Support vector machines regression for estimation of forest parameters from airborne laser scanning data

Jean-Matthieu Monnet, Frédéric Berger, Jocelyn Chanussot

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

ABSTRACT

Estimation of forest stand parameters from airborne laser scanning data relies on the selection of laser metrics sets and numerous field plots for model calibration. In mountainous areas, forest is highly heterogeneous and field data collection labour-intensive hence the need for robust prediction methods. The aim of this paper is to compare stand parameters prediction accuracies of support vector machines regression and multiple regression models. Sensitivity of these techniques to the number and type of laser metrics, and use of dimension reduction techniques such as principal component and independent component analyses are also tested. Results show that support vector regression was less accurate but more stable than multiple regression for the prediction of forest parameters.

***Index Terms*—** Support vector regression, airborne laser scanning, forest parameters estimation

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, such as multilayer perceptron, self-organizing map and support vectors regression [4]. Support vector machines are a training approach based on the framework of statistical learning theory. They have proved their robustness to dimensionality and generalization abilities [5] and thanks to the kernel trick non-linear relationships can be accounted for. Mainly used for the purpose of hyperspectral images classification, they have also been successful for continuous parameters estimation [6].

The main objective of this paper is to compare 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 (town of Saint Paul de Varces, 45°04'17"N, 05°38'25"E). 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 inventoried. All trees with diameter at breast height larger than 5 cm and located within 10 m radius from the plot center were calipered. Maples (mainly *Acer opalus*), downy oak (*Quercus pubescens*) and common whitebeam (*Sorbus aria*) represented nearly 60 % of the stems. Ten tree heights were sampled on each plot. The following forest parameters were then computed for each plot: dominant height (H_{dom}), basal area (G), stem density (N) and mean diameter at breast height ($\overline{dbh}$) (Table 1). Plot centers were georeferenced with a Trimble GPS Pathfinder Pro XRS receiver. After differential correction horizontal position accuracies (95% confidence interval) ranged from 0.6 to 1.5 m.

Laser data was acquired with an airborne RIEGL LMS-Q560 scanner on August 27th, 2009. Laser footprint was 0.3 m and scan angle ±30°. Average scanning density was 2.8 pulses.m^{-2} with 50% overlap between adjacent flight

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

Parameters	H_{dom}	G	N_s	$\overline{dbh}$
Unit	(m)	($m^2.ha^{-1}$)	(ha^{-1})	(cm)
Mean	17.8	34.8	1735	14.5
Min	8.1	4.6	764	8.3
Max	28.5	59.7	2833	22.7
σ	5.3	11.4	577	3.6

Table 1. Forest stand parameters statistics (31 field plots)

strips. The resulting point cloud was classified by the contractor into ground and non-ground echoes using the TerraScan software. Final echo density was 10 m^{-2}.

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 two 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. For the whole point cloud, entropy metrics were calculated as the entropy of the orthometric distribution of points included in height bins of equal width.

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:

$$y = b + \sum_{i=1}^{n_v} a_i \times v_i \tag{1}$$

by ordinary least squares, with $(v_i)_{i\in\{1,...,n_v\}}$ a set of predictors and $((a_i)_{i\in\{1,...,n_v\}}, b)$ the model parameters. 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:

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

where $((\alpha_j)_{j\in\{1,...,n\}}, \beta)$ are parameters determined during the training process, $(x_j)_{j\in\{1,...,n\}}$ samples from the training set, 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 and its coefficient of variation:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{3}$$

$$CV_{RMSE} = \frac{RMSE}{\bar{y}} \text{ with } \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{4}$$

where y_i and $\hat{y}_i$ are the observed and predicted values, and N the number of observations.

To evaluate the effect of the number and type of laser metrics on prediction accuracy, we tested predictors sets obtained by combination of $(n_h, n_d, n_e) \in (\{6, 8\}\times\{0, 1, 3\}\times\{0, 2\})$. For each of these predictors sets, derived sets were computed by extracting subsets of components obtained by ICA [7] and PCA.

