Comparison of stratified and non-stratified most similar neighbour imputation for estimating stand tables

BIANCA N. I. ESKELSON[1], HAILEMARIAM TEMESGEN[1]* AND TARA M. BARRETT[2]

[1]Department of Forest Resources, Oregon State University, 280 Peavy Hall, Corvallis, OR 97331-5703, USA
[2]USDA Forest Service, Pacific Northwest Research Station, Forestry Sciences Lab, 3301 "C" Street, Suite 200, Anchorage, AK 99503-3954, USA
*Corresponding author. E-mail: hailemariam.temesgen@oregonstate.edu

Summary

Many growth and yield simulators require a stand table or tree-list to set the initial condition for projections in time. Most similar neighbour (MSN) approaches can be used for estimating stand tables from information commonly available on forest cover maps (e.g. height, volume, per cent canopy cover and species composition). Simulations were used to compare MSN (using an entire database) with two stratified MSN approaches. The first stratified MSN approach used species composition to partition the population into two inventory type strata, while the second stratified MSN approach used average stand age to partition the data into two stand development stages (strata). The MSN approach was used within the whole population and within each stratum to select a reference stand and to impute the ground variables of the reference stand to each target stand. Observed *vs* estimated stand tables were then compared for the stratified and non-stratified simulations. The imputation within a stratum did not result in better estimates than using the MSN approach within the whole population. Possible reasons for poor performance of stratified MSN are provided.

Introduction

Increasingly, resource management plans must consider stand-, landscape- and forest-level attributes to mimic the complex interactions at different levels of ecosystem management, and thus, there is a need to link and integrate these attributes across scales. To realize this linkage, analysts require detailed data about every stand (land parcel) within a management area (Temesgen *et al.*, 2007).

Often auxiliary information (e.g. derived from air photo interpretation) is available for every stand, supplemented by detailed ground information for selected sample stands. Auxiliary attributes commonly include the following: species composition (per cent by crown closure), crown closure (per cent), height class, age class, site class/site index and elevation from forest cover and elevation maps. For sampled stands, additional information is available, including either a tree-by-tree record (tree-list) specifying the species, breast

Forestry, doi:10.1093/forestry/cpn003
Advance Access publication date 24 April 2008

height diameter (diameter outside bark measured at 1.3 m above ground; d.b.h.) and possibly tree height of each tree, or a stand table (trees per hectare (TPH) by species and d.b.h. class), based on the compiled ground sample data. Tree-lists or stand tables are often needed to initiate growth models, which may be used to update inventories, to develop landscape and forest management plans and to assess the dynamic development of structure and diversity at the stand level for use in habitat analyses. Since tree-lists are commonly available only for a portion of the land area, approaches that generate tree-lists and other detailed information from auxiliary attributes are beneficial.

Most research on stand table generation uses stand-level information measured on the ground and is done on simple stands with few species and unimodal diameter distributions. The approaches used can be broadly categorized into diameter distribution (frequency by species and diameter) modelling and imputation methods (Temesgen, 2003).

The diameter distribution modelling approach involves fitting a diameter distribution for each stand and then predicting the parameters of the distribution using stand variables. Depending upon how the parameters are predicted, the diameter distribution modelling approaches in the literature have been classified as parameter prediction (e.g. Biging *et al.*, 1994), parameter recovery (e.g. Hyink and Moser, 1983) and percentile prediction (e.g. Maltamo *et al.*, 2000). These approaches are based on unimodal distributions and do not accurately describe the diameter distributions of complex stands that have multimodal and irregular diameter distributions.

Imputation methods have been used to generate stand tables from stand-level variables (e.g. Moeur and Stage, 1995, Maltamo and Kangas, 1998, Temesgen *et al.*, 2003, LeMay and Temesgen, 2005). Imputing stand tables involves locating the 'neighbouring' reference stands with the most similar stand-level variables and using their known stand tables to impute the stand table for the target stand which lacks this detailed information. Most similar neighbour (MSN) approaches offer several potential advantages over classical estimation procedures, such as predictions using regression analysis, since many variables are predicted at once (multivariate) and the variability across stands is maintained. Also, estimates will be within the bounds of biological reality. The accuracy of these methods is dependent on (1) the size and representativeness of the sample, (2) the degree of similarity between the target and reference stands and (3) how well the attributes for which information is available serve as a predictor for the attributes for which information is missing.

Obtaining good estimates of stand tables is particularly difficult in complex stands with many species and high variability in tree sizes. Temesgen *et al.* (2003) found that for complex stands the MSN approach (by Moeur and Stage, 1995) was marginally better in estimating tree-lists than the other approaches that they examined in their study. However, the authors found the distributions by species not as accurately estimated as might be desired.

