# Regeneration imputation models for complex stands of southeastern British Columbia

by Badre T. Hassani[1], Valerie LeMay[2], Peter L. Marshall[3], H. Temesgen[4] and Abdel-Azim Zumrawi[5]

Two imputation techniques for predicting natural regeneration in complex stands prevalent in southeastern British Columbia (BC) were compared using data from the Interior Cedar-Hemlock moist warm subzone variant 2 (ICHmw2) in the vicinity of Nelson, BC. Imputation approaches offer advantages over other modeling approaches in that they provide estimates of many variables at one time (multivariate) and there are no assumptions regarding the probability distributions of the variables to be predicted. For the tabular imputation, the average regeneration per ha was calculated for each combination of five site groups, two residual density classes, five time-since-disturbance intervals, species, and height classes. For Most Similar Neighbour (MSN) imputation, data with both regeneration information, and overstory trees and site information (called reference plots) were used to impute regeneration of plots with only overstory trees and site information (called target plots), by selecting the most similar plot. Of the two approaches studied, the MSN approach gave better results than tabular imputation. The tabular imputation approach is simpler to implement, since tables of results can be published and made available for use. However, the MSN software has been made freely available, resulting in greater ease of access.

**Key words**: multi-species, multi-cohort, nonparametric imputation, multivariate prediction, regeneration estimation

Deux techniques d'imputation pour la prédiction de la régénération forestière dans des peuplements complexes retrouvés dans le sud-est de la Colombie-Britannique (CB) ont té comparées en utilisant les données de la variante 2 de la sous-zone chaude et humide Thuya-Pruche de l'Intérieur (ICHmw2) située à proximité de Nelson en CB. Les approches d'imputation offrent l'avantage par rapport aux autres approches par modèle d'estimer plusieurs variables simultanément (multi-variables) et qu'il n'y a pas d'hypothèse relativement à la probabilité de distribution des variables à être prédites. Dans le cas de l'imputation tabulaire, la régénération moyenne par hectare a été calculée pour chaque combinaison des groupes de cinq stations, pour deux classes de densité résiduelle, pour cinq intervalles écoulés depuis la perturbation, par espèces et par classes de hauteur. Dans le cas de l'imputation selon le voisinage le plus semblable (IVS), les données d'information sur la régénération et d'information sur les arbres formant le couvert forestier et sur la station (dénommées parcelles de référence) ont été utilisées pour imputer la régénération des parcelles d'arbres formant le couvert forestier et l'information sur la station (dénommées parcelles cibles), en choisissant la parcelle la plus semblable. Des deux approches étudiées, l'approche IVS a donné de meilleurs résultats que l'imputation tabulaire. L'approche selon l'imputation tabulaire est plus simple à implanter, puisque les tableaux de résultats peuvent être publiés et disponibles pour utilisation. Cependant, le logiciel IVS est disponible gratuitement ce qui permet une facilité d'accès plus grande.

**Mots-clés** : multi-espèces, multi-cohorte, imputation non paramétrique, prédiction pour multi-variables, estimation de la régénération

## Introduction

Complex (multi-species and multi-cohort) stands are created by natural small-scale disturbances and are regularly distributed throughout the stand as mosaics of small single-cohorts (Oliver and Larson 1996). Because they provide high aesthetic quality and wildlife habitat, protect community watersheds, and have greater biological and structural diversity, the understanding of the dynamic of complex stands has recently become a priority and is the subject of a number of studies (Smith *et al*. 1996). Regeneration in these stands is difficult to precisely predict (Oliver and Larson 1996, Ek *et al.* 1997). Manipulation of regeneration constitutes a critical action for managers striving to predict and control secondary succession after harvesting.

Partial cutting is assumed to mimic small-scale natural disturbances that have created many multi-cohort stands in western conifer forests (Oliver and Larson 1996, Smith *et al.* 1996), and is widely replacing clearcutting. While ensuring a successful natural regeneration, this silvicultural system allows also extraction of wood for commercial uses.

The growth and yield model, Prognosis (Stage 1973), has been adapted to conditions in southeastern forests of British Columbia (termed Prognosis$^{BC}$). This model can be used to forecast future stand conditions following partial cutting. However, the lack of an acceptable regeneration component has limited the use of Prognosis$^{BC}$ for making long-term projections to evaluate the impact of different silvicultural treatments and partial cutting scenarios.

As an alternative to regression approaches for modeling regeneration, imputation approaches could be used. Imputation involves substituting plausible measurements from one or more selected units with similar characteristics to units lacking these measures (Rubin 1987, Ek *et al.* 1997, Van Deusen 1997, McRoberts 2001). Data with all variables measured

[1]Research Associate, Department of Forest Resources Management, University of British Columbia, 2045–2424 Main Mall, Vancouver, British Columbia V6T 1Z4. E-mail: badre@interchange.ubc.ca
[2]Associate Professor, Department of Forest Resources Management, University of British Columbia, 2045–2424 Main Mall, Vancouver, British Columbia V6T 1Z4. E-mail: lemay@interchange.ubc.ca. Author to whom all correspondence should be addressed.
[3]Professor, Department of Forest Resources Management, University of British Columbia, 2045–2424 Main Mall, Vancouver, British Columbia V6T 1Z4. E-mail: marshall@interchange.ubc.ca
[4]Assistant Professor, Department of Forest Resources, 237 Peavy Hall Oregon State University, Corvallis, Oregon 97331–5703. E-mail: hailemariam.temesgen@oregonstate.edu
[5]Research Scientist, BC Ministry of Forests Research Branch, Department of Forest Resources Management, University of British Columbia, 2045–2424 Main Mall, Vancouver, British Columbia V6T 1Z4. E-mail: zumrawi@interchange.ubc.ca.

