Comparison of Nearest Neighbor Methods for Estimating Basal Area and Stems per Hectare Using Aerial Auxiliary Variables

Valerie LeMay, and Hailemariam Temesgen

Abstract: Simulations were used to compare variable-space nearest neighbor methods for imputing stems per ha and basal area per ha (ground measured) for complex stands (many species and sizes) of southeastern British Columbia, Canada. Species composition and other characteristics obtained for every stand by interpreting aerial photography were used as auxiliary variables. The simulations included three distance metrics (squared Euclidean distance, most similar neighbor, and absolute distance), three intensities of stands with full information (20%, 50%, and 80%), two sets of aerial variables (mixed versus moderately high correlations with ground variables), and three averaging methods (nearest neighbor, average of three nearest neighbors, and distance weighted average of three nearest neighbors weighted). Increasing the number of stands with full information to 50% from 20% resulted in increased accuracy, with no noticeable improvement with a further increase to 80%. Of the three distance metrics, the most similar neighbor measure gave good results in imputing stems per ha and basal area per ha, particularly when there was a mixture of correlations, high and moderate, between the auxiliary (aerial) variables, and the ground variables. No large gain was noted in using the average of three neighbors rather than a single neighbor. FOR. SCI. 51(2):109–119.

Key Words: nearest neighbor, *k*-NN, MSN, imputation simulations, forest inventory, forest sampling.

COMMONLY, FORESTED LANDS ARE DIVIDED into polygons based on forest type. Information for each polygon often includes variables that are measured on aerial photographs (e.g., height class), and additional variables derived from the aerial variables using yield or other models (e.g., estimated volume per ha). Ground sampling of every polygon for detailed information, such as the amount of coarse woody debris, stand structure, or tree lists (stems per ha by species and diameter) is usually not possible. However, this information would be useful to represent the current inventory, and as model inputs to project future conditions.

Imputation methods have been used to estimate the variables of interest from auxiliary variables and spatial positions. For geostatistical data, where each observation represents a point in continuous space (Cressie 1993, Schabenberger and Pierce 2002), a wide variety of kriging methods (e.g., universal kriging, regression kriging; see Odeh et al. 1995 and Nalder and Wein 1998 for comparisons of methods) and spatial nearest neighbor methods have been used to estimate variables of interest for unsampled points in space from sampled points in space (Cressie 1993, Schabenberger and Pierce 2002). However, forest polygons (stands) represent an aggregate of space, termed irregular lattice data (Cressie 1993, Schabenberger and Pierce 2002). Unlike climate and soils data, forests across stand boundaries are not continuous over space (Tuominen et al. 2003 [1]). Finally, there may be many variables of interest that need to be imputed for each stand. For these types of data, kriging and spatial nearest neighbor (NN) methods are not strictly appropriate (Cressie 1993, Schabenberger and Pierce 2002). Instead, variable-space NN methods (also called feature-space or spectral-space NN methods when remote sensing data are used [e.g., Hardin 1994] or covariate space, [e.g., McRoberts et al. 2002]) have been demonstrated by many authors to estimate the variables of interest (*Y*) for unsampled polygons via similarities in auxiliary (*X*) variable-space.

Variable-space NN methods for forest polygon data involve choosing a substitute for stands without detailed information (target stands) from a pool of stands that have detailed tree and stand data (reference stands), based on stand (or plot) level characteristics (*X* variables) that are available for every polygon. Variable-space NN methods include NN (e.g., Moeur 2000), most similar neighbor (Moeur and Stage 1995), *k*-NN (e.g., Maltamo and Kangas 1998), and tabular imputation methods (Ek et al. 1997, Hassani et al. 2004). The distance metric used to measure similarity, and the number of neighbors selected and averaged to obtain the imputed values, differ among applications.

In comparison to other estimation methods using auxiliary variables (e.g., regression or double sampling methods), for variable-space NN methods:

Valerie LeMay, Department of Forest Resources Management, University of British Columbia, 2045-2424 Main Mall, Vancouver, BC V6T 1Z4, Canada—Phone: (604) 822-4770; Fax: (604) 822-9106; Valerie.LeMay@ubc.ca. Hailemariam Temesgen, Department of Forest Resources, 237 Peavy Hall, Oregon State University, Corvallis, OR 97331-5703—Phone: (541) 737-8549; Fax: (541) 737-3039; hailemariam.tenesgen@orgegonstate.edu.

Acknowledgments: We gratefully acknowledge the cooperation and support provided by Ralph Winter, Barry Snowdon, Barry Phillips, and Abdel-Azim Zumrawi, all of the British Columbia Ministry of Forests, and Jon Vivian of BC Ministry of Sustainable Resources Management. This project was partially funded by the National Science and Engineering Research Council of Canada, the BC Ministry of Forests, and Forest Renewal BC.

Manuscript received September 3, 2002, accepted October 12, 2004

1. The averages calculated using the imputed and measured *Y* variables will not be unbiased estimates of the population mean (see example in Appendix 1).
2. As sample size increases (more stands with full information), there is no guarantee that the mean of the *Y*s (measured and estimated values) will approach the true mean (not consistent; see simple example given in Appendix 1). Better matches would be found as the sample size increases, because the *X* values for the selected reference stand would be closer (or the same distance) to their target stand-measured *X* values. However, there is no guarantee that the estimated *Y* values would be closer to their estimated values.
3. If more than one neighbor is averaged and used to estimate the missing *Y* values, some estimates may not be within the realm of real values (e.g., averaging one reference stand with species 1 and 2, with another of species 2 and 3, results in an average with three species, a combination that may not exist; Temesgen et al. 2004). Also, the variance in the estimates declines as more neighbors are averaged.
4. There is no guarantee that increasing the number of *X* variables will improve the results (McRoberts et al. 2002).
5. If there are several *Y* variables or "rare" polygons (stands), a good match will be very difficult to find (McRoberts et al. 2002).

Despite the disadvantages of NN methods, they are very attractive, because they:

1. can retain attribute variance structures of the data (Moeur and Stage 1995, Ek et al. 1997);
2. will result in estimates that are within the bounds of biological reality (Moeur and Stage 1995, Haara et al. 1997), and the logical relationships among *Y* variables will be maintained. Because a match is found from the sampled polygons, the estimate will, necessarily, exist in the population, if a single neighbor is used in the imputation;
3. are distribution-free (nonparametric [2]), in that there is no assumption of distributional characteristics for the auxiliary variables nor for the variables of interest (Haara et al. 1997; Katila and Tomppo 2002); and
4. are multivariate, in that many variables of interest can be estimated at once for each polygon (Katila and Tomppo 2002).

