Estimating soil organic carbon stocks using machine learning methods in the semi-arid rangelands of New South Wales

B. Wang[a], C. Waters[b], S. Orgill[a], A. Clark[c], D. L. Liu[a], M. Simpson[c], A. Cowie[d], I. McGowen[c] and T. Sides[a]

[a] NSW Department of Primary Industries, Wagga Wagga Agricultural Institute, NSW 2650, Australia
[b] NSW Department of Primary Industries, PMB 19, Trangie, NSW 2823, Australia
[c] NSW Department of Primary Industries, Orange Agricultural Institute, NSW 2800, Australia
[d] NSW Department of Primary Industries, Trevenna Rd, Armidale, NSW 2351, Australia
Email: bin.a.wang@dpi.nsw.gov.au

Abstract: Soil organic carbon (SOC) is pivotal for biological, chemical and physical processes and provides vital information on changes in soil fertility and land degradation. Rangelands, accounting for about 81% of Australian land area, represent considerable carbon storage potential. Efficient modelling techniques to evaluate the potential for rangeland SOC stocks are vitally important in the assessment for the global carbon cycle and quantum abatement. This study aimed to evaluate boosted regression trees (BRT) and random forest (RF) in predicting SOC stocks from ground measured and remotely-sensed variables using two feature selection techniques to identify the dominant variables that affect SOC stocks in the rangelands. Using field-based measurement of SOC stock collected from 564 sites across the study area and 28 of GIS-based environmental variables including climate, topography, radiometry, vegetation and land fractional cover data, we employed stepwise regression (SR, linear approach) and genetic algorithm (GA, nonlinear approach) to select the most informative variables. These selected predictors were then used to train the BRT and RF models. In all, four models were evaluated; BRT using stepwise selection of predictors (SR_BRT); RF using stepwise (SR_RF); BRT using GA selection of predictors (GA_BRT) and RF using GA (GA_RF). In addition, BRT using all predictors (All_BRT) and the RF using all predictors (All_RF) were used as benchmarks to test the performance of the four models. Of the field-based data, 75% was used to train the model ("calibration dataset") and the remaining 25% was used to validate the prediction of SOC stocks ("validation dataset"). The results indicate that the RF exhibited a better performance in predicting SOC stocks than the BRT regardless of input variables. The two models explained ~45% of the total SOC stocks. In addition, we verified that feature selection for both machine learning techniques is necessary for estimating SOC stocks, even though BRT was relatively insensitive to the input features selected by SR. The GA_RF was the most promising model with reliable predictors to predict SOC stocks, with the lowest root mean square error (RMSE) and the highest R^2 values (7.44 Mg C ha^{-1} and 0.48, respectively), suggesting that the proposed methodology may provide a cost effective method to predict SOC stocks in the rangelands. The important variables for explaining the observed SOC stocks were rainfall, elevation, prescott index (PI), and land fractional cover (bare ground fraction).

Keywords: *Soil organic carbon stocks, random forest, boosted regression tree, genetic algorithm, stepwise regression*

1. INTRODUCTION

The rangelands are extensive grazing areas, characterised by low, erratic rainfall and account for about 81% of the Australian land area (http://www.environment.gov.au/land/rangelands) (Allen et al., 2013). It is estimated that Australian rangeland soils contain between 34 and 48 Gt of carbon, representing a sequestration potential of 78 Mt C per year (Keating et al., 2009). Soil organic carbon (SOC) plays a vital role in a range of soil processes including the recovery of degraded soil and provides information about soil fertility. Therefore, accurately assessing the stock and distribution of SOC is essential to enhance this resource.

The estimation of SOC stocks using statistical models has been achieved by using the relationship between environmental variables (such as climate, soil properties and topography) and SOC stocks (Akpa et al., 2016; Badgery et al., 2013; Bonfatti et al., 2016). Identifying and understanding the factors influencing the amount and variability in SOC across different landscapes are of prime importance in quantifying the role of increases in rangeland SOC stocks to meet emissions reduction targets globally. In south eastern Australia, field surveys have demonstrated a significant influence of environmental variables on SOC stocks in agricultural systems with a small but varied influence of land management (Rabbi et al., 2014). Few studies have examined the role of environmental variables on SOC stocks in Australian rangelands.

