From plots to landscape: A *k*-NN-based method for estimating stand-level merchantable volume in the Province of Québec, Canada

by P.Y. Bernier[1,2], G. Daigle[3], L.-P. Rivest[3], C.-H. Ung[1], F. Labbé[4], C. Bergeron[4] and A. Patry[4]

ABSTRACT

Estimation of forest attributes at the stand or polygon level across the forest domain is a basic component of forest inventory programs. We tested a "*k*-Nearest Neighbours" (*k*-NN)-based method for imputing merchantable volume. Our target dataset consisted of a discrete set of forest polygons within two large forest management units, and our reference dataset was a large historical database of temporary sample plots measured over the past three decades. The linkage between the target and reference datasets was provided by polygon-level photo-interpreted forest attributes. Measurements in temporary sample plots located in all target polygons enabled us to estimate fit statistics between imputed and measured merchantable volumes. A parallel imputation exercise was also done using the current operational method used by the Province of Québec to map forest attributes over the publicly owned forest lands. Results show that the volumes estimated using the historical *k*-NN method show fit statistics similar to those of the operational method, with a slightly higher bias that is largely within the error term of the estimates. For both methods, the coefficient of determination between measured and imputed merchantable volume is between 0.16 and 0.19 for total volume, increases substantially when the volume is partitioned between hardwoods and softwoods, but then decreases when the volume is further distributed among species. The results underline the importance of photo-interpretation uncertainties in bounding the accuracy of volume imputation as well as the value of the *k*-NN procedure for imputation purposes in the context of natural forests.

Key words: Forest inventory; non-parametric methods; photo-interpretation; pre-stratification; natural forests

RÉSUMÉ

L'estimation des attributs forestiers à l'échelle du peuplement ou du polygone pour un territoire forestier est une composante de base des programmes d'inventaire forestier. Nous avons testé une méthode « *k*-plus proches voisins » (*k*-NN) pour imputer la valeur du volume marchand. Notre jeu de données cible était un ensemble de polygones forestiers dans deux unités d'aménagement forestier. Notre jeu de référence était un vaste ensemble de données de placettes temporaires mesurées au cours des derniers trente ans. Le lien entre les deux jeux a été fait à l'aide des attributs photo-interprétés à l'échelle de tous les polygones. Nous avons utilisé les mesures dans les placettes temporaires localisées dans les polygones cibles pour estimer les statistiques d'ajustement. Nous avons aussi effectué un exercice parallèle d'imputation au moyen de la méthode opérationnelle de la Province de Québec pour les terres publiques. Les résultats démontrent que l'ajustement des volumes estimés au moyen de la méthode *k*-NN historique est comparable à celui des estimés obtenus par la méthode opérationnelle, avec un biais légèrement plus élevé, mais en deçà du terme d'erreur. Le coefficient de détermination entre les volumes marchands totaux mesurés et imputés est de 0,16 à 0,19 pour les deux méthodes, est plus élevé quand le volume est réparti entre feuillus et résineux, mais décroît substantiellement quand le volume est réparti entre les essences. Les résultats montrent comment les incertitudes liées à la photo-interprétation limitent la précision des estimés de volume et la valeur de la procédure *k*-NN pour fins d'imputation dans un contexte de forêts naturelles.

Mots clés: Inventaire forestier; méthodes non-paramétriques; photo-interprétation; pré-stratification; forêts naturelles

Introduction

Systematic inventories of the public forests (85% of the forested area) have been carried out since the 1970s in the Province of Québec using a two-stage procedure. The first stage involves complete coverage of the forest domain with 1:15 000 aerial photos, followed by delineation of forest stands, or polygons, on these photos and the estimation of selected attributes by skilled photo-interpreters. The second stage involves deployment of temporary sampling plots (TSP) in a stratified field sampling program, with the objective of adequately sampling the most common forest types in a given sampling unit, based on aggregating most similar forest polygons into forest strata. Over the four 10-year forest inventory programs to date, the general trend has been to increase the number of TSPs that were deployed each year. Over the past few years, this number has plateaued at about 12 000 TSPs installed yearly.

There are significant costs associated with such a large inventory program, and questions have been increasingly raised about the actual benefits that arise from the establishment of additional TSPs. Ideally, estimation of benefits from a more accurate forest inventory can be translated into increased net present value of forest products (Borders *et al.* 2008). However, the methodology currently used to impute values of merchantable volumes over forest management units (FMUs) is not easily amenable to the attribution and

[1] Natural Resources Canada, Canadian Forest Service, Laurentian Forestry Centre, P.O. Box 10380, Stn Ste-Foy, Québec, Québec G1V 4C7.
[2] Corresponding author. E-mail: pbernier@nrcan.gc.ca
[3] Département de mathématiques et de statistique, Université Laval, Québec, Québec G1V 0A6.
[4] Ministère des Ressources naturelles et de la Faune du Québec, 880 Chemin Ste-Foy, Québec, Québec G1S 4X4.

quantification of errors and the subsequent derivation of costs versus benefits. In addition, this methodology is increasingly deemed inadequate considering the recent emphasis on optimization within the calculation of annual allowable cut, which would require a more accurate location of wood volumes on the landscape. Finally, a significant portion of the current procedure is heuristic in nature and is therefore largely irreproducible, a property that results in an overall lack of transparency.