Choices in using variable-space NN methods include: (1) What should be used as the distance metric in locating similar polygons? (2) How does the strength of the relationships between the auxiliary variables and the detailed information affect the choice of the distance metric? (3) What proportion of stands with full information is needed to obtain a good result? (4) Does averaging more than one polygon result in a better substitute? and (5) What fit statistics give useful measures of whether the substitute is suitable, because unbiasedness and consistency of estimates is not assured? Research concerning these choices has been presented in the literature, including previously referenced authors. However, much of this research has been conducted in stands with few species, and one or two vertical strata. Also, little was found on the sample size needed to obtain good matches.

In this article, simulations were used to examine these issues in the case of imputing stems per ha and basal area per ha (ground measured) to every polygon for complex stands (many species and sizes) in southeastern British Columbia (BC), Canada. Auxiliary variables were species composition and other characteristics obtained by interpreting aerial photography. The simulations included three distance metrics, three averaging methods, two sets of aerial variables, and three intensities of stands with full information (ground and aerial data; 20%, 50%, and 80%).

Simulations

Data Description

Ground and aerial data for 96 complex stands collected in 1996 for southeastern BC were used for this study. The stands included several tree species: Douglas-fir (*Pseudotsuga menziesii* (Mirb.) Franco), lodgepole pine (*Pinus contorta* Dougl. ex Loud.), western white pine (*Pinus monticola* Dougl. ex D.Don), Ponderosa pine (*Pinus ponderosa* Dougl. ex Laws.), whitebark pine (*Pinus albicaulis* Englem.), western larch (*Larix occidentalis* Nutt.), trembling aspen (*Populus tremuloides* Michx.), subalpine fir (*Abies lasiocarpa* (Hook.) Nutt.), spruce (*Picea glauca* (Moench) Voss and *P. engelmannii* Parry ex Engelm. and hybrids), black cottonwood (*Populus trichocarpa* Torr. & Gray), western hemlock (*Tsuga heterophylla* (Raf.) Sarg.), white birch (*Betula paperifera* Marsh.), and western redcedar (*Thuja plicata* Donn. ex D. Don). The aerial variables available from inventory databases for each stand were percent crown closure class, percent composition by species, height class, age class, and site index. Height classes 2, 3, and 4 were assigned heights of 15, 25, and 35 meters, respectively, and age classes 4, 5, 6, 7, and 8 were assigned ages of 70, 90, 110, 130 and 180 years, respectively, based on the class midpoints. Using the aerial information, a stand level growth and yield model, Variable Density Yield Projection model (VDYP) (BC Ministry of Forests 1995), was used to obtain estimates of volume per ha, average height, and quadratic mean diameter for each stand. For the ground data, four variable radius plots were randomly located in each stand, and the species, dbh (1.3 m aboveground), and status (i.e., live or dead) for all trees with a dbh of 12.5 cm or greater were recorded, along with other tree variables (BC Ministry of Forests 1998). The live trees for all ground data were compiled, and, for each stand, the average volume per ha, stems per ha, and basal area per ha were obtained for all species combined, and by species. The variable ranges in the data set were quite wide, including predicted volumes from 84.3 to 938.0 m^3/ha (Table 1). The 96 sampled stands were randomly divided into reference and target stands for simulations.

Table 1. Descriptive statistics for ground and aerial variables for the 96 polygons (FD = Douglas fir, PL = pine, Predicted volume/ha = predicted volume per ha from a stand level model)

	Mean	Standard Deviation	Minimum	Maximum	Correlation With Stems/ha	Correlation With Basal area/ha
Ground						
Stems/ha	807.7	358.6	220.1	2413.6		
Basal area/ha	37.2	14.4	15.0	87.5		
Aerial						
Age Class	116.6	43.5	70.0	180.0	−0.2643	0.2362
Crown closure Class	48.4	15.1	20.0	80.0	0.4242	0.3172
Height Class	23.3	6.9	15.0	35.0	−0.1843	0.3790
Percent FD	21.5	28.3	0.0	100.0	−0.1666	−0.0198
Percent PL	25.9	33.3	0.0	100.0	0.2816	−0.1854
Predicted Volume/ha	373.4	184.0	84.3	938.0	−0.0464	0.5336

Range of Simulations

Distance Metrics

For all distance metrics, standardized values are often used to remove the effects of scale of the X variables. For imputation using continuous variables, commonly used distance metrics include:

1. Squared Euclidean distance for standardized X variables, calculated as

$$d_{ij}^2 = (X_i - X_j)'(X_i - X_j),$$

where, X_i is a vector of standardized values of the stand level variables for the ith target polygon, and X_j is a vector of standardized values of the aerial variable for the jth reference polygon. The square root of this distance metric was used by Korhonen and Kangas (1997).

2. Most similar neighbor (Moeur and Stage 1995),

$$d_{ij}^2 = (X_i - X_j)'\Gamma\Lambda^2\Gamma'(X_i - X_j),$$

based on a canonical correlation of the X and Y variable sets, where Γ is a matrix of standardized canonical coefficients for the X variables, and Λ^2 is a diagonal matrix of squared canonical correlations.

3. Absolute distance or weighted absolute distance (Maltamo and Kangas 1998),

$$d_{ij} = \sum_{l=1}^{p} c_l |x_{il} - x_{jl}|,$$

where l is one of the p stand characteristics, and c_l is the weight for stand variable l.

For imputations using categorical variables, or a mixture of categorical and continuous variables, other measures could be used, including similarity (or distance measures) that use the number of matches for class data, such as the City Block distance (Dillon and Goldstein 1984) and the generalized distance for discrete variables (Kurczynski 1970).

The variables selected for use in these simulations were all continuous variables. Therefore, squared Euclidean distance, most similar neighbor, and absolute distance (with equal weights) were used. The squared Euclidean distance (Equation 1) gives equal weight to each of the X variables, but larger distances are given more emphasis because these are squared. The absolute value of the distance (Equation 3) reduces the emphasis of larger differences. The most similar neighbor distance metric (Equation 2) uses the strength of the relationship between the X and Y variable sets to provide weights; stronger correlations result in higher weights for a particular X variable. Equation 2 is intuitively appealing, because higher weights on X variables that have high correlations with the Y variables might be expected to result in better matches for the variables of interest, the Y set. However, because the selection of neighbors is multivariate in nature, this might result in a particularly good match for one Y variable, and not for other Y variables. Equal weighting of the X variables might result in a better match for overall Y variables, combined, and was used with Equation 3.

