Institute of Urban Environment,
Chinese Academy of Sciences,
CN 361021, China

Dr. Yin Ren
Email: yren@iue.ac.cn
5th Jun 2020

Biogeosciences Discussions
Open Access
EGU

Interactive comment on "Improving maps of forest aboveground biomass: A combined approach using machine learning with a spatial statistical model" *by* Shaoqing Dai et al. *(Additional Response to Anonymous Referee #1)*

Shaoqing Dai et al.

yren@iue.ac.cn

Dear Editor,

Thank you for your email and the reviewer's comments concerning our manuscript entitled "Improving maps of forest aboveground biomass: A combined approach using machine learning with a spatial statistical model" (ID: bg-2020-36). We thank the reviewer for the very helpful comments.

Additional reply to reviewer's comments:

Dear reviewer,

Anonymous Reviewer #2 recommended that we re-evaluate the performance of the machine learning models with additional predictors (at least latitude). For your convenience, the comments of anonymous reviewer #2 are listed below:

“b) They claim that the joint model combines the advantages of ML and the P-BSHADE model, the predictive non-linearity advantage of ML and the ability of the P-BSHADE to capture spatial relationships. However, if they are given the chance I think that ML models are also capable of detecting and using spatial relationships, that means, you have to provide them not only longitude but also latitude as predictor! Based on correlation with AGB, the authors selected only longitude, however, I would assume that an interaction of longitude and latitude would be a good predictor of spatial relationships (two variables of an interaction can show by themselves low correlation). Moreover, ML models such as RF are outstanding in detecting interactions and higher-order interactions (if they are given the chance). Also, hyper-parameter tuning is important in ML to improve predictive performance, even for RF! (e.g. see Probest et al., 2019 https://doi.org/10.1002/widm.1301). I recommend that the authors re-evaluate the performance of the ML models with hyper-parameter tuning, nested cross-validation, and additional predictors (at least latitude).”

We think this comment is relate to your General Comment 2). For your convenience, your comment is presented here:

“2) The cross-validation strategy that you used is not suitable if you have spatially clustered data (as you obviously have looking at the map). This is shown by several studies (see references below, to mention just a few). What would be appropriate is a spatial cross-validation that is testing the ability of your model to make predictions for spatially new samples. At least you should take care that you never use data points from the same forest patch for both training and testing. Otherwise, it is not possible to evaluate the ability of your models for regional mapping. Including coordinates as predictors when the data are spatially clustered is very dangerous (see Meyer et al 2019) and can lead to high overfitting which can only be revealed with spatial cross-validation. So I recommend that in addition to spatial cross-validation, to perform a spatial variable selection (i.e. can the predictors be used to make predictions for new locations?)

1. Meyer, H., Reudenbach, C., Wöllauer, S., Nauss, T., 2019. Importance of spatial predictor variable selection in machine learning applications - Moving from data reproduction to spatial prediction. Ecological Modelling. 411, 108815.

2. Pohjankukka, J., Pahikkala, T., Nevalainen, P., Heikkonen, J., 2017. Estimating the prediction performance of spatial models via spatial k-fold cross validation. Int. J. Geogr. Inform. Sci. 31, 2001–2019. https://doi.org/10.1080/13658816.2017. 1346255.
3. Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Hartig, F., Dormann, C.F., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography.
4. Valavi, R., Elith, J., Lahoz-Monfort, J.J., Guillera-Arroita, G., 2018. blockcv: an r package for generating spatially or environmentally separated folds for k-fold cross validation of species distribution models. BioRxiv."

To test whether include or exclude latitude, and whether leave-one out cross-validation or spatial block cross-validation would have impact on models performance in the forest AGB estimation, we re-build the model and evaluate the results again. We found that there has a error on Figure 1 to 3 in the previous response to your comments (https://www.biogeosciences-discuss.net/bg-2020-36/bg-2020-36-AC3-supplement.pdf). All the results were correct in this response. Therefore, we revised and presented the final results including six situations in Table 1:

The results from comparing situation 2 and 5 showed that cross validation strategy has significant impact on the ML models and P-BSHADE model, but showed limited impact on the three combination models. We agree with your comments and suggestions and decided to adopt spatial block cross-validation.

