Computers and Electronics in Agriculture 142 (2017) 91–100

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

Comparison of regression methods for spatial downscaling of soil organic carbon stocks maps

P. Roudier [a,b,*], B.P. Malone [c], C.B. Hedley [a], B. Minasny [c], A.B. McBratney [c]

[a] Landcare Research, Private Bag 11052, Manawatū Mail Centre, Palmerston North 4442, New Zealand
[b] Te Pūnaha Matatini, A New Zealand Centre of Research Excellence, Private Bag 92019, Auckland 1142, New Zealand
[c] Sydney Institute of Agriculture, The University of Sydney, Eveleigh, NSW 2015, Australia

ARTICLE INFO

Article history:
Received 29 March 2017
Received in revised form 10 August 2017
Accepted 21 August 2017

Keywords:
Spatial downscaling
Digital soil mapping
Machine learning
R

ABSTRACT

This paper presents a refinement of the `dissever` algorithm, a framework for downscaling spatial information based on available environmental covariates proposed by Malone et al. (2012). While the original algorithm models the relationships between the target variable and the covariates using a general additive model (GAM), the modified procedure presented in this paper allows the user to choose between a wide range of regression methods.

These developments have been implemented in an open-source package for the R statistical environment, and tested by downscaling soil organic carbon stocks (SOCS) maps available on two study sites in Australia and New Zealand using 4 different regression methods: linear model (LM), GAM, random forest (RF), and Cubist (CU). In this study, the spatial resolution of a set of reference maps were degraded to a coarser resolution, so to assess the performance of the different downscaling methods. On the Australian site, the 1-km SOCS coarse resolution map has been downscaled to a 90-m resolution. The best results were achieved using either CU or RF ($R^2 = 0.91$ and 0.94 respectively). On the New Zealand site, the 250-m SOCS coarse resolution map has been downscaled to a 10-m resolution. The best results were achieved using GAM ($R^2 = 0.90$). The results illustrate that the optimal regression methods for downscaling spatial information using `dissever` vary on a case-by-case basis. In particular, simpler approaches such as LM or GAM outperformed more complex approaches in cases where only a limited number of pixels are available to train the downscaling algorithm. This demonstrate the value of an implementation that facilitates testing of different regression strategies.

1. Introduction

The selection of a relevant spatial resolution is a central question for digital soil mapping (DSM) (Behrens et al., 2010; Malone et al., 2013; Smith et al., 2006; Taylor et al., 2013). Most DSM approaches require environmental predictors to be available on a unique prediction grid (McBratney et al., 2003). While upscaling (matching a fine resolution covariate to a coarser resolution grid) can be easily solved using approaches such as block averaging or block kriging, the opposite situation, downscaling (matching a coarse resolution covariate onto a finer resolution grid) is a more challenging task. While various interpolation methods can be tested, it often results in the prediction grid being limited to the resolution of the coarsest covariate. Another reason why downscaling of spatial information is of current interest in DSM is to increase the value of national digital soil maps that are becoming increasingly available through initiatives such as GlobalSoilMap (Arrouays et al., 2014). To increase their value to the primary sector and match the resolution of farm-scale management decisions (which are getting finer with the advent of precision agriculture techniques) these coarse resolution maps (resolution of between 1-km and 100-m) need to be downscaled to a finer resolution. Downscaling such national datasets also provides a useful tool to stratify soil sampling for estimating soil organic carbon stocks, as required by carbon farming initiatives (e.g. de Gruijter et al., 2016).

The `dissever` method for downscaling spatial information has been proposed by Malone et al. (2012). It is mass-preserving, and based on using a suite of covariates to reconstruct the signal of a coarse variable at a finer resolution. The current context in soil science is a favourable one to such an approach driven by covariates, since it has been recently disrupted by the emergence of various sensing technogies that allow information to be recorded that

* Corresponding author at: Landcare Research, Private Bag 11052, Manawatū Mail Centre, Palmerston North 4442, New Zealand.
E-mail address: roudierp@landcareresearch.co.nz (P. Roudier).

http://dx.doi.org/10.1016/j.compag.2017.08.021
0168-1699/

relates to soil-forming factors at a fine spatial scale (Roudier et al., 2015; Stockmann et al., 2015). Remote sensing methods such as LiDAR, mounted on an aircraft to record elevation data at very fine resolution, allow derivation of terrain parameters such as slope, aspect, or wetness index (DeGloria et al., 2014; Fink and Drohan, 2016). Additionally, proximal soil sensors can be mounted directly on a mobile platform, such as a tractor or a quad bike, and can record a range of physical properties such as soil electrical resistivity and conductivity (electromagnetic sensors, EM), and natural gamma emissions (gamma radiometric sensors, Viscarra Rossel et al., 2011).