Imputation techniques for assigning tree-lists to unsampled forest polygons can be seen as a specialized application for handling incomplete multivariate data. When a design-based forest inventory is used to randomly sample tree-lists (i.e. measure field plots), those polygons without tree-lists can be viewed as the result of a missing data mechanism that is Missing Completely at Random (MCAR) (Little and Rubin, 2002). If auxiliary information is used for double or stratified sampling, the probability of missingness is dependent on observed data, and the missing data mechanism is Missing at Random (MAR) (Little and Rubin, 2002). In either case, imputation has the potential to improve estimation through the use of all available information.

When a forest inventory is MAR or MCAR and imputation is appropriate, stratification prior to imputation can potentially improve prediction. This will happen when the best model for the relationship between ground variables and auxiliary variables differs among groups of observations. However, a thorough comparison to examine the performance of MSN within a non-stratified population *vs* its performance within a stratified population is lacking.

In this article, simulations were performed to investigate stratification into inventory type groups (ITGs) and age groups prior to applying the MSN approach. We examined whether the use of stratification improved the estimates of species within stand tables compared with the alternative of using the MSN approach within the whole population.

Methods

Data

Ground and auxiliary data collected in 134 complex stands (land parcels or polygons) in south-eastern British Columbia (BC) were used for this study. The stands were located in the Interior Cedar-Hemlock, Englemann Spruce/Subalpine-Fir, Interior Douglas-fir, Montane Spruce and pockets of Alpine Tundra biogeoclimatic ecological classification (BEC) zones (Braumandl and Curran, 1992) and included several tree species: Douglas-fir (*Pseudotsuga menziesii* (Beissn.) Franco), lodgepole pine (*Pinus contorta* var. *contorta* Dougl.), western white pine (*Pinus monticola* Dougl.), pondeosa pine (*Pinus ponderosa* Laws.), western larch (*Larix occidentalis* Nutt., L.), trembling aspen (*Populus tremuloides* Michx.), subalpine-fir (*Abies lasiocarpa* (Hook.) Nutt.), spruce (*Picea glauca* (Moench) Voss and *Picea engelmannii* Parry and hybrids), black cottonwood (*Populus trichocarpa* Torr. & Gray), western hemlock (*Tsuga heterophylla* (Raf.) Sarg.), white birch (*Betula paperifera* Marsh.) and western red cedar (*Thuja plicata* Donn.).

For each sampled polygon, a point was randomly selected from a grid and a cluster of nine plots (four full measure and five count plots) was located at the selected grid point. For the ground data, four variable-radius plots were randomly located in each stand, and the species, d.b.h. and status (i.e. live or dead) for all trees with a d.b.h. of $\geq$12.5 cm were recorded, along with other tree variables (BC Ministry of Forests, 1995a, b, 1998). The live trees for all ground data were compiled, and for each stand the stand table was calculated. In addition, average volume per hectare, TPH and basal area per hectare were calculated and compiled within the Variable Density Yield Projection model (VDYP) (BC Ministry of Forests, 1995a) for all species combined, and by species. The remaining plots were measured for basal area per hectare only and were not included in this study (BC Ministry of Forests, 1998). For each of the four plots, stand attributes were calculated including the TPH by species and 10-cm d.b.h. class, the species proportions by basal area and TPH. These variables were then averaged over the four plots to obtain the polygon estimates.

For each sampled polygon, auxiliary attributes were extracted from forest inventory records including crown closure (CC) class; tree species composition and proportions (per cent by CC class); site index in m height class and age class. In addition, the ITG, representing species composition groups, and the BEC zone (Meidinger and Pojar, 1991) were available for all polygons.

Using the auxiliary information (i.e. height, crown closure, site index and species composition), the stand-level growth and yield model VDYP was used to obtain estimates of volume per hectare, average height and quadratic mean diameter for each stand. Crown closure (CC) classes were assigned class midpoints of 20, 30, 40, 50, 60, 70 or 80 per cent to represent CC class. Table 1 provides the number of stands in each CC class.

