Biogeosciences Discuss., 7, C4623–C4629, 2011
www.biogeosciences-discuss.net/7/C4623/2011/

Biogeosciences Discussions

Interactive comment on "Spatial distribution of soil organic carbon stocks in France" *by* M. P. Martin et al.

M. P. Martin et al.

manuel.martin@orleans.inra.fr

Received and published: 7 January 2011

We would like to thank the referee#2 for the helpful and constructive comments. We give below responses to the comments and suggestions and hope it will satisfy the criticisms raised. A revised version of the manuscript is joined to this comment, which addresses the remarks of both referees requiring additions or changes (written in blue in the text). However, this revised version should not be considered as final as it is, at the same time, being edited for improving the general standard of English of the manuscript.

Referee's comments are italicized.

This paper is an analysis of the French soil monitoring data to predict soil carbon stocks on a 12km grid – rather than the 16km original sampling. I am concerned that by relating each point to a number of soil, landuse and climate variables in space and then treating this dataset as non spatial they are throwing away a lot of information.
At this point, we would like to emphasize the fact that within the dataset used for fitting the models, the soil properties and land use properties are directly observed on each point (i.e. the sampling sites) as explained in section 2.1.1. Only climatic and Net Primary Productivity (NPP) data are related to the points using a spatial joining. See also below for remarks about the non-spatial treatment.

They have analysed this dataset using Boosted Regression Trees. One of the drawbacks of regression trees is that the dependant variable, in this case organic carbon stocks, will be put in bins so the size and range of those bins will influence how good the predictions will be.
The number of individuals within each terminal bin (refereed in the text as min.obs) is one of the parameters which was adjusted during the tuning procedure. From our experience, varying this parameter within a reasonable range (5-15 individuals) resulted in comparable results. The optimal value for the RMQS datasets and the tested models was 8. The relatively low sensitivity of the Boosted Regression Tree (BRT) to this parameter might result from the fact that the resulting models are a combination of many single trees. The grouping of individuals in bins may vary from one tree to another and thus we think that throughout the algorithm the elementary unit is the individual instead of bins whose size consequently does not greatly influence the fit.

I suspect the authors only used 10% of their data to validate the models as using more would have shown how poor the model is. I am concerned that they appear to be overfitting – something that boosted regression trees are well known for.
Indeed, we performed a variant of the classical k-fold cross-validation: a repeated

10-fold cross-validation, also sometimes called 10-folds Monte Carlo cross-validation (Xu, Q. Liang, Y., 2001) resulting in the use of 10% of the data for validation. For the classical k-folds cross-validation, the size of the validation vs. calibration subsets can be debated. As explained by Hastie et al. (2001), using a small fraction of the individuals for fitting the models and thus validating on a great fraction might overestimate the prediction error (that is, obtained on the validation subset). On the opposite, increasing the fraction of individuals used for the fit reduces this overestimation but increases the variance of the prediction error estimate: the models seem to predict well but it very much depends on the small subset used for validating the model; it would then be misleading, as mentioned by the referee#2, to give the results of the cross-validation without referring to the prediction error variance. Consequently, one willing to use cross-validation has to look for a compromise between prediction error overestimation and high error variance. 5-fold or 10-fold (as we did in the manuscript) cross-validation is usually recommended as a good compromise (Hastie et al. 2001, section 7.10). We might add that of course, the optimal proportion for dataset splitting (calibration vs. validation) depends on the modeled property and the dataset.
We would like also to refer to a previous work involving a similar cross-validation procedure applied to boosted regression trees (Martin et al. 2009) and soil bulk density. There, different proportions for the dataset splitting were tested. The results obtained for a repeated 5-folds cross-validation were not drastically different than the results obtained for the 10-fold case. The repeated 10-folds cross-validation of the present paper (10% of the data used for validation) was chosen based on this previous experience. Additionally, in our case a 10% validation dataset is made of about 140 individuals. We think that it is already challenge to predict SOC stocks for these 140 independent samples with a reasonable prediction error variance. Furthermore, the prediction error variance, mentioned above in this comment, was given in the manuscript via a 95% confidence interval on RMSPE.ext values, in table 1. This allows the reader to judge whether the statistics on the prediction error are good or not. Based on these results, we think that we did find a good compromise between

prediction error overestimation and prediction error variance: the highest boundary of the predicting error confidence interval (that is the worst case scenario) is comparable to the prediction error found in the literature. For instance Mishra et al. 2008 obtained 2.89 kg/m2 using a single 20% validation dataset.
However, the referee's question regarding possible lower cross-validation results with larger proportion of independent samples remains interesting. Following this, we performed a test with a 20% validation dataset, corresponding to a repeated 5-fold cross-validation, and compared it to results obtained with 10-folds, i.e. 10% validation dataset proportion, as used in the manuscript. The test model was the Extra model, with the following parameterization ts=12, lr=0.01, min.obs=8, bf=0.75 and 100 replicates for this Monte Carlo Cross validation. For a 20% proportion, the mean RMSPE.ext value was 2.331 kg/m^2 (CI95% = [2.028, 2.635]) and for the 10% proportion the mean RMSPE.ext value was 2.316 kg/m^2 (CI95% = [1.944, 2.688]). First, we observe here what was expected from the literature: larger prediction error and lower prediction error variance with the 20% proportion. Most importantly, the performance of the model estimated with this 20% procedure is i) comparable to the performance assessed through the 10% validation dataset procedure and ii) with the models found in the literature on SOC stocks modeling at this scale (see also section 4.1.). As a conclusion and in that context, even using a larger validation dataset, the presented model could hardly be described as a poor model.