In parallel to this increase in available data, the field of machine learning has driven the development of many prediction techniques. Making use of the increasing computer power available, such advanced regression techniques have found applications in many domains, and are able to handle complex relations between covariates. A significant range of these prediction techniques have been succesfully used in digital soil mapping (Heung et al., 2016; Viscarra Rossel et al., 2015). The aim of this study was to modify the `dissever` algorithm so that it can use different regression methods. The performance of four different regression methods were tested and compared in downscaling coarse resolution soil organic carbon stocks (SOCS) maps using a suite of fine scale covariates, at two different study sites.

2. Material and methods

2.1. The `dissever` algorithm

The `dissever` algorithm, initially proposed by Malone et al. (2012), is a method to downscale a coarse resolution raster map using a suite of finer resolution environmental covariates. To do so, a relationship between the fine resolution covariates and the coarse resolution base map is built using a generalised additive model (GAM). The GAM is used in an iterative process to converge towards a solution that is mass-preserving, i.e. the mean of fine scale predictions is equivalent to the associated value of their encapsulating coarse scale pixel. The algorithm, implemented as follows, is detailed in Malone et al. (2012):

1. Interpolate the coarse resolution map of the target variable onto the grid used by the fine resolution covariates using nearest neighbour resampling.
2. Regress the fine gridded values of the target variable against the suite of covariates.
3. Upscale the predictions of this regression model by block averaging to the original base map resolution.
4. If the iteration number is greater than one, check whether upscaled estimates are changed from previous iteration. If estimated change is greater than some pre-defined threshold proceed to next step, otherwise stop. In Malone et al. (2012) an averaged absolute difference between the upscaled map from the present iteration and previous iteration was used. An arbitrarily selected threshold of 0.001 was used to determine if iteration should proceed or not.
5. Compute the deviation from mass balance for each coarse grid pixel, i.e. the difference between the mean of downscaled predictions and the original value of each pixel, and use it to correct the fine gridded estimates with deviation factor.
6. Go back to step 2.

2.2. Modification of the original algorithm

The original `dissever` method has been extended so that different regression methods can be used to build the best relationship between the coarse resolution target variable and the fine resolution environmental covariates. At the initialisation stage of the disseveration, for parametric regression methods, k-fold cross-validation is used to choose the optimal parameter values. In this case, the set of parameters that minimise the cross-validated root mean squared error (RMSE) are selected. This optimal set of parameters is then used for the iterative stage of the disseveration. For non-parametric methods, this step is skipped, and an initial model is simply fitted between the coarse resolution target variable and the environmental covariates.

The modified `dissever` procedure has been implemented using the R statistical environment (R Core Team, 2015). The modified procedure leverages the `caret` predictive modelling package for R (Kuhn, 2008), which provides a unified interface to 192 different regression methods. Additionally, the `caret` provides numerical methods to optimally choose parameters, and allows for parallel processing. The resulting code has been integrated in a dedicated R package, and has been made publicly available on Github.[1]

2.3. Regression methods tested

In this study, four different regression methods have been tested and compared for the downscaling of coarse scale maps. Linear models (LM), as implemented in base R (R Core Team, 2015), were chosen since they represent a simple yet robust predictive technique. Generalised additive models (GAM), used in the original `dissever` procedure, as implemented in R by the `gam` package (Hastie, 2015), have been used as a reference method. Also, random forest (RF), as implemented in R by the `randomForest` package (Liaw and Wiener, 2002), and Cubist (CB), as implemented in R by the `Cubist` package (Kuhn et al., 2014), were tested. These latter two methods are more recent data mining techniques and have received a great deal of attention in the digital soil mapping literature (Heung et al., 2016).

2.4. Comparison of the downscaled outputs

Fig. 1 shows the workflow that has been used to assess the downscaling performance using the `dissever` algorithm with different regression methods. The base map was the coarse resolution map to be downscaled. It was created by block-averaging a reference map, available at the same fine resolution as the environmental predictors. The downscaled map resulted from the `dissever` procedure, and was compared to the reference map. It was also block-averaged back to the coarse resolution support to create the restored map. This restored map was compared to the base map in order to assess the respect of the mass-conservation constraint of the algorithm.