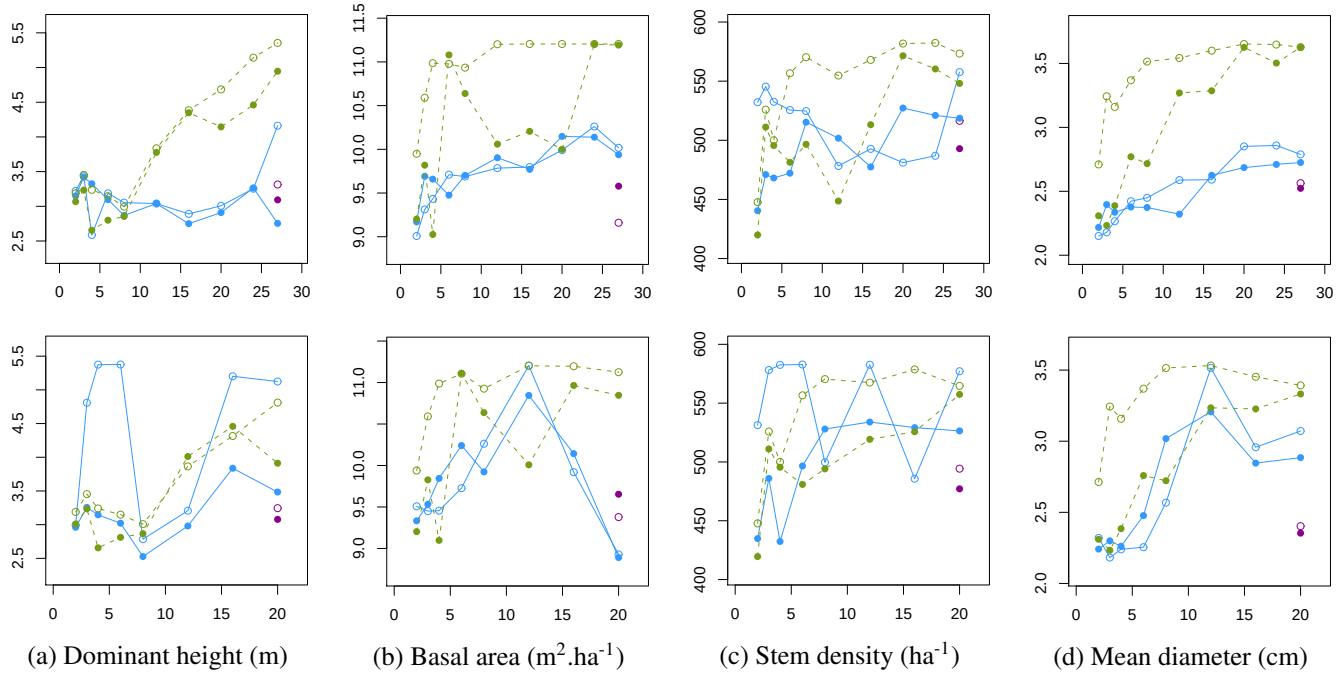
4. RESULTS AND DISCUSSION

Table 2 summarizes the best results obtained with multiple and ϵ-SV regressions for the predictors sets derived from the 27 laser metrics with $(n_h, n_d, n_e) = (6, 3, 0)$. Prediction estimates by multiple linear regression yield satisfactory results. The coefficient of variation of the RMSE ranges from 13.9 to 21.2%. The highest and lowest accuracies are respectively achieved for dominant height and stem density. Mean diameter and basal area yield 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) [8]. Dimension reduction slightly improves the accuracy for dominant height only ($CV_{RMSE} = 13.5\%$ with 12 components from PCA). Apart from mean diameter, multiple regression performs better than ϵ-SVR. However values are rather close, except for basal area.

Figure 1 illustrates the effect of dimension reduction and kernel selection on ϵ-SVR accuracy for the predictors sets derived from $(n_h, n_d, n_e) \in \{(6, 3, 0), (6, 0, 2)\}$. Dimension reduction always benefits to prediction accuracy, and the best

