Gaussian Process Regression for Airborne Laser Scanning Based Forest Inventory: Validation and Parameter Selection

P. Varvia[1], J. Räty[2], L. Korhonen[1], P. Packalen[1]

[1]School of Forest Sciences, University of Eastern Finland, Joensuu, Finland
Email: {petri.varvia; lauri.korhonen; petteri.packalen}@uef.fi

[2]Division of Forest and Forest Resources, Norwegian Institute of Bioeconomy Research, Ås, Norway
Email: janne.raty@nibio.no

1. Introduction

Airborne laser scanning-based (ALS) forest inventories that utilize the so-called area based approach (ABA) are of great practical importance. While ABA can be considered a mature problem, with many well-established approaches to predict the forest attributes, there is still be room for improved prediction methods.

Gaussian process regression (GPR) (e.g. Rasmussen and Williams 2006) is a popular machine learning method related to kriging that is based on modelling the forest attributes and the ALS predictors jointly as a Gaussian process. The main advantages of GPR are the capability to accurately represent highly nonlinear relations with a modest number of tuneable parameters, ability to effectively use large number of predictors, and that it produces uncertainty estimates for the predictions.

GPR has shown promise in providing slightly better prediction accuracy than established methods (Varvia 2019). However, the previous results on GPR were limited by 1) using data from only one study area, 2) using cross-validation instead of a separate test set. The aim of this work is to rectify these limitations and additionally test automatic tuning of GPR parameters.

To benchmark the GPR performance, random forests (RF) were chosen as a reference method. RF was chosen because it has produced excellent results in ABA (Cosenza et al. 2021) and it can also handle large number of predictors.

2. Data and Methods

2.1 Materials

The data consist of field measurements from three sites in Finland, Nummi-Pusula, Kurikka-Seinäjoki, and Pokka and corresponding ALS data produced by Finnish Forest Center in 2019. The study sites represent forests from Southern, Western, and Northern Finland, respectively. ALS data are openly available on the download service of the National Land Survey of Finland.

The field data consist of 1125, 830, and 763 circular field plots with a radius of either 9 m or 12.62 m in Nummi-Pusula, Kurikka-Seinäjoki, and Pokka, respectively. To evaluate the performance of the two prediction methods rigorously, each data set was randomly split to separate training, validation, and test sets in a 40%/20%/40% fashion. Of these, the validation set was used to choose optimal model parameters and only the test set to evaluate final prediction performance.

The corresponding ALS data had a nominal pulse density of 0.8 m^{-2}. After height normalization using ground echoes, large number of predictors, including height quantiles, other height metrics and canopy densities were computed separately from first of many and only echoes, and last of many and only echoes. Intensity metrics were also calculated. Predictors that did not show appreciable variation between plots were removed. Final set contained 45 predictor variables, with same variables in every site. No further variable selection was done.

Published in: Markus Hollaus, Norbert Pfeifer (Eds.): Proceedings of the SilviLaser Conference 2021, Vienna, Austria, 28–30 September 2021.
Technische Universität Wien, 2021. DOI: 10.34726/wim.1861
This paper was peer-reviewed. DOI of this paper: 10.34726/wim.1928

2.2 Methods

In this study, the total stem volume is predicted in the ABA framework. Gaussian process regression was implemented using an R package under development by the authors. In GPR, the choice of the so-called covariance function or kernel is one of the principal aspects that affects the predictions. As in our previous studies, isotropic Matérn 3/2 covariance function was used with Euclidean distance metric. This results in three tuneable parameters: length scale l, kernel variance σ_k^2, and error variance σ_e^2. The separate validation set was the used to choose the optimal values for these parameters by minimizing the sum of squared prediction errors in the validation set. The optimization was done using simulated annealing with the R *optimization* package.

As a reference method, random forest (RF) was used. For RF, we used the popular implementation in the R *randomForest* package. While it is common practice to use the default values for RF parameters, such as the number of decision trees, to facilitate honest comparison, the number of predictor candidates per split (i.e. *mtry*) and the number of trees were optimized using the validation set as in GPR.

3. Results and Discussion

The RMSE and bias of the total volume predictions evaluated using the separate test set for the three study sites are presented in Table 1.In all three sites, GPR produced slightly more accurate predictions, with relative RMSE being consistently better by 0.3-0.9 percentage points. Both methods showed small negative bias in the predictions, with GPR being slightly less biased in Nummi-Pusula and Kurikka-Seinäjoki, while RF is slightly better in Pokka.

Table 1. Prediction performance in the test set. Units are in m^3/ha, relative metrics are shown inside parentheses.

	RMSE (%)	Bias (%)
Nummi-Pusula	n_{test}=450	
GPR	42.3 (21.4%)	-2.4 (-1.2%)
RF	42.9 (21.7%)	-4.3 (-2.2%)
Kurikka-Seinäjoki	n_{test}=332	
GPR	33.1 (20.8%)	-1.0 (-0.6%)
RF	34.5 (21.7%)	-2.4 (-1.5%)
Pokka	n_{test}=305	
GPR	21.3 (24.2%)	-0.6 (-0.7%)
RF	22.1 (25.1%)	-0.4 (-0.5%)