2.4.1. Downscaling performance

Different metrics quantified the performance of the downscaling process, including the root mean squared error of downscaling (RMSEd), R^2, concordance correlation coefficient (CCC Lin, 1989), and bias. The RMSEd indicates the uncertainty of the downscaled map, while the bias gives an indication about its accuracy. The standard error (SE) was also reported. The CCC quantified the agreement between the downscaled map and the reference map as a value between 0 (absolute disagreement) and 1 (absolute agreement).

$$RMSEd = \sqrt{\sum_i^n \frac{(x_i - X_i)^2}{n}} \quad (1)$$

[1] https://github.com/pierreroudier/dissever.

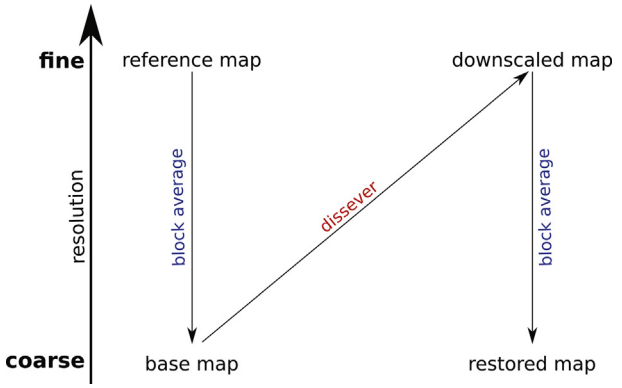

Fig. 1. Workflow and naming conventions. A reference map, available at fine resolution, is block-average to create the coarse resolution base map. This base map is being downscaled into using the `dissever` approach. The downscaled result is block-averaged to the coarse scale to form the restored map, which can be compared to the base map.

$$CCC = \frac{2 \cdot r \cdot \sigma \cdot \hat{\sigma}}{\sigma^2 + \hat{\sigma}^2 + (M - m)^2} \quad (2)$$

$$Bias = m - M \quad (3)$$

where X is the value of the reference data, x is the value of the downscaled data, n is the number of pixels, r is the correlation coefficient between the reference and the downscaled data, σ and $\hat{\sigma}$ are the standard deviations of the reference and downscaled data, and M and m are the means of the reference and downscaled data.

2.4.2. Restoration performance

The respect of the mass-preserving constraint is assessed by comparing the base map with the restored map, i.e. the downscaled map re-aggregated back onto the coarse scale spatial support using block average. The root mean squared error of restoration (RMSEr) was used to quantify the respect of this constraint:

$$RMSEr = \sqrt{\sum_i^n \frac{(Y_i - X_i)^2}{n}} \quad (4)$$

where X is the base map, Y is the restored map, and n is the number of pixels.

2.4.3. Spatial structure

Experimental semi-variograms were computed in order to compare the spatial structure of the downscaled maps with that of the reference map. Then, a variogram model was fitted to the experimental semi-variograms. In this study we tested a range of different variogram models: exponential, spherical and Matérn (Minasny and McBratney, 2005). These computations were done in the R statistical environment using the `gstat` package (Pebesma, 2004).

2.5. Case studies

2.5.1. Case study 1

The first case study is a 163,891-ha farm located in the Edgeroi District, NSW, Australia. Edgeroi is an intensive cropping area upon

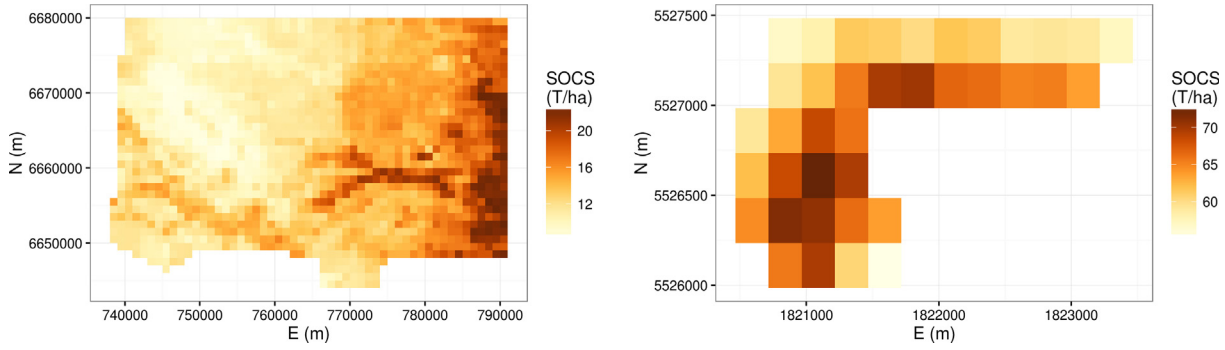

Fig. 2. Base maps for the Edgeroi and Massey sites, obtained by block-averaging the reference maps.