Several studies have demonstrated that machine learning algorithms are more accurate than traditional statistical methods such as stepwise linear regression, principal component regression and partial least squares regression (Guo et al., 2015; Mouazen et al., 2010), especially in complex ecological systems. Of these techniques, tree models such as the boosted regression tree (BRT) and random forest (RF) have been widely used to estimate SOC stocks because of advantages including fewer parameters and an ability to investigate non-linear and hierarchical relationships between the predictors and the response (Everingham et al., 2016). Although various machine learning algorithms have been widely used, it is still difficult to produce a robust model to predict SOC stocks due to high levels of SOC variation and complex relationships with environmental variables. A number of other studies have considered different methods to improve model performance in predicting soil properties (e.g. SOC stocks) (Guo et al., 2015; Ließ et al., 2016; Xie et al., 2015). From these studies, selection of predictor variables (i.e. feature selection) which elucidated the most relevant or informative input variables can minimize errors and develop the most robust models. Among the numerous feature selection methods available, genetic algorithms (GA) have been demonstrated to have superior performance in recognising differences in soil types (Xie et al., 2015).

Recent studies in the semi-aid rangelands have shown clear relationships between ground cover (perennial and, litter) and SOC stock (Orgill et al., 2017; Waters et al., 2016). These relationships suggest suitable satellite-derived covariates such as fractional cover data may be useful in the estimation of SOC stocks. The objective of this study was to evaluate alternative methods to derive predictions of SOC stocks from environmental variables. Specifically, this study aimed to (1) evaluate feature selection techniques to identify the dominant variables that affect SOC stocks and (2) compare RF and BRT methods to determine the most reliable and accurate model to predict SOC stocks in the surface (0.30 m) of the soil profile in the semi-arid rangelands of eastern Australia.

2. MATERIALS AND METHODS

2.1. Study area

The model training area (referred to as the study area), is located in the western New South Wales, Australia (Fig. 1). It lies between the latitude of 29.64° and 32.28° South and the longitude of 145.54° and 146.06° East. The dominant land use is extensive grazing. The climate is classified as BSh (arid-steppe-hot arid) in the world of Köppen-Geiger climate (http://koeppen-geiger.vu-wien.ac.at/present.htm) with an average annual rainfall and temperature, 379 mm and 18.9 °C, respectively. The elevation ranged from 105 to 398 m and the study area is dominated by small patches of shrubs,

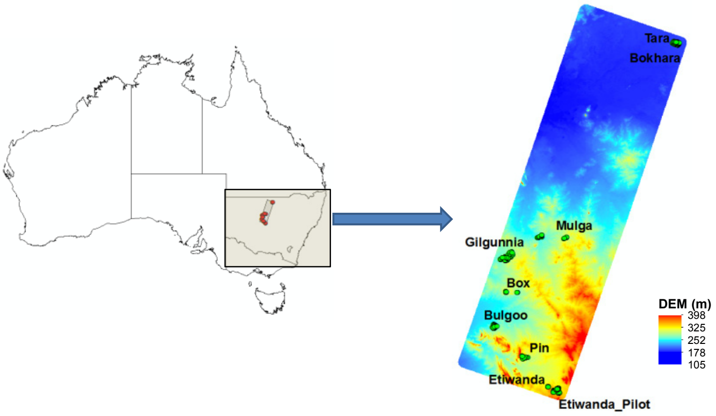

Figure 1. Location of the study area and 564 sampling sites.

scattered trees and large areas of open grasslands. A total of 564 data points within the sample area were used to calibrate and validate the models.

Table 1. A total of 28 environmental variables used in the prediction of SOC stocks in the study area.

