essd-2022-418

Responses to Reviewer 3

This study aims to predict and map lithium (Li) concentration in soil across Australia using a digital soil mapping framework and environmental covariates. The model was developed using a Cubist regression tree algorithm and validated on an independent Northern Australia Geochemical Survey dataset, showing good prediction for the top depth. The importance of variables indicates that Landsat 30+ Barest Earth bands and gamma radiometric dose have a strong impact on Li prediction.

My overall impression is that despite the not relay convincing prediction power and out-of-sample verification that needs to be extended authors rigorously planned their work and did their best. For example, the set of statistics chosen to evaluate prediction performance was chosen wisely, the methodology seems appropriate (although I have questions about that), but the MS is quite worthy of being published in ESSD after the questions of all reviewers are answered. Here I underline that in my review I primarily evaluated the work from a methodological aspect.

We thank the reviewer for the constructive comments, and will address the questions accordingly.

Major comments

1. I am curious why the Cubist model was chosen. There is no literature review on which other approach could have been used in this particular exercise. Did you check other tree-based machine learning algorithms like Random forest for example. Since Random Forest can capture complex non-linear relationships between input variables and output variables by creating multiple decision trees and combining them, whereas Cubist uses linear models to estimate the output values for each leaf node of the decision trees, which may not be able to capture complex non-linear relationships.
   I work in isotope hydrology and before conducting the prediction of isoscapes we conducted through research on which approach would be most suitable for the task keeping in mind the number of predictors, the drivers of the parameter etc. Such a comparison would be necessary to be:
   - cited
   - conducted and places in supplement
   - or published in another study.

   See for example: https://doi.org/10.1016/j.jhydrol.2023.129129. In addition, a flowchart should also be added to the MS, possibly in supplement to help reproduce the steps taken.

We did not compare the performance of Cubist model to other machine learning models. Both Cubist and random forest utilise decision tree approaches. Furthermore, Cubist has been shown to show similar performance / outperform random forest; and the results are more interpretable compared to random forest.

Khaledian, Y. and Miller, B. A.: Selecting appropriate machine learning methods for digital soil mapping, Applied Mathematical Modelling, 81, 401-418, https://doi.org/10.1016/j.apm.2019.12.016, 2020.
Pouladi, N., Møller, A. B., Tabatabai, S., and Greve, M. H.: Mapping soil organic matter contents at field level with Cubist, Random Forest and kriging, Geoderma, 342, 85-92, https://doi.org/10.1016/j.geoderma.2019.02.019, 2019.

We have added a flowchart into the MS.

2. How was the preprocessing conducted, outlyers, extreme values? See for example the ultimate two paragraphs of Sect. 2.1 in https://doi.org/10.1016/j.jhydrol.2023.129129 . Did you check the outliers in the input data, I'm not sure how the Cubist model can handle them, as it uses linear models to estimate output values for each leaf node of the decision trees. Outliers can have a large impact on the estimated output values of the linear models, which can lead to inaccurate predictions.

   We did not remove any high extreme values. These values are used to identify areas with anomalous concentration of Li. The values below detection limit (<0.1) are replaced with half of the detection limit (0.05). Furthermore, NGSA dataset underwent a thorough data quality assessment, explained in:

   de Caritat, P. and Cooper, M.: National Geochemical Survey of Australia: Data Quality Assessment. Record 2011/021, 2011b.

3. A more detailed description of the used metrics is requires, since all of these account for different kind of errors. E.g. the Lin's CCC measures both the correlation and the bias between the measured and predicted values and it provides a measure of the strength of the linear relationship between the two value sets, while accounting for the magnitude of the differences.
   In addition, references should be inlcuded, e.g. Lin, 1989 https://www.jstor.org/stable/2532051

   We have updated the description of the used metric and added the relevance citation to Lin's CCC.

4. The argument in L412-413 is acceptable, but isn't there a possibility to validate the results with data from other regions or conduct a pilot study

from elsewhere? In a study in a journal as ESSD (upper 1 percentile in SJR) it would be expected to provide an even broader validation scheme, or give an extensive explanation on why this is not possible.

The trained model could not be used to validate data from other regions, as it was only trained using environmental covariates within Australia, in particular as the soil may be quite different here.

Minor comments

1. It might be more appropriate to categorize the predictors according to which ones were considered static (do not change over time) and which ones were considered dynamic (can change over time).

   In this case, we are following the scorpan concept, and hence the covariates were grouped following the scorpan method.

2. It was not discussed earlier, is a linear relationship (measured by Pearson r) required, or is there a nonlinear relationship expected between the predictors and Li content. Please elaborate on this.

   Linear relationship between the predictors and Li content are not necessarily needed. It is part of the exploratory data analysis. The machine learning model is used to develop the relationship between the predictors and Li content. The linear correlation can potentially be used to explain the usage of predictors within a model.

3. L262: What were these correlation values for Al, B, Fe..., a table should be included e.g. in the supplement.

   The correlation tables have been added in the supplementary materials.

4. The significance values should be reported and all the statistics in APA style. https://www.socscistatistics.com/tutorials/correlation/default.aspx

   We have added statistical significance for the Pearson's correlation (see Table SM1).

5. L407: This limitation is very important and must be mentioned in the abstract, in addition, it could even be incorporated into the title.

   We mentioned that our samples are collected from the catchment outlet floodplain sediments in abstract and within the Materials section.

Miscellaneous

1. Add spaces before and after mathematical operators.

   We have updated the text.

2. L20 and all other places use superscript for measurement units kg^{-1}

   We have updated the text.

3. Variables should be in italics.

   We have italised the parameters used to tune the model.

4. A more detailed description is needed on the boxplots explaining what is in the figure, see e.g. caption of Fig. 3 in https://doi.org/10.1016/j.jhydrol.2022.128925

   We have added more description in the caption.