157. Asheville, NC: pp. 201-208.

Draft

Tomppo, E. 1988. Standwise forest variate estimation by means of satellite images. University of Helsinki, Department of Forest Mensuration and Management, Research Notes 21. pp. 103–111.

Tomppo, E., and Czaplewski, R.L. 2002. Potential for a remote-sensing aided forest resource survey for the whole globe. Unasylva. **53**(210):16-18.

Tomppo, E., Haakana, M., Kaitla, M., and Perasaari, J. 2008a. Multi-source national forest inventory. *In* Multi-source national forest inventory: methods and applications. Springer: New York, NY, USA. pp. 1 – 62.

Tomppo, E., Olsson, H., Ståhl, G., Nilsson, M., Hagner, O., and Katila, M. 2008b. Combining national forest inventory field plots and remote sensing data for forest databases. Remote Sens. Environ. **112**(5):1982-1999. doi: 10.1016/j.rse.2007.03.032.

Tomppo, E. 2010. Chapter 1: Introduction. *In* National Forest Inventories: Pathways for common reporting. *Edited by* E. Tomppo, Th. Gschwantner, M. Lawrence, and R. McRoberts. Springer. New York. pp. 1-18. doi: 10.1007/978-90-481-3233-1.

Tuominen, S., Fish, S., and Poso, S. 2003. Combining remote sensing, data from earlier inventories, and geostatistical interpolation in multi-source forest inventory. Can. J. For. Res. **33**(4):624-634. doi: 10.1139/x02-199.

USGS Earth Explorer 2013. Climate data record [online]. Available from https://remotesensing.usgs.gov/ecv/CDR_1.php [accessed 4 December 2013].

van den Boogaart, K. G., and Tolosana-Delgado, R. 2008. "Compositions": a unified R package to analyze compositional data. Comput. Geosci. **34**(4):320-338. doi: 10.1016/j.cageo.2006.11.017.

Ver Hoef, J. M., and Temesgen, H. 2013. A comparison of the spatial linear model to nearest neighbor (k-NN) methods for forestry applications [online]. PLoS One. **8**(3). doi:10.1371/journal.pone.0059129.

Draft

Vidal, C., Sallnas, O., Redmond, J., Alberdi, I., Barreiro, S., Hernandez, L., and Schadauer, K. 2016. Introduction. *In* National Forest Inventory: Assessment of Wood Availability and Use. *Edited by* C. Vidal, I. Alberdi, L. Hernández, and J.J. Redmond. Springer International Publishing pp. 1-23. doi: 10.1007/978-3-319-44015-6.

Wang, T., Hamann, A., Spittlehouse, D.L., and Murdock, T.Q. 2012. ClimateWNA – High-resolution spatial climate data for western North America. J. Appl. Meteorol. Climatol. **51**(1):16-29. doi: 10.1175/JAMC-D-11-043.1.

Weed, A.S., Ayres, M.P., and Hicke, J.A. 2013. Consequences of climate change for biotic disturbances in North American forests. Ecol. Mono. **83**(4): 441-470.doi: 10.1890/13-0160.1.

Zhao-gang, L., and Li, Fr. 2003. The generalized Chapman-Richards function and applications to tree and stand growth. J. For. Res. **14**(1):19-26. doi: 10.1007/BF02856757.

Zhu, Z., and Woodcock, C.E. 2012. Object-based cloud and cloud shadow detection in Landsat imagery. Remote Sens. Environ. **118**(1):83-94. doi: 10.1016/j.rse.2011.10.028.

Draft

Tables

Table 1 Characteristics of the X-variables used as possible predictors for estimating multiple forest attributes. Values were averaged for the 90 m pixel when the spatial resolution was < 90 m.