Variables	Definition and formula	Resolution
Topography		
Slope	The inclination of the land surface from the horizontal	30m
Elevation	The height of a location above the Earth's sea level	30m
Aspect	The direction in which a land surface slope faces	30m
Topographic Wetness Index (TWI)	The relative wetness within moist catchments, but is more commonly used as a measure of position on the slope with larger values indicating a lower slope position	30m
Partial Contributing Area (CA)	Contributing area in m^2 computed using multiple flow directions on hillslopes and ANUDEM-derived flow directions in channels	30m
MrVBF	Measure of flatness and up-ness	30m
MrRTF	Identify high flat areas at a range of scales	
Plan Curvature (PLC)	The rate of change of aspect (across the slope) and represents topographic convergence or divergence	30m
Profile Curvature (PRC)	The rate of change of potential gradient down a flow line and represents the changes in flow velocity down a slope.	30m
Prescott Index (PI)	Measure of water balance	30m
Topographic Position Index (TPI)	Topographic position classification identifying upper, middle and lower parts of the landscape	30m
Vegetation/anthropogenic factors		
NDVI	Normalized Difference Vegetation Index _NDVI=(NIR[a]-R[b])/(NIR+R)_	30m
EVI	Enhanced Vegetation Index _EVI=2.5*((NIR-R)/(NIR+6*R-7.5*B[c]+1))_	30m
SAVI	Soil Adjusted Vegetation Index _SAVI=((NIR-R)/(NIR+R+0.5))*(1+0.5)_	30m
MSAVI	Modified Soil Adjusted Vegetation Index _MSAVI=(2*NIR+1-sqrt ((2*NIR+1)^2-8*(NIR-R)))/2_	30m
NDMI	Normalized Difference Moisture Index _NDMI=(NIR-SWIR1[d])/(NIR+SWIR1)_	30m
NBR	Normalized Burn Ratio _NBR=(NIR-SWIR2[e])/(NIR+SWIR2)_	30m
NBR2	Normalized Burn Ratio 2 _NBR2=(SWIR1-SWIR2)/(SWIR1+SWIR2)_	30m
Fractional cover data		
Band 1	Bare ground fraction (bare ground, rock, disturbed)	30m
Band 2	Green vegetation fraction	30m
Band 3	Non-green vegetation fraction (litter, dead leaf and branches)	30m
Band 4	Model fitting error	30m
Climate		
Rainfall	Mean rainfall	-
Temperature	Mean temperature	-
Radiometrics		100m
Potassium	Concentrations of the radioelements potassium	100m
Uranium	Concentrations of the radioelements uranium	100m
Thorium	Concentrations of the radioelements thorium	100m
Dose	Terrestrial gamma-ray dose rate	100m

[a]NIR: near infrared; [b]R: red; [c]B: blue; [d]SWIR1: shortwave infrared-1; [e]SWIR2: shortwave infrared-2.

2.2 Climate, topography and GIS-based environmental variables

A total of 28 environmental variables (predictor variables) that could be related to SOC stocks are provided in Table 1. Predictors shown in Table 1 in bold were selected by stepwise regression and those underlined were selected by genetic algorithm.

2.3 Feature selection

Since some variables among the 28 predictor variables are inter-related (redundant) (Fig. 2), they are better to be avoided when developing optimal model in predicting SOC stocks. A linear approach (stepwise linear regression; SR) was used to find redundant predictors and select the smallest set of predictors giving the best linear regression results. In addition, a nonlinear approach (GA) was used to select the most informative predictor variables. The linear approach revealed that 11 of the 28 environmental variables were significant for estimating SOC stocks ($P<0.05$), and retained in the SR model. These significant variables included: elevation, rainfall, Band 1, NBR, PI, temperature, potassium, MrVBF, Band 4, uranium and MrRTF. In GA model, we used 10-fold cross-validation with 100 iterations. The GA model with SOC stocks as the target variable and all ancillary variables as inputs, resulted in 11 predictors including: elevation, rainfall, PI, Band 1, MSAVI, MrVBF, Band 4, Band 3, MrRTF, TWI and TPI.

