SUPPORT VECTOR MACHINES REGRESSION FOR ESTIMATION OF FOREST PARAMETERS FROM AIRBORNE LASER SCANNING DATA

J.-M. Monnet[*]*, F. Berger*

Cemagref, UR EMGR
2 rue de la Papeterie-BP 76
F-38402 St-Martin-d'Hères, France

J. Chanussot

GIPSA-Lab
Grenoble Institute of Technology, BP 46
38402 Saint Martin D'Heres, France

1. INTRODUCTION

Numerous studies have shown the accuracy and efficiency of airborne laser scanning (ALS) for estimation of forest stand parameters [1]. One of the widely-used processing method is the so-called area-based method. It consists in relating forest parameters to several height and density metrics derived from the laser point cloud in fixed areas [2]. Whatever the forestry context, most of the studies relied on ordinary least squares to establish relationships between laser metrics and forest parameters. However parametric methods reach their limits when dealing with a small number of field observations combined with high dimensional data. Such cases tend to occur frequently when laser scanning data is acquired over mountainous forests. Indeed, the lack of accessibility hamper field inventories whereas numerous laser metrics may be extracted from the point cloud. k-most similar neighbor method has been successfully tested for species-specific stand attributes estimation from laser data [3], opening ways to investigate the potential of other non parametric methods.
Support vector machines are a training approach based on the framework of statistical learning theory. They have proven their robustness to dimensionality and generalization abilities [4] and thanks to the kernel trick non-linear relationships can be accounted for. In this paper we aim at comparing accuracies of forest parameters estimates obtained with ordinary least squares multiple regression and support vector regression (SVR). The sensitivity of these techniques to the number of laser metrics combined with dimension reduction by principal component analysis (PCA) or individual component analysis (ICA) has also been investigated.

2. MATERIAL

The study area is a 4 km^2 hillside situated in the French Alps. The forest is mainly constituted of coppice stands and deciduous stands on poor quality sites. From September to November 2009, 31 circular field plots were georeferenced and inventoried. All trees with diameter at breast height larger than 5 cm and located within 10 m radius from the plot center were calipered. 10 tree heights were sampled on each plot. The following forest parameters were then computed for each plot: mean diameter at breast height (dbh), basal area (G), stem density (N) and dominant height (H_{dom}).
The laser data was acquired with an airborne RIEGL LMS-Q560 scanner on August 27th, 2009. Final average scanning density was 2.8 pulses.m^{-2}. The point cloud was classified into ground and non-ground echoes using the TerraScan software.

3. METHODS

For each plot, laser points within 10 m horizontal distance form the plot center were extracted. Their relative heights were computed by subtracting the terrain height at their orthometric coordinates. Terrain surface was estimated by bilinear interpolation of points classified as ground points. Points with relative height lower than 2 m were excluded. Three point groups were then constituted according to the return position of the echoes: single echoes (only one echo for a given pulse), first echoes and last echoes. For each group three types of laser metrics were calculated. Height metrics correspond to breakpoints of height bins containing an equal number of points, plus mean height. Density metrics were computed as the values of the cumulative density in height bins of equal width. Entropy metrics were calculated as the entropy of the orthometric distribution of points included

[*]Thanks the Région Rhône-Alpes for doctoral fellowship.

in height bins of equal width. All calculations were performed with R 2.10.0 statistical software.
A set of independent predictors $(v_i)_{i\in\{1,...,n_v\}}$ is thus composed of $n_v = 3 \times (n_h + n_d + n_e)$ laser metrics, where n_h is the number of height breakpoints plus one (for mean height), n_d the number of density bins and n_e the number of entropy bins. When the number of observations $N = 31$ was greater than the number of variables n_v, PCA and ICA were performed to reduce dimension. The obtained principal and independent components were also used as sets of predictors. For each dependent variable $y \in \{dbh, G, N_s, H_{dom}\}$ and each predictors set (v_i), the resulting training data $\{(x_1, y_1), ..., (x_N, y_N)\} \subset \mathbf{R}^{n_v} \times \mathbf{R}$ was used to fit a multiple regression model by ordinary least squares:

$$y = b + \sum_{i=1}^{n_v} a_i \times v_i \quad \text{with } (v_i)_{i\in\{1,...,n_v\}} \text{ a set of predictors and } ((a_i)_{i\in\{1,...,n_v\}}, b) \text{ the model parameters.} \tag{1}$$

Models including a maximum of four predictors were tested by exhaustive search. Models which did not fulfill the linear model assumptions or including a predictor with a partial p-value greater than 0.05 were discarded. For each predictors set the model with the highest adjusted coefficient of determination (adjusted R^2) was selected. The data sets were also used to train an ϵ-SVR. The algorithm approximates a function $f : y = f(v)$ with a solution of the form:

$$f(v) = \sum_{j=1}^{n} \alpha_j k(v, x_j) + \beta \tag{2}$$

with $(x_j)_{j\in\{1,...,n\}}$ samples from the training set, $((\alpha_j)_{j\in\{1,...,n\}}, \beta)$ parameters determined during the training process and k a kernel function. Linear and radial basis kernels were tested. Hyperparameters were selected by tuning over a range of a priori values. Multiple and ϵ-SV regression accuracies were evaluated in leave-one-out cross validation by computing the root mean square error ($RMSE$) and its coefficient of variation (CV_{RMSE}):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} o(y_i - \hat{y}_i)^2}{N}} \ , \ CV_{RMSE} = \frac{RMSE}{\bar{y}} \quad \text{with } \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{3}$$

where y_i and $\hat{y}_i$ are the observed and predicted values, and N the number of observations. Table 1 summarizes all tested combinations of parameters and methods.