Single or Weighting of Many Reference Polygons

Once NNs are found for the target stand, the imputation of detailed information from the Y set of variables has been based on:

1. using the information for the NN as the substitute (e.g., Moeur and Stage 1995);
2. using an average of the Y variables over the kth nearest neighbors (KNN; e.g., Korhonen and Kangas 1997, Maltamo and Kangas 1998); or
3. using a weighted average of the kth nearest neighbors, often based on distance from (or similarity to) the target stand (WKNN; e.g., Maltamo and Kangas 1998).

The choice of how many neighbors to use and what weight to use in calculating the average values is not clear, and is sometimes chosen to meet an objective criterion (e.g., small root mean squared error used by McRoberts et al. 2002 for pixel classification). Tuominen et al. 2003 noted that "The higher the value of k, the more averaging that occurs in the estimates. Thus, the optimal value of k is a trade-off between the accuracy of estimates and the variation retained in the estimates." McRoberts et al. 2002 also noted that, as k increases, the biases (average difference between observed and predicted values) rise for extreme

values of the variables of interest. Using one neighbor would likely provide the best results if there were a high proportion of stands with full information, because these reference stands would represent the population well. Conversely, if there were a low proportion of stands with full information, using more than one neighbor might give better results because averaging would provide a wider variety of *Y* variables. Using too many neighbors might result in less variability in the imputed values than was present in the population because of averaging of values. Also, as noted earlier, values that do not exist in the population may also result by means of averaging.

For each simulation, imputed values were repeated using the first nearest neighbor (NN), average of three nearest neighbors (KNN), and weighted average of three nearest neighbors (WKNN). The *k* value of three was selected because the stands are very variable in species and sizes. Using a larger value of *k* would most likely result in averages that were not biologically possible.

Choice of Variables

The choice of variables used as the *X* set depends on what information is available for every plot or stand, and how these variables relate to the *Y* set of variables. Examples of variables that have been used include:

1. basal area of growing stock, location, altitude, site class, soil type, dominant tree species, mean diameter, and age of growing stock to impute sample tree information (Korhonen and Kangas 1997);
2. aerial variables such as crown closure class, height class, age class, species composition by crown closure class, and site index to impute tree-lists (LeMay and Temesgen 2001);
3. stand records, maps, and aerial photographs to impute volume variables (Moeur and Stage 1995);
4. forest cover type, based on relative stocking of tree species to impute postharvest regeneration (Ek et al. 1997);
5. basal area per ha and basal area median diameter to impute basal-area diameter distribution; and
6. stand age, basal area, basal area median diameter, and stand height to impute diameter distributions (Maltamo and Kangas 1998).

If the *X* and *Y* variables are very well correlated, then a good match on the *X* variables (as measured by the distance metrics) would result in a good match on the *Y* variables. More often, there is a mixture of high and moderate or low correlations between the two sets of variables. The results will be dependent on the strength of the correlations, but may also be confounded by the choice of distance metric, and the proportion of stands with full information.

For the simulations, the variables of interest (*Y*) were basal area per ha and stems per ha. Two sets of *X* variables were used. For the first variable set, the *X* variables were predicted volume per ha (m^3/ha), crown closure class (%), percent Douglas-fir (FD) by crown closure, and percent

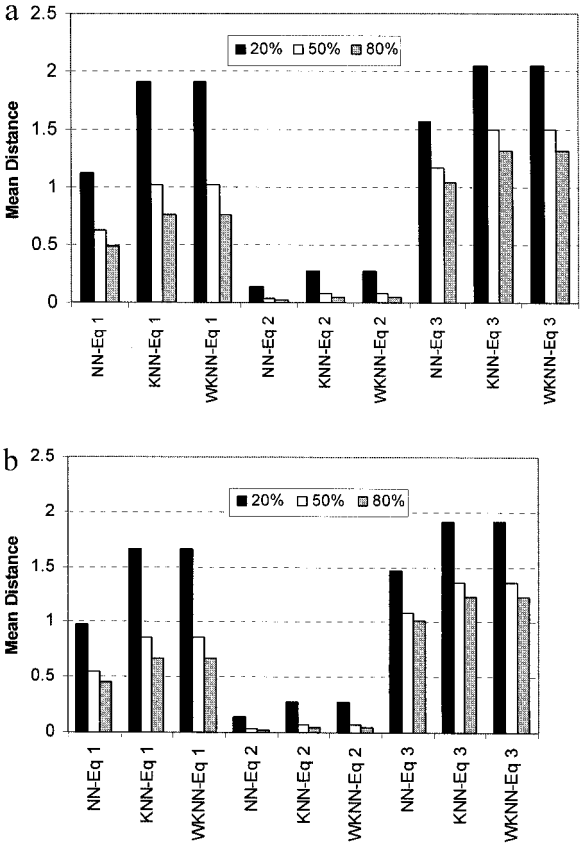

Figure 1. Mean distance over the 1,000 sampling replications using (a) variable set 1 (mixed correlations) and (b) variable set 2 (moderately high correlations).

lodgepole pine (PL) by crown closure, representing a mixture of high and low correlations with the *Y* set (Table 1). Correlations between the aerial and ground variables were low to moderately high, with the highest correlation (0.5336) for ground basal area per ha with predicted volume per ha. Correlations between the ground variables and percent FD and percent PL were the lowest. The second variable set included variables that have moderately high correlation with the *Y* set, and were predicted volume per ha, height class (m), crown closure class (%), and age class (years).

Proportion of Stands with Full Information

The stands (or plots) with information on all variables form the reference set. If the reference set were based on a simple random selection of stands (or plots) from the population, a larger proportion should result in better imputation results, as least in matching the *X* variables, because the reference set would better represent the variability in the population. If the reference set represents the variability present in the population for the *Y* set of variables, and the *X* and *Y* sets of variables are well correlated, then the imputation should work well. The proportion (or number of observations) needed to obtain a "good" representation of the population will increase with increasing variability of