Table 1. Local and scientific names, and species codes for trees found in ICHmw2

| Local Name | Scientific Name |
|---|---|
| Black cottonwood | *Populus trichocarpa* Torr. & Gray |
| Douglas-fir | *Pseudotsuga menziesii* var. *glauca* (Beissn.) Franco |
| Grand fir | *Abies grandis* (Doug.) Lindl. |
| Hybrid spruce | *Picea engelmannii* Parry x *glauca* (Moench) Voss |
| Lodgepole pine | *Pinus contorta* Dougl. Var. *latifolia* |
| Paper birch | *Betula papyrifera* Marsh. |
| Subalpine fir | *Abies lasiocarpa* (Hook.) Nutt. |
| Trembling aspen | *Populus tremuloides* Michx. |
| Water birch | *Betula occidentalis* |
| Willow | *Salix* sp. |
| Western hemlock | *Tsuga heterophylla* (Raf.) Sarg. |
| Western larch | *Larix occidentalis* Nutt. |
| Western redcedar | *Thuja plicata* Donn |
| Western white pine | *Pinus monticola* Dougl. |
| Western yew | *Taxus brevifolia* |

are termed "reference data," whereas data with some variables missing are termed "target data." If only one selected unit is used in the substitution, the variability of the missing variables as represented in the reference data will be preserved in the estimates imputed to the target data (Moeur *et al.* 1995, Moeur and Stage 1995, Ek *et al.* 1997, Haara *et al.* 1997, Van Deusen 1997, Maltamo and Kangas 1998, Moeur and Hershey 1998, Moeur 2000, Temesgen and LeMay 2000). This differs from regression approaches where averages, conditional on the values of the predictor variables, are used as the estimates for the missing variables in the target data. For example, if all trees have measured diameters outside bark at breast height (1.3 m above ground; dbh) and heights, but only some trees have measured volume, regression would involve fitting an equation with volume as the dependent variable and dbh and height as the predictor variables. To obtain the estimated volumes for the target trees, the dbh and height would be input into the fitted equation, and the estimated volume obtained, which would be an estimated average volume for all trees of that particular dbh and height (conditional average). Using imputation, the volume of the tree in the reference dataset that has a dbh and height most similar to the target tree would become the estimated volume for the target tree. This results in higher variability of volume estimates. However, if the average of a number of similar units is used as the estimate for the target unit, imputation results approach regression results as the number of units used in the averaging increases. Also, imputation can provide estimates of several missing variables at one time (multivariate), whereas regression allows for one dependent variable in each fitted equation, and no distributional assumptions for the variables of interest are needed for imputation.

The objective of this study was to explore the use of two imputation techniques, tabular and most similar neighbour (MSN), to predict regeneration in the complex stands prevalent in the Interior Cedar Hemlock moist warm zone variant 2 (ICHmw2) of BC (Meidinger and Pojar 1991). Imputation methods were considered as an attractive alternative to other prediction methods, since the regeneration per ha by species was of interest and there are often more than five species in these complex stands.

# Methods

## Study sites

Study sites were located in the Columbia-Shuswap moist warm Biogeoclimatic Ecosystem Classification (BEC) variant of the Interior Cedar Hemlock (ICHmw2) in the vicinity of Nelson, BC. The Interior Cedar Hemlock (ICH) zone is the largest and the most productive zone in the interior of British Columbia (BC) (Meidinger and Pojar 1991). This zone occurs on lower to middle elevations and ranges from 500 to 1450 m in the northern part of its range, and from 1200 to 1450 m in the southern part (Braumandl and Curran 1992). These forests have a continental climate characterized by cool wet winters and warm dry summers. Morainal soils with loamy or silty surface textures predominate in the area. Forests consist of complex stands (mixed-species, multi-cohort), where western hemlock (*Tsuga_heterophylla* (Raf.) Sarg.) and western redcedar (*Thuja plicata* Donn) constitute the climax tree species. A list of tree species present is shown in Table 1, based on Ketcheson *et al.* (1991) and Braumandl and Curran (1992). The ICHmw2 also includes a diverse composition of shrubs, herbs, mosses, and lichens.

## Sampling design and data collection

All the ICHmw2 polygons that were accessible and partially harvested between two and 25 years ago were identified from the Ministry of Forests' silvicultural surveys database (ISIS) and topographic maps of Arrow and Kootenay Lake Forest Districts, and were included in the sampling frame. Undisturbed stands found in the neighbourhood of harvested stands were also included in the sampling population, but not used in the imputation analyses. Sampling sites were purposively spaced throughout the ICHmw2 to cover ranges of regeneration methods, site preparation, aspect, slope, and elevation. Once these sites (polygons) were selected, plots were established using systematic sampling with a random start. To avoid getting confounded results due to edge effects, plots were established at least 50 m from the roads or any other openings and at a random bearing. The number of plots established in a selected polygon and the distance between plots depended on the size of the polygon and the degree of structure variability present. Structurally-variable sites were sampled more heavily than more homogenous sites. At least two plots were established on each polygon and, for most polygons, plots were 100 m apart.