X-variable	Description	Spatial Resolution
Spectral		
Landsat bands	B_1-Blue (0.45 - 0.52 µm); B_2-Green (0.52 - 0.60 µm); B_3-Red (0.63 - 0.69 µm); B_4-Infrared (0.77 - 0.90 µm); B_5-Infrared (1.55 - 1.75 µm); B_7-Mid-Infrared (2.08 - 2.35 µm)	30 m
Landsat indices		
NDVI	Normalized difference vegetation index $(B_4-B_3)/(B_4+B_3)$ (Rouse et al. 1974)	30 m
NDMI	Normalized difference moisture index $(B_5-B_4)/(B_5+B_4)$ (Jin and Sader 2005)	30 m
NLI	Nonlinear Index $(B_4{}^2 - B_3)/(B_4{}^2 + B_3)$ (Goel and Qin 1994)	30 m
NBR	Normalized burn ratio $(B_4+B_7)/(B_4+B_7)$ (Key and Benson 2006)	30 m
NSI	Normalized soil index $[(B_5+B_3) - (B_1+B_4)]/[(B_5+B_3) + (B_1+B_4)]$ (Roy et al. 1996)	30 m
Albedo	Albedo, $\sum_{i=1, i\neq 6}^{7} B_i$, the sum of reflectances between 0.45-2.35 µm (Lu et al. 2004)	30 m
Climatic*		
Precipitation		
MAP	Mean annual precipitation (mm)	1 000 m
PPT_{sm}	Summer (June to August) precipitation (mm)	1 000 m
PPT_{wt}	Winter (December to February) precipitation (mm)	1 000 m
CMD	Climatic moisture deficit	1 000 m
Temperature		
MAT	Mean annual temperature (°C)	1 000 m
MT_{sm}	Summer (June to August) mean temperature (°C)	1 000 m
MT_{wt}	Winter (December to February) mean temperature (°C)	1 000 m
MCMT	Mean temperature of the coldest month (°C)	1 000 m
MWMT	Mean temperature of the warmest month (°C)	1 000 m
FFP	Length of the frost-free period (days)	1 000 m
Degree days		
DD5	Degree-days above 5°C (growing degree days)	1 000 m
Topographic**		
Elv	Elevation above sea level (m)	30 m
Slp	Slope angle in degrees	30 m
Asp	Angle from north in degrees	30 m
CTI	Compound topographic index, ln[(AC + 1)/Slp], where AC is the accumulation value of all cells flowing into each downslope cell with each cell weight equal to 1 (Tarboton 1997)	30 m
Vector***		
SS	Canvec+ dataset of saturated soil polygons (0=not saturated; 1=saturated).	.
Coordinates		
AlbX	Albers X coordinate (m)	.
AlbY	Albers Y coordinate (m)	

* Seasonal and annual climatic variables were accessed from ClimateNA (Wang et al. 2015).

** Accessed from Canada Digital Elevation Data (Geogratis 2013).Eleven variables describing interactions of Elv, Slp and Asp were calculated as per Stage and Salas (2007).

*** Accessed from the Natural Resource Canada CanVec+ dataset (Geogratis 2013).

Draft

792 Table 2. Statistics for forest attributes (Y-variables) using all data for CC (63 428 and 15 025 records for the reference and target datasets,
793 respectively), but using only treed records for the other Y-variables (52 807 and 12 347 for the reference and target datasets, respectively).

Attribute	Description	Reference				Target			
		Mean	**Min.**	**Max.**	**Std. Dev.**	**Mean**	**Min.**	**Max.**	**Std. Dev.**
CC (%)	Percent of ground area covered by the vertical projection of tree crown areas.	45. 6	0.00	100.0	28.1	45.3	0.00	100.0	29.1
Species (%)	Separation of the CC% into species groups (sum to 100%)								
Aw	*Populus* spp. + *Betula* spp.	25.4	0.0	100.0	35.9	25.1	0.0	100.0	35.8
Pj	*Pinus* spp.	14.1	0.0	100.0	28.0	14.4	0.0	100.0	28.5
Sb	*Picea mariana* + *Larix* spp.	46.7	0.0	100.0	41.6	46.5	0.0	100.0	42.0
Sw	*Picea glauca* + *Abies* spp.	13.2	0.0	100.0	23.7	13.4	0.0	100.0	25.8
Other	Remaining spp.	0.6	0.0	100.0	4.5	0.6	0.0	100.0	5.0
Ht (m)	Average height of dominant trees	12.5	0.2	42.5	6.3	12.4	0.2	36.4	6.4
Age (Years)	Average age of the leading tree species	77.1	1.0	304.0	37.5	75.3	1.0	290.0	37.2
Vol (m^3ha^{-1})	Total stem volume (live + dead) in for all trees > 1.3 m tall	107.2	0.0	649.0	81.7	106.2	0.0	609.0	82.9

794 Note: Min. is the minimum, Max. is the maximum and Std. Dev. is the standard deviation.

Draft

Table 3. Accuracies of SNLM, KED and VSNN (*k*=2) methods. VSNN with *k* =15 was added for comparison (shaded grey). Statistics were computed using all of the target data and also using the 0 to 10th and the 90 to 100th percentiles of the corresponding Y-variable. Bold indicates a more accurate method (e.g., a lower MD).

