Application of the k-NN Estimation Technique to the Quebec Forest Inventory

Gaétan Daigle[1], Louis-Paul Rivest[1], Carl Bergeron[2], Sylvain Bernier[2], Pierre Bernier[5], Lévis Coté[4], Rémi Gagnon[3], François Labbé[2], Anick Patry[2] and Chhun-Huor Ung[5]

[1] Department of Mathematics and Statistics, Université Laval, 1045 rue de la médecine Québec (Québec) Canada G1V 0A6 Gaetan.Daigle@mat.ulaval.ca ,Louis-Paul.Rivest@mat.ulaval.ca

[2] Direction des inventaires forestiers, Ministère des Ressources naturelles et de la Faune 880, chemin Sainte-Foy, 5e étage Québec (Québec) G1S 4X4 Sylvain.Bernier@mrnf.gouv.qc.ca Carl.Bergeron@mrnf.gouv.qc.ca, Francois.Labbe@mrnf.gouv.qc.ca, Anick.Patry@mrnf.gouv.qc.ca

[3] Direction de l'aménagement des forêts publiques et privées, Ministère des Ressources naturelles et de la Faune 880, chemin Sainte-Foy, 5e étage Québec (Québec) G1S 4X4 Remi.Gagnon@mrnf.gouv.qc.ca

[4] Direction générale région Bas-Saint-Laurent, Ministère des Ressources naturelles et de la Faune, 92, 2e rue Ouest, Rimouski (Québec) G5L 8B3 Levis.Cote@mrnf.gouv.qc.ca,

[5]Natural Resources Canada, P.O. Box 10380, Québec, Québec, CanadaG1V 4C7 pbernier@rncan.gc.ca , Chhun-Huor.Ung@RNCan-NRCan.gc.ca

Introduction

The Quebec forest inventory is currently carried out using a stratified sampling design. Aerial photos of a target Forest Management Unit (FMU) are first taken. The areas of such FMUs vary between 160 km^2 and 27 000 km^2. Photo-interpreters divide the territory into homogeneous polygons and, for each of those polygons, estimate the value of specific biological and physical variables. The most pertinent of these for the current exercise are listed in Table 1. Once the photo-interpretation is completed, polygons with similar photo-interpreted characteristics are grouped into strata. There can be more than 1000 strata in a FMU. The stratification is used to direct the field inventory in which 400 m^2 plots are established in stratum and in which forest properties such as basal area and merchantable volume per species are measured. The target sample size is of 15 per stratum, but only a fraction of the plots required can be established within the target FMU. Several plots are therefore recruited from other FMUs or from previous inventories through heuristic rules. Field measurements are taken in the sampled plots such as the basal area and wood volume by species. Plots are assigned to strata, and strata-level forest properties are imputed from plot averages.

Table 1. Photo interpreted variables collected on a FMU

The dominant species or groups of species in % (23 possible classifications)
The average height (in 5 meters class)
The average age of the stand (a class variable)
A dichotomic variable recording whether the polygon is disturbed
The percentage of forest cover (a class variable)

As can be seen from the synthetic description of the current forest inventory practice in Québec, the photo-interpreted variables are used only for constructing the strata. Can they be used more efficiently? Can the k-NN method be applied in this context?

The k-NN technique, as reviewed by McRoberts & al. (2007), is a statistical imputation method that can be used to integrate satellite or photo image data in the estimation of field characteristics in a forest inventory. The relatively large satellite or photo image sample is regarded as the target set while the smaller sample of field plots is the reference set. Field measurements in the target set are predicted using nearest neighbors which are selected using satellite or photo image variables, in the reference set. Thus, in a typical k-NN application, the reference set contains the sampled plots where field measurements are taken.

This work investigates nearest neighbors techniques as a way to incorporate historical plot data in the estimation of the field characteristics of a FMU. More than 100 000 field plots have been visited through successive forest inventories, and all forest polygons in which those plots were located have been re-photo-interpreted to current standards. The polygons for which photo-interpreted variables and field measurements are available are contained in the Quebec file. The aim of this paper is to investigate the quality of the predictions of the field measurements for a target photo-interpreted polygon obtained by selecting nearest neighbors in the Quebec file.

Distance Metrics

The reference Quebec file contains data from plots, and associated polygon-level photo-interpreted properties, sampled throughout the 570 000 km^2 area covered by the intensive forest inventory in Québec. The development of suitable metrics is crucial for getting good nearest neighbors. Besides the photo-interpreted variables presented in Table 1, geophysical variables on soil composition and drainage, and climatic variables on precipitation and temperature (Ung et al., 2001) are available for constructing a distance. For each polygon in the Quebec file, the climatic variables were calculated using BioSIM (http://dsp-psd.pwgsc.gc.ca/collection_2008/nrcan/Fo113-3-134E.pdf). A hierarchical ecological classification of the territory with 8 levels described in Saucier et al. (1998) is also considered; a description can be found at http://www.mrnf.gouv.qc.ca/publications/forets/connaissances/Systeme.pdf.