For each established plot, trees were sub-divided into regeneration, small trees, and large trees. Regeneration was defined as being at least 15 and 30 cm tall for shade-tolerant and shade-intolerant species, respectively, and less than 7.5 cm dbh, based on the sampling design used by Ferguson *et al.* (1986) and by Fergu-

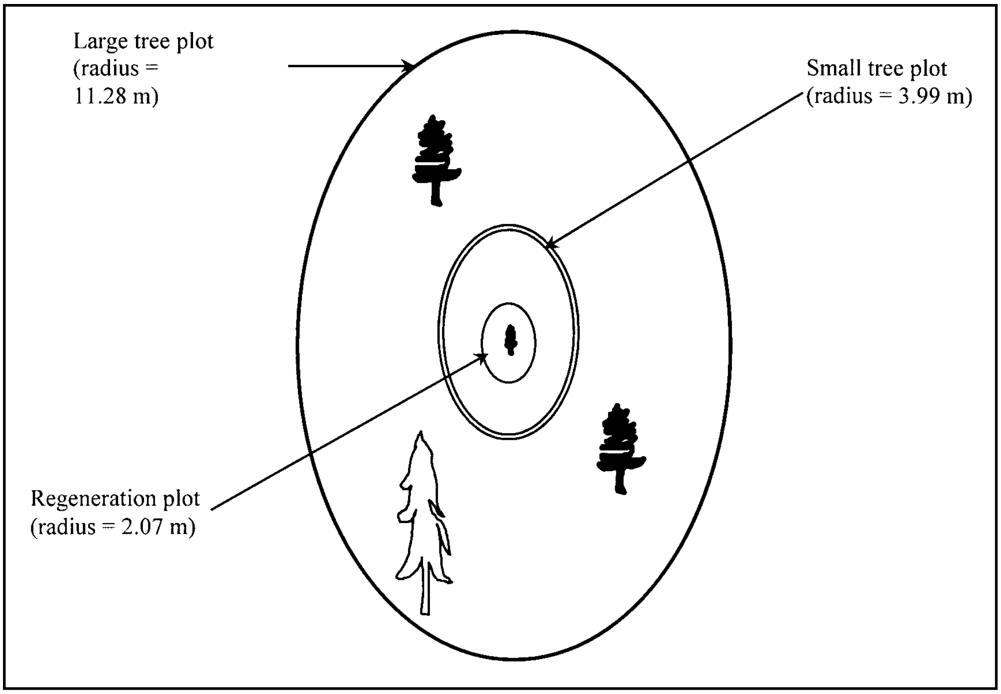


**Fig. 1.** Plot layout for sampling regeneration, small, and large trees.

son and Carlson (1993). Small trees were defined as having dbhs between 2.0 and 7.5 cm, and large trees were greater than 7.5 cm.

The three tree size ranges were sampled within three concentric fixed area plots (Fig. 1). For the 2.07-m center plot (0.00135 ha), established regeneration was tallied by four height classes: (1) 15–49.9 cm; (2) 50–99.9 cm; (3) 100–129.9 cm; and (4) > 129.9 cm. Next, "best trees" were selected and height and total age were measured. Following Ferguson and Carlson (1993), the criteria for "best trees" were: (1) the two tallest trees, regardless of species; (2) the one tallest tree of each additional species present; and (3) the tallest of the remaining trees until at least four trees were sampled. If only one species was present on the plot, measurements on the four tallest trees of that species were selected. For determinant species, these measurements were made on standing trees whenever possible. Non-determinant species were destructively sub-sampled for total age and height. Any evidence of tree damage, disease, or insects, was also noted for each sub-sampled regeneration tree.

Small trees were sampled using a central 3.99-m radius (0.005 ha) plot (Fig. 1). Dbhs and heights were measured for all trees. Some small trees were also counted as regeneration; these were noted to avoid the overlap and double counting on plot summaries. When more than two species were present, two small trees for each species were selected randomly for total age and five-year height increment measurement; four trees were randomly selected and sub-sampled when only two or less species were tallied. To ensure that all trees sampled reflected the same growing period, the current growing season was not included.

Large trees were sampled using an 11.28-m radius (0.04 ha) plot (Fig. 1). Dbh and species were recorded for all trees. Whenever the number of trees allowed, two trees for each species present were chosen randomly and measured for heights. The number of trees and the species composition were used to identify overstory composition, to estimate retention level and residual basal area, and to study the resultant impact of residual cover on regeneration establishment and subsequent growth. Other information, such as presence of scars, diseases, fire signs and any other physical deformation, was recorded as well.

Site information was collected *in-situ* for each established plot. This information included the BEC site series, elevation, slope angle (percent), slope position, aspect (degrees), and site preparation.