It is within this context that we were mandated to investigate if a more formal (i.e., less heuristic) imputation procedure could be used for estimating the merchantable volume of Québec's public forests. Over the years, the forest inventory program has generated a distinctively rich database of TSPs, as well as complete decadal photo coverage of the public commercial forest domain. We therefore explored how these two unique sources of information could be used to produce estimates of forest attributes while improving on the more negative aspects of the current system (high costs and coarseness, and low reproducibility).

Over the past 15 years, forest biometricians have been developing and using *k*-Nearest Neighbours, or *k*-NN, for estimating forest attributes by combining multi-spectral imagery and field plot measurements. In this method, desired attributes of a forest stand are estimated as an average of values from *k* neighbouring stands for which these attributes are known, and where the neighbourhood is defined as a space whose multiple dimensions are defined by variables measured in all stands. The method, first proposed by Moeur and Stage (1995), was further developed in Finland (Tomppo 1997), and has since been applied to forest inventories in a number of different countries (e.g., Gjertsen *et al.* 1999, Tomppo *et al.* 1999). Since then, a significant body of literature has developed on this method, covering *inter alia* comparisons with other methods (e.g., LeMay and Temesgen 2005), use of information with different scales of resolution (Tomppo and Halme 2004), or from different sources (Tuominen *et al.* 2003, LeMay *et al.* 2008) and estimation of forest structure (Temesgen *et al.* 2003, McRoberts 2009).

One of the key features of non-parametric methods such as *k*-NN is that they allow simultaneous estimation of multiple variables and thus preserve the covariance structure among these variables (Moeur and Stage 1995). Because of such features, and of the wealth of expertise present within the expert community, we chose to investigate the potential use of *k*-NN as a formal imputation tool. In this application, the historical TSP dataset from the full forest domain covered by the forest inventory would be used as the reference dataset, and the photo-interpreted attributes would replace the multi-spectral information used in traditional *k*-NN inventory applications. We refer to this method as the "historical *k*-NN". The objectives of this work were therefore to generate estimates of forest attributes using the historical *k*-NN approach for landscapes of natural, mixed conifer-deciduous forests in Québec, and to compare the accuracy and precision of volume estimates produced by the historical *k*-NN with the estimates produced by the current operational method. We present below the most recent results of this ongoing project.

Methods

The analysis was carried out in two forest management units that contrast in forest composition (Fig. 1, Table 1). FMU 071-51 is located in a warmer portion of the province, and is dominated by temperate hardwoods and white pine (*Pinus strobus* L.) growing in generally uneven-aged stands of mixed composition. By contrast, FMU 031-51 is in a colder portion of the forest domain and is dominated by the most northern temperate hardwood, yellow birch (*Betula alleghaniensis* Britt.), growing in association with balsam fir (*Abies balsamea* [L.]), a boreal tree species. The size of both FMUs is typical of that of all other FMUs and represents the area over which forest inventories are compiled for forest management purposes.

As mentioned above, the Québec forest inventory is a two-stage process involving photo-interpretation and stratified field sampling. For a given decadal forest inventory program, the forest information gathered from photo-interpretation and field sampling is combined within geographically defined compilation units to estimate specific forest attributes of the

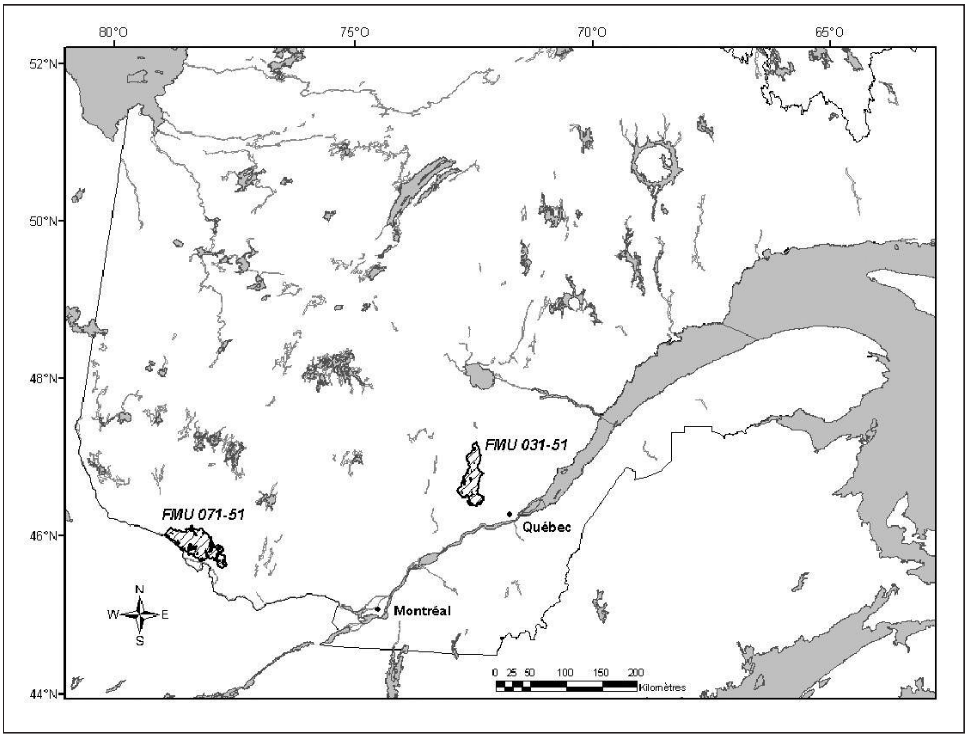

Fig. 1. Location of the two forest management units used in this study.