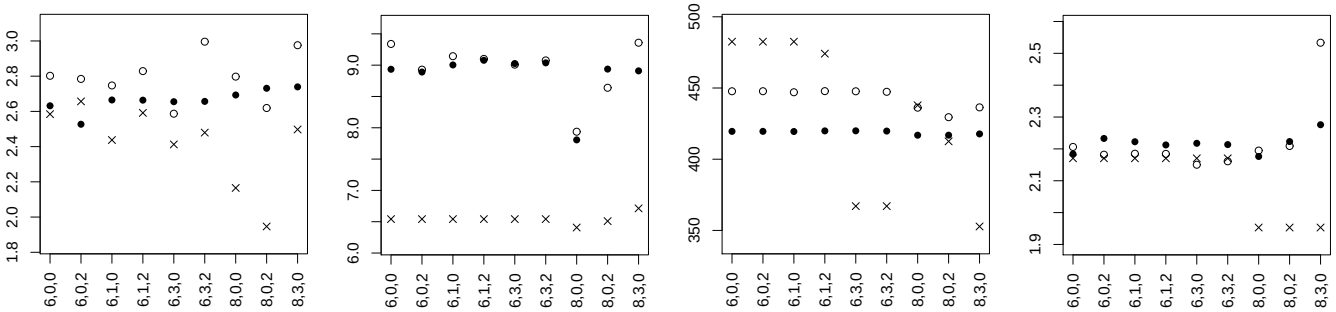
	Multiple regression			ϵ-SVR		
	CV_{RMSE} (%)	Model predictors	Dimension reduction and number of components	CV_{RMSE} (%)	Kernel	Dimension reduction and number of components
H_{dom}	13.5	4	PCA-12	14.5	linear	PCA-4
G	18.8	4	none	25.9	linear	PCA-2
N_s	21.2	3	none	24.2	radial	ICA-2
$\overline{dbh}$	15.0	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 (○) and radial (●) kernels, plotted against the number of predictors. Line types and colors refer to the method used for dimension reduction: PCA (**blue** solid lines), ICA (**green** dashed lines) or none (**dark magenta** single symbols). Predictors sets are derived from $(n_h, n_d, n_e) = (6, 3, 0)$ (top row) and $(6, 0, 2)$ (bottom row).



(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 (×) and ϵ-SVR with linear (○) and radial (●) 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.

results are obtained with less than eight components, except for basal area with predictors set $(6, 0, 2)$. PCA performs better than ICA, except for stem density. Accuracy tends to decrease when the number of predictors increases further than ten. However, ϵ-SVR appears less sensitive to the number of components when PCA is employed instead of ICA. On the whole, radial kernel appears to be more robust regarding the type and number of components included in the predictors sets. Stem density turns 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, 8]. ϵ-SVR best accuracies are quite similar for predictors sets $(6, 0, 3)$ and $(6, 0, 2)$: $CV_{RMSE} = 14.5$ and 14.2% for dominant height, 25.9 and 25.5% for basal area, 24.2% in both cases for stem density, and 14.8 and 15.1% for mean diameter. However multiple regression models for dominant height and stem density were less precise when density metrics were replaced by entropy values (from 13.5 to 14.9% and from 21.2 to 27.8% respectively).

Figure 2 depicts the influence of the number and type of laser metrics included in the predictors sets on 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 six to eight 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. These findings are consistent with multiple regression predictive models obtained for coniferous stands [2], which always included density metrics for stem density models whereas height models did not.

5. CONCLUSION

The results of the area-based method applied in this study to predict forest parameters from airborne laser scanning data showed that ordinary least squares multiple regression performs slightly better than ϵ-SVR. However, multiple regression accuracy is highly sensitive to the number and type of laser variables included in the training sets, whereas ϵ-SVR displays greater stability. Besides, the effect of addition or removal of laser metrics depends on the predicted forest parameter. Regarding dimension reduction effects, PCA improves the ϵ-SVR accuracy, whereas multiple regression performs better on raw laser metrics.

Further research should focus on factors that may improve support vector regression, such as finer tuning of hyperparameters or use of other kernels or algorithms (ν-SVR). 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.

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," *Int. J. Remote Sens.*, 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 Sens. Environ.*, 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 Sens. Environ.*, vol. 109, no. 3, pp. 328 – 341, 2007.

[4] H. Niska, J. P. Skön, P. Packalén, T. Tokola, M. Maltamo, and M. Kolehmainen, "Neural networks for the prediction of species-specific plot volumes using airborne laser scanning and aerial photographs," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 3, pp. 1076–1085, 2010.

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

[6] D. Y. Sun, Y. M. Li, and Q. Wang, "A unified model for remotely estimating chlorophyll a in lake Taihu, China, based on SVM and in situ hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 8, pp. 2957–2965, Aug. 2009.

[7] A. Hyvarinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. Neural Netw.*, vol. 10, no. 3, pp. 626 –634, May 1999.

[8] 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.