### Data preparation

Due to the high number of species in the ICHmw2 subzone variant, the species were grouped into four ecological guilds (three levels of shade tolerance plus a hardwood group). The shade-tolerant species group was composed of grand fir, subalpine fir, western redcedar, hemlock, and spruce; the shade semi-tolerant species group included Douglas-fir and white pine; and the shade-intolerant species group included lodgepole pine and larch. The hardwood species group included cottonwood, trembling aspen, white birch, Douglas maple, and willow.

### Imputation approaches

Two imputation approaches were used to estimate regeneration stems per ha. The tabular approach used average regeneration per ha for a similar group of plots, whereas MSN used the most similar plot to estimate regeneration in the target plot.

Data splitting (Snee 1977) was used to test the accuracy of each approach. Plots were randomly and evenly divided into five sets of data. Then, one set was reserved as test data, and this approach was repeated for a total of five tests. Each test set represented target plots that supposedly lacked regeneration.

*Tabular Imputation Approach*

Tabular imputation models were developed by generating tables of regeneration per ha averages for specific stand conditions at some time following disturbance. Preliminary analyses using a generalized linear models procedure and correlation analysis, as well as empirical knowledge showed that site series, years since disturbance, and basal area per ha were most

Table 2. Variables used in the MSN analyses

| Variables of Interest (Y) | Overstory and Site Variables (X) |
|---|---|
| Regeneration stems per ha | Number of years since disturbance |
| for four height classes: | Site series |
| 1: 15–49.9 cm | Aspect (radians) |
| 2: 50–99.9 cm | Elevation (m) |
| 3: 100–129.9 cm | Slope (%) |
| 4: >129.9 cm; | Residual trees per ha (TPH) |
| by four species group: | Residual basal area per ha (BA) |
| Shade-intolerant species | Crown competition factor (CCF) |
| Shade-semi-tolerant species | Slope position: lower, level, middle, plateau, and upper |
| Shade tolerant species | Site preparation: none, burning (burn), brushing (brush), |
| Hardwood species | brushing and burning (bbrush), and mechanical (mech) |
| (16 variables of interest). | |

related to regeneration establishment and survival. Plot and individual tree information was used to produce tables that showed average regeneration per hectare for each site, years since disturbance, and residual basal area class. For site, similar site series were grouped into: dry (site series 02 and 03), slightly dry (04), mesic (01), slightly wet (05), and wet (06, 07, and 08). The years since last disturbance classes were: 1 (2–5 yrs), 2 (6–10 yrs), 3 (11–15 yrs), 4 (16–20 yrs), and 5 (21–25 yrs). Plots were stratified into two residual basal area (BA) classes: open ($\leq 5.0$ m$^2$) and dense ($> 5.0$ m$^2$). This allowed for an approximately even distribution of plots among the two basal area classes. For each table, the average regeneration stems per ha was calculated for each of 16 cells, defined by the four species groups previously defined (i.e., shade-tolerant species, shade semi-tolerant species, shade-intolerant species, and hardwood species) and the four height classes used in sampling.

Data from 80% of the plots were used to produce regeneration tables (reference data), which were then used to impute regeneration on the reserved 20% of data (target data). Observed versus estimated regeneration values for the target data were then compared. This process was repeated five times, reserving a different set of data each time.

### The MSN Approach

For the MSN analysis (Moeur and Stage 1995, Moeur 2000), the software provided by Moeur (2000) (MSN version 1.0) was used. For each of the five datasets used in the tabular imputation, imputed values for the target data were obtained using the variables given in Table 2. Variables (X set) used to estimate regeneration variables (Y set) were obtained from different sources, including the BC Ministry of Forests' silviculture database (ISIS), maps, stand records, and the information recorded during the data collection phase. The number of years since disturbance was obtained from ISIS database and from stand records, whereas site series, aspect, elevation, slope position, and site preparation were recorded on the field. The stand density variables, residual trees per ha (TPH), residual basal area per ha (BA), and crown competition factor (CCF), were derived from the information collected on trees (species, dbh, and total number of trees (small and large)) within each plot. A set of ten indicator variables represented site preparation and slope position categories. The similarity measure used to find the most similar neighbour was:

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

where $D_{ij}^2$ is the squared distance between target plot i and the reference plot j, $X_i$ is the vector of the standardized values of the target plots, $X_j$ is the vector of the standardized values of the reference plots, $\Gamma$ is the matrix of standardized canonical coefficients for the X variables, and $\Lambda^2$ is the diagonal matrix of squared canonical correlations between X and Y variables. Together, $\Gamma\Lambda^2\Gamma'$ comprise a weight matrix. These weights were selected in order to emphasize X variables that have higher correlations with the 16 regeneration variables (Y; regeneration stems per ha by species group and height class). As with tabular imputation, observed versus estimated regeneration values were compared for the target data of each of the five datasets.