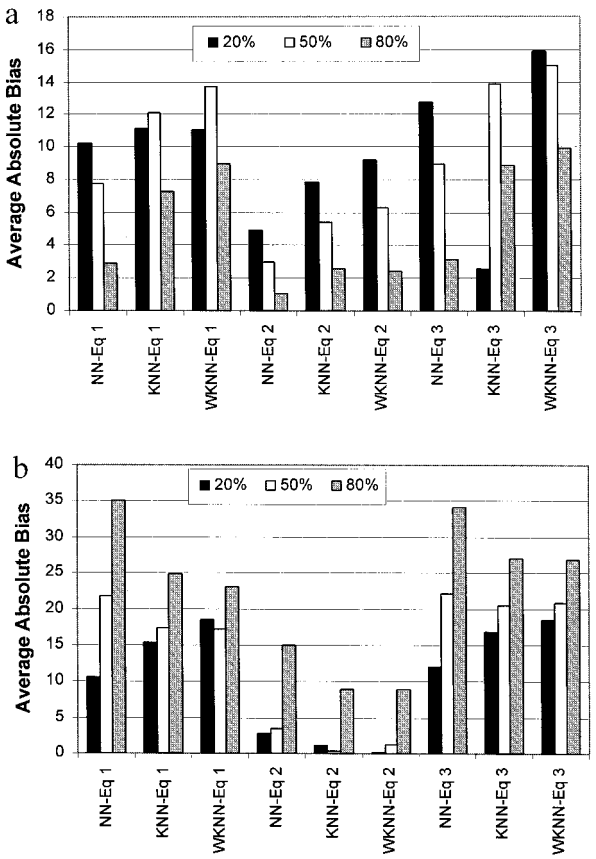

Figure 2. Absolute value for mean of biases over 1,000 sampling replications for (a) all four *X* variables combined using variable set 1 (mixed correlations), and (b) both *Y* variables using variable set 1.

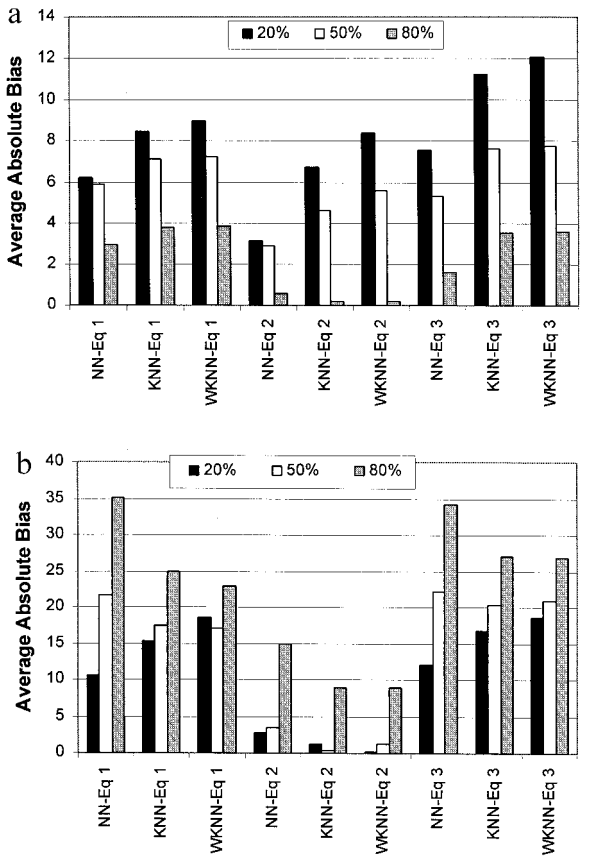

Figure 3. Absolute value for mean of biases over 1,000 sampling replications for (a) all four *X* variables using variable set 2 (moderately high correlations), and (b) both *Y* variables using variable set 2.

the *Y* variables in the population, and decrease with increasing correlations between *X* and *Y* variables. Also, a greater proportion would be needed as the number of *Y* variables increases, because it will be difficult to find a match that is similar for all *Y* variables.

For this simulation, three sampling intensities were selected: 20%, 50%, and 80%. Moeur (2000) indicated that 20% sampling intensity is likely sufficient for estimating stand level variables. LeMay and Temesgen (2001) used simulations to compare the use of different proportions for imputing tree lists from aerial variables. They noted that there was an improvement in results when the proportion of stands with full information was increased from 20% to 50%, but little gain in extending to 80%.

Measures of Accuracy

For each of the 54 combinations (two variable sets, three distance metrics, three sample sizes, and three weightings), the random separation of the data into target versus reference stands was repeated 1,000 times. Fit statistics commonly used by other authors are based on comparing observed with estimated values in the simulated target data set, and in particular, the average difference (often called bias) and root mean squared error (square root of the average squared difference; RMSE) are often calculated.

As the example in the introduction illustrates, the average of the measured and estimated *Y* values will not be an unbiased estimator of the population average, even if there is only one variable of interest in the variable-space NN imputation. For more than one variable of interest, a small average difference in one variable could be compensated by small average difference in another variable. Also, large negative and positive differences would result in an average difference of zero. The RMSE gives a better indication of the imputation results, because differences are squared before averaging. Moeur and Stage (1995) suggested that the distance metrics could be used to assess the adequacy of results. If distance metrics were high for some stands, then no suitable match was found in the reference set.

To evaluate the results for each simulation, bias (average difference) and RMSE were calculated for each replicate, as follows:

1. Bias for each variable in the *X* and *Y* sets of the target data, as shown for the *l*th *X* variable:

$$\text{bias} = \sum_{i=1}^{n} (x_{il} - x_{jl})/n$$

2. RMSE for each variable in the *X* and *Y* sets, as shown for the *l*th *X* variable,

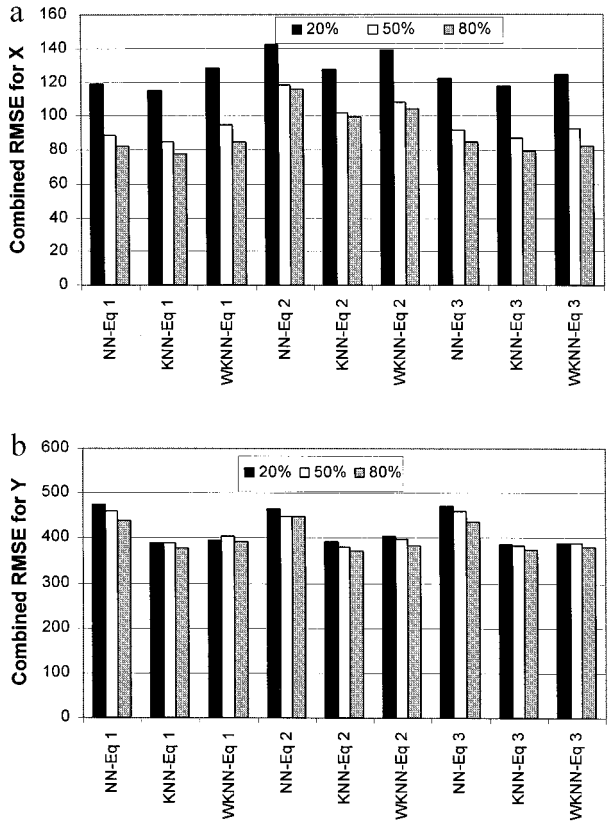

Figure 4. Mean RMSE over 1,000 sampling replications for (a) all four *X* variables combined using variable set 1 (mixed correlations), and (b) both *Y* variables using variable set 1.