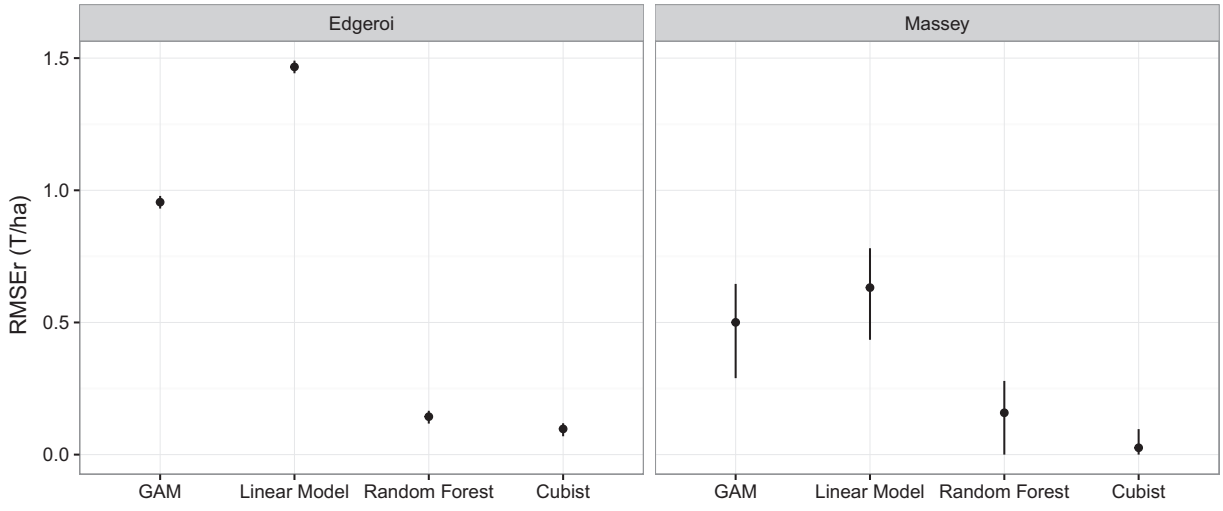

Fig. 3. Error associated with the mass-preservation constraint for each regression methods tested for downscaling. The solid lines show the 90% confidence intervals around each RMSE value.

Table 1
Summary statistics of the reference map and the maps downscaled using the four different regression methods on the Edgeroi and Massey study sites. Min.: Minimum. Pct.: Percentile. Max.: Maximum. Std. Dev.: Standard deviation. Skew.: Skewness.

Site	Model	Min.	2.5% Pct.	25% Pct.	50% Pct.	Mean	75% Pct.	97.5% Pct.	Max.	Std. Dev.	Skew.
Edgeroi	Linear model	−3.33	8.21	11.17	12.93	13.32	14.98	21.58	35.05	3.36	0.93
Edgeroi	GAM	3.57	8.82	10.91	12.77	13.32	15.04	21.71	26.89	3.24	0.94
Edgeroi	Cubist	2.26	8.32	10.65	12.72	13.32	15.33	22.08	27.54	3.55	0.80
Edgeroi	Random forest	7.58	8.35	10.73	12.78	13.32	15.25	21.99	24.50	3.48	0.78
Edgeroi	Reference	4.09	8.32	10.64	12.76	13.32	15.63	22.51	24.98	3.61	0.82
Massey	Linear model	53.17	55.78	60.92	63.82	64.67	68.26	75.47	77.86	5.39	0.38
Massey	GAM	53.32	55.85	60.84	63.94	64.67	68.25	75.24	77.31	5.30	0.36
Massey	Cubist	43.82	57.84	61.03	64.16	64.67	68.87	73.01	76.95	4.54	0.23
Massey	Random forest	52.52	57.85	60.34	64.36	64.67	68.17	73.43	74.70	4.77	0.30
Massey	Reference	47.49	56.48	60.35	63.98	64.58	68.54	74.98	77.76	5.34	0.34

