SOIL Discuss.,
https://doi.org/10.5194/soil-2018-28-AC1, 2018

Interactive comment on "Using deep learning for Digital Soil Mapping" *by* José Padarian et al.

José Padarian et al.

jose.padarian@sydney.edu.au

Received and published: 8 October 2018

Thanks for your comments to our manuscript.

Data

We will add extra information about the dataset in our revised manuscript.

Printer-friendly version

Discussion paper

Data augmentation

Data augmentation can certainly be used with other machine learning techniques, not only CNNs. It is reasonable to think that we should use data augmentation in the comparison but the fact is that the previous study (Padarian et al., 2017) but most researchers use point information (a vector of covariates) as input for the Cubist model. Transforming that vector into the same 3D structure that we used in this study generates a 1x1 image with n channels, where n is the number of covariates. Rotating that image by 90, 180 and 270° generates exact copies of the image, which is oversampling and not data augmentation.

A possible alternative is to use some context (pixels from the vicinity), but Cubist is not designed to handle that type of information, thus making the comparison even less level. If the 3D array is flattened as a vector, adding more context actually increases the prediction error.

In terms of the autocorrelation of the data, we assumed that there is no variance when distance=0 by adding samples with exactly the same SOC content. That is theoretically true if we consider that the distance is exactly equal to 0. In practical terms, when calculating the semivariogram, the semivariance value of the first bin will be lower, but that does not significantly affect the final model. We will expand the discussion to clarify this point in the revised manuscript.

Dataset names

Test dataset: We selected 10% of the original data (n=49) We augmented the remaining 90% (n=436) obtaining 1,744 training data. With those samples, we ran a bootstrapping routine. At each repetition, we sampled with replacement (n= 1,744) which we used as a training dataset. A sampling with replacement usually draws around 2/3

(66.66%) of the samples, hence excluding 33.33% of the samples (which we used as a validation set and to select hyperparameters). That is the meaning of the 1/3. Just to clarify, 1/3 is not a 1:3 ratio between the size of the training and validation sets.

Results

Figure 4: Because we are using a multi-task CNN, we excluded the samples with depths lower than 100cm. In consequence, all depth have the same number of observations (we will add a clarification to make it more explicit). It is reasonable to assume that there is a greater effect of data augmentation due to the lower range of SOC content values in depth, but it is something that has to be confirmed with future studies.

Figure 5: Actually, it is perfectly possible to have a lower error in the test dataset. That would be the case if the SOC content of the samples are relatively low (which is the case for this test dataset) because the error is higher in samples with larger SOC content. A lower error in the test dataset is not an indication of overfitting. A larger error would be an indication of overfitting.

Section 5.6 We will re-phrase the text to make it clearer. "... higher areas of the landscape" is in terms of elevation, which is where we saw high uncertainty in Padarian et al., (2017). About the proportion of the landscape that has a great reduction of intervals width, we think that a map with a proper colour palette shows exactly that, and much better than a non-spatial calculation.

Interactive comment on SOIL Discuss., https://doi.org/10.5194/soil-2018-28, 2018.

Printer-friendly version

Discussion paper