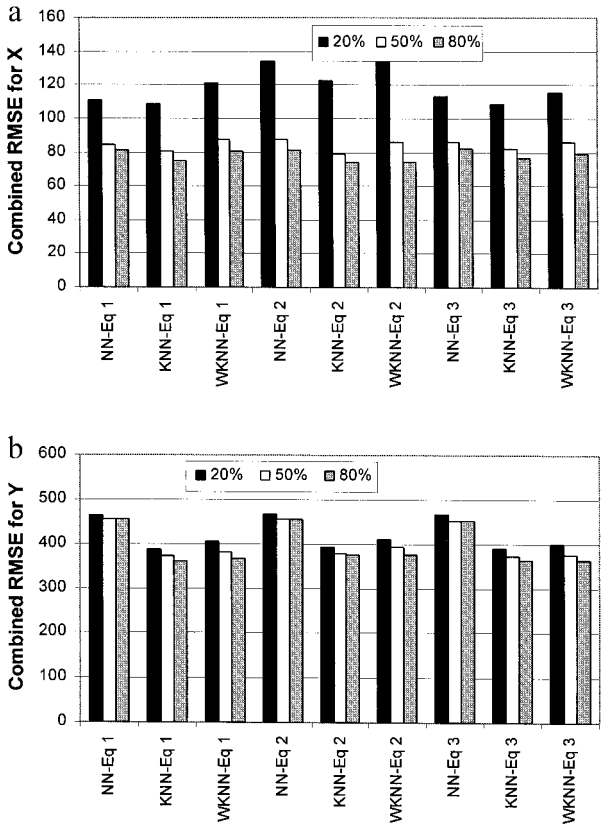

Figure 5. Mean RMSE over 1,000 sampling replications for (a) all four *X* variables using variable set 2 (moderately high correlations), and (b) both *Y* variables using variable set 2.

$$\text{RMSE}_l = \sqrt{\sum_{i}^{n} (x_{il} - x_{jl})^2 / n},$$

where n is the number of stands with missing ground data (target stands) for the replicate. Because a large value for one variable might be compensated by a small value for another variable, these two statistics were also obtained by replicate for all Y variables combined and for all X variables combined, as shown for the Y variables,

3. Bias for all Y variables combined:

$$\text{bias} = \left[\sum_{i=1}^{n} \sum_{j=1}^{\text{all } Y} (y_{il} - y_{jl}) / n \right]$$

4. RMSE for all Y variables combined,

$$\text{RMSE} = \sqrt{\left[\sum_{i=1}^{n} \sum_{j=1}^{\text{all } Y} (y_{il} - y_{jl}) \right]^2 / n}.$$

The mean, minimum, maximum, and range of each of these two statistics were summarized over the 1,000 sampling replications. In addition, the mean distance was also calculated for each simulation and then averaged over the 1,000 sampling replications. Although the motivation for imputation is to obtain estimates for the Y set of variables, statistics for the X variables were also calculated using only the imputed (target) polygons. Because the X set is used to find a polygon match, these fit statistics could give insights into what polygons are not well matched in terms of the variables that are known for every polygon. This information could be used to inform the sampling designer on possible changes to the sampling design (e.g., stratification) to improve the subsequent imputation results for a given sampling cost.

Results and Discussion

Mean Distances

Distances cannot be compared across the three distance equations. However, comparisons across the three methods (NN, KNN, and WKNN), variable sets, and proportions of stands with full information were made for each distance equation.

As expected, the mean distances (mean of the average distances over all 1,000 replicates) decreased with increasing sample size, because the availability of more polygons with the larger reference set must result in the same or lower distances (Figure 1). Distances were necessarily lowest for NN that uses the nearest neighbor, and were the same for KNN and WKNN because the same three neighbors were selected. Distances were lower for variable set 2 versus set 1, when Equations 1 and 3 were used. Likely, this occurred

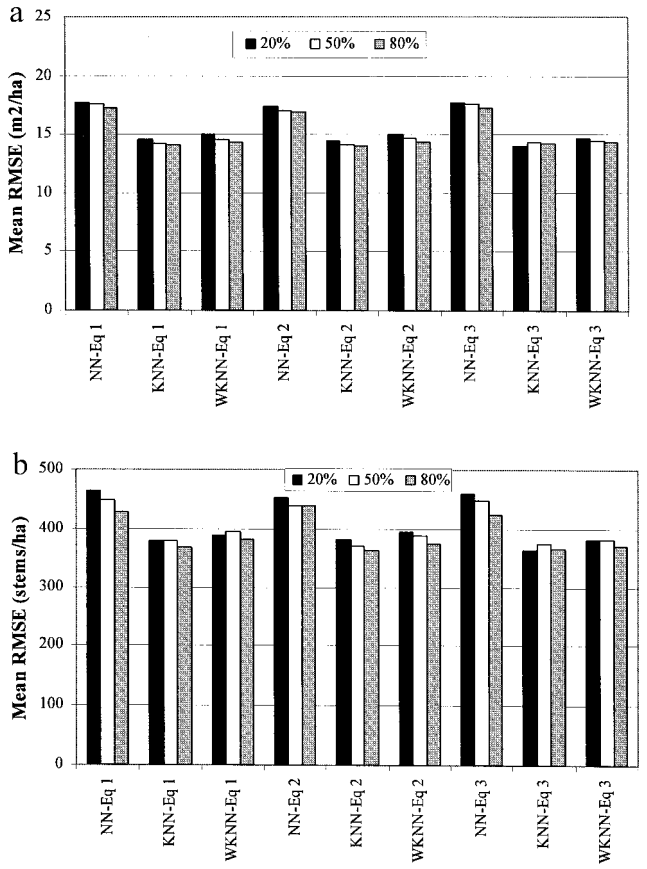

Figure 6. Mean RMSE over 1,000 sampling replications for *Y* variables using variable set 1 (mixed correlations): (a) basal area per ha, and (b) stems per ha.