### Comparison of the two approaches

For each imputation method and for each of the five datasets, bias, mean absolute deviation (MAD), and root mean-squared error (RMSE) values were calculated over the 16 regeneration stems per ha variables as follows:

$$Bias = \sum_{j=1}^{m}\sum_{i=1}^{n}\left[\frac{\left(y_{ij} - \hat{y}_{ij}\right)}{nm}\right]$$

$$MAD = \sum_{j=1}^{m}\sum_{i=1}^{n}\left[\frac{\left|y_{ij} - \hat{y}_{ij}\right|}{nm}\right]$$

$$RMSE = \sum_{j=1}^{m}\sqrt{\sum_{i=1}^{n}\left[\frac{\left(y_{ij} - \hat{y}_{ij}\right)^2}{n}\right]} / m$$

where $y_{ij}$ is the observed regeneration stems per ha for each target plot $i$ and regeneration variable $j$, $\hat{y}_{ij}$ is the estimated regeneration stems per ha, $n$ is the number of target plots, and $m$ is the number of regeneration variables (16). These three statistics were used to examine the performance and to compare the predictive capability of the two approaches.

In addition, observed and estimated regeneration values were summarized into 16 cells represented by the four height classes and four species groups used in tabular imputation. The number of cells with non-zero observed and estimated regeneration (presence of regeneration), or zero observed and estimated regeneration (absence of regeneration) were counted. A good match was defined as correctly predicting the presence or absence of

Table 3. Number of plots summarized by variable classes

| Site Serie | No. Plots | Years since class[1] | No. Plots | Site preparation | No. Plots | Slope percent | No. Plots |
|---|---|---|---|---|---|---|---|
| 02 | 0 | 2–5 | 56 | Brushing | 22 | 0–10 | 56 |
| 03 | 114 | 6–10 | 108 | Burning | 88 | 10–20 | 66 |
| 04 | 88 | 11–15 | 77 | Mixed[2] | 16 | 20–30 | 72 |
| 01 | 54 | 16–20 | 55 | Mechanical | 13 | 30–40 | 62 |
| 05 | 54 | 20–25 | 37 | None | 194 | 40–50 | 40 |
| 06 | 7 | | | | | 50–60 | 19 |
| 07 | 13 | | | | | > 60 | 18 |
| 08 | 3 | | | | | | |
| **BA[3] ($m^2$/ha)** | **No. Plots** | **Slope position** | **No. Plots** | **Elevation (m)** | **No. Plots** | **Aspect** | **No. Plots** |
| 0 | 33 | Crest | 7 | < 800 | 34 | Flat | 6 |
| 1–5 | 134 | Lower | 30 | 8–9 | 24 | E | 57 |
| 5–10 | 55 | Depression | 11 | 9–1 | 33 | N | 36 |
| 10–15 | 32 | Middle | 220 | 10–11 | 56 | NE | 30 |
| 15–20 | 28 | Plateau | 15 | 11–12 | 52 | NW | 22 |
| 20–25 | 9 | Toe | 6 | 12–13 | 46 | S | 57 |
| 25–30 | 12 | Top | 4 | 13–14 | 58 | SE | 35 |
| 30–35 | 6 | Upper | 28 | 14–15 | 27 | SW | 59 |
| 35–40 | 7 | Level | 10 | 15–16 | 3 | W | 31 |
| > 40 | 17 | Blank | 2 | | | | |

[1]Number of years since last disturbance.
[2]Burning and brushing.
[3]Residual basal area per ha.

regeneration in at least 15 out of the 16 regeneration cells. A moderate match was defined as having eight to 14 agreements between the actual and the predicted regeneration cells. Finally, a poor match was considered to be those plots that had less than eight agreements. The number of plots that had low (< 1000 regeneration stems/ha), medium (1000–2000), and high (> 2000) RMSE were also determined, and cross-tabulated with good, medium, and poor matches. An accurate prediction would then be good match and low RMSE, whereas an inaccurate prediction would be poor match and high RMSE.

For the final comparison, regeneration predictions using MSN or tabular imputation were used as input data for Prognosis$^{BC}$ for a limited number of plots, and differences between 50-year volume-yield projections using observed versus estimated regeneration were determined. To increase the amplitude of these tests, plots were grouped by a combination of match and RMSE classification, and a few plots were randomly selected from each group. For each imputation method, four plots were selected from good match-low RMSE (good accuracy) and from moderate match-medium RMSE (medium accuracy) groupings. Eight plots were selected from poor match-high RMSE (poor accuracy) grouping. For each selected plot, an observed tree list using overstory (large and small trees) and regenerated trees was input to Prognosis$^{BC}$ and projected for 50 years. This process was repeated using the observed overstory and the estimated regeneration (either tabular or MSN) as the input tree lists. For each selected plot, yield differences and standard deviations of these differences were calculated and summarized for each imputation method and regeneration prediction category (good, medium, or poor accuracy).

## Results

Eighty percent of 138 sampled polygons were partially harvested, representing a variety of cutting intensities. Over 50% of the 333 measured plots were disturbed during the last decade, occurred on southerly exposures, and had no site preparation (Table 3). Residual basal areas ranged from 0 to 92 $m^2$/ha; about 50% of the plots had residual basal area per ha less or equal to 5 $m^2$/ha. The regeneration was highly variable, ranging from 0 to 124 081 stems per ha and averaging 9518 stems/ha overall. The averages of regeneration per hectare were 4643, 2376, 1410, and 1089 for shade-tolerant, shade semi-tolerant, hardwood, and shade-intolerant species, respectively. The species composition of the overstory included 15 species, and over 66% of the large trees (by stems/ha) were shade-tolerant.