Parameters	Number of height metrics n_h	Number of density metrics n_d	Number of entropy metrics n_e	Variables transformation	Number of extracted components
Values	$\{6, 8\}$	$\{0, 1, 3\}$	$\{0, 2\}$	none {PCA, ICA}*	- $\{2, 3, 4, 6, 8, 12, 16, 20, 24, all\}$*

Table 1. Independent variable sets used to fit the regressions (*when relevant).

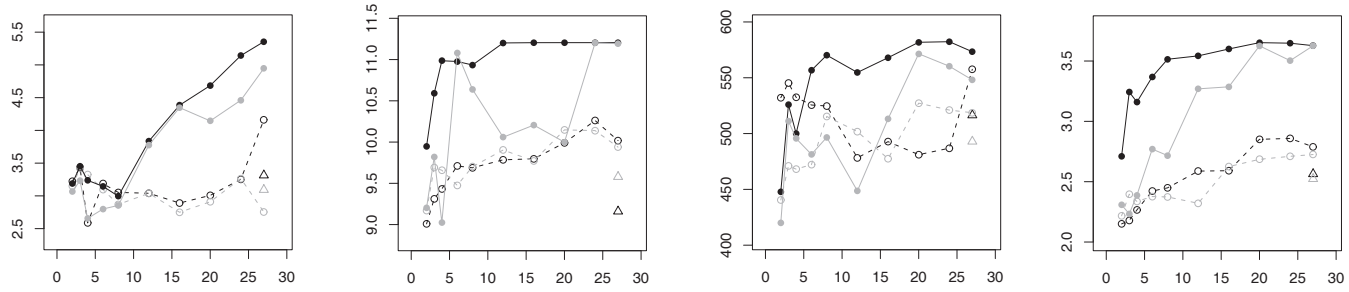
4. RESULTS AND DISCUSSION

Prediction estimates by multiple linear regression yielded satisfactory results. For the variables set with 27 laser metrics without dimension reduction $(n_h, n_d, n_e) = (6, 3, 0)$, the coefficient of variation of the RMSE ranged from 13.9 to 21.2%. The best result is achieved for dominant height whereas stem density performed poorly. Mean diameter and basal area obtained intermediate values (18.8 and 21.2% respectively). These results are similar to those obtained in a study carried on 34 deciduous plots located in the Bavarian Forest National Park (Germany) [5]. Dimension reduction improved slightly the accuracy for dominant height only ($CV_{RMSE} = 13.5\%$ with 12 components from PCA). Table 2 summarizes the best results obtained with multiple and ϵ-SV regressions for the predictors sets derived from $(n_h, n_d, n_e) = (6, 3, 0)$. Apart from mean diameter, multiple regression performed better than ϵ-SVR. However obtained values were rather close, except for basal area.
Figure 1 illustrates the effect of dimension reduction and kernel selection on ϵ-SVR accuracy for the same predictors sets $(n_h, n_d, n_e) = (6, 3, 0)$. Drastic dimension reduction (number of components less than 5) yielded better accuracies than with the original predictors set. Best results were obtained with PCA, except for stem density. Accuracy tends to decrease when the number of predictor increases further than 5. However, ϵ-SVR seems less sensitive to the number of components when PCA is employed instead of ICA. Even tough the best individual results were mostly obtained with linear kernel, radial kernel seems

more robust regarding the type and number of components included in the predictors sets. Stem density turned out to be the most complex case to interpret, as well as the most difficult parameter to estimate, as pointed out in other studies [2][5].
Figure 2 depicts the influence of the number and type of laser metrics on the prediction accuracy. ϵ-SVR is generally less accurate than multiple regression. However its results tend to be more stable, in particular with radial kernel. An improvement in basal area estimation by ϵ-SVR can be observed when the number of height metrics increases from 6 to 8 but it is mitigated when other metrics are added. Stem density prediction by multiple regression improves when density metrics are added to predictors sets. So does the accuracy of mean diameter estimates when the number of height metrics is increased. Besides, accuracy values for basal area, stem density and basal area are quite stable. Dominant height estimates display no particular trend, except that the increase in height metrics number combined with the inclusion of entropy metrics yields better accuracy with multiple regression.

