SOIL Discuss., 1, C522–C526, 2015
www.soil-discuss.net/1/C522/2015/

SOIL Discussions Open Access

Interactive comment on "Comparison of spatial association approaches for landscape mapping of soil organic carbon stocks" *by* B. A. Miller et al.

B. A. Miller et al.

miller@zalf.de

Received and published: 30 January 2015

We thank the reviewer for their comments and suggestions for improving the paper. The following is our response to the specific comments.

Comment-1: I wonder about the effect of the 10 closely-paired samples. . .in addition to ensuring the estimated errors capture random variation, could it induce a bias towards these sample points?

Response-1: The closely-paired samples were spread across the feature space, which should help them be representative of the random variation across the spectrum of conditions and minimize emphasis on a particular set of conditions. Some text was added to explain this part of the sampling strategy.

The Cubist software does have an option for dealing with data sets that are biased towards a grouping of target variable values. Our experimentation with this feature did not produce regression equations or estimated errors that were very different from those produced without that feature activated. This gives us confidence that the closely-paired samples were distributed in a way to not bias the models built.

Comment-2: I am a little bit worried about the large number of potential predictors in the pool compared to the number of data on the target variable (117). I appreciate that the authors use the discussion to suggest explanations of why particular predictors were selected, and thus partly validate their selection, but am still not totally convinced that the same could not be done even with junk data and this many potential predictors. I wonder if some acknowledgement of the potential of data-mining software to overfit should be included and commented on. I don't think Cubist does anything to deal with the size of predictor pool (in a multiple hypothesis testing kind of way): : :a comment on this issue could be useful.

Response-2: Text has been added to provide more detail about how we used Cubist to reduce the predictor pool. Specifically, multiple passes were used so that the final models were built from predictor pools equal to the size of the sets used in the models (3-19 predictors divided between multiple rules). This is explained in greater detail in the referenced paper Miller et al. (2015).

Comment-3: As you state in the methods section, for the propagation of error in the indirect approach, the variances and covariances should be those of the residuals from the fitted models, not of the data themselves. It seems that this is what was done, but lines 5 and 6 of page 770 made me wonder if the variances and covariances of the raw data had been used. Could you clarify this, as this could be an alternative explanation of the larger uncertainties resulting from the indirect approach?

Response-3: The text has been corrected to specify the use of the residual's covariance.

Comment-4: Define f in Equations 2 and 3, and explain exactly what |f| is.

Response-4: |f| is now defined with an explanation for why it is used. Equation 3 was also corrected.

Comment-5: In the cross validation (Table 5), I am not sure why the results for predictions of SOC stock by the indirect approach are omitted. I think the table should include these.

Response-5: Cross-validation was only conducted on the products of the Cubist models themselves. To cross-validate the SOC stock predictions from the indirect approach would be the cumulative cross-validations of the component models. We argue that cross-validations only offer a measure of the models' robustness (how much it would change if different points were used), but does not measure the quality of the best model's performance. For these reasons, calculating a cross-validation of the SOC stock from cross-validations of the component models would be intensive without a large gain in information.

Comment-6: I think it would also be good to provide some validation of the uncertainties. . .I appreciate the difficulties of validating with a small, clustered dataset such as this, but I think it would be worth including some measure of the adequacy of uncertainty assessment in the cross validation. One possibility is the mean of the theta-statistic, which should be close to 1 (see e.g. Lark, RM. 2000. A comparison of some robust estimators of the variogram for use in soil survey. European Journal of Soil Science, 51, 137-157).

Response-6: We contend that the cross-validation statistics are not really a validation of the model based on all of the sample points, especially for MLR models of soil systems. For this reason, our focus in this paper is to compare the results of the two modelling approaches with each other. Nonetheless, the change in the MAE between the models used (based on all points) and the cross-validation models are compared. Additional text has been added to point out that the stability of the MAEs suggests that

the estimated uncertainties are also robust.

Comment-7: It is quite interesting that although the ME for all subsoil component variables was <1, the resulting predictions of SOC stock gave a ME of 1.67. This is worth commenting on in Section 3.1.1.

Response-7: We agree this is an intriguing result and text has been added to section 3.1.1. to highlight it.

Comment-8: I think that the residuals for all variables are assumed normal. . .however, depending on the dataset, it may be more appropriate to model log SOC % as normally distributed. Some comment about this, and about the effect that this could have on predictions and uncertainties in the indirect approach could be useful.

Response-8: This is indeed an important point. Text, along with a table of residual skewness coefficients, has been added to describe the distribution of the residuals and their potential effect on the results.

Comment-9: Is a conservative estimate of the spatial distribution the best thing? The most conservative would be to use the mean across the entire study area, but this would not be very useful. I am not sure whether the paper is recommending that the more conservative approach should be used, or just saying that the direct approach is more conservative than the indirect approach.

Response-9: Recognizing that different situations will have different needs, we were careful to chose the term 'conservative' to avoid judging which approach may be best for a given set of goals or purpose. Specifically, sometimes representing variation can be more useful than minimizing the amount of error, and vice versa. A sentence has been added to the conclusions to emphasize this point.

Comment-10: What exactly is meant by the 'spatial association approach'?

Response-10: Spatial association is a term parallel to spatial autocorrelation. Like the specific method of kriging is often described when applying a spatial autocorrelation

approach, the term spatial regression is a popular method for applying the concept of spatial association. We use the term spatial association frequently to emphasize the difference between our study and a similar study that compared different spatial autocorrelation methods. Additional text has been added to clarify spatial association for those who may not be familiar with the term.

Comment-11: Were all soil profiles deeper than 2 m?

Response-11: Thank you for catching this omission. Indeed field logistics prevented the full 2 m from being sampled for some of the profiles and some assumptions needed to be made. Text explaining this has been added.

Comment-12: Page 767, sentence starting on line 27: 'correlation. . .of R2 = 0.59'. Correlation should be measured by r, not R2. . .reword this sentence.

Response-12: Wording has been corrected.

Comment-13: Page 768, line 6: direct R2 = 0.14, but in Table 4 is 0.19. . .is this correct?

Response-13: The typographical error has been fixed.

Comment-14: Figures 2, 3 and 4: I am not sure that the hillshade effect helps. I found it difficult to distinguish between the effect of the hillshade and the SOC stock differences. I would suggest removing this effect.

Response-14: Agreed. The hillshade effect has been removed from the respective figures.

Interactive comment on SOIL Discuss., 1, 757, 2014.