Assessment of Digital Soil Mapping products: independent ground-truthing is essential

Elisabeth Bui[A]

[A]CSIRO Land and Water, GPO Box 1666, Canberra ACT 2601, Australia

Abstract

Models and maps for predicted soil properties produced over agricultural areas of Australia using legacy soil survey data have been viewed with suspicion by many, yet these are key to the success of projects such as the GlobalSoilMap.net. While the modelling procedure encompassed a statistical model uncertainty assessment, that assessment is short of an accuracy assessment of the predicted maps. Here models and predicted maps for topsoil (0-30 cm) soil organic C, total N, and total P are presented and assessed against new independent data that serves as ground-truth for an accuracy assessment of the maps. The map of predicted SOC is credible, more so than the map of total N that consistently over-estimates N.

Keywords

Accuracy, error propagation, digital soil mapping, legacy data, soil organic C, total N, total P.

Introduction

During the first phase of the National Land and Water Resources Audit (NLWRA) the Australian Soil Resources Information Systems (ASRIS) project in 2001, a relatively large point database of soil properties was created by collating various legacy databases into a single Oracle database (Johnston *et al*. 2003). Depending on the soil property, 5,000 to 24,000 points had useful data. Using this point database linked to national environmental data for climate (19 continuous variables), geology (23 discrete classes), land use (14 discrete classes), 4 Landsat MSS bands, and topography (14 continuous terrain variables), rule induction using Cubist (http://www.rulequest.com) decision trees was used to predict the spatial distribution of soil properties across the intensively used agricultural areas of Australia (Henderson *et al*. 2001). In this talk, models and predicted maps for soil organic C (SOC), total N and P in the 0-30 cm depth interval will be presented and assessed. The maps have been produced at 0.01° resolution (~1.1 km) and are part of the Australian Natural Resources Atlas, available from: http://www.nlwra.gov.au/national-land-and-water-resources-audit/atlas.

Modelling and statistical diagnostic assessment

Cubist models are presented as a series of rules, each with starting with conditional if statement that subsets the data. Continuous predictor variables can feature as splitting criteria in conditional statements and in the linear regressions at each leaf of the piecewise linear decision trees but categorical predictors can only be used to subset the data. Models were constructed with a 70:30 training to test data split: 70% of the observations were randomly selected to construct the model in the model development stage; 30% were held back in order to assess the performance of each model. Once the strongest possible model according to performance on the test data was identified, it was refitted using all the data to maximize the use of the relatively sparse data over Australia, with the same model form and options. The performance of the model on the full data set was assessed by 10-fold cross validation. The data were randomly split into 10 partitions or folds; at each step, nine of these partitions were used to fit the model and the performance assessed on the remaining partition held back as the test data. This procedure was repeated for each partition sequentially. The performance, averaged over all 10 partitions held back, delivers the cross-validated performance assessment. The performance of models was also assessed in terms of a number of key indicators: the number of points used in the model, the R^2 between measured and predicted values, the (rank) correlation, the RMSE (root mean square error), which gives an estimate of the standard deviation of the errors, the average error, and the relative error. The average error gives the average absolute difference between the observed and predicted values, i.e

$$\text{average error} = \frac{1}{m}\sum_{j=1}^{m}\left|y_j - \hat{y}_j\right|.$$

Lower average errors imply that the predicted values are closer to the observed values more often. The average error is also known as the mean absolute deviation. The relative error is defined as the ratio of the average absolute error magnitude to the average error magnitude that would result from predicting the mean value:

$$\text{relative error} = \frac{\frac{1}{m}\sum_{j=1}^{m}\left|y_j - \hat{y}_j\right|}{\frac{1}{m}\sum_{j=1}^{m}\left|y_j - \bar{y}\right|}$$

If there is little improvement on the mean, the environmental variables have little predictive capacity and the relative error is close to 1. Generally, the smaller the relative error, the better the model. These model diagnostic statistics were reported for the model test subset and for discrete regions of Australia in (Henderson *et al*. 2001). They are summarized in Table 1 for the Cubist models used to map SOC and P.

