Earth Syst. Sci. Data Discuss.,
https://doi.org/10.5194/essd-2018-103-SC2, 2018

Earth System Science Data Discussions

Interactive comment on "Depth-to-Bedrock Map of China at a Spatial Resolution of 100 Meters" *by* Fapeng Yan et al.

W. Shangguan

shgwei@mail.sysu.edu.cn

Received and published: 29 October 2018

Thanks for your valuable advices:

(1) The resolution of the produced DTB map should be much coarser than 100m. The authors declared that its spatial resolution is 100m. But I notice that most environemntal variables used in the models are at 1km resolution. Fig5 clearly shows contributions of the used variables to the DTB model prediction. The first four covariates are the dominant contributors to the DTB prediction, and they are TWI, landform units, openness index and slope all at a 1km resolution. Two variables including 100m elevation and 30m land cover are only minor contributors to the prediction. In addition to the two variables, all other variables used in the prediction are at 1km resolution. So, it is not

appropriate to say the resolution of the DTB map arrives 100m.

Reply: We agree that the produced DTB map may be considered less than 100m. However, the resolution is not only determined by the resolution of most environmental covariates (1km). As we also used covariates at 100m or finer, a map produced at a resolution coarser than 100 m may also lose some information from these covariates. Considering the balance of covariates at the coarser and finer resolutions (30m to 1km), plus the computation cost, we choose to produce the map at 100 m. We plan to update the map with covariates at finer resolution due to the suggestion of the reviewer. In our study, the most four important covariates are topographic witness index(TWI), topographic landform units(TLU), topographic openness index(TOI) and slope. Among them, the TWI, TOI and slope can be obtained from DEM which is at 100 meter. Combined with the DEM, there will be four covariates at 100 meters at the top five dominant contributors. We also will derive more topographic covariates, such as curvature, surface roughness, valley depth and so on, based on the 100 m DEM. This will further make the map with more spatial details. Of course, we expect the importance of covariates will change in Fig. 5. The new map will be recalculated based on our fully automated framework.

(2) For the list of covariates in the supplement file, I notice that although over 130 covariates (too many) were considered, only a small part of these covariates have true contribution to the final predictions. It is not necessay to list so many unuseful covariates. I suggest the authors only list the covariates showed in Fig5 or a little more. The advantage of a brief list would be easy to convey our readers a clear and brief knowledge or understanding on the relationship of the DTB and its covariates. Besides, quite a few of the covariates in the list have high correlation. Removing redundant variables would simplify the models and reduce the computation. In addition, some of the covariates may have data quality and consistence problem. For example, the layer of Landsat TM band3(red) of year 2000 was produced by mosaicing many scenes of TM images of different months and seasons in the year of 2000. The inconsistence

would introduce error to prediction.

Reply: we agree to use only the relatively important covariates in the final model. In this study we considered more than 130 covariates but we only use a part of covariates in the final model to reduce the noise as much as possible, this has been explained in our manuscript and supplement file. We also agree that it is better to reduce the number of covariates to simplify the models and reduce the computation. Fig5 shows the top 20 important covariates. It should be aware that the importance of covariates based on random forests model in Fig5 should be seen as a relative measurement of their contributions. A feature's importance is the increase in the model's prediction error after we permuted the feature's values. This measure has some limitations (see https://christophm.github.io/interpretable-ml-book/feature-importance.html#disadvantages-7). When features are correlated, the permutation feature importance measure can be biased by unrealistic data instances. Adding a correlated feature can decrease the importance of the associated feature, by splitting up the importance of both features. As a result, we can not drop a covariate only based on the low importance, and the change of the performance of the model should be considered. We described the modified process of choosing covariates as follows: We first used all the covariates to fit a model, then some of covariates with low importance were dropped in the final model. Covariates with no or weak relations with DTB may produce noise in fitted models. This noise results in higher error of predictions. Our results based on modeling with different covariates showed that the noise has a certain degree of influence on the accuracy of the models, especially for the gradient boosting tree model. In addition, some of the covariates may have data quality and consistence problem, which would introduce error to the prediction. Therefore, we removed some covariates with low importance based on the random forests model to reduce prediction errors, model complexity and computation time. Because there are some limitations of the importance and the importance of correlated covariates is underestimated (Molnar, 2018), a covariate was removed only when it does not make the model without this covariate significantly worse, i.e., when the R2 of the model decreased more than 0.01 or

increased. In this way, we kept the balance between the model complexity (i.e. number of covariates) and model accuracy. The covariates we used are listed in Supplement File A. The last two column shows the choice of covariates for the RF and GBT model, where a value of one indicates that the corresponding covariate was used in the final model. The final RF and GBT model used ?? and ?? covariates, respectively. (The number of covariates will be determined after the recalculation)

(3) The method framework of the prediction is almost the same to Shangguan et al. (2017) and Hengl et al (2017?). It would be good to clearly refer to these previous work in the method part and the figure3.

Reply: The method framework is the fundamental framework of "scopan" method in digital soil mapping. We used the method partly based on the work of Hengl et al (2017) and Shangguan et al. (2017) as well as our practical work. We modified the first sentence of section 2.4 as: The framework of our research is shown in Fig. 3, which is based on the work of Hengl et al (2017) and Shangguan et al. (2017).

(4) The authors used RF and GBT models to produce a map of DTB but used another model 'quantile regression forest' to estimate uncertainty of the DTB map. This is not consistent. The problem is that the resulting uncertainty estimation may not actually reflect true uncertainty of the DTB map.

Reply: We were aware of the inconsistency between the prediction by RF and GBT models and the uncertainty derived by quantile regression forest. Due to the reviewer's comment, we will offer two sets of data in the next revision. One is the prediction by the ensemble of RF and GBT models, which avoids the overshooting effect (Sollich and Krogh, 1996) and provides a more robust estimation. The other is the prediction and the uncertainty by quantile regression forest. Because most users do not need an uncertainty map in their applications, it is recommended to use the ensemble prediction and take the uncertainty map as a reference. In cases where a consistent prediction and uncertainty are needed, it is recommended to use the estimation by quantile re-

gression forest. Although the mean prediction (quantile is 0.5) is somewhat different from the prediction of ensemble model, it was also validated and will be a better choice for users who need both final prediction and uncertainty estimation. We will also show the map and its accuracy by quantile regression in the revised manuscript. We added the following sentenced in section 2.4: Two sets of data are provided for users. One is the prediction by the ensemble of RF and GBT models and the other is the prediction and the uncertainty by quantile regression forest. Because most users do not need an uncertainty map in their applications, it is recommended to use the ensemble prediction and take the uncertainty map by quantile regression forest as a reference. In cases where a consistent prediction and uncertainty are needed, it is recommended to use the estimation by quantile regression forest.

(5) Line52-57: two problems in thses words: 1) In soil survery, when soil thickness is greater than 2m but the observed depth is less than 2m, the surveyors never record the soil thickness as a value lower than 2m BUT record it as a censored data '>2m'. This is standard recording in soil surveys. 2) The reason why soil survey generally does not reach bedrock for some very thick soils is NOT the equipment and technological constraints.This should come to the purpose of soil survey, it is just not necessary to reach bedrock.

Reply: According to this comment, we will revise Line52-57 to the following contents: The observed depth of a soil profile is generally less than 2 m, and the thickness of the soil is therefore recorded as a value less than 2 m or a censored value (>2 m) when the soil thickness is greater than 2 m. For the purpose of soil surveys, it is not necessary to dig deeper than 2 m and reach bedrock in most cases.

Interactive comment on Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2018-103, 2018.