Biogeosciences Discuss.,
https://doi.org/10.5194/bg-2020-36-RC1, 2020
© Author(s) 2020. This work is distributed under
the Creative Commons Attribution 4.0 License.

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.

Anonymous Referee #1

Received and published: 25 March 2020

The study presented in this manuscript compares the suitability of machine learning as well as geostatistical modelling approaches to retrieve maps of tree AGB on a regional scale based on plot-based surveys. The paper is reasonably written, however it is lacking a lot of important details that will be outlined in the more specific comments below. Apart from the need to improve the writing there is a clear need to improve the methodology.

GENERAL COMMENTS

1) Important information on the methodology are not given. In the method section it’s not even mentioned which predictors were used for the machine learning model

training. But this essential as the success of the models is only marginally depending on the choice of the algorithm but in the first place on the ability of the variables being used to serve as predictors for AGB! Only in the result section we get an idea on the variables (longitude, DBH, H, and forest age). I’m surprised about these variables as remote sensing information (especially NDVI) would present much more obvious predictors for AGB when the aim is to model AGB on a regional scale. With the selected variables, how could you upscale the results to a regional scale? I guess neither DBH nor H are not available in a spatial continuous way. So your model cannot be used for regional mapping! However, your motivation is to use it for regional modelling so my question is a) can you really do it with your approach and b) if yes, why are you not doing so and also show the results in the manuscript?

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?)

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.

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.

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.

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.

3) The concerns outlined above can be improved, however, I have also doubts about the general value of the paper. Relying on 30 plots only is very very limited for machine learning application (the spatial CV will probably reveal this). So I doubt that the results will produce results that allow for general conclusions on the value of combining machine learning with geostatistical modelling.

SPECIFIC COMMENTS

Line 25: I disagree that longitutde and latitude on this scale affect AGB. Even on a large scale they don't but are just proxies for e.g. climate but they are certainly not underlying factors for AGB on your small study area.

Line 49-51: One important thing is missing: The model might also fail because the predictor variables are not sufficient to estimate AGB.

Line 54: "An estimated 18%–103% of the uncertainty in AGB mapping can be attributed to model-dependent uncertainty". In fact between nearly nothing (∼18) and everything (>100). That sounds unreasonable, consider taking that sentence out.

Line 60-62: This differentiation between allometric models and statistical models does not seem to make sense. E.g. Allometric models can be based on linear relationships as well. Please improve the logical structure here.

Line 67-71: Be careful with the logical structure here as well: The major advantage is that machine learning is able to fit complex relationships which e.g. linear models

don't. And THEREFORE they might be advantageous in predictions (not "in addition" as you write in Line 72).

Line 74-81: The fundamental difference between the approaches is not getting clear here but this is important because combining the two approaches is the objective of the paper. In contrast to the statistical (including machine learning) approaches explained above, the spatial statistical approaches have the major assumption that "near things are more related that distant things". I think the general idea should be made clear and it should be explained why you expect that a combination might be the way forward.

Line 85-88 "studies that used machine learning methods have not considered the spatial heterogeneity of multiple environmental covariates (such as longitude, latitude, and forest structure)". I disagree. Most approaches use environmental covariates which of course have been heterogeneous as well.

No information on model tuning is given. Also please state which software implementations and settings of the algorithms you used.

Line 157-158: please explain why you tested for spatial autocorrelation etc. Why is this information relevant for the modelling?

Line 205:"Because of the Law of Large Numbers, RF does not overfit." That's wrong! Maybe random forest is robust to overfitting in terms of hyperparameter selection but it is not the case if you have data that are not independent. See e.g. the references mentioned above.

Line 206-207: Accurate predictions of random forest do NOT in the first place originate from injecting randomness. E.g. If the predictors are not sufficient to estimate a response variable, random forest will fail (and so will other algorithms)!

Fig. 6 : Is this based on the cros-validation?

Line 502:503: "The assumption is that estimated AGB is accurate in all sampling plots except the target samplig plot. In other words, the premise behind using only the P-

BSHADE model is that the reference AGB data is accurate or strongly correlated with AGB. “ I don’t understand that. The same reference data were used for all modedliung approaches and for sure we assume that the reference data are accurate for both types of models.

Line 578: “We used FMPI data to upscale the optimal plot-level AGB model from plot level to region scale.” Did you? We don’t get to see the results for the regional upscaling.

I wonder: Is your model really better than simply using the average measured AGB from each forest site as estimate for AGB for the entire patch?

Interactive comment on Biogeosciences Discuss., https://doi.org/10.5194/bg-2020-36, 2020.