A variation of 'leave-one-out' data-splitting analysis (Mosteller and Tukey, 1968) was used to split the data into target and reference stands (as in Moeur and Stage, 1995). The 134 sampled polygons in the complete dataset were randomly assigned, without replacement, to 33 groups of four stands each. Then, each group of four stands was omitted one group at a time constituting the target polygons. The remaining 130 stands were used as reference polygons. This is referred to as 'leave-four-out' data splitting, as was done in Moeur and Stage (1995). The target polygons were assumed to be non-sampled polygons, lacking ground inventory data. The reference polygons constituted the pool of potential polygons with ground and auxiliary data, which could be selected to impute the stand table for target polygons. The target polygons were used to validate the accuracy of the imputation methods by comparing the known (observed) stand table with the imputed (expected) stand table selected from the reference polygons.

MSN method

All reference polygons formed the pool of potential similar neighbours that could be selected to impute the stand tables on to target polygons without stratification. For MSN, the distance metric used is based on the distance between the variables measured on both the reference and target polygons (X variables), weighted by

Table 1: Distribution of the 134 sampled stands by crown closure and species mixture class.

	CC class (%)							Species mixture class	
	20	30	40	50	60	70	80	1	2
No. of stands	6	23	23	30	26	23	3	52	82

the correlations between the X variables and the variables available only for the reference data (Y variables) (Moeur and Stage, 1995). First, canonical correlation analysis between ground (Y variables) and auxiliary data (X variables) using the reference data was used to determine the weights. Second, the 'most similar' reference polygon was selected based on similarity of the auxiliary data, weighted by the correlations to the ground data. The distance measure used was

$$D_{uj}^2 = (X_u - X_j)\Gamma\Lambda^2\Gamma'(X_u - X_j)', \qquad [1]$$

where Γ is the matrix of standardized canonical coefficients for the auxiliary variables; Λ^2 is the diagonal matrix of squared canonical correlations between auxiliary attributes and ground variables; $\Gamma\Lambda^2\Gamma'$ is the matrix of the weights; X_u is a vector of standardized values of the auxiliary variables for the uth target stand and X_j is a vector of standardized values of auxiliary variables for the jth reference stand. The ground data for the reference polygon with the smallest distance (MSN based on the auxiliary data) was then imputed to the target polygon.

When the diagonal matrix of squared canonical correlations was non-singular, the Moore–Penrose inverse was used to find a unique definition of the inverse matrix (Draper and Smith, 1998; SAS Institute Inc., 2004).

Stratified nearest neighbour

For the stratified nearest neighbour method, the ITG variable was used to separate the 134 sample polygons into the following two strata based on the leading (by photo-estimated volume per hectare) tree species: (1) Douglas-fir, subalpine-fir or western hemlock leading (64 stands) and (2) spruce, pine or larch leading (70 stands). To account for differences in species composition within each stratum, two species mixture classes were established and each stand was assigned to one of the following species mixture classes (see Table 1): Class 1, where the composition of the leading species is >70 per cent (pure-species stands), and Class 2, where the composition of the leading species is ≤70 per cent (mixed-species stands). The stand age was used to separate the 134 sample polygons into two age strata: (1) stands with age ≤90 (73 stands) and (2) stands with age >90 (61 stands). The MSN approach, as described above, was used to impute ground variables for a group of four target stands at a time using only reference stands from the same stratum. Target stands for each stratum were determined by using the 'leave-four-out' data-splitting method as described above. Table 2 provides the variables used in the MSN analysis.

Comparison of approaches

The separation of the data into target *vs* reference stands using the leave-four-out data-splitting method was repeated 1000 times for all 134 sample polygons and for each of the inventory type and age strata, respectively. To evaluate the results for each simulation, bias (equation 2) and root mean square difference (RMSD, after Stage and Crookston, 2007; equation 3) were calculated for each variable in the Y set for each replicate as follows:

$$\text{Bias} = \frac{\sum_{i=1}^{n}(\text{observed}_i - \text{imputed}_i)}{n}, \qquad [2]$$

$$\text{RMSD} = \sqrt{\frac{\sum_{i=1}^{n}(\text{observed}_i - \text{imputed}_i)^2}{n}}, \qquad [3]$$

where n is the number of target stands.

The mean, minimum and maximum of each of these two statistics were summarized over the 1000 sampling replications of each simulation. The mean of the distance measure was calculated for each simulation and then averaged over the 1000 sampling replications.

Results and discussion

The dataset covered a wide range of age (36–300 years), average height (12.8–37.9 m), TPH by

Table 2: Ground and auxiliary variables used in the imputations. Ground variables were selected as a proxy for the tree-list.