The results showed that using both longitude and latitude as predictors did not improve the performance of RF, SVM, or RBF-ANN. However, this does not mean that machine learning cannot capture the spatial relationships. We also show that the combination of three MLs and P-BSHADE can all improve the prediction accuracy compared to the single ML in all of the situations. We think that this further illustrates the advantages of the combined models.

Table 1 Performance of models in different predictor situations

Model	MAE	MRE	RMSE	nRMSE
Situation 1: includes longitude and latitude with spatial block cross-validation				
SVM	18.260	2.260	25.929	0.548
RBF-ANN	14.665	0.324	21.429	0.453
RF	16.435	1.292	27.419	0.579
P-BSHADE	12.069	0.251	19.490	0.412
SVM & P-BSHADE	5.075	0.108	6.495	0.137
RBF-ANN & P-BSHADE	12.164	0.244	17.546	0.371
RF & P-BSHADE	6.344	0.153	12.279	0.259
Allometric Model	16.412	0.577	25.080	0.530
Situation 2: includes longitude but excludes latitude with spatial block cross-validation				
SVM	13.659	1.382	20.251	0.428
RBF-ANN	14.884	0.324	20.719	0.438
RF	14.919	1.107	21.371	0.451
P-BSHADE	12.069	0.251	19.490	0.412
SVM & P-BSHADE	6.879	0.128	10.757	0.227
RBF-ANN & P-BSHADE	12.597	0.260	17.819	0.376
RF & P-BSHADE	5.816	0.137	9.520	0.201
Allometric Model	16.412	0.577	25.080	0.530
Situation 3: excludes longitude and latitude with spatial block cross-validation				
SVM	15.824	2.034	21.609	0.456
RBF-ANN	14.175	0.373	20.354	0.430
RF	15.201	1.092	21.330	0.451
P-BSHADE	12.069	0.251	19.490	0.412
SVM & P-BSHADE	9.679	0.173	16.055	0.339
RBF-ANN & P-BSHADE	12.227	0.232	18.124	0.383
RF & P-BSHADE	6.104	0.136	9.911	0.209
Allometric Model	16.412	0.577	25.080	0.530
Situation 4: includes longitude and latitude with leave-one out cross-validation				
SVM	8.208	0.220	11.064	0.232
RBF-ANN	12.710	0.268	19.466	0.409
RF	10.696	0.289	20.012	0.420
P-BSHADE	12.045	0.279	19.781	0.415
SVM & P-BSHADE	5.214	0.110	6.668	0.140
RBF-ANN & P-BSHADE	11.876	0.224	17.479	0.367
RF & P-BSHADE	5.853	0.145	10.963	0.230

Situation 5: includes longitude and excludes latitude with leave-one out cross-validation(in submitted article)				
SVM	11.168	0.248	10.388	0.218
RBF-ANN	12.149	0.267	10.388	0.218
RF	10.155	0.259	9.428	0.198
P-BSHADE	18.371	0.391	14.077	0.296
SVM & P-BSHADE	6.883	0.125	6.304	0.132
RBF-ANN & P-BSHADE	10.136	0.205	9.633	0.202
RF & P-BSHADE	5.679	0.130	5.299	0.111
Situation 6: excludes longitude and latitude with leave-one out cross-validation				
SVM	11.473	0.240	18.426	0.387
RBF-ANN	12.674	0.269	18.584	0.390
RF	10.759	0.232	17.043	0.358
P-BSHADE	12.045	0.279	19.781	0.415
SVM & P-BSHADE	9.203	0.157	15.977	0.336
RBF-ANN & P-BSHADE	16.282	0.395	27.770	0.583
RF & P-BSHADE	5.933	0.122	9.765	0.205

Note: MAE: mean absolute error; MRE: mean relative error; RMSE: root mean square error; nRMSE: normalized root mean square error

We thank the reviewer and remain at your disposal for any further questions.

Yours sincerely,
Yin Ren