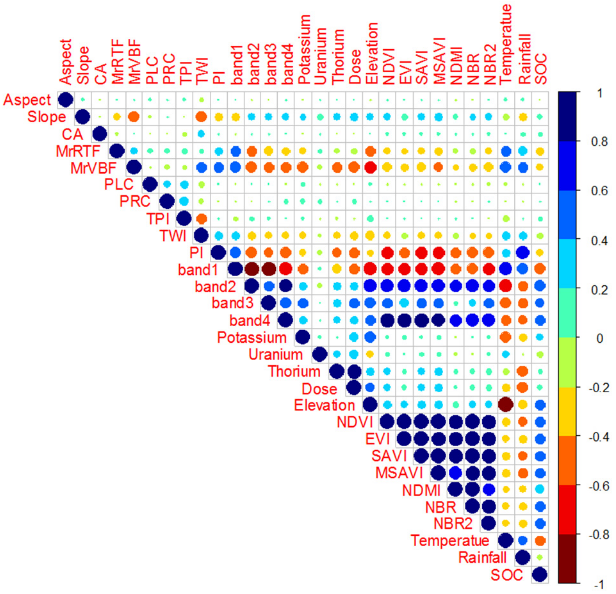

Figure 2. Pearson correlation coefficients for the relation between SOC stocks (0-0.30 m) and 28 predictor variables used in this study.

2.4 Model evaluation

Based on previous studies (Bonfatti et al., 2016; Román-Sánchez et al., 2016), we used 75% of total amount of 564 randomly selected field-collected soil data points for training ("calibration dataset") while the remaining 25% of field data was used as the "validation dataset" to validate the prediction of SOC stocks. To ensure model stability and increase reliability, the procedure was repeated 50 times applying a sampling with replacement method, to obtain 50 random sub-samples of the data, each one with its own calibration and validation dataset. The performance of each model with optimal parameters (identified during the feature selection approaches outlined above) was evaluated using the difference between the observed and the predicted response variable. To do this, four statistical indices were considered: Regression Coefficients of determination (R^2) which measures the percentage of variation explained by each model; Mean Absolute Error (MAE), indicating how close the prediction is to observation; Root Mean Square Error (RMSE), measuring the overall accuracy of the prediction and Lin's Concordance Correlation Coefficient (LCCC) which provides a measure of the agreement between predicted and observed values that follow the 45° line using the following equations:

$$R^2 = \frac{\sum_{i=1}^{n}(P_i - \bar{O})^2}{\sum_{i=1}^{n}(O_i - \bar{O})^2} \qquad (1)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|P_i - O_i| \qquad (2)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - O_i)^2} \qquad (3)$$

$$LCCC = \frac{2r\sigma_o\sigma_p}{\sigma_o^2 + \sigma_p^2 + (\bar{P} - \bar{O})^2} \qquad (4)$$

Where P_i and O_i are the predicted and observed SOC stocks; n is the number of samples; $\bar{P}$ and $\bar{O}$ are the means for the predicted and observed SOC stocks; σ_o^2 and σ_p^2 are the variances of predicted and observed values and r is the Pearson correlation coefficient between the predicted and observed values. A good model will have $LCCC$ and R^2 close to 1 and $RMSE$ and MAE of almost 0.

3. RESULTS

We used the selected predictors to train the RF and BRT models. Four models were evaluated: BRT using stepwise selection of predictors (SR_BRT), RF using stepwise (SR_RF), BRT using GA selection of predictors (GA_BRT) and RF using GA (GA_RF). In addition, BRT using all predictors (All_BRT) and the RF using all predictors (All_RF) were used as benchmarks to test the performance of the four models. The independent validation datasets were used to validate the model performance. Figure 3 shows the performance of four indicators R^2 (Eqn. 1), MAE (Eqn. 2), RMSE (Eqn. 3) and LCCC (Eqn.4) of the 50 trials using the RF and BRT based on different input features.