I am concerned that they appear to be overfitting – something that boosted regression trees are well known for.
Our use of the Extra model (see above) suggests otherwise. Could please the referee give more detail regarding the over-fitting of our models in the present paper? It is true that the performance of the models when estimated on the dataset used for fitting is much better than the performance obtained through cross-validation. However, as mentioned above, the performance of the BRT models presented here, assessed by cross-validation was good and comparable to what was obtained in studies found

in the literature. It might be a proof of non-overfitting as overfitted models are poor predictors of independent data. Anyway, throughout the paper, we have consistently used the results of the cross-validation and not of the fit (which might have presented the model's performance too favorably) for assessing the model and the validity of the SOC stock estimate.
Additionally, the gbm implementation of BRT involves a stochastic gradient boosting procedure: at each iteration of the boosting procedure, the dataset is subsampled and the model fitted on the sampled individuals. This procedure was in fact introduced for minimizing the risk of overfitting (Friedman 2002, Lawrence et al. 2008). Furthermore, the gbm package used in this paper includes built-in procedures for identifying the optimal number of iterations in the boosting algorithm. These procedures aim at stopping before the model actually overfits, and it is true that without such procedures there would be a great risk of overfitting. We used the cv method proposed by gbm, as is it recommended by Leathwick et al., 2006 and Ridgeway, 2006. We have taken every opportunity to avoid overfitting and have the tested the consequences of a different proportion of validation data to demonstrate that the model is not overfitted.

Had they used some of the variables to fit a General linear model and then used some kind of spatial residuals they would have been able to use the spatial structure of the data to predict and use the actual data points to control that prediction as well as the underlying variables on the smaller grid. I would have thought an approach such as Marchant et al 2010 – which applied a robust spatial prediction method to different soil properties measured in the same survey would have generated a more robust methodology.
The approach proposed by marchant et al. 2010 is indeed a very promising one, as it enables researchers to mix regression models relating environmental factors to the studied property and spatial dependency among the observations. Furthermore, this approach includes a robust methodology for setting aside the outliers.
Nevertheless, the method proposed by marchant et al. 2010 do not yet include some

novel features included in the statistical models presented here, i.e. handling nonlinear relationships between qualitative and quantitative predictors and independent variable, nonlinear interactions between the predictors, in an automated manner.
Both approaches share the robustness to the presence of outliers in the dataset. We might add that using marchant et al. 2010 like approaches, but with BRT like models instead of General linear model, is our ultimate goal. However, to our knowledge this has not been done yet and there is, as yet, no such "ready to use" method.
Lastly, many recent studies in the literature about SOC spatial estimation, involved models without any spatial dependency terms (for instance Leifeld et al., 2005, Grimm et al. 2008, Yang et al. 2008, Meersmans et al. 2008, Bui et al. 2009). We do not claim that this is a good thing, with regard to the fact that the above-mentioned studies could possibly be improved by including a spatial dimension within the models. However, they were validated, as is, by the scientific community. Omitting spatial dependency terms is often justified by the fact that the SOC driving factors are site specific, contrary to diffuse processes that influence cadmium distributions in soils (Marchant et al. 2010) for instance. Even if some factors influencing SOC may have themselves a spatial structure, this is accounted for via. their inclusion in the prediction models.
Regarding the resolution of the final map (12x12km^2), it was constrained by the climatic grid, which provided spatial distribution of among the most influential predictors for SOC (rainfall, temperature). For the French territory, a finer resolution exists but only at 9x9km^2 and results from complex interpolation models developed by the French meteorological institute. Unfortunately, it was not available for this study and in any case, would not improve the resolution greatly.

The actual variables that are included in the BRT are not well described and I find it hard to understand how they generated all these variables across the whole of France. In particular what is the difference between wregime and wlogging?
This remark was also made by referee#1. We revised the variables description (includ-

ing wregime and wlogging) and the sections where the description of the generation of the variables across the whole of France is done. For wregime and wlogging we quoted “wregime indicates soil moisture regime resulting from field observations (by soil scientists). It is described according to 6 classes, depending on the saturation, degree and its timing : ”permanently saturated ”, ”saturated every day”, ”saturated for some part of the year ”, ”continuously moist”, ”dry for some part of the year”, ” continuously dry”. wlogging describes the origin of waterlogging (perched water table, groundwater, springs and resurgences, submersion), according to field observations”.

The general standard of English needs to be improved throughout the paper.
See the introducing remark.

References not already cited in the manuscript :

Xu, Q. Liang, Y. (2001). Monte Carlo cross-validation. Chemometrics and Intelligent Laboratory Systems 56, 1, 1 - 11. Friedman, J.H. (2002). Stochastic gradient boosting. Comput Stat Data Anal 38, 4, 367-378. Leifeld, J., Bassin, S. Fuhrer, J. (2005). Carbon stocks in Swiss agricultural soils predicted by land-use, soil characteristics, and altitude. Agr Ecosyst Environ 105, 1-2, 255-266. Leathwick, J.R., Elith, J., Francis, M.P., Hastie, T. Taylor, P. (2006). Variation in demersal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees. Mar Ecol-Progr Ser 321, , 267-281.

Please also note the supplement to this comment:
http://www.biogeosciences-discuss.net/7/C4623/2011/bgd-7-C4623-2011-supplement.pdf

---

Interactive comment on Biogeosciences Discuss., 7, 8409, 2010.