Table 1. Biophysical description of the two forest management units used in the present study.

Forest Management Unit (FMU)	031-51	071-51
Total area (km^2)	1590	2536
Number of temporary sample plots	1238	2468
Dominant species (% of volume)	balsam fir (26.5) yellow birch (24.7)	maples (19.8) white pine (18.6) poplars (16.6)
Mean temperature (°C)	1.5	3.7
Mean total precipitation (mm)	1243	914

strata to be used for forest management planning, including merchantable volume by species and tree age. This exercise is essentially one of imputation since only a very small fraction of forest polygons can be sampled. Polygons are therefore aggregated into strata, and small or uncommon strata into aggregated strata, until the number of plots that have been established in polygons within each stratum reaches a target number, currently set at 15. Since the number of TSPs established during the particular sampling program within a given compilation unit rarely meets this target, various heuristic procedures are called upon to select additional plots, either from similar polygons located outside the compilation unit, or from the previous forest inventory, with the latter plots first being aged to current time through an empirical modelling process. In the remainder of this text, this method is called the "operational method".

The *k*-NN technique is a statistical, non-parametric imputation method in which the desired forest attributes of a land unit are estimated as the mean of values of these desired attributes from the most similar land units for which these attributes are known (Moeur and Stage 1995, McRoberts *et al.* 2007). The land units can be forest stands or image pixels. The set of land units for which we want to estimate attributes is regarded as the target set while the smaller sample of land units within which desired tree-level and site-level attributes are known is the reference set. The degree of similarity is based on an analysis of proximity of target and reference units in a multivariate space in which the variables defining this space, and available for both target and reference sets, can be any mix of values such as spectral reflectance and forest cover attributes (Tomppo *et al.*1999, Tuominen *et al.* 2003, LeMay *et al.* 2008). The degree of similarity between selected land units and the individual target land unit is used to weigh the averaging of the desired attributes. Eskelson *et al.* (2009) provide a review of *k*-NN usage, limitations and possibilities in forest inventory applications.

In our application of the "historical *k*-NN", the reference dataset was composed of 119 449 TSPs established over the last three inventory programs since the early 1980s, accompanied by the photo-interpreted attributes of each polygon within which the plots were located. Plots measured during the first inventory program carried out in the 1970s could not be used because of irreconcilable differences in field measurement protocols. The target set was composed of photo-interpreted forest polygons within the two FMUs (Table 1), the set being limited to polygons within which we also had TSP measurements of the desired variables. The TSPs from the fourth (most recent) inventory program from these FMUs were excluded from the reference set and were used as "ground truth", providing estimates of the imputation error. In the current exercise, the desired variables were merchantable wood volume by species, by softwood and hardwood categories, and as a total of all merchantable trees.

Photo-interpretation standards have evolved over the 30 years covering the three forest inventory programs. In an earlier effort, the forest inventory branch of the MRNFQ re-photo-interpreted to current standards all forest polygons associated with each TSP, using the historical aerial photos taken at the time of plot measurement. Also, the current forest cover classification system used by photo-interpreters in the Province of Québec is related to, but not directly interpreted as, percent cover by dominant species. The photo-interpreted forest cover nomenclature was therefore re-interpreted to generate a value of percent forest cover by dominant species.

Distance metrics

The *k*-NN method is based on the quantification of similarity or distance in a given attribute domain between a geographical unit of interest (a pixel, forest stand or forest polygon) in the target dataset and the geographical units that contain the plots in the reference set. We used photo-interpreted attributes to link the historical reference dataset to the target dataset. The variables used in the distance metrics were: the dominant species or groups of species (in % as extracted from 23 possible classifications); a dichotomous variable recording whether or not the polygon has been recently disturbed (e.g., harvest, fire, windthrow); as well as class variables representing average tree height, stand age, and percentage of forest cover.

Following preliminary analyses, two distance components were retained. The first is a species composition distance (as in Legendre and Legendre 1998: 279) between polygon *i* of the reference set and polygon *p* of the target FMU:

$$[1] \quad D_{1,(i,p)} = \left[1 - \sum_{j=1}^{23} \frac{q_j^{(i)}}{\sqrt{\sum_{j=1}^{23} (q_j^{(i)})^2}} \frac{q_j^{(p)}}{\sqrt{\sum_{j=1}^{23} (q_j^{(p)})^2}} \right]$$

where variable q_j refers to the proportion q of species j or of group of species j in the polygon. The statistic *LL* in equation (7.37) of Legendre and Legendre (1998) reaches a maximum value of $\sqrt{(2)}$. The species composition distance D_1 of our metric is $D_1 = LL^2/2$, and is therefore a normalized version of *LL* that has a maximum value of 1. The second component is an abundance distance defined as:

$$[2] \quad D_{2,(i,p)} = \frac{(\hat{B}_i - \hat{B}_p)^2}{Var(\hat{B})}$$

where $\hat{B}$ is the total basal area in a plot, estimated using a regression model with linear and quadratic terms. The predictor variables were selected from a pool of 35 variables including the photo-interpreted variables (e.g., stand height, stand density, species composition, drainage index), climatic variables (degree-day temperature, mean annual total precipitation and aridity index), geographic variables (longitude, latitude, altitude, slope, aspect, growth potential and wind exposure index) and an ecological classification variable (bioclimatic sub-domain). All quadratic terms and pairwise interactions were also considered in model selection. The final model was obtained using a stepwise approach with a significance level of 0.1%, which was carried out with the SAS *glmselect* procedure (SAS Institute, Cary, NC, USA).

The complete distance metric (D_c) was calculated as:

$$[3] \quad D_{c,\,(i,p)} = D_{1,\,(i,p)} + D_{2,\,(i,p)}$$

The rationale for this metric is heuristic. Two components, species composition and abundance, were selected a priori to identify nearest neighbours. We elected to measure abundance with a prediction of the basal area, while a modified Legendre and Legendre metric was used for species composition. The two components of the metric were given the same weight; this provided the best results among all the trials that were carried out.

The total merchantable wood volume ($\hat{V}$) in the target polygon *p* was estimated as:

$$[4] \quad \hat{V}_p = \frac{\sum_{k\ neighbours} V_i / \max\left[0.001, D_{c,\,(i,p)}\right]}{\sum_{k\ neighbours} 1/\max\left[0.001, D_{c,\,(i,p)}\right]}$$

where V_i is the merchantable volume from the *i*th nearest TSP and *k* is the total number of nearest TSPs retained, the nearest TSPs being those with the smallest value of D_c with respect to the target polygon. Considering the large size of the reference set, we used $k = 30$ nearest neighbours in this initial round of tests for calculating merchantable wood volume in eq. 4. Also, the value of 0.001 used in eq. 4 prevented perfect fits ($D_c = 0$) from dominating the estimation of volumes.

Imputed values of total merchantable volume obtained using both the current operational method and the *k*-NN method were compared with actual field measurements in TSPs located in polygons within each of the two FMUs (n = 1238 in FMU 031-51 and n = 2468 in FMU 071-51). The imputations under the current operational method were verified using a leave-one-out method. The imputed value for a plot is the stratum mean value calculated without the plot used for verification. We used the bias (measured - estimated), the root mean square of the error (RMSE) and the coefficient of determination (r^2) between the predicted and observed volumes as comparison statistics:

$$[5] \quad RMSE_{FMU} = \sqrt{\frac{1}{N_{FMU}} \sum_{FMU} (\hat{V} - V)^2} \ ,$$

and

$$[6] \quad r^2_{FMU} = \left(cor(\hat{V}, V)\right)^2 \ ,$$

where *N* represents the total number of target polygons within each FMU, and $\hat{V}$ and *V* represent the predicted and measured merchantable volumes of target polygons within a given FMU.

Results and Discussion

In general, the mean and RMSE of merchantable volume estimates obtained using the historical *k*-NN compared with measured values were similar to those from the current operational method. Biases in particular are very small in both methods for the hardwood-dominated FMU 071-51 (Table 2). For the conifer-dominated FMU 031-51, the historical *k*-NN yields a somewhat greater bias for total volume than the operational approach. In all cases, however, the RMSE is much larger than the biases, meaning that these biases are within the uncertainty of the prediction. These results are very promising and show that it is possible to use data from plots measured in past decades to estimate the current attributes of forest polygons, as long as the variables (photo-interpreted attributes in our case) used to link the reference and target datasets are coherent over time.

Table 2. Observed and estimated merchantable volumes for all species combined, for hardwoods and for softwoods, for the two forest managements units, along with the root mean squared error (RMSE) and the coefficient of determination (r^2) for each imputation method.

		Mean (m^3/ha)			RMSE		r^2	
FMU	Species	Observed	Operational	*k*-NN	Operational	*k*-NN	Operational	*k*-NN
071-51	all species	183	181	180	77	79	0.18	0.16
	hardwoods	117	115	114	65	69	0.33	0.29
	softwoods	66	65	65	63	65	0.35	0.32
	red maple	36	36	38	36	33	0.27	0.26
	white pine	33	33	28	56	53	0.30	0.27
	Populus spp.	31	30	30	51	52	0.26	0.28
031-51	all species	128	127	117	48	50	0.19	0.17
	hardwoods	68	68	63	38	41	0.51	0.46
	softwoods	60	60	55	39	40	0.36	0.35
	balsam fir	34	34	25	31	33	0.26	0.22
	yellow birch	31	31	23	35	36	0.38	0.35
	white birch	18	18	18	20	20	0.31	0.31