Ground variables (*Y* set)	Auxiliary variables (*X* set)
TPH for seven species or species groups: Douglas-fir, western red cedar, lodgepole pine, spruce, western larch, subalpine-fir and hardwoods*	Species mixture class: (1) pure-species stands and (2) mixed-species stands
Basal area per hectare for all species in m^2 ha^{-1}	Age (years) (not used in age strata)
	Site index (m)
	CC class (%)
	Average height (m), model estimated
	Quadratic mean diameter (cm), model estimated
	Volume in m^3 ha^{-1}, model estimated

* Hardwoods included aspen, cottonwood and birch.

species (0–1415) and total basal area (15–87.5 m^2 ha^{-1}). ITG 2 (spruce, pine or larch leading) includes the oldest stands (300 years). However, ITG 1 (Douglas-fir, subalpine-fir or western hemlock leading) also includes old stands (280 years), and the stands with largest total basal area and volume are found in this stratum. The largest TPH occurs for lodgepole pine (1415) followed by Douglas-fir (1058) (Tables 3 and 4).

Moeur and Stage (1995) suggested the use of the distance metric to assess the adequacy of results. The variability of the mean distance (mean of the average distances over all 1000 replicates) differs among the five simulations (Figure 1). Table 5 displays minimum, maximum and mean of the mean distance for each simulation. The range and variability of the mean distance is larger for all of the individual strata compared with the whole population.

Moeur and Stage (1995) indicated that MSN resulted in poor estimates of species composition. It was expected that the stratification into inventory types representing species composition groups would result in closer matches of species composition. The evaluation of the simulation in terms of bias and RMSD (Tables 6 and 7) shows, however, that the stratification into inventory types did not improve the imputation results. Only in five cases is the range of bias or RMSD smaller for inventory type strata than for the simulation using all stands. For ITG 1, the range of the mean bias is smaller than that for the whole population for TPH of western larch and lodgepole pine, and the range of the mean RMSD is smaller than that for the whole population for TPH of lodgepole pine. For ITG 2, both mean bias and mean RMSD are smaller for TPH of western hemlock. It is surprising that ITG 1, which includes Douglas-fir, subalpine-fir and western hemlock leading stands, gives better results for TPH of larch and pine than ITG 2, which comprises stands in which larch, pine and spruce are the leading tree species. On the other hand, ITG 2 gives better results for TPH of western hemlock which is one of the leading species in ITG 1. For both age strata the range of the mean RMSD is larger for all *Y* variables compared with the simulation using all stands. For age stratum 1, the mean bias for all *Y* variables is also larger compared with the simulation using all stands. Age stratum 2 has a smaller range of mean bias for TPH of western hemlock, lodgepole pine and hardwoods compared with the simulation using all stands.

The stratification into ITGs and age groups prior to using MSN did not improve the estimates of species within stand tables compared with the alternative of using the MSN approach within the whole population. This can be ascribed to the following.

1 The number of stands used in the analysis. The data used only included 134 stands of southern BC. Additional stand data would increase the number of stands in each stratum. A larger number of stands within each stratum might improve the results substantially. It can be suspected that creating strata decreased the range of the reference data and/or created gaps along the space spanned by the *X* variables. The increase of the mean distance of the stratified MSN approach compared with the non-stratified

Table 3: Minimum, maximum and mean values for the *Y* variables for all polygons and by inventory type and age class

Variable		All, $N = 134$	Inventory type: ITG 1, $N = 64$	Inventory type: ITG 2, $N = 70$	Age class: Age 1, $N = 73$	Age class: Age 2, $N = 61$
Basal area (m^2 ha^{-1})	Min	15.00	18.00	15.00	15.00	18.00
	Max	87.50	87.50	72.50	72.50	87.50
	Mean	38.68	40.48	37.03	35.59	42.37
Number of western red cedars per hectare	Min	0	0	0	0	0
	Max	703	703	703	703	703
	Mean	70	82	60	89	48
Number of Douglas-firs per hectare	Min	0	0	0	0	0
	Max	1058	1058	888	888	1058
	Mean	282	372	201	194	388
Number of western hemlocks per hectare	Min	0	0	0	0	0
	Max	803	803	392	803	545
	Mean	56	81	34	58	55
Number of western larches per hectare	Min	0	0	0	0	0
	Max	437	240	437	437	262
	Mean	46	19	70	68	20
Number of hardwoods per hectare	Min	0	0	0	0	0
	Max	482	482	323	482	283
	Mean	25	26	24	37	11
Number of lodgepole pines per hectare	Min	0	0	0	0	0
	Max	1415	538	1415	1415	1150
	Mean	249	67	415	372	99
Number of spruces per hectare	Min	0	0	0	0	0
	Max	753	398	753	753	539
	Mean	68	71	65	46	95

approach might indicate that the stratification created gaps in the space spanned by the *X* variables. Imputation does not extrapolate nor interpolate (Crookston *et al.*, 2002) which results in imputation error whenever the target data are not within the span of the reference data or whenever the density of the *X* variables is low (Stage and Crookston, 2007). Small strata that are prone to exhibit large gaps in the space spanned by the *X* variables can therefore be expected to show large imputation error and with that worse results than non-stratified MSN imputation. The complete non-stratified dataset tends to have a similar range for target and reference datasets and smaller gaps in the space spanned by the *X* variables which will result in less imputation error.