the fertile alluvial Namoi River plain. Different fine-scale covariates (all rescaled to a 90-m resolution by block averaging) were collected on this farm. A digital elevation model (DEM) provided information about elevation, and terrain derivatives such as slope and topographic wetness index (TWI). Data from the Landsat ETM + satellite were used to derive normalised difference vegetation index (NDVI), along with a suite of band ratios: band 5/band 7, band 3/ band 7, band 3/ band 2. Finally, potassium and thorium abundance estimates were derived from an airborne gamma radiometrics survey over the region. Those covariates were used to map SOCS to a depth of 30 cm at a 90-m resolution, as detailed in Malone et al. (2012). This reference map (202334 pixels) was upscaled to a 1-km resolution using block averaging to create the base map to be downscaled (1689 pixels, Fig. 2).

2.5.2. Case study 2

The second case study is the 129-ha Massey University Farm Number 1 located in Palmerston North, New Zealand. A suite of fine resolution covariates, collated at a 10-m resolution, were collected on the farm. Two proximal soil sensors mounted on a quad bike, electromagnetic (EM) and gamma radiometrics, were used to survey the farm. Their outputs were then kriged to match the 10-m resolution grid. A DEM was created from an aerial LiDAR survey collected by the Horizons Regional Council. It was then rescaled