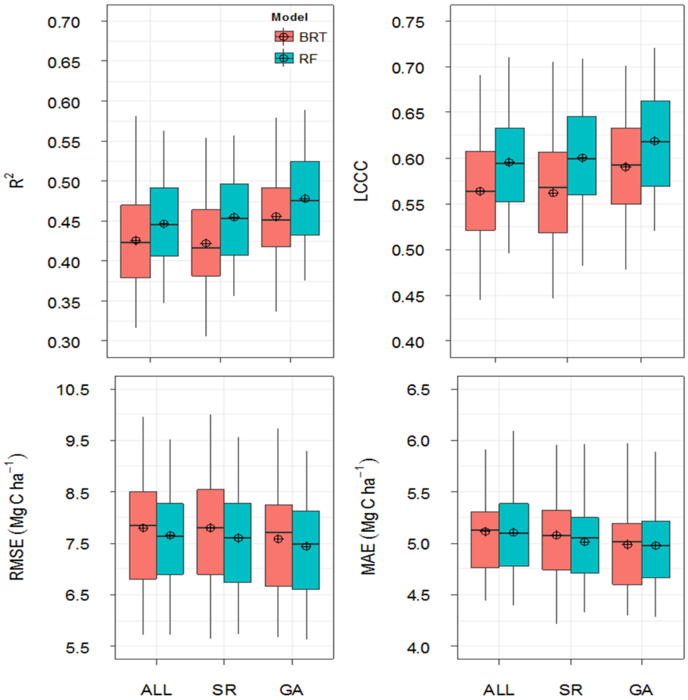

Figure 3. Results of model evaluation criteria for prediction of soil organic carbon stocks (Mg C ha^{-1}) using boosted regression tree (BRT) and random forest (RF) models with 50 runs for the different input selections (ALL: 28 predictors; SR: 11 out of 28 predictors selected by Stepwise Regression method ($P < 0.05$); GA: 11 of 28 predictors selected by Genetic Algorithms). The coefficient of determination (R^2), Lin's concordance correlation coefficient (LCCC), root mean squared error (RMSE), and mean absolute error (MAE) are used to evaluate accuracy. The black lines within the box indicate the medians with 50 runs while crosshairs indicate means. Box boundaries indicate the 25th and 75th percentiles, whiskers below and above the box indicate the 10th and 90th percentiles.

When using the validation samples, the results show that all models predicted SOC stocks moderately well based on the range of average values of R^2 from 0.42 to 0.48 (LCCC ranged from 0.56 to 0.62) and RMSE between 7.44 and 7.80 Mg C ha^{-1} (MAE ranged from 4.98 to 5.11 Mg C ha^{-1}). Overall, the RF model performed better than the BRT model based on the four validation measurements regardless of input features. Specifically, when using all of 28 environmental variables as input predictors, the model of All_RF had a higher R^2 (0.45) and LCCC (0.60) with a lower RMSE (7.66 Mg C ha^{-1}) and MAE (5.11 Mg C ha^{-1}) than the All_BRT model (R^2=0.43, LCCC=0.56, RMSE=7.79 Mg C ha^{-1} and MAE=5.11 Mg C ha^{-1}). In terms of the importance of the variables contributing to SOC stocks, All_RF and All_BRT showed similar patterns (Fig. 4a). The top four most important variables explaining the SOC stock variation were elevation, Band 1, Band 2 and PI, though the order of variable importance varied among models. For example, the predictor rainfall in the RF model was the most important variable followed by elevation, PI, Band 1, and Band 2. In contrast, the majority of the topographic variables showed a very low contribution to the model, as well as uranium concentration and terrestrial gamma-ray dose rate. The result of BRT indicated that Band 1 was the most important variable affecting SOC stocks, followed by PI, elevation, Band 2 and EVI.

Due to the minor importance attributed to some of the predictor variables, SR_RF and SR_BRT used the selected variables by stepwise regression (in total 11 predictors), to examine whether prediction accuracy would remain unchanged or increased when using a smaller number of variables. The R^2 obtained by SR_RF increased compared to All_RF and SR_RF performed better in terms of prediction error, indicating that a more parsimonious model did not impact its capability in predicting SOC stocks (Fig. 3). However, a smaller increase can be found in BRT model. The order of the variables according to their relative contribution to the model did not change substantially, with elevation being the most important, followed by rainfall, Band 1, NBR and PI in the RF model (Fig. 4b). For the BRT model, the top five variables (in order of importance) were Band 1, PI, NBR, elevation and potassium concentration.

However, applying the condensed dataset (in total 11 variables) selected by GA to RF resulted in the highest prediction accuracy with R^2 being equal to 0.48 (RMSE=7.44 Mg C ha^{-1}, MAE=4.98 Mg C ha^{-1} and LCCC=0.62). When GA_RF was compared with All_RF, the GA_RF model improved the predictive performance by increasing the R^2 and LCCC by 7.3% and 4.0%, respectively, while also reducing the RMSE and MAE by 2.8% and 2.5% respectively. For the GA_RF model, rainfall, elevation, PI, Band 1 and MSAVI were the most important variables to predict SOC stock while TPI was identified as being of minor importance to SOC. Similarly, the BRT model also predominately identified Band 1, PI, MSAVI and elevation as the top 4 most important variables.