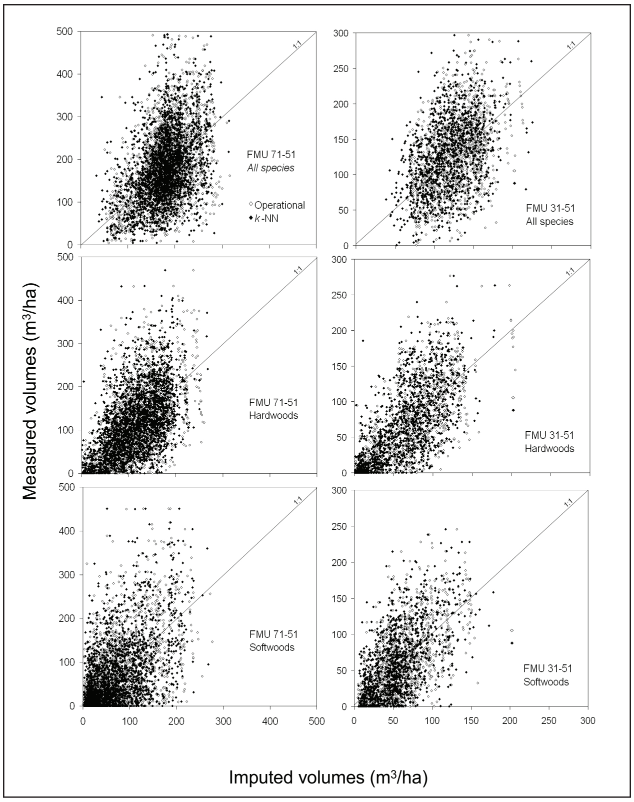

Fig. 2. Measured versus estimated merchantable volumes (m^3/ha) within each FMU for all species, for hardwoods and for softwoods, and for specific dominant species. Estimates were done using the operational method and the historical *k*-NN.

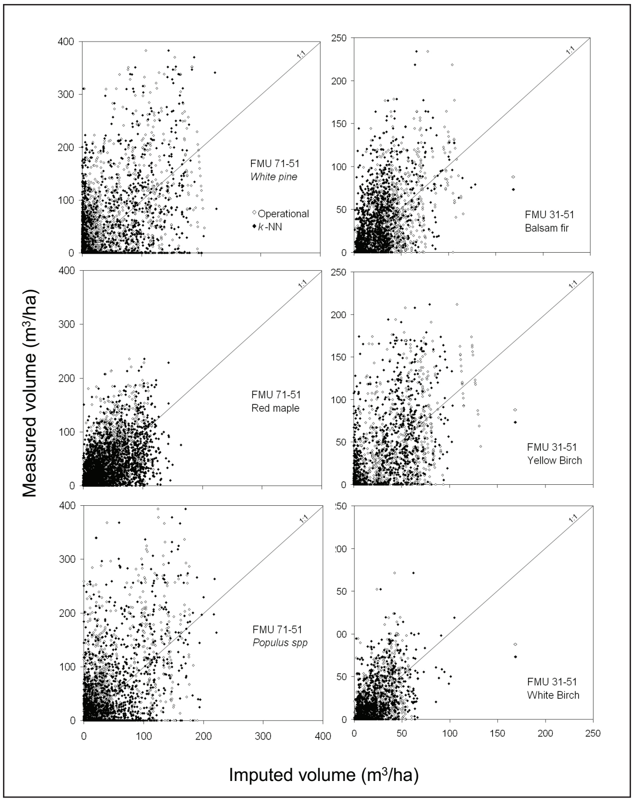

Fig. 3. Measured versus estimated merchantable volumes (m^3/ha) within each forest management unit, for white pine (*Pinus strobus*), red maple (*Acer rubrum*) and poplars and aspen (*Populus* spp.) for FMU 71-51, and for balsam fir (*Abies balsamea*), yellow birch (*Betula alleghaniensis*) and white birch (*Betula papyrifera*) for FMU 31-51. Estimates were done using the operational method and the historical *k*-NN.

Both methods show a generally low predictive capacity for total volume, all species combined, with r^2 values ranging from 0.16 to 0.19 (Fig. 2, Table 2). Their predictive capacity generally improves when the analysis is done for hardwoods and softwoods as aggregated species, with r^2 values reaching 0.29 to 0.51 in both FMUs, but declines with further splitting into individual species (Fig. 3, Table 2). The low values of r^2 for total volume reflect the relatively narrow range of observed total volumes as TSPs were selectively not established in stands with heights <7 m in past forest inventories. Splitting volume into hardwood and softwood components stretches the range of observed volumes for these individual components, thereby increasing the value of r^2. In addition, the partitioning of forest cover into its hardwood and softwood components is an easy and low-error task in photo-interpretation. However, photo-interpreting the cover of individual species likely carries a larger uncertainty, which is reflected in a lower coefficient of determination in the comparison between measured and estimated merchantable volume (Fig. 3; Table 2). Prediction errors are particularly large for secondary species with low average volumes on the landscape, a pattern that is found in both methods (Fig. 4). As mentioned above, the translation of the forest cover types into species cover density may have generated an additional source of error. However, results from a recent trial (unreported results) on a different dataset in which the photo-interpretation was redone in terms of percent cover density by species failed to show a large improvement in volume estimate. These elements suggest that the ability to estimate merchantable volume of single species or of combination of species is strongly limited by uncertainties in photo-interpreted variables.