Three distance components were identified and a measure for each one was developed. First a species composition distance due to Legendre and Legendre (1998, p.279) was retained. The distance between polygon *i* of the Quebec file and polygon k of the target FMU is

$$D_1(i,k) = \left[1 - \sum_{j=1}^{23} \frac{p_j^{(i)}}{\sqrt{\sum_1^{23} (p_\ell^{(i)})^2}} \frac{p_j^{(k)}}{\sqrt{\sum_1^{23} (p_\ell^{(k)})^2}} \right],$$

where the p_j's refer to the proportion of species, or of group of species *j* in the polygon. Second, an abundance distance was defined as

$$D_2(i,k) = \frac{(\widehat{BA}_i - \widehat{BA}_k)^2}{\mathrm{Var}(\widehat{BA})},$$

where $\widehat{BA}$ is a prediction, using a regression model on photo-interpreted variables, of the basal area in a plot, and $\mathrm{Var}(\widehat{BA})$ is the variance of these predictions in the Quebec file. Finally, an ecological distance using six levels of Saucier et al. (1998) classification was calculated using Jaccard metric, see Legendre and Legendre (1998, p. 256), which is defined as

$$D_3(i,k) = 1 - \frac{\#\text{levels where i and k are equal}}{6}.$$

Observe that the distances D_1 and D_3 vary in [0,1] while D_2 takes positive values that are typically less than 1. The findings presented in the result section are obtained with the distance $dist(i,k) = D_1(i,k) + D_2(i,k) + D_3(i,k)$. Considering the large size of the Quebec file, *k=30* nearest neighbors were used for calculating k-NN predictions of species volume using weights inversely proportional to the distance. The formula used for predicting a wood volume in target polygon *k* is

$$\widehat{Vol_k} = \frac{\sum_{30\text{ neighbors}} Vol_i / \max[0.001, dist(i,k)]}{\sum_{30\text{ neighbors}} 1/\max[0.001, dist(i,k)]}.$$

Table 2. Evaluation, in terms of bias and correlation, of the k-NN predictions obtained with the D_1+D_2+D_3 distance and *k*=30 nearest neighbors in three FMUs.

		Mixedwood			Softwood			Hardwood		
	SPECIES	$\overline{VOL}$	bias	Cor	$\overline{VOL}$	bias	cor	$\overline{VOL}$	bias	cor
Hardwood	Yellow Birch	30.9	-4.8	0.5685	6.2	-0.9	0.6850	10.6	-0.3	0.4391
	White Birch	17.8	-2.6	0.5514	13.3	1.7	0.5600	8.6	1.5	0.1665
	Red Oak	0.0	0.0	NA.	0.0	0.0	NA	14.5	-0.9	0.4639
	Red Maple	5.0	-0.1	0.3737	3.6	-1.2	0.5245	14.5	-0.6	0.2979
	Sugar Maple	5.8	-1.0	0.5765	1.2	-0.4	0.3966	21.0	2.8	0.4784
	Beech	1.6	-0.3	0.6751	0.0	0.2	NA.	8.0	-1.4	0.2407
Softwood	White Spruce	5.8	-0.9	0.1958	7.7	-1.5	0.1105	7.9	-0.1	0.2387
	Black Spruce	7.3	3.0	0.6093	19.4	-0.3	0.6133	4.2	-0.4	0.6039
	White Pine	0.0	0.4	NA	1.1	-0.7	0.3430	33.7	-1.8	0.5534
	Red Pine	0.0	0.0	NA	0.3	-0.2	0.0608	3.5	0.3	0.3327
	Eastern Hemlock	0.4	0.0	0.6211	0.0	0.0	NA	3.7	-2.1	0.2487
	Balsam Fir	34.3	0.0	0.4661	17.9	4.1	0.4082	7.6	0.8	0.2870
	Eastern White Cedar	0.0	0.8	NA	1.2	-0.3	0.2125	4.7	1.0	0.3777
Combi.	Maple	10.8	-1.1	0.6124	4.7	-1.6	0.5569	35.5	2.2	0.4982
	Poplar	6.9	-1.3	0.5747	12.1	-1.3	0.6129	31.3	-1.8	0.5494
	Spruce	19.3	-3.9	0.5250	19.6	0.9	0.6097	4.6	0.1	0.6156
Association	Hardwood on humid station	35.9	-4.6	0.5637	10.6	-2.6	0.6314	27.1	-1.2	0.4633
	Hardwood intolerant to shade	24.7	-3.9	0.6648	25.5	0.4	0.6517	40.0	-0.4	0.5531
	Hardwood tolerant to shade	43.3	-6.1	0.7094	10.9	-2.3	0.7186	74.7	-2.4	0.6082
	HARDWOOD	68.0	-9.9	0.6258	36.5	-2.0	0.6973	116.4	-3.1	0.5427
	SOFTWOOD	60.1	-3.5	0.5742	48.5	2.8	0.4410	67.0	-1.6	0.5913
	TOTAL	128.1	-13.4	0.4513	85.0	0.7	0.4000	183.4	-4.7	0.4448