Y	Percentiles (Ranges)	n_{targ}	SNLM		KED		VSNN (*k*=2)		VSNN (*k*=15)	
			MD	RMSPE	MD	RMSPE	MD	RMSPE	MD	RMSPE
CC (%)	0-10th (0-21)	1 250	**10.7**	23.8	13.2	**23.1**	18.3	29.4	25.0	30.2
	90-100th (>85)	1 567	-30.4	35.5	**-23.9**	**29.8**	-36.1	41.0	-35.8	37.8
	All (0-100)	15 025	-2.0	26.4	2.9	**24.3**	**-0.6**	29.4	-0.6	24.9
Ht (m)	0-10th (0-3.2)	1 240	**2.8**	6.4	3.0	**5.8**	3.5	6.4	7.0	8.7
	90-100th (>20.9)	1 234	**-6.4**	**9.6**	-8.2	10.4	-7.7	9.8	-7.3	8.5
	All (0-36.4)	15 025	1.2	7.4	**0.4**	**6.4**	0.79	7.0	1.4	6.1
Age (Years)	0-10th (0-27)	1 234	**17.5**	**31.5**	27.3	38.9	28.1	44.3	48.4	56.8
	90-100th (>120)	1 975	-44.7	61.9	-39.9	56.1	**-36.9**	**54.5**	-34.4	45.5
	All (0-290)	15 025	-3.6	43.1	**0.9**	**38.9**	6.0	44.3	10.9	38.8
Vol (m^3ha^{-1})	0-10th (0-9)	1 279	32.3	59.8	**30.8**	**56.9**	37.5	70.4	53.2	72.2
	90-100th (>216)	1 249	-131.7	155.8	-112.9	**141.6**	**-111.9**	146.2	-111.0	131.6
	All (0-609)	15 025	**-0.1**	76.8	2.7	**70.6**	5.1	79.6	9.8	68.7

Note: RMSE and MD are defined in Eq. [10] and [12], respectively.

Draft

Table 4. Confusion matrix of broad class species groups for each multivariate estimation method. Classes include: NT (non-treed); D-Aw (>80% Aw); DC (mixed but dominated by Aw); CD (mixed but dominated by coniferous); C-Pj (> 80% conifer, Pj leading); C-Sw (> 80% conifer, Sw leading); C-Sb (> 80% conifer , Sb leading); D-Ot (50 %<Aw < 80% and >20 % other species groups); and C-Ot (50% <conifer < 80% and > 20% other species groups. OA is the overall accuracy. Bold indicates a more accurate method.

		Actual										Users
Estimated	Method	C-Ot	C-Pj	C-Sb	C-Sw	CD	D-Aw	D-Ot	DC	NT	Total	Accuracy
C-Ot	SNLM	0	0	0	0	0	0	0	0	0	0	0%
	KED	0	0	0	0	0	0	0	0	0	0	0%
	VSNN (k=2)	**4**	3	15	2	10	1	0	12	6	53	**8%**
C-Pj	SNLM	0	270	200	70	59	21	0	30	105	755	36%
	KED	0	**450**	292	77	79	30	0	46	148	1 122	**40%**
	VSNN(k=2)	1	422	298	87	53	24	0	25	158	1 068	**40%**
C-Sb	SNLM	31	902	4 109	352	535	211	0	415	851	7 406	55%
	KED	32	785	**4 266**	303	503	185	0	390	896	7 360	58%
	VSNN(k=2)	16	621	3 769	271	308	131	0	165	896	6 177	**61%**
C-Sw	SNLM	8	163	400	429	119	45	0	91	69	1 324	32%
	KED	7	159	459	**488**	133	44	0	100	74	1 464	**33%**
	VSNN(k=2)	0	85	196	203	36	25	0	40	50	635	32%
CD	SNLM	4	78	197	65	111	267	0	204	135	1 061	10%
	KED	6	74	223	64	110	258	0	209	155	1 099	10%
	VSNN(k=2)	19	219	578	160	**252**	214	0	280	279	2 001	**13%**
D-Aw	SNLM	0	22	21	12	31	945	0	87	271	1 389	**68%**
	KED	0	22	27	13	32	1 027	0	98	312	1 531	67%
	VSNN(k=2)	0	30	66	38	55	**1 169**	0	162	276	1 796	65%
D-Ot	SNLM	0	0	0	0	0	0	**0**	0	0	0	**100%**
	KED	0	0	0	0	0	0	**0**	0	0	0	**100%**
	VSNN(k=2)	0	0	5	0	2	4	**0**	6	6	23	**100%**
DC	SNLM	4	33	79	27	46	477	0	150	123	939	16%
	KED	3	31	83	36	50	451	0	146	129	929	14%
	VSNN(k=2)	7	94	200	197	178	376	0	**281**	215	1 548	**18%**
NT	SNLM	2	141	669	55	30	82		37	**1 135**	2 151	53%
	KED	1	88	325	29	24	53	0	25	975	1 520	**64%**
	VSNN(k=2)	2	135	548	52	37	104	0	43	803	1 724	47%
	Total	*49*	*1 609*	*5 675*	*1 010*	*931*	*2 048*	*0*	*1 014*	*2 689*	*15 025*	
Producers Accuracy	SNLM	0%	17%	72%	42%	12%	46%	**100%**	15%	**42%**	OA:48%	
	KED	0%	**28%**	**75%**	**48%**	12%	50%	**100%**	14%	36%	**OA:50%**	
	VSNN(k=2)	**8%**	26%	66%	20%	**27%**	**57%**	**100%**	**28%**	30%	OA:46%	