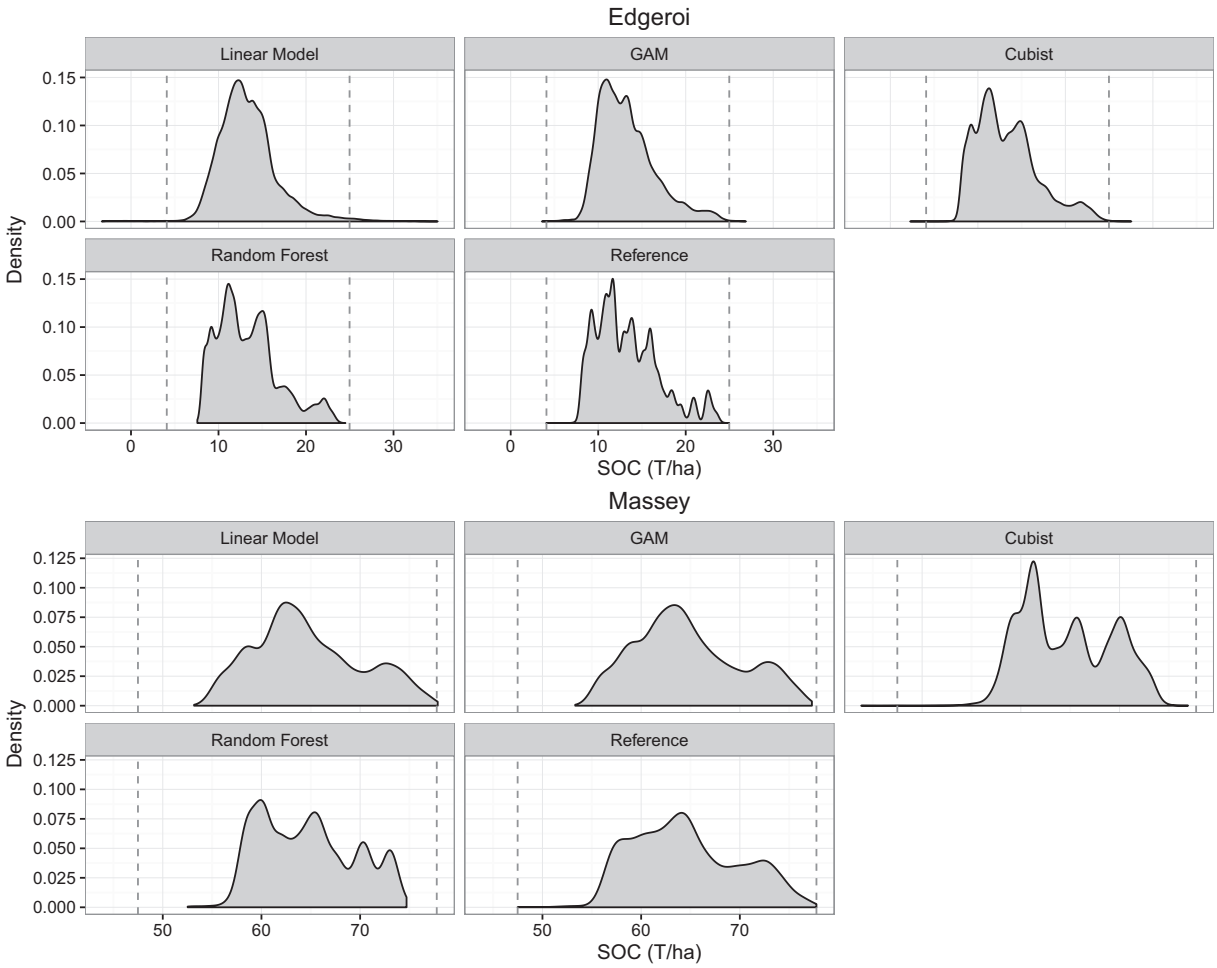

Fig. 4. Probability density functions of the disaggregated and reference SOC maps for the Edgeroi and Massey sites. The range of the reference data is indicated using broken vertical lines.

to the 10-m resolution grid using block averaging. Terrain derivatives such as slope and SAGA wetness index (SWI) were derived from the DEM using SAGA GIS. The distance to the river was also mapped since those soils are regularly affected by fluvial deposits following flood events. Finally, a legacy soil map was also used as a covariate layer. As detailed for the first case study above, those covariates were used alongside a collection of 100 soil core samples to create a SOCS map to 30 cm at a 10-m resolution. Because the farm is much smaller than the Australian study site, the reference map (12,901 pixels) was block-averaged to a 250-m resolution grid to create the base map to be downscaled (38 pixels, Fig. 2).

3. Results

For each downscaled map, the respect of the mass-preservation constraint is illustrated in Fig. 3. The figure shows the errors asociated with the restoration process by comparing the base map and the restored map. In both study cases, RF and CU out-perform GAM and LM, and the restored data downscaled using these two regression methods shows very little errors (RMSEr < 0.5 T/ha). Results using the GAM model show a RMSEr of 0.96 T/ha at the Edgeroi site, and 0.5 T/ha at the Massey site. Results using LM show the highest RMSEr (1.47 T/ha at the Edgeroi site, 0.63 T/ha at the Massey site).

The summary statistics of the downscaled and reference maps are presented on Table 1. In general, the mean value is not distorted by the downscaling operations, which shows the mass-preserving constraint has been successfully observed for all regression methods tested on this case study. Moreover, at both study sites, the interquartile range (the difference between the 25th and the 75th percentiles) remains more or less identical to the reference map for all downscaled maps.

However, significant differences can be observed in the tendency of each method to extrapolate data outside the reference map boundaries or not. This is illustrated in the probability density functions of the downscaled maps (Fig. 4). As a consequence of this, at the Edgeroi site, both LM and GAM are positively skewing the distribution of the downscaled outputs compared to that of the reference map. It can be observed that on one hand the shape of the probability distributions for the linear model and the GAM model are very similar, while on the other hand RF and CU have similar distributions too. A potential explanation is the similarities