Table 1. Performance of final Cubist models used to make predictions and map the soil properties

Model diagnostics	SOC	Total P
Number of points used		
	11483	7377
R^2 (predicted vs observed)	0.49	0.83
Average error	0.38	0.61
Relative error	0.64	0.49

Because of a strong correlation between SOC and total N and the poor spatial distribution of points with N measurements, N was predicted as a function of SOC: log N= -2.6589 + 0.8761 log SOC (Henderson *et al*. 2001). The simple linear regression model for total N appeared to over-predict N at the low end but its statistical performance assessment was good: RMSE was relatively low (0.42 on the log scale, R^2 = 0.75) (Henderson *et al*. 2001).

Knowledge-based assessment

The modelling is not explicitly spatial, i.e., it does not use geographical coordinates as predictors, rather, spatial structure is introduced implicitly by reliance on predictors that are available spatially extensively. In the absence of newly collected, independent ground-truth data, evaluation of the models in a spatial context can proceed via an evaluation of the spatial distribution of the predictors in the context of model structure (Bui *et al*. 2006). ASRIS maps were thus assessed against expert knowledge in natural sciences using visualization of model rules and of patterns of usage of predictor variables—what variables were important in models, whether consistent patterns emerged in their thresholds, and the spatial pattern defined by these thresholds (Bui *et al*. 2006).

The Cubist model for soil organic C had 29 rules whereas the model for total P had 18 rules—the smaller number of rules for the P model suggests that the environmental correlation patterns are more evident in that dataset. The model for total P performed better than that for SOC in terms of model evaluation statistics (Table 1) however both appeared reasonable in terms of their predicted spatial patterns and what is known about the soil processes driving these soil nutrient patterns (Bui *et al*. 2006). Climatic variables alone were the most important predictors in the SOC model whereas lithology was also important in the total P model. Visualization of model rules showed a spatial correspondence between extent of rules and bioregions of Australia, as independently determined by the Interim BioRegionalization of Australia expert committee. A major spatial pattern in climatic thresholds seemed to correspond to soils with SOC > 2% and to the distribution of rainforests and *Eucalyptus* forests along the Australian coast.

Independent assessment/Validation

The diagnostic performance evaluation gives an estimate of the uncertainty associated with the models. However the accuracy of the predictions from the models still needs to be assessed—in remote sensing research, this is referred to as ‘validation’ and is usually performed by collecting independent ground-truth data. The dataset reported in the Auxiliary Material of Wynn *et al*. (2006) has been used as an independent dataset for validation of the predictions in ASRIS. The data of Wynn *et al*. (2006) were collected over 1999-2002 using a sampling design spatially stratified across the range of Australian native vegetation formations, and analysed by a single laboratory procedure (LECO furnace) for SOC and total N for depth 0-30 cm, near and away from trees. Unfortunately, no P data are reported. A total of 25 points overlap with the ASRIS extent.

While it appears that the SOC model for the topsoil layer (0-30 cm) is relatively poor based on the Cubist statistical model evaluation (Table 1), validation against independent data collected by Wynn *et al*. (2006) suggests that the predictions are better than suggested by the Cubist model diagnostic statistics (Figure 1). This discrepancy is likely due to the laboratory measurement errors associated with different SOC determination procedures pooled together in the ASRIS database (Henderson *et al*. 2001; Johnston *et al*. 2003): the ASRIS point database used to build and test the Cubist models contains a lot of errors. Nevertheless the Cubist algorithm was able to identify meaningful structure under fairly low signal to noise conditions to generate a credible model for topsoil SOC.