Draft

Figures

Figure 1. The boreal zone in Canada (Brandt 2009) showing the study area (south of 60° N; boundary is bolded in black) with black squares representing the 2 km by 2 km photo-plots (n=3 298). The dark gray shows areas where forest companies operate. The insert provides a hypothetical photo-plot with delineated polygons and 90 m by 90 m pixel windows. Crossed markings are intersections of major latitudes and longitudes.

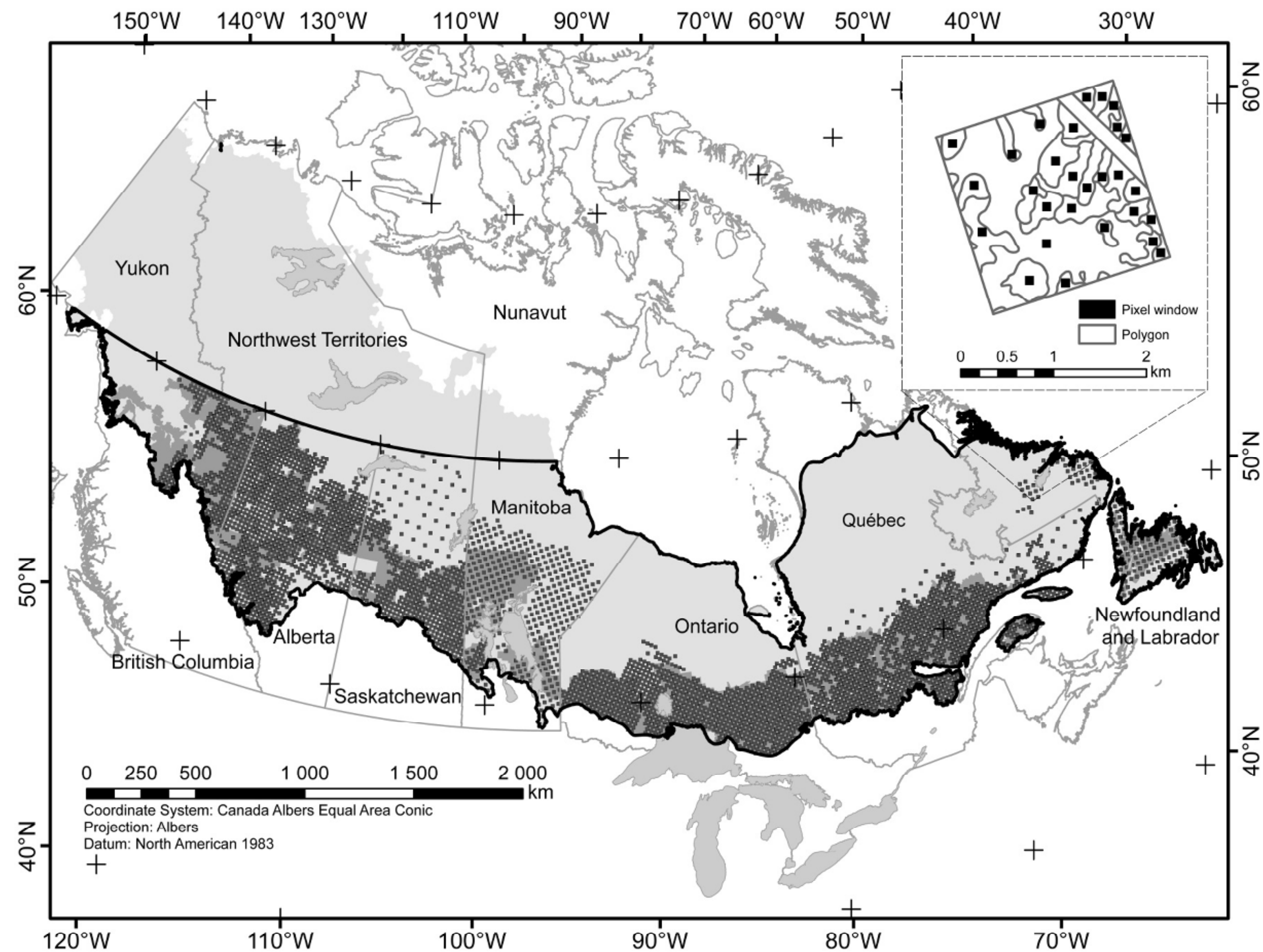
Figure 2. Percent root mean square prediction errors (%RMSE; Eq.11) averaged over CC, Ht, Age, Vol and Hotelling's paired T^2 (Eq. 9) by multivariate estimation method using the target dataset (n_{targ}=15 025). The *k* refers to the number of neighbours in VSNN.