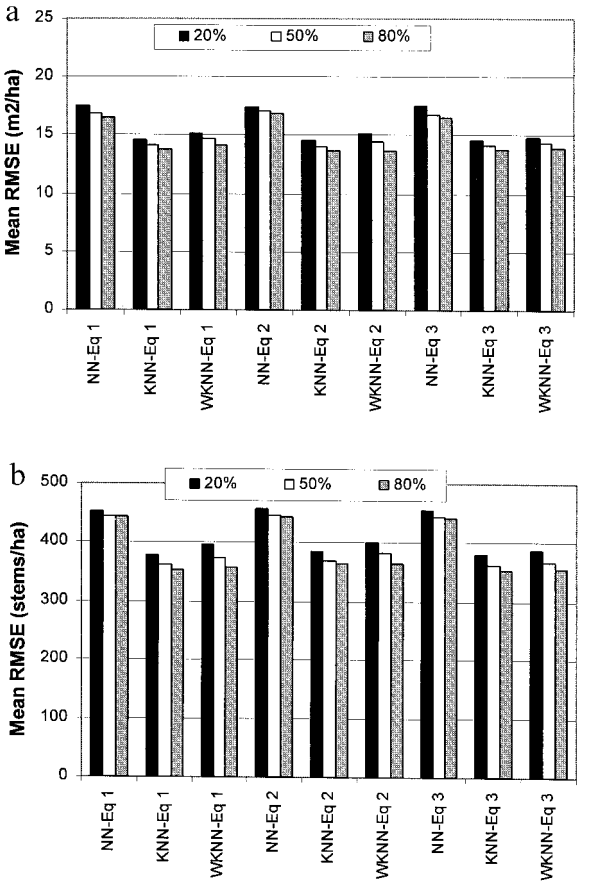

Figure 7. Mean RMSE over 1,000 sampling replications for *Y* variables using variable set 2 (moderately high correlations): (a) basal area per ha, and (b) stems per ha.

because the percent FD and percent PL used in variable set 1 (mixed correlations) are quite variable, whereas height class and age class used in variable set 2 (moderately high correlations) are less variable. This resulted in better matches and lower squared Euclidean distances using variable set 2. Conversely, the distances using variable set 1 were very similar to those for variable set 2, using Equation 2. Since the percent FD and percent PL had lower correlations, these would have been given less weight in Equation 2 distance calculation. As a result, there were no real differences in distances between variable sets 1 and 2 when Equation 2 was used.

Distances decreased with the increasing proportion of stands with full information, with the greatest gain from 20% to 50%, and less gain from 50% to 80%. This result is similar to that noted by LeMay and Temesgen (2001) for imputing tree lists. Conversely, Moeur (2000) suggested that 20% might be sufficient. However, the stands used in this current study are more variable in structure, particularly numbers of species, than those used by Moeur. As a result, 50% of the stands with full information would be preferred for imputation in complex stands of southeastern BC.

Average Differences

For variable set 1, the bias (average difference) for the combined *X* variables averaged over the 1,000 sampling replications (mean bias) was close to zero for the 80% reference sets using Equation 2, all methods, and using Equations 1 and 3 for NN only (Figure 2a [3]). For the 50% and 80% reference sets, Equation 2 gave lower (absolute value) mean biases using all three methods, indicating an improvement resulting from weighting the *X* variables using correlations with the *Y* variables. For variable set 2, the mean bias was close to zero for all methods and equations using the 80% reference sets, indicating that it was easier to find good matches for these four variables versus those in variable set 1 (Figure 3a). For both variable sets, there was no noticeable reduction in the absolute value of mean bias in using the KNN or WKNN methods over using a single neighbor (NN), nor in using the 50% over the 20% reference sets (Figures 2a and 3a).

For the variables of interest, the *Y* set, the mean biases were not close to zero (Figures 2b and 3b). In their simulations, Moeur and Stage (1995) showed percent biases of −4.0 to 0.6% using Equation 2 and NN to impute volume variables from land classification variables. For imputing diameter distributions, Maltamo and Kangas (1998) also obtained nonzero biases. Mean biases in the *Y* variables tended to increase with an increase in the proportion represented in the reference sets, using either set of *X* variables (Figures 2b and 3b). For variable set 1, mean biases were

nearly zero for the 20% reference sets, for all methods and equations (Figure 2b). Also, mean biases tended to be lower using Equation 2, as might be expected, because the canonical correlations between the *X* and *Y* variables are used to weight the distance metric. Using Equation 2, some reduction in mean bias was obtained through using the KNN and WKNN approaches, over the NN approach. For variable set 2, mean biases were again lower using Equation 2, and some reduction in mean bias occurred using the KNN and WKNN approaches (Figure 3b).

Mean, Minimum, and Maximum RMSEs

The RMSE for the *X* variables averaged over all 1,000 sampling replications (mean RMSE) consistently decreased or remained the same with increasing proportions of data with full information (Figures 4a and 5a). There was a large reduction in mean RMSE from 20 to 50% reference sets, but very little reduction in using the 80% reference sets, corresponding to the decreases in mean distances.

For the *Y* variables, generally there were some reductions in mean RMSE values, although this varied with the method, equation, and variable set (Figures 4b and 5b). Generally, lower mean RMSE values were obtained using KNN and WKNN over NN for both variable sets, with no large difference between KNN and WKNN. For variable set 1 with mixed correlations, Equation 2 was slightly better, particularly for KNN and WKNN. For variable set 2 with moderately high correlations between *X* and *Y* variables, Equations 1 and 3 performed slightly better than Equation 2. The weighting using correlations (Equation 2) only provided better results when the correlations varied more greatly among variables. Because the stems per ha values are very large compared to the basal area per ha values, the mean RMSE combined for the *Y* variables, is mostly reflective of the stems per ha variable (Figures 6 and 7). However, a similar pattern of results was obtained for each of the individual *Y* variables.

Generally, the minimum of the 1,000 RMSE values for each *Y* variable decreased with increasing proportions of stands with full information for all simulations, indicating improvements with an increase in the number of reference stands (Tables 2 and 3). However, the maximum of the 1,000 RMSE values decreased for 50% over 20% of stands with full information, but then increased for the 80%. Because the RMSE is calculated using target stands only, a large squared difference between observed and estimated values for one target stand would have more impact on the RMSE for the 80% proportion. For example, given volume per ha values of 300, 300, 350, 350, 400, 400, 375, 375, 800, and 800 m^3/ha, eight stands would be selected as the reference set and two stands as the target stands using the 80% proportion. RMSE values would be small, except when the two stands of 800 m^3/ha were selected as the target stands in the simulation. Using 50%, the large maximum RMSE value obtained with the 80% proportion would not occur, because the RMSE values would be averaged over five stands. In application, this would translate into a situation where the number of target stands is few, and these differ greatly from the reference stands. As noted by Moeur and Stage (1995), the distance metric should indicate this problem. In this case, approaches other than nearest neighbor methods might give better results. Although the minimum RMSE may be useful in comparing methods, the maximum RMSE value must, therefore, be interpreted with caution.