2 In stratified sampling, samples are allocated to strata in a balanced manner. If data are collected for the purpose of stratified MSN and the data used for imputation within each stratum cover the range of values without any large gaps in the *X* variable space, good imputation results can be expected. In the given example, however, data splitting for MSN was performed. Due to this the data reserved for target stands may have caused the *Y* and *X* variables to distribute themselves in a far from balanced manner over the strata (e.g. reference stands have a smaller range than target stands or gaps along the *X* variable space are created). Thus, the gains of stratification can be lost in estimating attributes for the target polygons.

3 The choice of classification variables might also have contributed to poor performance of stratified MSN. We examined the impacts of stratification in imputing stand tables using age and inventory type only. Stratification prior to imputation may improve prediction if the relationship between *Y* and *X* variables

Table 4: Minimum, maximum and mean values for the *X* variables for all polygons and by inventory type and age class

Variable		All, $N = 134$	Inventory type		Age class	
			ITG 1, $N = 64$	ITG 2, $N = 70$	Age 1, $N = 73$	Age 2, $N = 61$
Species mixture class	No. in 1	52	24	28	33	19
	No. in 2	82	40	42	40	42
Age (years)	Min	36	50	36	36	95
	Max	300	280	300	90	300
	Mean	109	128	92	69	158
Site index (m)	Min	6.6	8.8	6.6	11.9	6.6
	Max	33.8	33.8	32.0	33.8	32.3
	Mean	17.8	18.6	17.1	19.7	15.5
Crown closure (%)	Min	20	20	20	20	20
	Max	80	80	80	80	80
	Mean	50	48	51	52	46
Average height (m)	Min	12.8	14.8	12.8	12.8	17.8
	Max	37.9	36.0	37.9	34.2	37.9
	Mean	23.0	24.6	21.5	20.5	25.9
Quadratic mean diameter (cm)	Min	16.3	19.9	16.3	16.3	22.1
	Max	53.1	53.1	44.6	31.4	53.1
	Mean	26.3	29.3	23.6	22.6	30.8
Volume ($m^3\ ha^{-1}$)	Min	84.3	99.8	84.3	84.3	119.1
	Max	938.0	938.0	629.5	938.0	856.1
	Mean	346.0	404.9	292.2	303.1	397.4

For the species mixture class the number of stands in each mixture class is given.

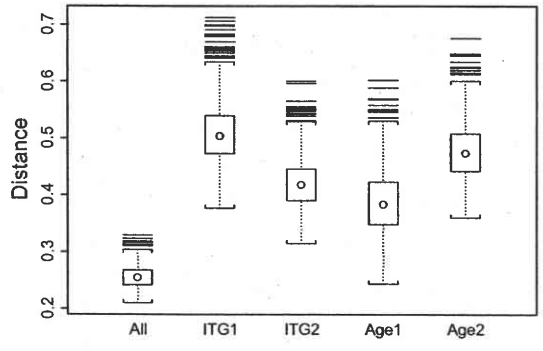

Figure 1. Box-and-whisker plots of the mean distance over 1000 sampling replications for all polygons and by inventory type and age class.

differs among the groups of observations. This is only the case if the groups are disparate enough which is not true for the *X* variables of the given inventory and age strata (see Table 4). Therefore, the results might have been better if more disparate strata had been chosen. It can be expected that more disparate stratification (e.g. stratification on land type: forest land *vs* grass land) improves imputation results because then the nature of the relationship between the *Y* and *X* variables differs between strata.

4 The assignment of inventory type and age groups. The cut-off point at age 90 for the age strata was chosen for creating two strata with approximately the same number of observations. This, however, might not have created strata with the lowest within-stratum variability and the highest variability among strata. Because of the small number of available stands, the two inventory type strata combined stands with three leading species each, so that each stratum was still very heterogeneous with few observations. The heterogeneity and the small number of observations of the strata probably created inadequate ranges of the target and reference datasets and gaps in the space spanned by the *X* variables which will result in the problems already discussed above. With a larger dataset, other possible stratifications, which