Figure 3. Actual versus estimated values by forest attribute variable for the target dataset. The grey dashed line represents a 1:1 relationship and 'r' is the Pearson's correlation coefficient. Contour lines depict the numbers of points from low (white) to high (black) densities (n_{targ}=15 025).

Draft

Figure 4. Estimated forest attributes using kriging with external drift (KED) for the areas within Canada's boreal forest where forest companies operate. The color ramp displays the minimum (yellow; 0 for all attributes) and the maximum (dark blue; 100 % for CC and species percentages, 45 m for Ht, 300 years for Age and 500 m^3 ha^{-1} for Vol).

Figure 5. Ternary diagrams of species percentages for wetland and upland ecological communities. The vertices of each triangle represent 100 % of the labeled species. Contour lines depict the numbers of points from low (white) to high (black) densities (n_{targ}=15 025).


150°W
140°W
130°W
110°W
90°W
70°W
60°W
50°W
40°W
30°W
60°N
50°N
40°N
120°W
110°W
100°W
90°W
80°W
70°W
Yukon
Northwest Territories
Nunavut
Manitoba
Québec
Ontario
Alberta
British Columbia
Saskatchewan
Newfoundland
and Labrador
Pixel window
Polygon
0 0.5 1 2
km
0 250 500 1 000 1 500 2 000
km
Coordinate System: Canada Albers Equal Area Conic
Projection: Albers
Datum: North American 1983