Another type of error in species-level volume estimates in the historical *k*-NN approach is the erroneous introduction of companion species from ecologically different environments. This type of error is inherent to a method in which plots from widely different ecological regions are pooled, and all have a non-zero possibility of being selected for the imputation. A large value of *k* (=30) also favours this undesirable effect, and a reduction in this value should reduce the possibility of introduction of companion species from other ecological environments. The operational method contains a number of heuristic rules designed specifically to address this problem and therefore performed better on that score.

The results also show a distinct volume-related bias in the polygon-level imputation results (Fig. 2). This is particularly apparent for FMU 071-51 where, in general, large polygon-level volumes are underestimated and low polygon-level volumes are overestimated. This trend reflects the averaging nature of both procedures, a property that tends to introduce biases in extreme values (McRoberts *et al.* 2002). We used k = 30 in this initial application of the *k*-NN method. Reducing

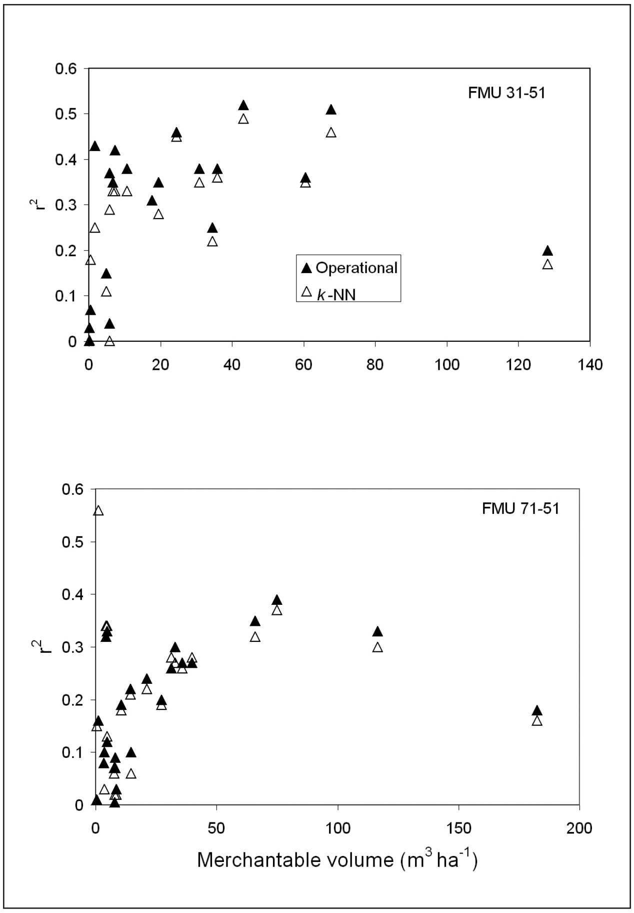

Fig. 4. Coefficients of determination between imputed and measured merchantable volumes for individual species, combination of species (e.g., conifer and deciduous), and total merchantable area (largest values to the right), plotted against measured merchantable volumes for both forest management areas (FMU) and imputation procedures.

this value would increase the retention of natural variability in volume estimates (Franco-Lopez *et al.* 2001) but may result in lower values of r^2. This averaging effect is less of a concern for the operational methods since the estimates are *de facto* produced for aggregated strata and are not intended to capture the polygon-level variability in merchantable volume.

Operational applications of *k*-NN to forest inventory have been mostly performed using pixel-level, multi-spectral data to link the target set of geographical units to the reference set of field plots (e.g., Tomppo 1997, McRoberts *et al.* 2007). However, a number of studies have also explored the use of forest cover information (Moeur and Stage 1995, Korhonen and Kangas 1997, Temesgen *et al.* 2003) for defining the variable space. Of particular interest is the study by LeMay and Temesgen (2005) who have also used a mixture of photo-interpreted and modeled variables as their spatial dataset within a comparison of nearest-neighbours imputation methods. In particular, they used estimates of merchantable volume as predicted from the provincial growth and yield model as a spatial variable because of the operational use of this variable as a stratification criterion in their jurisdiction. The historical approach adopted in the present work was very similar, with the use of both photo-interpreted (e.g., forest composition) and modeled (basal area) variables. Also, our choice to forfeit multi-spectral data was based on the decision to exploit the large historical pool of TSPs as a reference dataset. The historical nature of this dataset made the retrospective acquisition of time-relevant multi-spectral data either very difficult or impossible.

Our results suggest that the accuracy of both historical *k*-NN and operational imputation methods is limited by the uncertainty in photo-interpreted attributes that we have used to link target and reference polygons, rather than by the lack of field plots. Comparing the polygon-level results of the *k*-NN method and the operational method (Fig. 2) shows that the ability to predict polygon volumes, even for common species and forest types for which abundant sample plots exist, remains equally limited. What is also striking is that the pattern of predicted versus measured volumes for both methods is very similar, although we can presume that the set of plots used to impute attributes to any given polygon would have been at least partially different. In addition, because of the large number of plots in the reference dataset, perfect matches were numerous (eq. 3, $D_c = 0$) between target and reference polygons for the *k*-NN approach, but yet still resulted in only moderate accuracy in the volume predictions. It is clear from these observations that gains in accuracy or cost reductions can be achieved through a revision of the roles of both photo-interpretation and of TSP establishment in the procedure for imputing forest attributes across large landscapes.