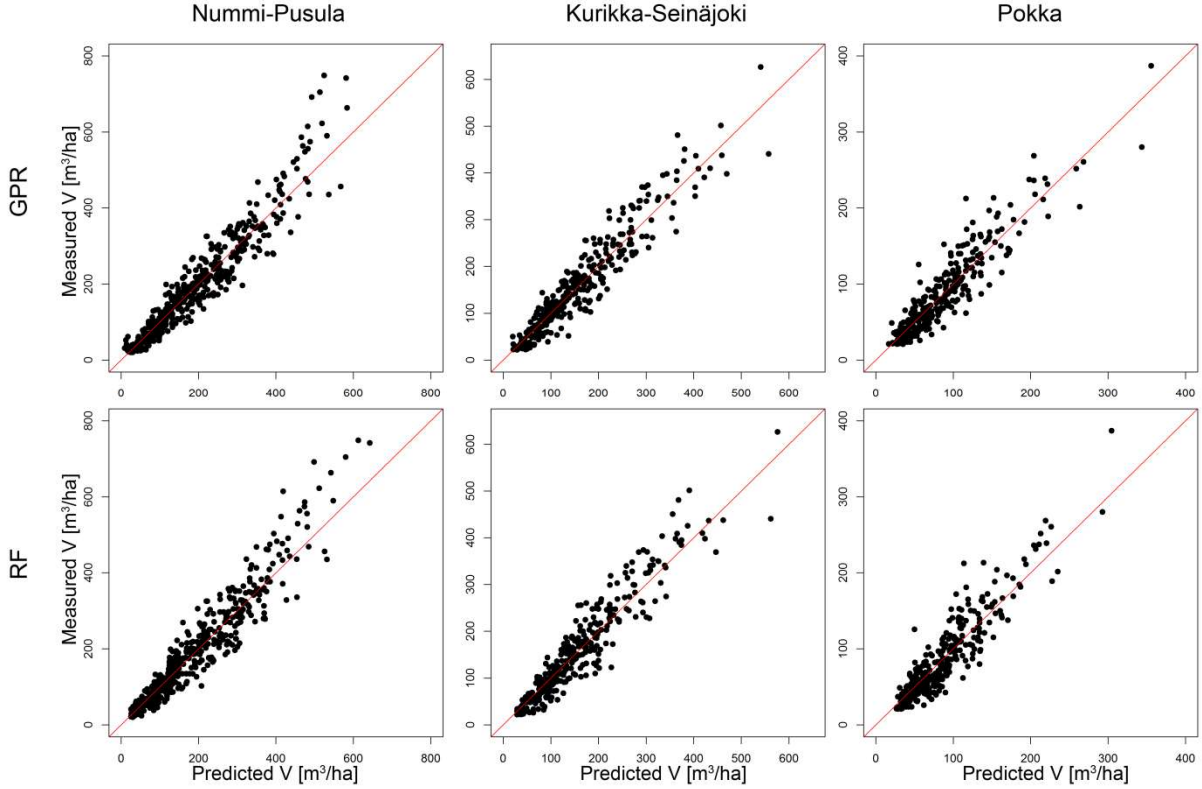
Model parameters were optimized by simulated annealing using the validation set are shown in Table 2. Both the GPR and RF variables show large variability by study site. Parameter selection problems are generally difficult to optimize, due to usually having multiple local minima. Simulated annealing was chosen to mitigate this, but as a method it gives no guarantee that the converged solution is the global optimum. Given the small number of parameters, grid search would be still feasible and guarantee an optimal solution. In RF, the default parameters (*mtry*=33%, *n*=500) are commonly used. The optimized values here were compared to the predictions using the default values and the difference in RMSE was negligible, supporting the common practice.

Table 2. Optimized parameter values by study area, σ_v^2 is the sample variance of total stem volume in the training set.

	GPR	RF
Nummi-Pusula	l=25.6, σ_k^2=2.7σ_v^2, σ_e^2=0.51σ_v^2	*mtry*=42%, *n*=494
Kurikka-Seinäjoki	l=29.6, σ_k^2=1.1σ_v^2, σ_e^2=0.03σ_v^2	*mtry*=17%, *n*=487
Pokka	l=16.3, σ_k^2=2.3σ_v^2, σ_e^2=0.17σ_v^2	*mtry*=29%, *n*=225

GPR also produces prediction variances from which credible intervals can be computed. The 95% credible intervals (CI) covered 99.8% of the field measured volumes in Nummi-Pusula, 82.5% in Kurikka-Seinäjoki, and 99.0% in Pokka. The values imply that the variances were severely overestimated in Nummi-Pusula and Pokka, and underestimated in Kurikka-Seinäjoki. The variance estimation aspect could be potentially improved by incorporating CI coverage in the cost function used to find optimal parameter values.

Figure 1. Scatter plots of the predictions. Identity line shown in red.



4. Conclusions

In this work, Gaussian process regression was rigorously validated at three study sites representing boreal forest in Southern, Western, and Northern Finland. The prediction performance of GPR was compared with RF. The performance of the two methods was quite similar, although GPR produced consistently slightly lower RMSEs. Compared to RF, GPR has the additional capability to also simultaneously produce variance/interval estimates for the predictions. In conclusion, the results support the previous studies on the potential of GPR as a prediction method in the area-based approach.

Acknowledgements

This study was funded by the Academy of Finland (grant number 332707).

References

Rasmussen C and Williams K, 2006, *Gaussian processes for machine learning*. MIT Press, Cambridge, MA, USA.

Varvia P, Lähivaara T, Maltamo M, Packalen P, Seppänen A, 2019, Gaussian process regression for forest attribute estimation from airborne laser scanning data. *IEEE Trans. Geosci. Remote Sens.*, 57(6):3361–3369

Cosenza D N, Korhonen L, Maltamo M, Packalen P, Strunk J L, Næsset E, Gobakken T, Soares P, Tomé M, 2021, Comparison of linear regression, k-nearest neighbour and random forest methods in airborne laser-scanning-based prediction of growing stock, *Forestry*, 94(2):311–323