Slightly smaller minimum and maximum RMSE values were consistently obtained using KNN and WKNN over NN. The smallest minimum RMSE values were obtained using variable set 2, WKNN, 80% of stands with full information, for all distance equations. Using the average of three stands to impute stems per ha and basal area per ha reduced the possibility of a poor match when the target stands were quite different from the remainder of the stands.

Table 2. Minimum and maximum RMSE for Y variables over 1,000 sampling replications of each nearest neighbor simulation for variable set 1, mixed correlations

Equation/		NN			KNN			WKNN		
Variable	Prop*	20%	50%	80%	20%	50%	80%	20%	50%	80%
Equation (1)										
Basal area/ha	Min	13.689	13.074	8.056	12.094	9.503	6.966	11.365	9.213	7.417
	Max	24.730	23.920	26.264	19.171	19.465	24.166	19.914	19.441	26.064
Stems/ha	Min	344.290	309.342	200.007	303.253	250.909	188.622	306.450	257.397	189.187
	Max.	726.462	685.784	711.780	486.376	496.113	698.063	511.793	495.226	717.084
Equation (2)										
Basal area/ha	Min	12.815	12.305	7.918	11.790	9.871	4.565	11.767	10.977	6.126
	Max	28.983	22.811	27.506	22.226	18.921	23.824	21.351	19.550	25.635
Stems/ha	Min	332.299	277.142	240.493	303.094	256.095	195.901	317.135	278.540	185.698
	Max.	831.757	631.285	752.968	507.025	484.019	603.543	539.509	509.546	606.263
Equation (3)										
Basal area/ha	Min	13.115	12.950	7.036	11.591	9.769	6.686	11.684	9.531	6.592
	Max	25.282	24.217	26.952	19.243	19.612	25.046	19.564	19.653	26.443
Stems/ha	Min	346.897	304.497	199.524	303.984	257.777	184.214	296.284	256.625	191.830
	Max.	742.898	724.463	728.854	489.377	490.706	706.974	500.438	485.665	720.809

* Proportion of stands with full information, ground and aerial attributes

Table 3. Minimum and maximum RMSE for Y variables over 1,000 sampling replications of each nearest neighbor simulation for variable set 2, moderately high correlations

		NN			KNN			WKNN		
Equation./Variable	Prop*	20%	50%	80%	20%	50%	80%	20%	50%	80%
Equation (1)										
Basal area/ha	Min	13.139	11.808	8.977	12.156	9.303	6.330	12.418	10.588	5.591
	Max	27.769	25.060	25.613	20.338	18.932	24.343	21.071	19.111	26.095
Stems/ha	Min	315.414	265.139	226.804	288.730	251.282	185.283	303.978	258.617	201.966
	Max.	799.791	769.366	802.733	529.476	473.300	644.147	577.189	490.768	637.017
Equation (2)										
Basal area/ha	Min	13.225	10.633	8.236	11.445	9.832	7.171	11.890	10.707	7.171
	Max	27.176	23.893	25.678	19.969	18.521	22.331	21.738	18.803	22.331
Stems/ha	Min	332.666	305.974	193.976	307.615	581.013	307.615	296.278	264.622	173.288
	Max.	889.533	696.139	831.283	43.832	274.088	43.832	598.385	508.844	716.671
Equation (3)										
Basal area/ha	Min	13.359	12.627	9.402	11.823	9.856	6.626	12.003	9.815	6.291
	Max	26.821	25.705	25.031	20.720	19.062	23.302	20.817	19.354	23.900
Stems/ha	Min	306.517	273.018	215.487	298.558	244.483	171.630	305.738	252.097	168.608
	Max.	815.663	758.400	823.389	544.633	480.664	625.804	548.735	484.392	615.435

* Proportion of stands with full information, ground and aerial attributes

Overall

The distance metric does indicate when matches are "poor," as noted by Moeur and Stage (1995). This metric would normally be available in applying NN methods, as would the distributions of actual and imputed values of the *X* variables. When testing alternative methods via simulation, a combination of distances, mean bias, and mean RMSE better indicate the "success" of the imputation.

For very large tracts of forested lands, such as those in Canada and other countries, the sampling proportion might be 20% or less. In this study, the distances indicated improvements in sampling 50% over 20%, but no great improvement was obtained in using the 80% versus the 50%. The data used represent complex stands with up to nine species, and a wide range of tree sizes. For forested areas with less variability, the 20% sampling intensity might be sufficient. For forested areas with even higher variability, such as tropical stands of more than 100 species, a very large sampling intensity would likely be needed.

The mean biases over the target stands were lower for the *X* variables when the 80% was used, but this did not translate into lower mean biases for the *Y* variables. The use of a more correlated set of *X* variables, represented by variable set 2, did not show the anticipated decrease in the mean bias for the *Y* variables. However, the use of Equation 2, incorporating the correlations between *X* and *Y* variables, did result in lower mean biases for the *Y* variables for both variable sets. The mean RMSE values reflected the mean distances, in that no real gain was obtained in using 80% over 50% of stands with full information. Generally, the KNN and WKNN methods showed some improvement over the NN method, but no real gain in accuracy was noted in using the WKNN over the KNN method. MacLeod et al. (1987) indicated that a weighted distance measure was better than an unweighted measure for imputing class data, given a carefully selected weight function. Korhonen and Kangas (1997) used a search to select weights to obtain the lowest MSE. Although this study indicated no improvement in using WKNN over KNN, a weight other than the inverse of the distance might have yielded different results, based on results from other studies.

Conclusions

Aerial data are commonly available for forested lands. These data can be used as auxiliary variables to impute variables of interest obtained from ground information, collected on a proportion of stands. For complex stands of southeastern BC, increasing the number of stands with full information to 50% from 20% resulted in increased accuracy, not noted in a further increase to 80%. For less complex stands, with one or two species, a smaller sample size would likely be sufficient to obtain good matches, and impute a few *Y* variables. Of the three equations tested, the most similar neighbor distance metric gave good results for estimating the variables of interest, stems per ha, and basal area per ha, particularly when there was a mixture of correlations, high and moderate, between the auxiliary variables, the aerial variables, and the variables of interest. A small decrease in mean RMSE was noted in using the average of three neighbors rather than a single neighbor. Greater advantage in averaging neighbors over using a single neighbor would likely occur if more than two *Y* variables were of interest.

Endnotes

[1] Tuominen et al.2003 stated that "geostatistical interpolation is likely to be futile when it is extended beyond the stand border."

[2] Note that nonparametric methods have been used in literature to indicate distribution-free methods or methods based on distributions other than the normal distribution.

[3] Because mean biases can be positive or negative, the absolute values of mean biases are shown in Figures 2 and 3.