First and foremost, the historical bank of TSPs now provides a foundation for basic imputation of forest attributes. Establishment of a more limited number of new TSPs could be done for correcting the imputation biases arising from a historical *k*-NN exercise, or evaluating the imputation errors within a given ecological unit. New TSPs could also be established to fill specific gaps, identified in the historical dataset through rigorous statistical analysis, for the characterization of less common forest types or of new forest structures created by forest management. However, this should be done while keeping in mind the limits placed on the imputation method by the use of photo-interpreted attributes. A corollary to this limitation is that significant gains in accuracy in estimated forest attributes could likely be generated by improving the quality and rigour of the variables extracted from current and historical aerial photos. Again, the use of the historical TSP database becomes a limiting factor as spectral band analysis can only be done with current numerical aerial photos. Nevertheless, other projects are showing that indices based on the image texture of aerial photos can yield quantitative information that could possibly improve *k*-NN estimates (Tuominen and Pekkarinen 2004, Miranda *et al.* 2009).

The results presented above stem from a preliminary project aimed at identifying the potential of the historical *k*-NN method in the context of the Québec inventory. One of the challenges of the *k*-NN method lies in the determination of the proper distance metrics. Those chosen for the present exercise are likely not optimal, but several additional tests (results not shown) carried out using variants of these metrics, or the introduction of different metrics based, for example, on drainage class or climate, have improved the results only to a limited degree. Further testing is ongoing with more advanced multivariate approaches that may eliminate some of the subjectivity in the choice of distance metrics components or formulation.

The preliminary results presented above already show that the historical *k*-NN method has a number of advantages over the current operational imputation method. In addition to potentially reducing the importance and costs of the field inventory program, the historical *k*-NN method circumvents the need for pre-stratification of polygons into strata, which is currently done only to facilitate the matching of plots to polygons, or strata in this case. Once polygons of identical photo-interpreted attributes have been grouped into strata, the limit to the number of available TSPs forces the analyst to further aggregate these strata into larger, more heterogeneous "aggregated" strata in order to get a match with enough TSPs for the imputation of attributes. Downstream users of this information must then live with the consequences of these largely heuristic aggregation choices, a situation that may affect their own use of the data.

The second and related advantage is that the historical *k*-NN approach can be used to develop a formal and rigorous imputation methodology, and can therefore help eliminate subjective decisions made by domain experts during individual imputation exercises. A move to a more quantitative imputation method improves the transparency and repeatability of the procedure, facilitates the diffusion and acceptance of the results, and greatly speeds up the whole imputation process by shifting the burden of the task from expert analysis to computer processing.

As discussed in many previous studies, the non-parametric *k*-NN method itself presents some distinct advantages as compared with parametric imputation methods such as those based on regression models. The first is that many variables can be estimated at once from the same weighted average of data from most similar land units, preserving the co-variance structure among these imputed variables (Moeur and Stage 1995, LeMay and Temesgen 2005). The second is that estimates are provided directly from observations, preventing *de facto* the possibility of extrapolation, as could be the case with parametric methods (Eskelson *et al.* 2009). A final advantage is that the types of variables that can be imputed go beyond traditional forestry interests, as is shown by Temesgen *et al.* (2008), who use the *k*-NN approach to estimate cavity tree abundance. This capacity to impute non-traditional forest attributes increases the usefulness of the method. For the benefit of potential users, Crookston and Finley (2008) have developed a freeware *k*-NN application written in R.

As stated above, one of the results of the analysis is that further investments in TSP establishment for feeding the primary imputation exercise would generate marginal gains in accuracy within the context of this historical application of the *k*-NN method. This statement will certainly be true for the most common forest types or forest structures, but may not hold completely for uncommon forest types or especially the novel forest structures that are being created by forest management. The quantification, attribution, and mapping of imputation errors from the *k*-NN procedure based on the estimated standard deviation of the neighbour attributes likely provide a realistic quantitative basis for determining forest inventory needs through a cost–benefits analysis of investments.

Conclusion

We believe that the historical *k*-NN approach presents a viable alternative to the current operational method used for imputing forest attributes in the Province of Québec. However, the analysis of polygon-level results clearly shows that the spatial accuracy of the results from both the operational and the historical *k*-NN methods is limited by the accuracy of photo-interpretation. Currently, tests are being carried out on the application of the *k*-NN method to data from airborne LiDAR and multi-spectral airborne scanners, and the preliminary results (not shown) reveal a large increase in accuracy for mapping merchantable volumes. It may be that the benefits of this increased accuracy may rapidly surpass those offered by the use of a large historical TSP database.