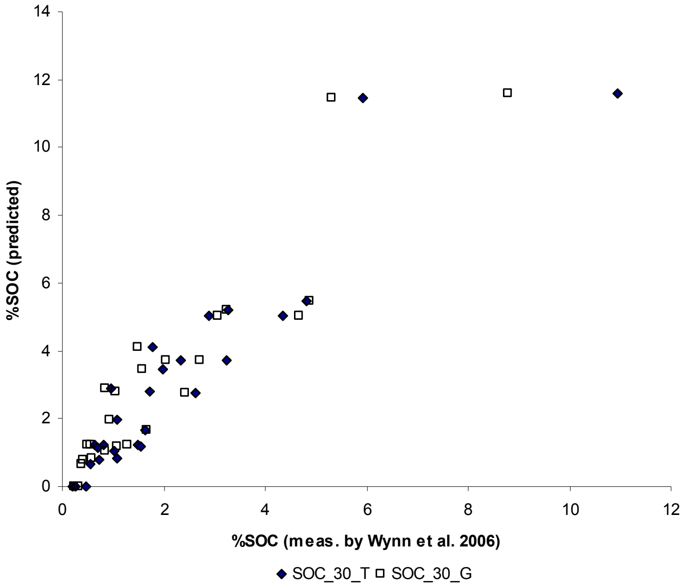

Figure 1. Relationship between SOC predicted with ASRIS data and data reported by Wynn et al. (2006). R^2 between predictions and SOC_30_T (near trees) is 0.84 and R^2 between predictions and SOC_30_G (away from trees, in grass) is 0.84. There is a tendency toward over-estimation of SOC.

Because of its reliance on the linear regression relationship with SOC, the total N map incorporates errors in the underlying SOC map. Validated against the data of Wynn *et al*. (2006), the modelled map for total N was found to be consistently over-estimating N throughout the range of N values (Figure 2), not only at the low end as suggested by Henderson *et al*. (2001). The likely error at high N content is much larger than at low N.

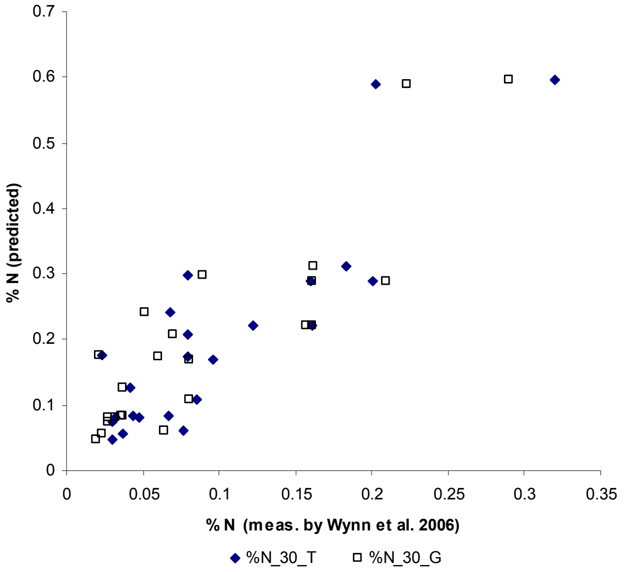

Figure 2. Relationship between total N predicted with ASRIS data and data reported by Wynn et al. (2006). R^2 between predictions and N_30_T (near trees) is 0.76 and R^2 between predictions and N_30_G (away from trees, in grass) is 0.77.

This problem starts with the tendency toward over-estimation in the SOC map but is also partially due to the logarithmic transformation of N and SOC data in the linear regression model used to produce the N map. Plotting SOC against total N on a linear graph shows that there are two sub-populations, a large one associated with a C:N ratio of ~12 and another smaller one associated with C:N ratio of > 12 (Figure 3); these are not so evident on a log-log graph. Soils with a high C:N ratio have a low N content and their N level may be over-estimated by the model used to make the total N map.