4. DISCUSSION AND CONCLUSIONS

This study evaluated alternative methods to determine the most reliable and accurate model to predict SOC stocks in the surface 0.30 m of the soil using field collected SOC data and remotely-sensed variables in the semi-arid rangelands of NSW. The results from our study suggest there is some opportunity to use remotely sensed vegetation indices to predict SOC stocks in rangelands. Two commonly used machine learning methods were applied to assess these existing data sets based on two different feature selection methods. The results suggest GA_RF is the most promising model for predicting SOC stocks in the semi-arid rangelands of NSW. The validation results (Fig. 3) show that the prediction accuracy of the GA_RF model was acceptable with explained variances of 48% for SOC stocks, which were comparable to most recent studies predicting SOC stocks. For example, in semi-arid areas, Wiesmeier et al. (2011) found the RF_CRAT model could explain 53.4% of variation in the model building process, whereas Román-Sánchez et al. (2016) also used RF but achieved much lower explained variance of 18% in a rocky, semi-arid landscape. This study shows the importance of feature selection prior to predicting SOC stocks, even if the BRT model is relatively insensitive to the input features selected by SR. The results have also shown that the bare ground fraction, elevation, PI and rainfall were important variables explaining the observed variability of SOC stocks in this semi-arid environment, and the contributions of the other environmental factors were only marginal. The approach proposed here can be extended in data-scarce areas (e.g. rangelands) to produce more detailed information about SOC stocks. As such, the results of this study are of particular importance to provide statistical and theoretical basis for producing digital SOC stocks maps based on readily available satellite products across rangelands.

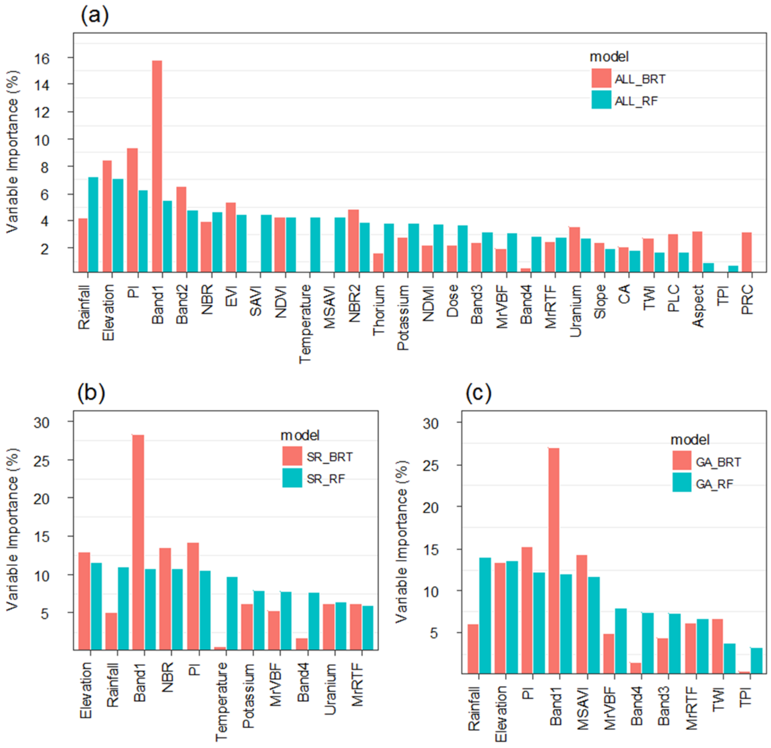

Figure 4. Patterns in the importance of each predictor variable used in RF and BRT models to predict SOC stocks were similar. Each variable was scaled to sum to 100%. (a) ALL: All 28 predictor variables; (b) SR: 11 out of 28 predictors selected by Stepwise Regression method and (c) GA: 11 of 28 predictors selected by Genetic Algorithms.