An example of the tables produced using tabular imputation is given in Table 4. Tables for other stand conditions, using all available data, can be found on the Web site www.forestry.ubc.ca/Prognosis.

For the MSN analysis, 16 canonical variates resulted from the canonical correlations analysis on the 16 regeneration variables representing the Y variables, and 18 plot level variables representing the X variables. Ninety percent of the inherent variance of the Y variables was explained by the first eight canonical variates; these were retained in calculating the distances between plots. The site, years since disturbance, and residual basal area per ha had the highest canonical coefficients on most of the retained variates.

Although RMSE obtained for both imputation methods were comparable, the MSN approach commonly produced lower bias and MAD values (Table 5). Results varied more widely over the five data sets for MSN versus tabular analyses, indicating that the particular set of reference data had more effect on MSN analysis. However, predicting regeneration from a single observed plot using MSN preserved inherent variance found in the data. Also, the tabular imputation approach consistently produced negative biases (overestimation). Since the reference plots did not always represent all possible stand conditions, a complete range of regeneration tables was not produced. As a result, some target plots did not have a corresponding regeneration table to use in estimating regeneration. In this case, a similar table was used. This occurred mostly where the target plots represented more extreme conditions, either very dry or very wet sites, and regeneration estimates were obtained from a table for mesic,

Table 4. Average regeneration per ha by height class and species for time-since-disturbance interval 1, basal area class "Dense" and "Dry" sites (n = 18 plots)

| | Height class (cm) | | | | |
|---|---|---|---|---|---|
| Species | 15–49.9 | 50–99.9 | 100–129.9 | > 129.9 | All Heights |
| Grand fir | 0 | 0 | 0 | 0 | 0 |
| Subalpine fir | 0 | 41 | 41 | 0 | 83 |
| Western redcedar | 991 | 743 | 165 | 248 | 2146 |
| Western hemlock | 1445 | 165 | 206 | 206 | 2023 |
| Hybrid spruce | 1486 | 83 | 41 | 41 | 1651 |
| TOLERANT | 3921 | 1032 | 454 | 495 | 5903 |
| Douglas-fir | 2064 | 660 | 330 | 413 | 3467 |
| Western white pine | 826 | 289 | 41 | 165 | 1321 |
| SEMI-TOLERANT | 2889 | 949 | 372 | 578 | 4788 |
| Lodgepole pine | 1032 | 0 | 41 | 0 | 1073 |
| western larch | 165 | 41 | 0 | 0 | 206 |
| INTOLERANT | 1195 | 41 | 41 | 0 | 1280 |
| HARDWOOD | 454 | 248 | 248 | 743 | 1692 |
| ALL SPECIES | 8462 | 2270 | 1115 | 1816 | 13663 |

Table 5. Bias, mean absolute deviation (MAD), and root mean-squared error (RMSE) averaged over the 16 regeneration stems per ha variables for MSN and tabular imputation methods and for each of the five test data sets

| | | | MSN | | | Tabular | |
|---|---|---|---|---|---|---|---|
| Dataset | Number of Target Plots | Bias | MAD | RMSE | Bias | MAD | RMSE |
| 1 | 68 | 36 | 638 | 1639 | –110 | 698 | 1441 |
| 2 | 65 | –35 | 666 | 1685 | –188 | 705 | 1302 |
| 3 | 65 | 154 | 576 | 1415 | –144 | 773 | 1567 |
| 4 | 64 | –143 | 792 | 2093 | –99 | 660 | 1298 |
| 5 | 71 | 58 | 594 | 1381 | –200 | 813 | 1646 |

slightly dry, or slightly wet plots, that commonly have more abundant regeneration. This led to an overestimation of regeneration occurring on these rarer sites.

Using the number of matches and the RMSE classes to compare results, the MSN approach consistently resulted in a higher number of matched plots, and also fewer plots that were poorly matched (Tables 6 and 7). For the five MSN runs, the percent of good match/low RMSE (good accuracy) ranged from 8 to 18% of all target plots, whereas the percent of poor match/high RMSE (poor accuracy) was lower than 5% for all five runs. In comparison, good accuracy resulted for only 0 to 3% of the plots using tabular imputation. Tabular imputation uses the averages of all plots with similar stand characteristics. Therefore, the imputation can result in species combinations that do not exist on the landbase. As a result, species was more often mismatched using tabular imputation. This cannot happen with the MSN approach since the regeneration species are predicted from one measured matched plot.

For the final comparison, the Prognosis$^{BC}$ 50-year volume–yield projections revealed only small average volume differences (less than 22 m$^3$ per ha) between the observed and predicted regeneration for the MSN approach (Table 8). Average differences of the projected merchantable volumes for the tabular imputation method were higher (absolute values greater than 19 m$^3$ per ha). The standard deviation of the poor predictions (poor match–high RMSE) was substantially higher than the two other predictions categories for both approaches, as might be expected. Unexpectedly, some of the plots that were categorized as having poor predictions yielded small difference between their projected volumes. The poorness of the predictions was due to stand density variables mismatches between target and selected plots for the MSN approach. For tabular approach, it was due to either the use of similar or less reliable tables as a pool for imputation.