822

Fig. 1

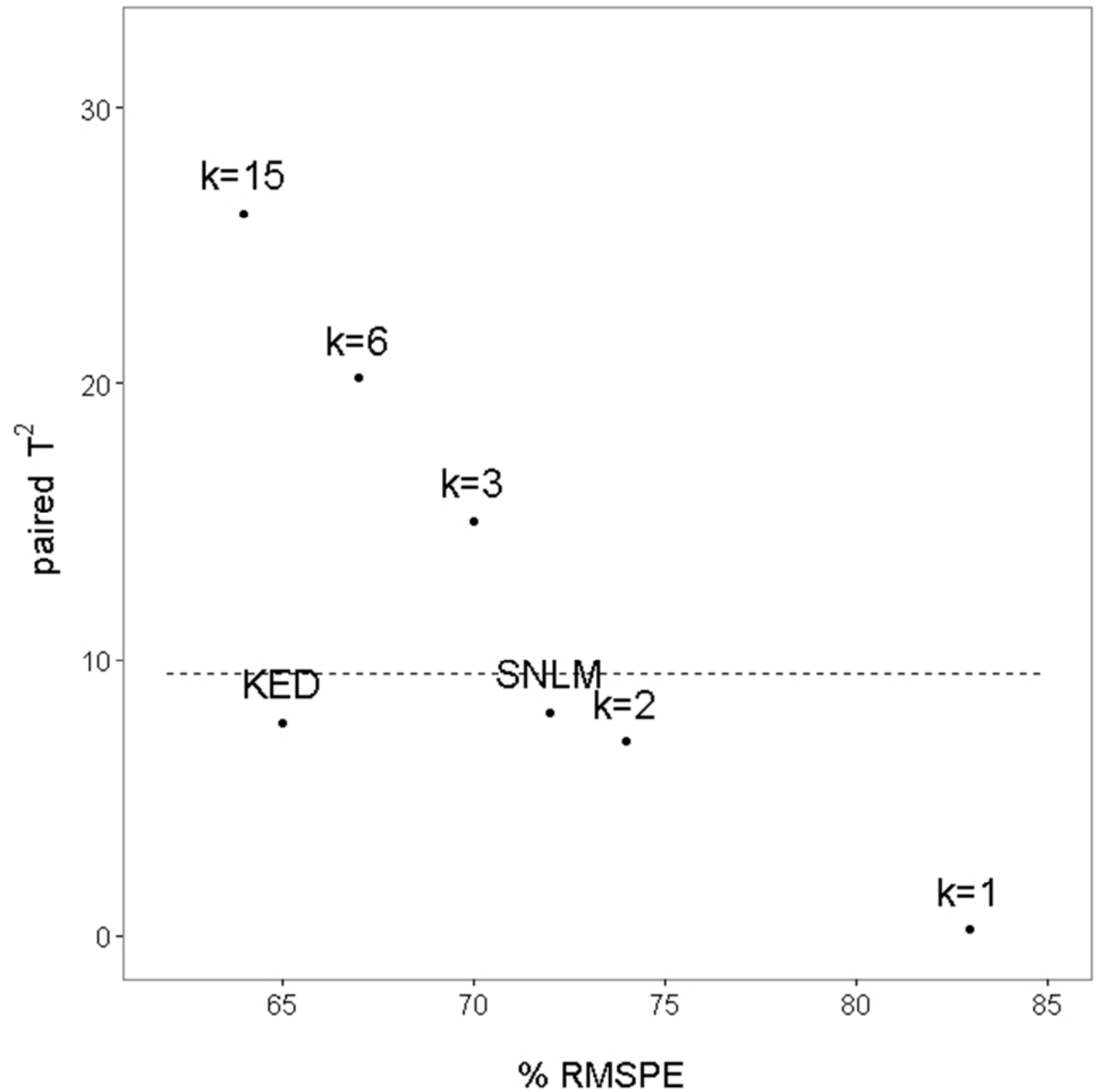


Fig. 2

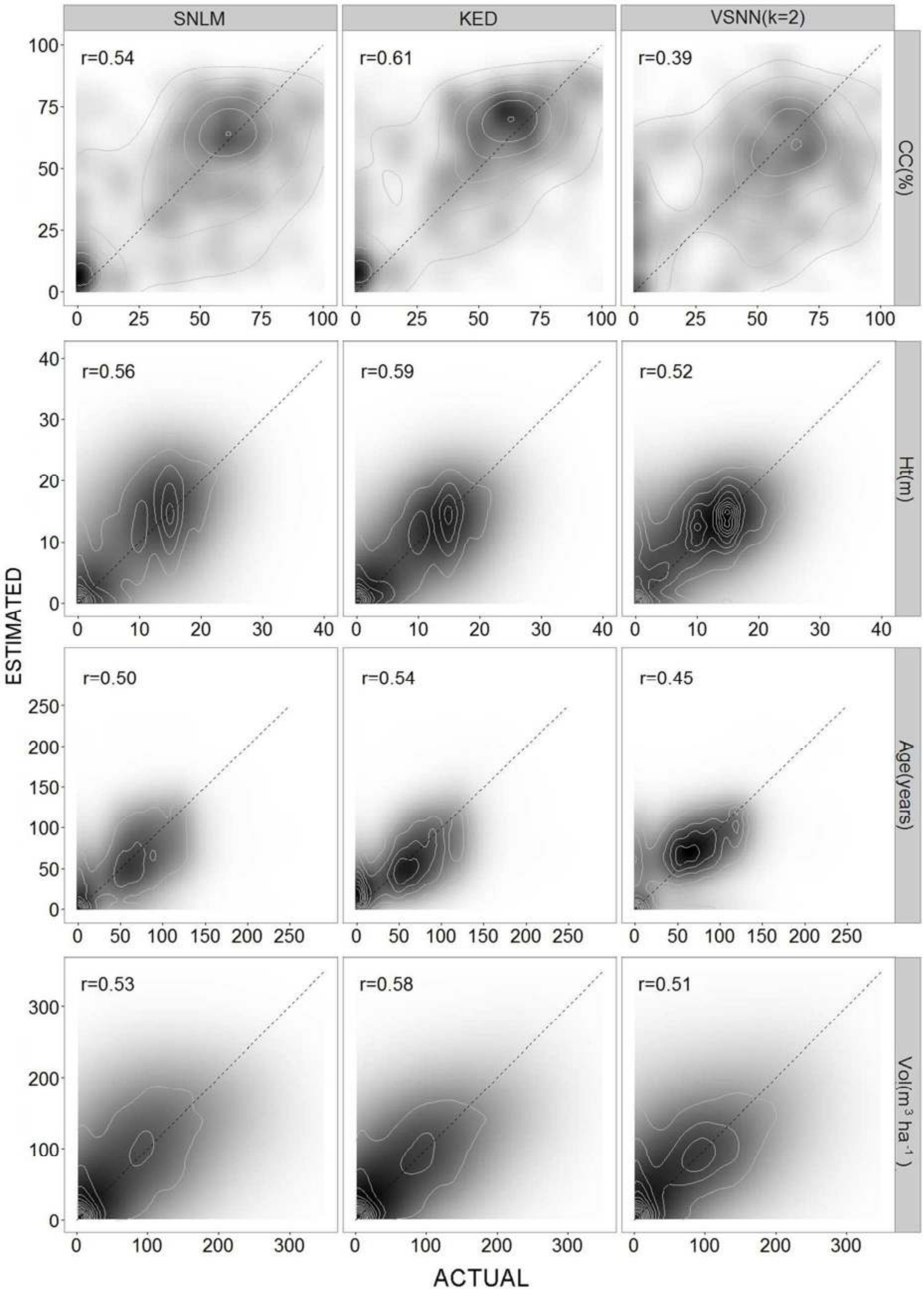

SNLM
KED
VSNN(k=2)
r=0.54
r=0.61
r=0.39
CC(%)
r=0.56
r=0.59
r=0.52
Ht(m)
r=0.50
r=0.54
r=0.45
Age(years)
r=0.53
r=0.58
r=0.51
Vol(m³ ha⁻¹)
ESTIMATED
ACTUAL

