Biogeosciences Discuss.,
https://doi.org/10.5194/bg-2020-409-RC2, 2021

Biogeosciences Discussions

Interactive comment on "Quantifying the Importance of Antecedent Fuel-Related Vegetation Properties for Burnt Area using Random Forests" *by* Alexander Kuhn-Régnier et al.

Anonymous Referee #2

Received and published: 4 March 2021

The authors use a machine learning approach (ML) to investigate the impact of biophysical and climatic variables on burned area over a five year period where there are available data on a global scale. FAPAR appears to be the most important predictor from their RF simulations. Variables were tested for their impact if they included the preceeding X number of months. This allowed for the investigation of the impact of lagged relationships, arguably very important for fire modelling.

I am concerned about the cross-validation strategy employed here (L137. By randomly choosing the validation dataset the potential for spatial autocorrelation issues arises. This is well known in the literature (see Roberts et al. 2017; Ploton et al. 2020; Kühn

and Dormann, 2012; Meyer et al. 2019). Here is a snippet from the abstract to the Roberts article: 'Ecological data often show temporal, spatial, hierarchical (random effects), or phylogenetic structure. Modern statistical approaches are increasingly accounting for such dependencies. However, when performing cross-validation, these structures are regularly ignored, resulting in serious underestimation of predictive error. One cause for the poor performance of uncorrected (random) cross-validation, noted often by modellers, are dependence structures in the data that persist as dependence structures in model residuals, violating the assumption of independence. Even more concerning, because often overlooked, is that structured data also provides ample opportunity for overfitting with non-causal predictors.'. Because the authors devote considerable space to discussion of these predictors, I think this issue is worth consideration. The authors also argue the the gap in R2 of the training-validation simulations gives an idea of the generalizability of the model - but that breaks down if there are spatial autocorrelation issues. Also there is some spatial structure in their biases (Fig S2) that could be coming from this issue. I would suggest adopting other CV strategies as outlined in the papers I list above.

As I am not yet convinced by their CV strategy, which is important as it impacts the results quite heavily, I suggest major revisions as I assume it will take a bit of work to demonstrate that the chosen CV strategy doesn't give misleading results.

Minor comments: - Line 78: I am not sure if I understand the DD calculation. If you had, say 5 days in one month below the precip threshold then 10 into the next. A brief precip event then the rest of the month below threshold. How does that rate? Would it be concatenated so the whole month is seen as being DD?

- L80: Soil moisture was done how? Was it percent of saturation? relative to field capacity? Or some sort of index since it is later refered to as SWI.

- L81 - The Kaplan and Lau reference is for the WGLC that is based on the WWLLN, not the WWLLN directly. It also says it provides the frequency of lightning flashes per

unit area but doesn't specify that these are ground strikes. Can you please check that these are indeed cloud to ground and not total (cloud to ground + cloud to cloud)

- Table 1: For the AGB datasets, was there any overlapping latitudes between the two datasets? If so, how was that dealt with?

- L105 - I am concerned about the gap-filling approach.So for SWI doesn't this mean that it would assume drought conditions? How often would you have this condition applied (outside of winter, L100)?

- L110: I don't follow what was done here. What do those numbers mean? This is the smallest area of herb or crop in a pixel?

- L125: I think X is referring to both the variables (FAPAR, etc.) and a single month, i.e. I think you are saying variable X at month, t, has the seasonal cycle for variable X (called in the text, X 12M) subtracted from it? Please reconsider how this is formulated in the text. This doesn't work as written.

-Table 2: To help choose which variables were useful for the ML algorithm (and which might only be contributing to overfitting), why did the authors not try applying one of the most standard techniques available such as recursive feature elimination with cross-validation (available in scikit-learn, which the authors are already using (L132); Pedregosa et al. 2011)? These techniques could help get around arbitrary choices about the number of predictors, e.g. why 15 and not 10 or 20?

- Why does Fig S2/2/3/7/etc. have straight line cutoffs? Top of Mexico, Australia, E of Iran. Plotting error? Real problem? I don't recall this being mentioned in the main text but it is in all figures. Fig 3-Also for v. low BA, what about grey instead of black? Would be easier to see the larger BA values...

- L 300- Are those the right refs? Both papers use Causality Analysis and not regression techniques as mentioned here. What aspects of those papers touches on inclusion of extra predictors and overfitting in a ML approach?

- p 12 has a lot of 'discussion' in the results section. Either have a 'results and discussion' section or keep them separate.

- Fig 6 - can you not use sci notation for the numbers? It makes it easier to read. I am not sure if this warrants main text inclusion as it is very simple with almost no interesting structure. I would put this in supplement and move S2 or S3 into the main.

- L310 - what about masking for areas with longer fire return intervals to see if it them pops out as more important?

Literature cited:

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. and Dormann, C. F.: Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure, Ecography , 40(8), 913–929, 2017.

Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C., Cornu, G., Viennois, G., Bayol, N., Lyapustin, A., Gourlet-Fleury, S. and Pélissier, R.: Spatial validation reveals poor predictive performance of large-scale ecological mapping models, Nat. Commun., 11(1), 4540, 2020.

Kühn, I. and Dormann, C. F.: Less than eight (and a half) misconceptions of spatial analysis, J. Biogeogr., 39(5), 995-998, 2012.

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

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J.: Scikit-learn: Machine Learning in Python, J. Mach. Learn. Res., 12, 2825–2830, 2011.

Interactive comment on Biogeosciences Discuss., https://doi.org/10.5194/bg-2020-409, 2020.