Evaluation of the Historical Nearest Neighbors Predictions

k-NN predictions are evaluated using two statistics, the mean bias and the correlation of the volume predicted by the nearest neighbors with the observed volume, calculated as follows

$$bias(FMU) = \sum\nolimits_{FMU} (\widehat{Vol} - Vol) / N(FMU) \text{ and } cor(FMU) = cor(Vol,\ \widehat{Vol}),$$

where *N(FMU)* is the number of target plots in a FMU. These comparisons use three FMUs surveyed in 2007 whose polygons have been excluded from the Quebec file. They are respectively, a mixedwood FMU composed of both hardwood and softwood species, with *N(FMU)*=1238 polygons where field measurements are available, a softwood FMU, with *N(FMU)*=1498 polygons, and a hardwood FMU with *N(FMU)*=2468 sampled polygons.

Results

Table 2 presents the average volume in m^3 per hectare ($\overline{VOL}$) for 21 wood species, or combinations of species, for each of the three FMUs. The biases and the correlations of the k-NN predictions are also presented. The biases are generally small except in the mixedwood FMU where the volume of hardwood is underestimated by more than 10%. The average correlations for species or combinations of species whose volume is larger than 5m^3/ha are respectively 0.54, 0.55 and 0.44 in the three FMUs. They are smaller in the hardwood FMU as this forest type is not represented as well as the others in the Quebec file.

k-NN predictions obtained with the D_1+D_2 distance, which excluded the ecology component, were also calculated and evaluated. The detailed results are not presented here. The D_1+D_2 distance yielded larger biases, especially in the softwood FMU where wood volume was over-estimated by about 15%. The average correlations were slightly smaller with the *D1+D2* distance, especially in the hardwood FMU where the average loss was 0.025. Thus, the addition of the ecology classification to the distance measures yields better predictions.

Discussion

Several empirical evaluations of the k-NN method for imputing ground level measurements using photo-interpreted variables have reported important biases and large root mean square errors, see for instance Temesgen, LeMay, Froese, and Marshall (2003) and Lemay and Temesgen (2005). Despite this fact, results from the present study (Table 2) shows that k-NN gives some pre-survey information on wood volume that could be useful for a forest inventory. The techniques presented in this paper can be applied to obtain k-NN predictions for all the polygons of a FMU, before taking ground measurements. They have the advantage of bypassing the pre-stratification of forest polygons required in the current forest inventory process. Following those predictions, an inventory would still be necessary to calibrate the pre-survey predictions.

Before applying this technique on a larger scale, alternatives to the distance presented in this paper can be investigated. Combining the composition and the abundance distances is an interesting idea suggested by V. Lemay. Calculating the distance using the first two or three principal components of all the available variables could also be tried. Moreover, the genetic algorithm of Tomppo and Halme (2004) could possibly be applied to optimize the k-NN results.

Conclusion

This work has shown that historical data provides consistent information on wood volume per species for the individual photo-interpreted polygons delimited during the Quebec forest inventory process. The k-NN method can be used to extract this information. The results of the preliminary investigation presented here are good enough for investigating the estimation of wood volumes in a FMU using k-NN predictions.

Literature Cited

Legendre, L. and Legendre, P. 1998. Numerical Ecology. Second Edition. Elsevier Science

Lemay, V. and Temesgen, H. 2005. Comparison of Nearest Neighbor Methods for Estimating Basal Area and Stems per Hectare Using Aerial Auxiliary Variables. Forest Science, 51,109-119

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

Saucier, J.-P., Bergeron, J.-F., Grondin, P., and Robitaille, A. 1998. Les régions écologiques du Québec méridional (3e version): un des éléments du système hiérarchique de classification écologique du territoire mis au point par le ministère des Ressources naturelles du Québec. L’Ordre des Ingénieurs forestiers du Québec, Québec, Qc. Aubelle 124.

Temesgen H., LeMay, V. M., Froese, K.L. and Marshall, P.L. 2003. Imputing tree-lists form aerial attributes for complex stands of south-eastern British Columbia. Forest Ecology and Management, 177, 277-285

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

Ung, C.-H., P. Y. Bernier, F. Raulier, R. A. Fournier, M.-C. Lambert and J. Régnière 2001. Biophysical site indices for shade tolerant and intolerant boreal species. Forest Science 47: 83-95.