ACKNOWLEDGEMENTS

Funding for this research was provided by NSW Treasury and NSW Department of Primary Industries.

REFERENCES

Akpa, S.I.C., Odeh, I.O.A., Bishop, T.F.A., Hartemink, A.E., Amapu, I.Y. (2016). Total soil organic carbon and carbon sequestration potential in Nigeria. *Geoderma* 271, 202-215.

Allen, D.E., Pringle, M.J., Bray, S., Hall, T.J., O'Reagain, P.O., Phelps, D., Cobon, D.H., Bloesch, P.M., Dalal, R.C. (2013). What determines soil organic carbon stocks in the grazing lands of north-eastern Australia? *Soil Research* 51(8), 695-706.

Badgery, W.B., Simmons, A.T., Murphy, B.M., Rawson, A., Andersson, K.O., Lonergan, V.E., van de Ven, R. (2013). Relationship between environmental and land-use variables on soil carbon levels at the regional scale in central New South Wales, Australia. Soil Research 51(8), 645-656.

Bonfatti, B.R., Hartemink, A.E., Giasson, E., Tornquist, C.G., Adhikari, K. (2016). Digital mapping of soil carbon in a viticultural region of Southern Brazil. *Geoderma* 261, 204-221.

Everingham, Y., Sexton, J., Skocaj, D., Inman-Bamber, G. (2016). Accurate prediction of sugarcane yield using a random forest algorithm. *Agronomy for Sustainable Development* 36(2), 1-9.

Guo, P.-T., Li, M.-F., Luo, W., Tang, Q.-F., Liu, Z.-W., Lin, Z.-M. (2015). Digital mapping of soil organic matter for rubber plantation at regional scale: An application of random forest plus residuals kriging approach. *Geoderma* 237–238, 49-59.

Keating, B., Grundy, M., Battaglia, M., Eady, S. (2009). An Analysis of greenhouse gas mitigation and carbon biosequestration opportunities from rural land use. St Lucia, QLD: CSIRO; 2009. changeme:822.

Ließ, M., Schmidt, J., Glaser, B. (2016). Improving the Spatial Prediction of Soil Organic Carbon Stocks in a Complex Tropical Mountain Landscape by Methodological Specifications in Machine Learning Approaches. *PLOS ONE* 11(4), e0153673.

Mouazen, A.M., Kuang, B., De Baerdemaeker, J., Ramon, H. (2010). Comparison among principal component, partial least squares and back propagation neural network analyses for accuracy of measurement of selected soil properties with visible and near infrared spectroscopy. *Geoderma* 158(1–2), 23-31.

Orgill, S.E., Waters, C.M., Melville, G., Toole, I., Alemseged, Y., Smith, W. (2017). Sensitivity of soil organic carbon to grazing management in the semi-arid rangelands of south-eastern Australia. *The Rangeland Journal* 39(2), 153-167.

Rabbi, S.M.F., Tighe, M., Cowie, A., Wilson, B.R., Schwenke, G., McLeod, M., Badgery, W., Baldock, J. (2014). The relationships between land uses, soil management practices, and soil carbon fractions in South Eastern Australia. *Agriculture, Ecosystems and Environment* 197, 41-52.

Román-Sánchez, A., Vanwalleghem, T., Peña, A., Laguna, A., Giráldez, J.V. (2016). Controls on soil carbon storage from topography and vegetation in a rocky, semi-arid landscapes. *Geoderma*.

Waters, C.M., Orgill, S.E., Melville, G.J., Toole, I.D., Smith, W.J. (2016). Management of Grazing Intensity in the Semi-Arid Rangelands of Southern Australia: Effects on Soil and Biodiversity. *Land Degradation & Development* 28, 1363–1375.

Wiesmeier, M., Barthold, F., Blank, B., Kögel-Knabner, I. (2011). Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem. *Plant and Soil* 340(1-2), 7-24.

Xie, H., Zhao, J., Wang, Q., Sui, Y., Wang, J., Yang, X., Zhang, X., Liang, C. (2015). Soil type recognition as improved by genetic algorithm-based variable selection using near infrared spectroscopy and partial least squares discriminant analysis. *Scientific Reports* 5, 10930.