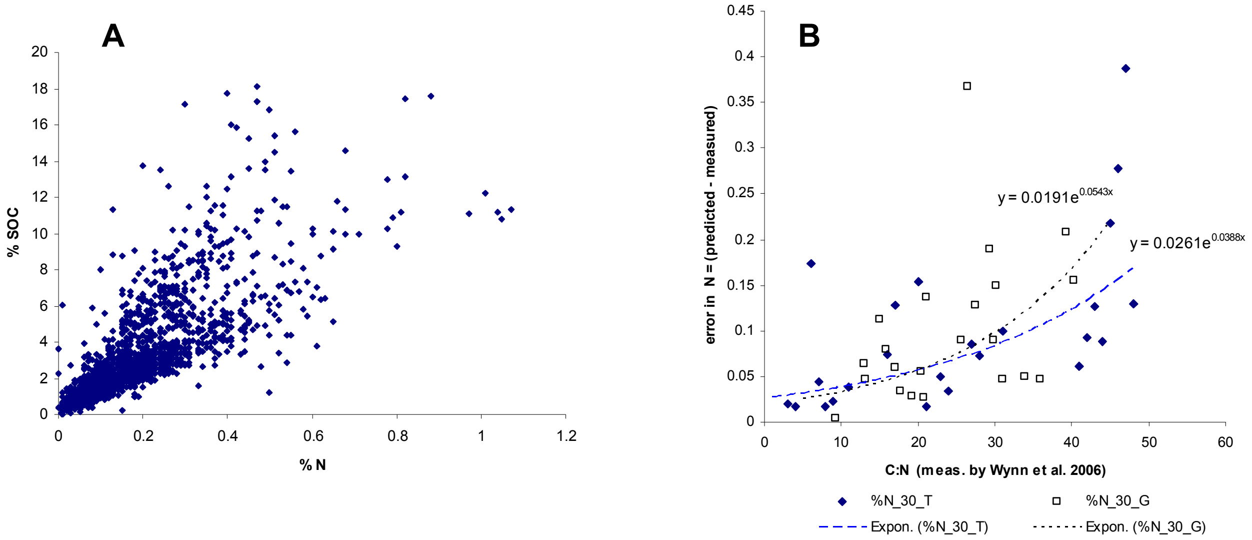

Figure 3. A) Relationship between topsoil SOC and total N on linear axes; B) Error in N predicted becomes exponentially larger as C:N increases.

Conclusion

Whereas uncertainty assessment of the models using statistical diagnostics appeared to suggest that the SOC map was not likely to be reliable, accuracy assessment against newly collected independent data suggests that the map is credible, although it shows a slight tendency toward over-estimation. The map of total N over-estimates N content consistently, especially at the upper end of the range—this shows how errors can be propagated and amplified during modelling. The P map could not be independently assessed for its accuracy. Although the independent dataset is small it demonstrates that ground-truth is essential for accuracy assessment of digital soil mapping predictions and that even a limited number of ground-truth points can be informative.

References

Johnston RM, Barry SJ, Bleys E, Bui EN, Moran CJ, Simon DAP, Carlile P, McKenzie NJ, Henderson B, Chapman G, Imhoff M, Maschmedt D, Howe D, Grose C, Schokneckt N, Powell B, Grundy M (2003) ASRIS: The database. *Australian Journal of Soil Research* **41**, 1021-1036.

Henderson BL, Bui EN, Moran CJ, Simon DAP, Carlile P (2001) ASRIS: Continental-scale soil property predictions from point data. *Technical Report 28/01*, CSIRO Land and Water, Canberra.

Bui EN, Henderson BL, Viergever K (2006) Knowledge discovery from models of soil properties developed through data mining. *Ecological Modelling* **191**, 431-446.

Wynn JG, Bird MI, Vellen L, Grand-Clement E, Carter J, Berry SL (2006) Continental-scale measurement of the soil organic carbon pool with climatic, edaphic, and biotic controls. *Global Biogeochemical Cycles* **20**, GB1007, doi:10.1029/2005GB2576.