Fig. 3

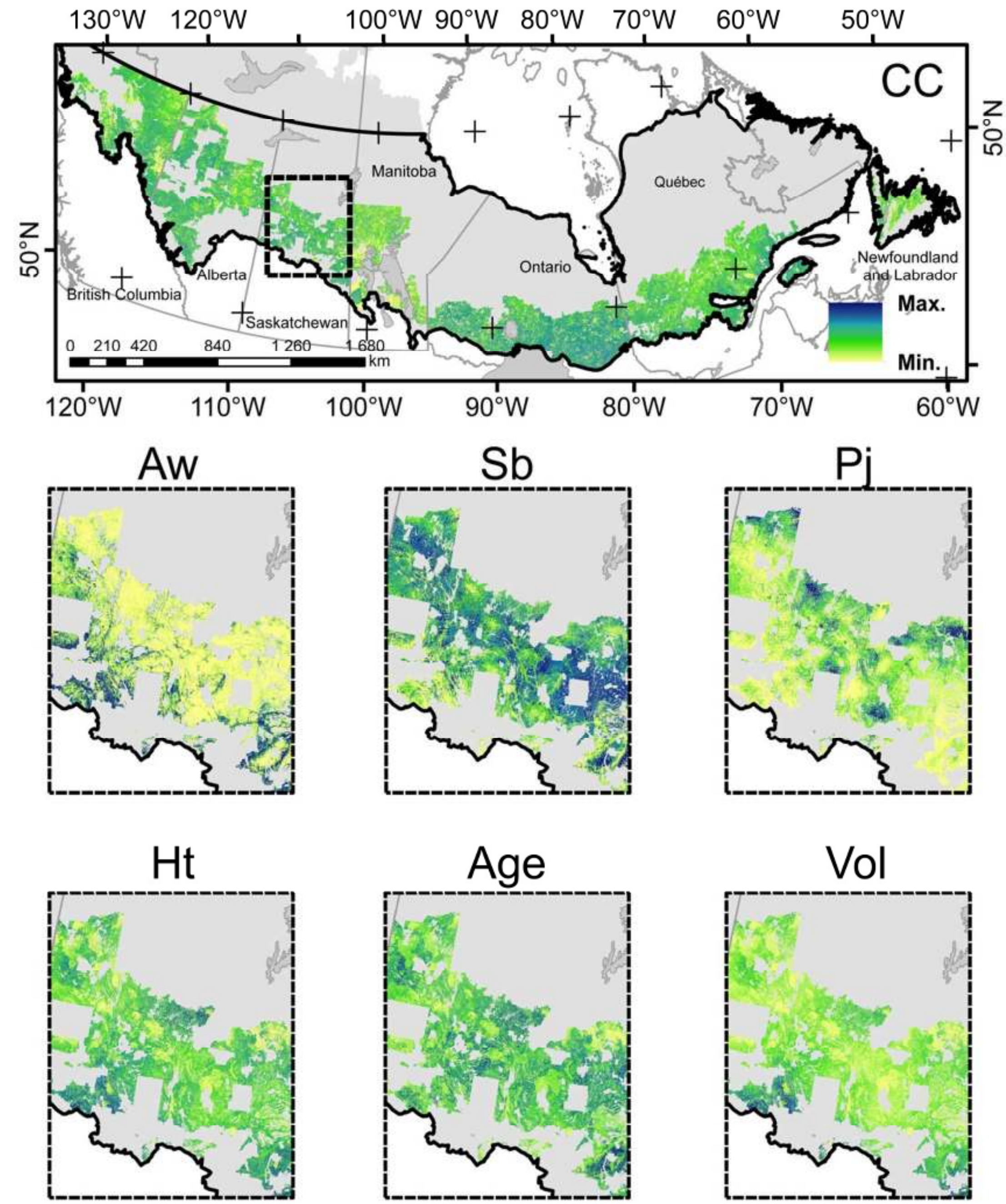

CC
130°W
120°W
100°W
90°W
80°W
70°W
60°W
50°W
50°N
Manitoba
Québec
Ontario
Alberta
British Columbia
Saskatchewan
Newfoundland and Labrador
Max.
Min.
0 210 420 840 1 260 1 680
km
110°W
Aw
Sb
Pj
Ht
Age
Vol

Fig. 4.