The work on the historical *k*-NN approach has demonstrated the possibility of avoiding the pre-stratification exercise of grouping photo-interpreted forest polygons into strata, thus making the final results more amenable to the needs of users. However, the segmentation of the forest cover by photo-interpreters is in itself a pre-stratification, and the heuristic rules on which it is based may vary from one photo-interpreter to the next. The possible move to a *k*-NN approach based on quantitative data from multi-spectral images and LiDAR (e.g., Falkowski *et al.* 2010) pushes the idea further. The availability of a continuous coverage of numerical variables opens the way to pixel-level imputation of forest attributes, and the production of maps that can be post-stratified to meet the user's demands. This is already done in some countries such as Finland, but the adoption of such an approach would be a major paradigm change in Canadian forest landscape mapping.

Acknowledgements

The authors wish to thank Jean-Gabriel Élie, Louis-Vincent Lemelin, Sylvain Bernier and David Pothier of Université Laval for their comments on early versions of the text, and Jean-Gabriel Élie for the preparation of Fig. 1. Lévis Côté, Rémi Gagnon and Robert Jobidon played crucial roles in the early orientation of this work. Special thanks are also extended to Dr. Erkki Tomppo for his much-appreciated interaction with our group, as well as to Valerie LeMay and an anonymous reviewer for their very helpful comments on an earlier version of this text. Funds for carrying out the *k*-NN analysis were provided by the *Ministère des Ressources naturelles et de la Faune du Québec*.

References

Borders, B.E., W.M. Harrison, M.L. Clutter, B.D. Shiver and R.A. Souter. 2008. The value of timber inventory information for management planning. Can. J. For. Res. 38: 2287–2294.

Crookston, N.L. and A.O. Finley. 2008. yaImpute: An R package for *k*NN imputation. J. Stat. Software 23: 1–16.

Eskelson, B.N.I., H. Temesgen, V. LeMay, T.M. Barrett, N.L. Crookston and A.T. Hudak. 2009. The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases. Scand. J. For. Res. 24: 235–246.

Falkowski, M.J., A.T. Hudak, N.L. Crookston, P.E. Gessler, E.H. Uebler and A.M.S. Smith. 2010. Landscape-scale parameterization of a tree-level forest growth model: A *k*-nearest neighbor imputation approach incorporating LiDAR data. Can. J. For. Res. 40: 184–199.

Franco-Lopez, H., A.R. Ek and M.E. Bauer. 2001. Estimation and mapping of forest stand density, volume, and cover type using the *k*-nearest neighbors method. Remote Sens. Environ. 77: 251–274.

Gjertsen, A.K., E. Tomppo and S. Tomter. 1999. National forest inventory in Norway: using sample plots, digital maps, and satellite images. Proc. International Geoscience and Remote Sensing Symposium (IGARSS) 2. pp. 729–731.

Korhonen, K.T. and A. Kangas. 1997. Application of nearest-neighbour regression for generalizing sample tree information. Scand. J. For. Res. 12: 97–101.

Legendre, P. and L. Legendre. 1998. Numerical ecology, 2nd English edition. Elsevier Science BV, Amsterdam. 853 p.

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

LeMay, V., J. Maedel and N.C. Coops. 2008. Estimating stand structural details using nearest neighbor analyses to link ground data, forest cover maps, and Landsat imagery. Remote Sens. Environ. 112: 2578–2591.

McRoberts, R.E. 2009. A two-step nearest neighbors algorithm using satellite imagery for predicting forest structure within species composition classes. Remote Sens. Environ. 113: 532–545.

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 Sens. Environ. 82: 457–468.

McRoberts, R.E., E.O. Tomppo, A.O. Finley and J. Heikkinen. 2007. Estimating areal means and variances of forest attributes using the *k*-Nearest Neighbors technique and satellite imagery. Remote Sens. Environ. 111: 466–480.

Miranda, M., C.-H. Ung, L. Guindon, A. Condal, P. Bernier, A. Beaudoin and A. Patry. 2009. A simple Mahanalobis distance in the kNN method using aerial photograph for predicting stand table in a boreal mixed forest. *In* R. Fournier and R. McRoberts (eds.). Extended forest inventory. Proceedings of the IUFRO- Division 4 conference, Quebec City, May 19–22, 2009. 5 p.

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

Temesgen, H., V.M. LeMay, K.L. Froese and P.L. Marshall. 2003. Imputing tree-lists from aerial attributes for complex stands of south-eastern British Columbia. For. Ecol. Manag. 177: 277–285.

Temesgen, H., T.M. Barrett and G. Latta. 2008. Estimating cavity tree abundance using nearest neighbor imputation methods for western Oregon and Washington forests. Silva Fenn. 42: 337–354.

Tomppo, E. 1997. Recent status and further development in the Finnish multi-source inventory. Lectures given at the 1997 Marcus Wallenberg Prize Symposium, Stockholm, Sweden, October 14, 1997. pp. 53–68.

Tomppo, E., C. Goulding and M. Katila. 1999. Adapting Finnish multi-source forest inventory techniques to the New Zealand preharvest inventory. Scand. J. For. Res. 14: 182–192.

Tomppo, E., and M. Halme. 2004. Using coarse scale forest variables as ancillary information and weighting of variables in the k-NN estimation: a genetic algorithm approach. Remote Sens. Environ. 92: 1–20.

Tuominen, S. and A. Pekkarinen. 2004. Performance of different spectral and textural aerial photograph features in multi-source forest inventory. Remote Sens. Environ. 94: 256–268.

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.