	Muliple regression			**ϵ-SVR**		
	CV_{RMSE} (%)	Number of predictors in the model	Dimension reduction and number of components	CV_{RMSE} (%)	Kernel	Dimension reduction and number of components
H_{dom}	13.5	4	PCA-12	14.5	linear	PCA-4
G	18.8	4	none	25.9	linear	PCA-2
N_s	21.2	3	none	24.2	radial	ICA-2
dbh	15.0	1	none	14.8	linear	PCA-2

Table 2. Best prediction accuracy obtained with multiple regression and ϵ-SVR with the predictors sets derived from laser metrics with $(n_h, n_d, n_e) = (6, 3, 0)$, and corresponding dimension reduction settings.

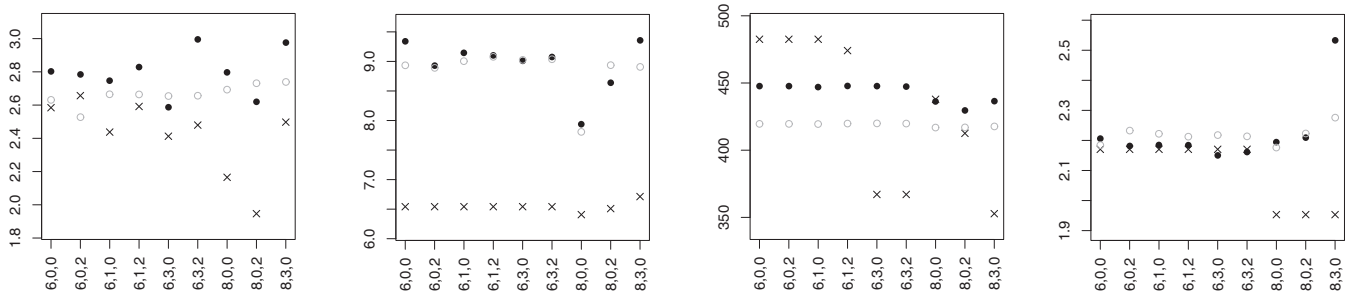


(a) Dominant height (m) (b) Basal area ($m^2.ha^{-1}$) (c) Stem density (ha^{-1}) (d) Mean diameter (cm)

Fig. 1. Accuracy of prediction (RMSE obtained by leave-one-out cross validation) of ϵ-SVR with linear (black symbols) and radial (gray symbols) kernels, plotted against the number of predictors. Symbol types refer to the method used for dimension reduction: PCA (dotted lines), ICA (solid lines) or none (triangles). Predictors sets are derived from $(n_h, n_d, n_e) = (6, 3, 0)$.

5. CONCLUSION

The results of the area-based method applied in this study to predict forest parameters from airborne laser scanning data showed that the accuracy of ϵ-SVR estimates are similar to or poorer than those obtained by ordinary least squares multiple regression. Dimension reduction of laser metrics by PCA improved the ϵ-SVR accuracy, whereas multiple regression performed better on raw laser metrics. On the whole, radial kernel turned out to be slightly more accurate and robust than linear kernel. Multiple regression was more sensitive to the number and type of laser variables included in the training sets than ϵ-SVR. Moreover the effect of addition or removal of laser metrics depended on the predicted forest parameter.
Further research should focus on factors that may improve support vector regression, such as other kernels or algorithms (ν-SVR), or inclusion of a larger number of chosen laser metrics. Besides advantage could be taken of SVR robustness when predicting parameters for forest stands or laser data different from those used to train the algorithm. The trade-off between accuracy of estimates and intensity of field campaign is indeed a major factor of concern when dealing with forest inventory at operational scale in mountainous areas.



(a) Dominant height (m) (b) Basal area ($m^2.ha^{-1}$) (c) Stem density (ha^{-1}) (d) Mean diameter (cm)

Fig. 2. Influence of the number and type of laser metrics on the accuracy of prediction (RMSE obtained by leave-one-out cross validation) of multiple regression ($\times$) and ϵ-SVR with linear ($\bullet$) and radial ($\circ$) kernels. Triplets on the x-axis refer to the number of laser height, density and entropy metrics (n_h, n_d, n_e) used to construct the predictors sets.

6. REFERENCES

[1] J. Hyyppä, H. Hyyppä, D. Leckie, F. Gougeon, X. Yu, and M. Maltamo, "Review of methods of small-footprint airborne laser scanning for extracting forest inventory data in boreal forests," *International Journal of Remote Sensing*, vol. 29, no. 5, pp. 1339–1366, 2008.

[2] E. Næsset, "Predicting forest stand characteristics with airborne scanning laser using a practical two-stage procedure and field data," *Remote Sensing of Environment*, vol. 80, no. 1, pp. 88–99, 2002.

[3] P. Packalén and M. Maltamo, "The k-msn method for the prediction of species-specific stand attributes using airborne laser scanning and aerial photographs," *Remote Sensing of Environment*, vol. 109, no. 3, pp. 328–341, 2007.

[4] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning*, Springer, corrected edition, July 2003.

[5] M. Heurich and F. Thoma, "Estimation of forestry stand parameters using laser scanning data in temperate, structurally rich natural european beech (fagus sylvatica) and norway spruce (picea abies) forests," *Forestry*, vol. 81, no. 5, pp. 645–661, 2008.