## Discussion

Low correlations between the regeneration and most of the predictor variables used in this study may indicate that either predictor variables are not useful or that the regeneration is highly variable. Continual fluctuation of standard error of the means with time-since-disturbance and site conditions for imputation tables that used all data and the descriptive analyses reported in this study provided substantial evidence for the second alternative.

For all three comparisons made, generally MSN gave better results than tabular imputation. Averaging the regeneration of several plots gave less accurate results. The variables used to stratify plots into tables were similar to those used by Ek *et al.* (1997), but separation of these variables into classes may not be optimal. However, more reference data would have been needed to produce more regeneration tables using more strata.

Results for the MSN analysis were more variable over the five datasets than for tabular imputation. Moeur and Stage (1995) reported that poor MSN performance has been noted mainly for under-representation of particular conditions.

Table 6. Number of plots by match and root mean-squared error (RMSE) using the MSN approach for each of the five datasets

| RMSE Class[1] | Number of Target Plots for Each Dataset | Match Category[2] | | | Total |
|---|---|---|---|---|---|
| | | Good | Moderate | Poor | |
| Low | 68 | 12 | 19 | 0 | 31 |
| | 65 | 6 | 28 | 0 | 34 |
| | 65 | 5 | 31 | 0 | 36 |
| | 64 | 8 | 17 | 0 | 25 |
| | 71 | 5 | 27 | 0 | 32 |
| Medium | 68 | 2 | 18 | 0 | 20 |
| | 65 | 3 | 16 | 1 | 20 |
| | 65 | 1 | 17 | 0 | 18 |
| | 64 | 1 | 20 | 1 | 22 |
| | 71 | 2 | 20 | 2 | 24 |
| High | 68 | 2 | 13 | 2 | 17 |
| | 65 | 1 | 10 | 0 | 11 |
| | 65 | 1 | 7 | 3 | 11 |
| | 64 | 0 | 14 | 3 | 17 |
| | 71 | 2 | 13 | 0 | 15 |

[1]Low RMSE had < 1000 regeneration stems/ha, medium RMSE had between 1000 and 2000, and high RMSE had > 2000.
[2]Good match was defined as correctly predicting the presence or absence of regeneration in at least 15 out of the 16 regeneration cells by plot, moderate match between 8 to 14, and poor match has less than 8 agreements.

Table 7. Number of plots by match and root mean-squared error (RMSE) using the tabular approach for each of the five datasets

| RMSE Class[1] | Number of Target Plots for Each Dataset | Match Category[2] | | | Total |
|---|---|---|---|---|---|
| | | Good | Moderate | Poor | |
| Low | 68 | 2 | 22 | 12 | 36 |
| | 65 | 1 | 15 | 16 | 32 |
| | 65 | 0 | 18 | 14 | 32 |
| | 64 | 0 | 21 | 12 | 33 |
| | 71 | 2 | 10 | 16 | 28 |
| Medium | 68 | 0 | 7 | 12 | 19 |
| | 65 | 2 | 6 | 13 | 21 |
| | 65 | 0 | 10 | 8 | 18 |
| | 64 | 0 | 12 | 7 | 19 |
| | 71 | 1 | 7 | 13 | 21 |
| High | 68 | 0 | 6 | 7 | 13 |
| 65 | 0 | 2 | 10 | 12 | |
| | 65 | 0 | 8 | 7 | 15 |
| | 64 | 0 | 6 | 6 | 12 |
| | 71 | 0 | 12 | 10 | 22 |

[1]Low RMSE had < 1000 regeneration stems/ha, medium RMSE had between 1000 and 2000, and high RMSE had > 2000.
[2]A good match was defined as correctly predicting the presence or absence of regeneration in at least 15 out of the 16 regeneration cells by plot, moderate match between 8 to 14, and poor match has less than 8 agreements.

In selecting observations to be used as reference data, consideration should be given to obtaining information for a very wide range of conditions, including unusual stand conditions, so that the reference data reflect the variability represented in the population. For complex stands, possibly more reference data would be needed since the variability in overstory and understory is likely to be higher than for stands with fewer species and sizes.

Both the MSN and the tabular imputation approaches resulted in reasonably accurate and close volume–yield predictions, based on measured small and large trees, and estimated regeneration. This may be partly due to low volumes represented by the regenerated trees, even after 50 years, relative to the volumes of the small and large trees. However, longer periods of simulation might reveal different results between the two approaches, particularly less accurate volumes using the tabular predictions.

Improvements in imputing regeneration may be possible. Geographic location, either broadly based by large latitudinal and longitudinal ranges or via considering plot spatial position, might improve results. Other information on stand structure from ground, aerial or other remotely sensed data might improve predictions. A more expensive option is to relate actual spatial positions of overstory trees to regenerated trees, particularly for stand conditions where constraints to regeneration are considered to be more local in nature.

## Conclusions

Relative to other modeling methods, such as regression analysis, imputation approaches do not require distributional assumptions, and are multivariate in nature. For regeneration prediction, imputation provides the advantage of predicting regeneration by species and height class in one step.