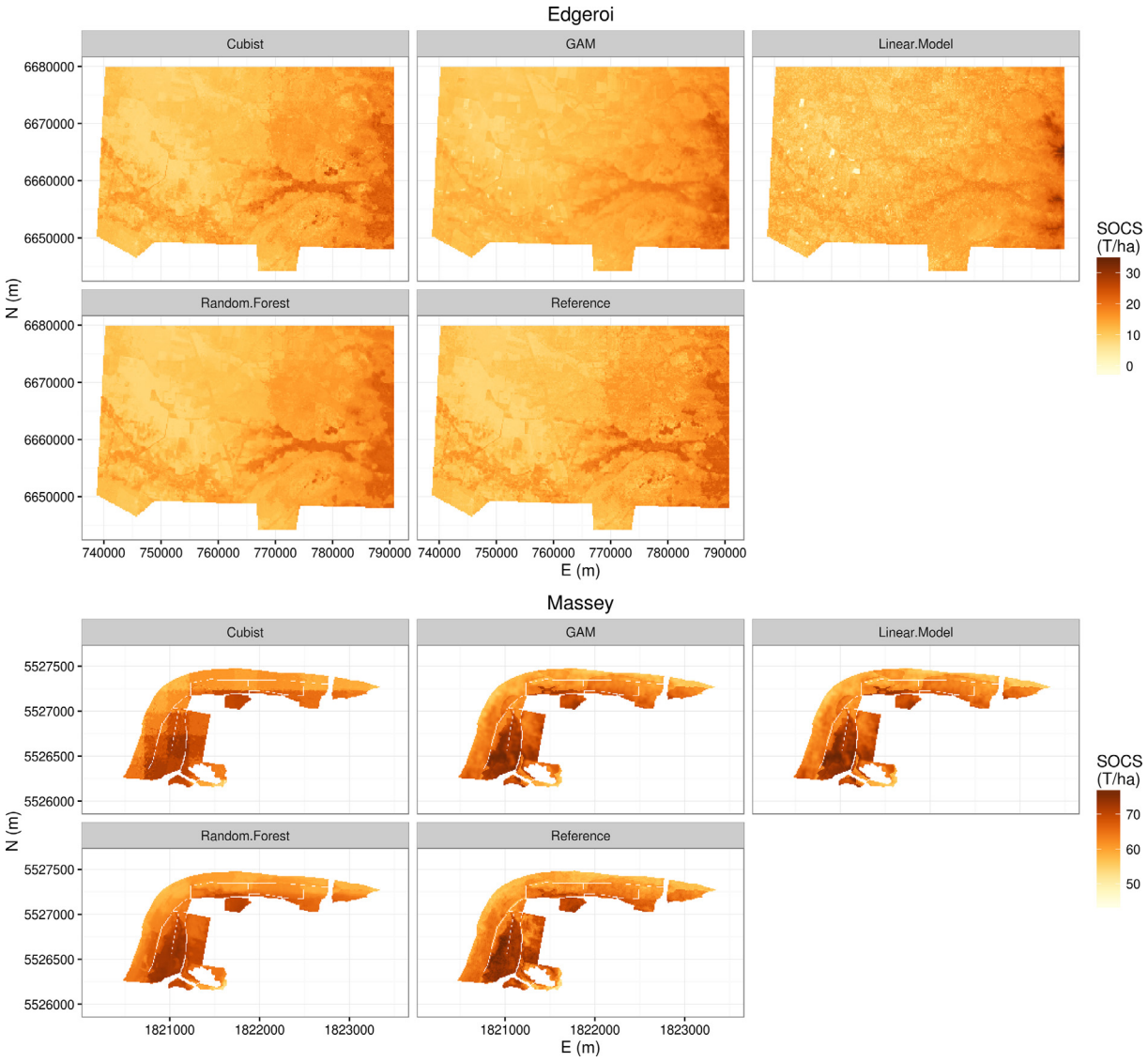

Fig. 5. Downscaled maps for the Edgeroi and Massey sites, along with the respective reference maps.

that exists between LM and GAM on one hand, and between CU and RF on the other hand. GAMs are linear predictors using smoothing functions on predictive variable, while both CU and RF are tree-based machine learning techniques. The distributions derived from RF and CU are closest to the reference map distribution, expecially at the Edgeroi site, as evidenced by the Kolmogorov-Smirnov statistic D computed between the reference population and that of each of the downscaled outputs (0.10, 0.08, 0.04, and 0.05 for LM, GAM, CU, and RF respectively).

Fig. 5 shows the downscaled maps for the different regression methods tested at the two study sites. The reference maps, from which the coarse scale map has been derived, are also reported. Results show some different spatial patterns depending on the regression method used in the disseveration. At the Edgeroi site, both CU and RF produce details that are more consistent with the original data than LM and GAM. In particular, both CU and RF appear to better reproduce the sharper variations in space. This is possibly due to the tree nature of those regression algorithms. While the overall pattern given by the GAM method is consistent with that of the reference map, the spatial variations appeared to have been smoothed out.

At the Massey site, the situation is different: maps downscaled using both LM and GAM appear to be very similar with the reference map. The RF map is reasonably similar too, but did not capture some of the finer variations observed on the reference map. Looking at the relative importance of the different fine-scale covariates revealed that the legacy soil map, in particular, was used by the model. The spatial pattern of the downscaled map is affected by the use of this map, which represents the broad patterns of soils in the farm, as opposed to the fine-scale variations recorded by other covariates. Finally, the map produced using the CU method is affected by a tiled pattern that matches the spatial support of the base map. The inspection of the regression rules used by the CU model shows that the distance from the river is the covariate that has been most prevalently used in the downscaling regression, at the expense of other covariates that explain the finer details of the reference map.