Literature Cited

BC MINISTRY OF FORESTS. 1995. VDYP interactive application user guide. Version 6.3. Victoria, BC, Canada. 39 pp.

BC MINISTRY OF FORESTS. 1998. Inventory audit sampling standards and procedures. Victoria, BC, Canada. 40 pp.

CRESSIE, N.A.C. 1993. Statistics for spatial data. Revised edition. John Wiley & Sons, Inc., Toronto. 887 pp.

DILLON, W.R., AND M. GOLDSTEIN. 1984. Multivariate analysis. John Wiley & Sons, Ltd., Toronto. 587 pp.

EK. A.R., A.P. ROBINSON, P.J. RADTKE, AND D.K. WALTERS. 1997. Development and testing of regeneration imputation models for forests in Minnesota. For. Ecol. Manage. 94: 129–140.

HAARA, A., M. MALTAMO, AND T. TOKOLA. 1997. The *k*-nearest neighbor method for estimating basal area diameter distribution. Scand. J. For. Res. 12:200–208.

HARDIN, P.J. 1994. Parametric and nearest neighbor methods for hybrid classification: A comparison of pixel assignment accuracy. Photogram. Eng. Remote Sensing 60(12):1439–1448.

HASSANI, B.T., V. LEMAY, P.L. MARSHALL, H. TEMESGEN, AND A.-A. ZUMRAWI. 2004. Regeneration imputation models for complex stands of southeastern British Columbia. For. Chron. 80(2):271–278.

KATILA, M, AND E. TOMPPO. 2002. Stratification by ancillary data in multisource forest inventories employing *k*-nearest neighbor estimation. Can. J. For. Res. 32:1548–1561.

KORHONEN, K.T., AND A. KANGAS. 1997. Application of nearest-neighbor regression for generalizing sample tree information. Scand. J. For. Res. 12:97–101.

KURCZYNSKI, T.W. 1970. Generalized distance and discrete variables. Biometrics 26:525–534.

LEMAY, V., AND H. TEMESGEN. 2001. Nearest neighbor methods for generating tree-lists from aerial data. Invited paper presented at the IUFRO 4.11 Forestry biometry, modelling, and information science conference, The University of Greenwich, School of Computing and Mathematical Sciences, June 26–29, 2001, Greenwich, England. 23 pp.

MACLEOD, J.E.S, A. LUK, AND D.M. TITTERINGTON. 1987. A re-examination of the distance-weighted *k*-nearest neighbor classification rule. IEEE Trans. Syst, Man, Cybern. 17(4): 689–696.

MALTAMO, M, AND A. KANGAS. 1998. Methods based on *k*-nearest neighbor regression in the prediction of basal area diameter distribution. Can. J. For. Res. 28:1107–1115.

MCROBERTS, R.E., M.D. NELSON, AND D.G. WENDT. 2002. Stratified estimation of forest area using satellite imagery, inventory data, and the *k*-nearest neighbors technique. Remote Sensing Environ. 82:457–468.

MOEUR, M. 2000. Extending stand exam data with most similar neighbor inference. P. 99–107 *in* Proc. of the Society of American Foresters National Convention. Sept. 11–15, 1999, Portland, Oregon.

MOEUR M., AND AR STAGE. 1995. Most similar neighbor: An improved sampling inference procedure for natural resource planning. For. Sci. 41:337–359.

NALDER, I.A., AND R.W. WEIN. 1998. Spatial interpolation of climatic normals: Test of a new method in the Canadian boreal forest. Agric. For. Meteorol. 92(4):211–225.

ODEH, I.O.A., A.B. MCBRATNEY, AND D.J. CHITTLEBOROUGH. 1995. Further results on prediction of soil properties from terrain attributes: Heterotropic cokriging and regression-kriging. Geoderma. 67:215–226.

SCHABENBERGER, O., AND F.J. PIERCE. 2002. Contemporary statistical models for the plant and soil sciences. CRC Press, New York. 738 pp.

TEMESGEN, H., V.M. LEMAY, K.L. FROESE, AND P.L. MARSHALL. 2004. Inputing tree-lists from aerial attributes for complex stands of south-eastern British Columbia. For. Ecol. Manage. 177:277–285.

TUOMINEN, S., S. FISH, AND S. POSO. 2003. Combining remote sensing, data from earlier inventories, and geostatistical interpolation in multisource forest inventory. Can. J. For. Res. 33:624–634.

Appendix 1. Example to Illustrate Bias and Inconsistency of Variable Space NN Methods

The averages calculated using the imputed and measured Y variables from variable space NN will not be unbiased estimates of the population mean. For example, if there are three polygons with X and Y values of (1, 10), (3, 30), and (4, 50) and $\mu_Y = 30$, two of these polygons are randomly sampled without replacement, and the third is imputed, there are three possible outcomes:

1. Polygons 1 and 2 are sampled and used to impute Polygon 3. Polygon 2 is closest, based on the X variable. The outcome is then Y values (measured and estimated) of $y_1 = 10$, $y_2 = 30$, and $\hat{y}_3 = 30$, with $\bar{y} = 23.3$.
2. Polygons 1 and 3 are sampled and used to impute Polygon 2. Polygon 3 is closest with the outcome of $y_1 = 10$, $\hat{y}_2 = 50$, $y_3 = 50$, with $\bar{y} = 36.7$.
3. If Polygons 2 and 3 are sampled and used to impute Polygon 1, Polygon 2 is closest. The outcome is $\hat{y}_1 = 30$, $y_2 = 30$, and $y_3 = 30$, with $\bar{y} = 36.7$.

The average of these three possible outcomes is 32.3, not equal to the true mean of 30.

Variable space NN methods will not necessarily provide consistent estimates. Using the same data and using a sample size of 1, the following results are obtained:

1. Polygon 1 is sampled and used to impute Polygons 2 and 3. The outcome is then Y values (measured and estimated) of $y_1 = 10$, $\hat{y}_2 = 10$, and $\hat{y}_3 = 10$, with $\bar{y} = 10.0$.
2. Polygon 2 is sampled and used to impute Polygons 1 and 3. The result is $\hat{y}_1 = 30$, $y_2 = 30$, and $\hat{y}_3 = 30$, with $\bar{y} = 30.0$.
3. If Polygon 3 is sampled and used to impute Polygons 1 and 2, the outcome is $\hat{y}_1 = 50$, $\hat{y}_2 = 50$, and $y_3 = 50$, with $\bar{y} = 50.0$.

The average of these three outcomes (expected value) is 30.0, equal to the true mean. The outcome, in this case, was closer with the smaller sample size. Larger populations would produce similar outcomes.