**Table 8. Summary of volume-yield estimates for the MSN and the tabular methods**

| Fit Category | Number of Plots | Average Merchantable Volume Difference ($m^3$/ha) | SD[1] ($m^3$/ha) |
|---|---|---|---|
| MSN | | | |
| Good Match & Low RMSE | 4 | –9.7 | 50.1 |
| Moderate Match & Medium RMSE | 4 | 7.5 | 74.8 |
| Poor Match & High RMSE | 8 | 21.5 | 91.6 |
| Tabular | | | |
| Good Match & Low RMSE | 4 | 25.8 | 46.5 |
| Moderate Match & Medium RMSE | 4 | –19.3 | 44.36 |
| Poor Match & High RMSE | 8 | –39.4 | 142.87 |

[1]SD: standard deviation of the volume differences.

Of the two approaches studied, the MSN approach gave better results. The tabular imputation approach is simpler to implement, since tables of results can be published and made available for use. However, the MSN software has been made freely available, resulting in greater ease of access.

As more data become available, the MSN and the tabular regeneration imputation models could be easily updated. Undoubtedly, more reference plots for unusual or under-represented conditions would increase the precision of estimates, particularly for the MSN approach. Also, either imputation approach could be easily linked to an existing growth and yield model.

## Acknowledgments

We are grateful to the Resource Inventory Branch, Research Branch, and Forest Practices Branch of the BC Ministry of Forests, via Forest Renewal BC funding, and to the National Engineering and Science Research Council for providing financial support for this project. We also thank Barry Snowden, Forest Practices Branch, Ivan Istar and Richard E. Logan, Nelson Forest Region, Jim Annunziello, Kootenay Lake Forest District, and Tom Johnston and Sandi Best, Arrow Forest District for their logistic support for field work. Cornel Lencar provided assistance with data collection.

## References

**Braumandl, T.F. and M.P. Curran. 1992.** A field guide for site identification in the Nelson Forest Region. Ministry of Forests, Nelson. 311 p.

**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. & Mgmt. 94: 129–140.

**Ferguson, D.E. and C.E. Carlson. 1993.** Predicting regeneration establishment with the Prognosis Model. USDA Forest Service. Intermountain Research Station Research Paper INT-467.

**Ferguson, D.E., A.R. Stage and R.J. Boyd. 1986.** Predicting regeneration in the grand fir-cedar-hemlock ecosystem of the northern Rocky Mountains. For. Sci. Monogr. 26. 41 p.

**Haara, A., M. Maltamo and T. Tokola. 1997.** The k-nearest neighbor method for estimating basal-area diameter distribution. Scan. J. For. Res. 12: 200–208.

**Ketcheson, M.V., T.F. Braumandl, D. Meidinger, G. Utzig, D.A. Demarchi and B.M. Wikeen. 1991.** Interior Cedar-Hemlock Zone. *In* Ecosystems of British Columbia. Special Report Series # 6. B.C. Ministry of Forests, Victoria.

**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. 2001.** Imputation and model-based updating techniques for annual forest inventories. For. Sci. 47 (3): 322–330.

**Meidinger, D. and J. Pojar. 1991.** Ecosystems of British Columbia. BC Ministry of Forests, Victoria.

**Moeur, M. 2000.** Extending stand exam data with most similar neighbor inference. In, Pioneering New Trails, Proceedings, Society of American Foresters National Convention, Sept. 11–15, 1999, Portland, OR.

**Moeur, M., N.L. Crookston and A. R. Stage. 1995.** Most similar neighbor analysis: a tool to support ecosystem management. *In* Analysis in Support of Ecosystem Management. J.E. Thompson (comp.). Analysis Workshop III, April 10–13, 1995, Fort Collins, CO. pp. 31–44. USDA, Forest Service, Ecosystem Management Analysis Center, Washington, D.C.

**Moeur, M. and R.R. Hershey. 1998.** Preserving spatial and attribute correlation in the interpretation of forest inventory data. Chapter 49. *In* K. Lowell (ed.). Spatial Accuracy Assessment: land information uncertainty in natural resources. Proceedings, 3rd International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental sciences, May 20–22, 1998, Quebec City, Quebec. pp. 419–430. Ann Arbor Press.

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

**Oliver, C.D. and B.C. Larson. 1996.** Forest stand dynamics. John Wiley & Sons, Toronto. 520 p.

**Rubin, D.B. 1987.** Multiple imputations for non-response in surveys. John Wiley & Sons, New York. 258 p.

**Smith, D.M., B.C. Larson, M.J. Kelty and P.M.S. Ashton. 1996.** The practice of silviculture. John Wiley & Sons, Toronto. 537 p.

**Snee, R.D. 1977.** Validation of regression models: methods and examples. Technometrics. 19: 415–428.

**Stage, A.R. 1973.** Prognosis model for stand development. USDA Forest Service, Res. Pap. Int – 137. 32 p.

**Temesgen, H. and V. LeMay. 2000.** Possible approaches for relating growth and yield models to aerial information. Part II: Example using existing audit data. A report submitted to the BC Ministry of Forests, Forest Practices Branch. 19 p.

**Van Deusen, P.C. 1997.** Annual forest inventory statistical concepts with emphasis on multiple imputations. Can. J. For. Res. 27: 379–384.