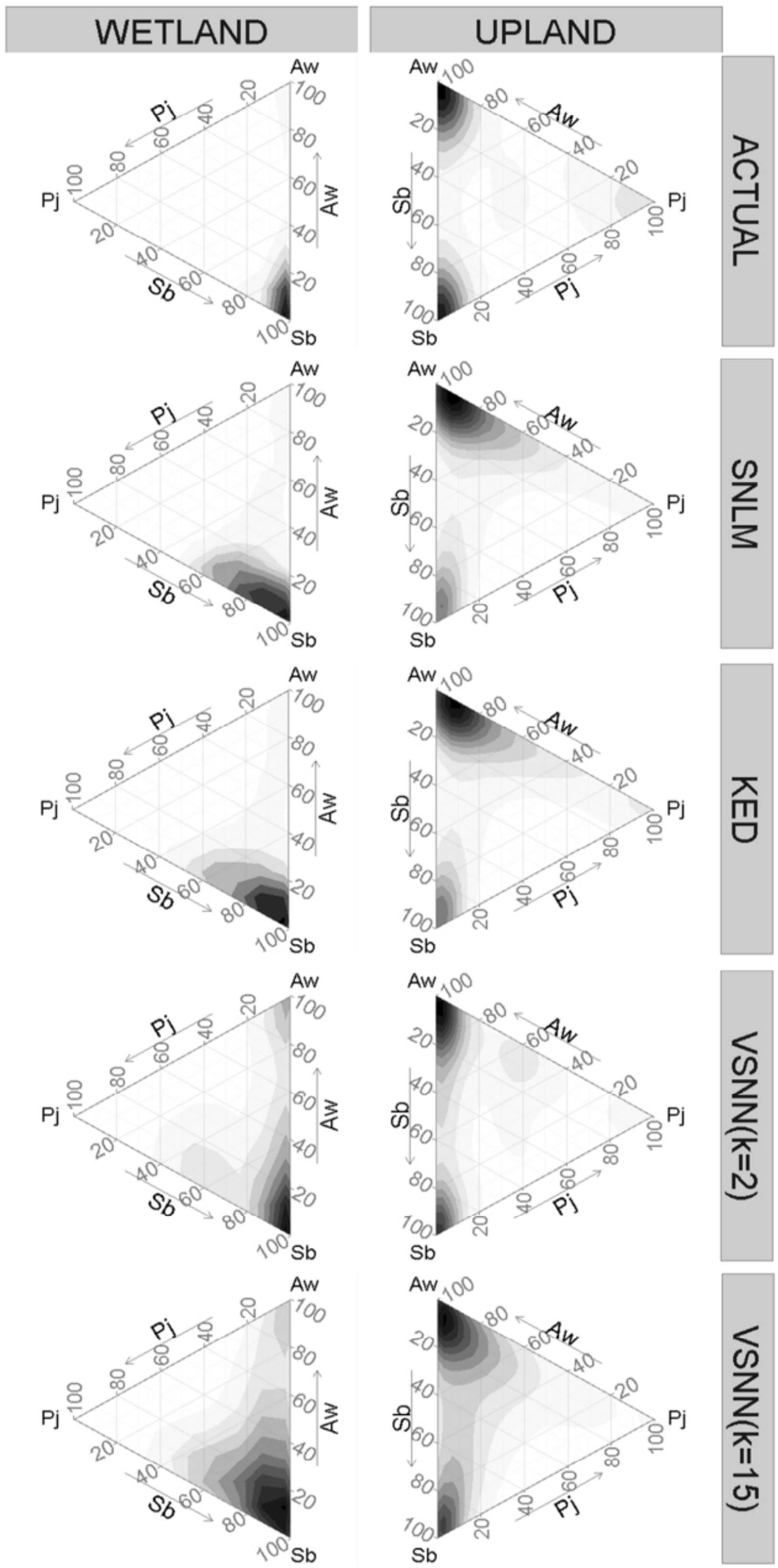

WETLAND
UPLAND
ACTUAL
SNLM
KED
VSNN(k=2)
VSNN(k=15)
Aw
Pj
Sb
20
40
60
80
100

Fig. 5.

Draft