Geostatical methods were used to quantify the spatial structure of the different downscaled results displayed on Fig. 5. Fig. 6 compares the semi-variograms of the downscaled maps with that of the reference maps. All variograms follow a similar model to that of the reference map (Matérn model for the Edgeroi maps, spherical model for the Massey maps). At both sites, a drop in the sill variance can be observed for some of the downscaled results (Table 2). This indicates that the some downscaled maps captured more of the variance of the reference map than others. At the Edgeroi site, this drop in sill is most pronounced for GAM, while LM and CU present the closest variograms to the reference. At the Massey site, both RF and CU are affected by a significant drop in sill variance, while GAM and LM show a very similar sill variance to the reference map. Looking at the range of the modelled variograms shows that at the Edgeroi site, GAM, RF and, to a lesser extent, CU, produced a downscaled map that does not capture all short range variabilities observed on the reference map. At the Massey site, a similar problem was observed for RF and CU, while the range of the variograms of the maps produced using LM and GAM were close to that of the reference map.

Fig. 7 maps the absolute error between the different downscaled results and the reference maps at both study sites. At the Edgeroi site, it shows smaller errors when using CU or RF as

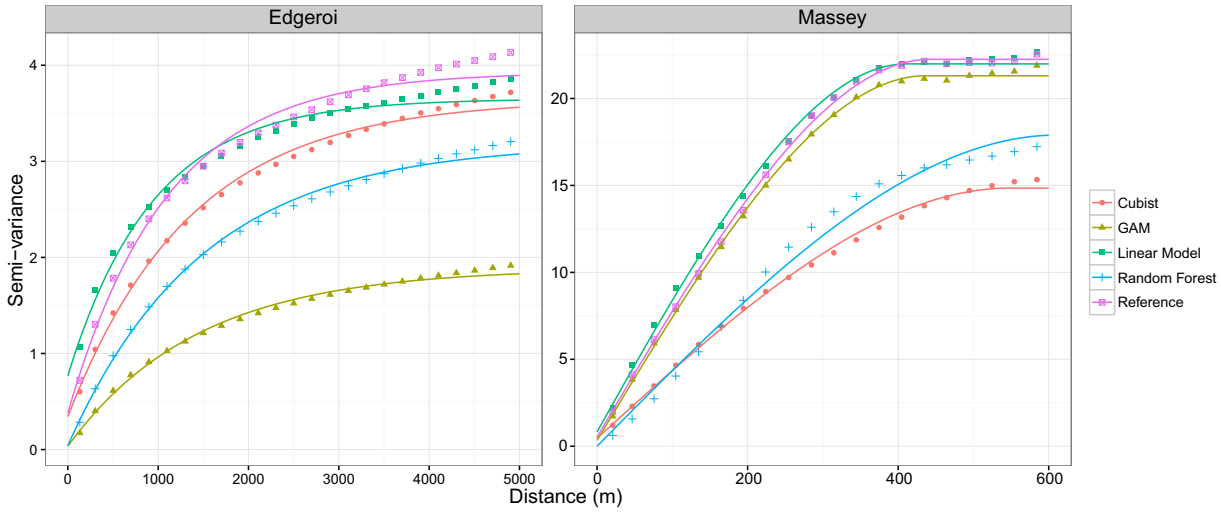

Fig. 6. Comparison of the semi-variograms of the different downscaled maps, along with the reference map.

Table 2
Variogram models of the semi-variograms of the different downscaled maps, along with the reference map.

Site	Reg. model	Var. model	Nugget	Sill	Range (m)
Edgeroi	Linear model	Ste	0.77	2.88	1336.93
Edgeroi	GAM	Ste	0.03	1.85	2033.14
Edgeroi	Cubist	Ste	0.34	3.30	1919.87
Edgeroi	Random forest	Ste	0.04	3.14	2104.06
Edgeroi	Reference	Ste	0.38	3.55	1539.40
Massey	Linear model	Sph	0.83	21.16	411.58
Massey	GAM	Sph	0.34	20.97	436.80
Massey	Cubist	Sph	0.46	14.39	548.35
Massey	Random forest	Sph	0.00	17.89	611.28
Massey	Reference	Sph	0.53	21.72	443.02

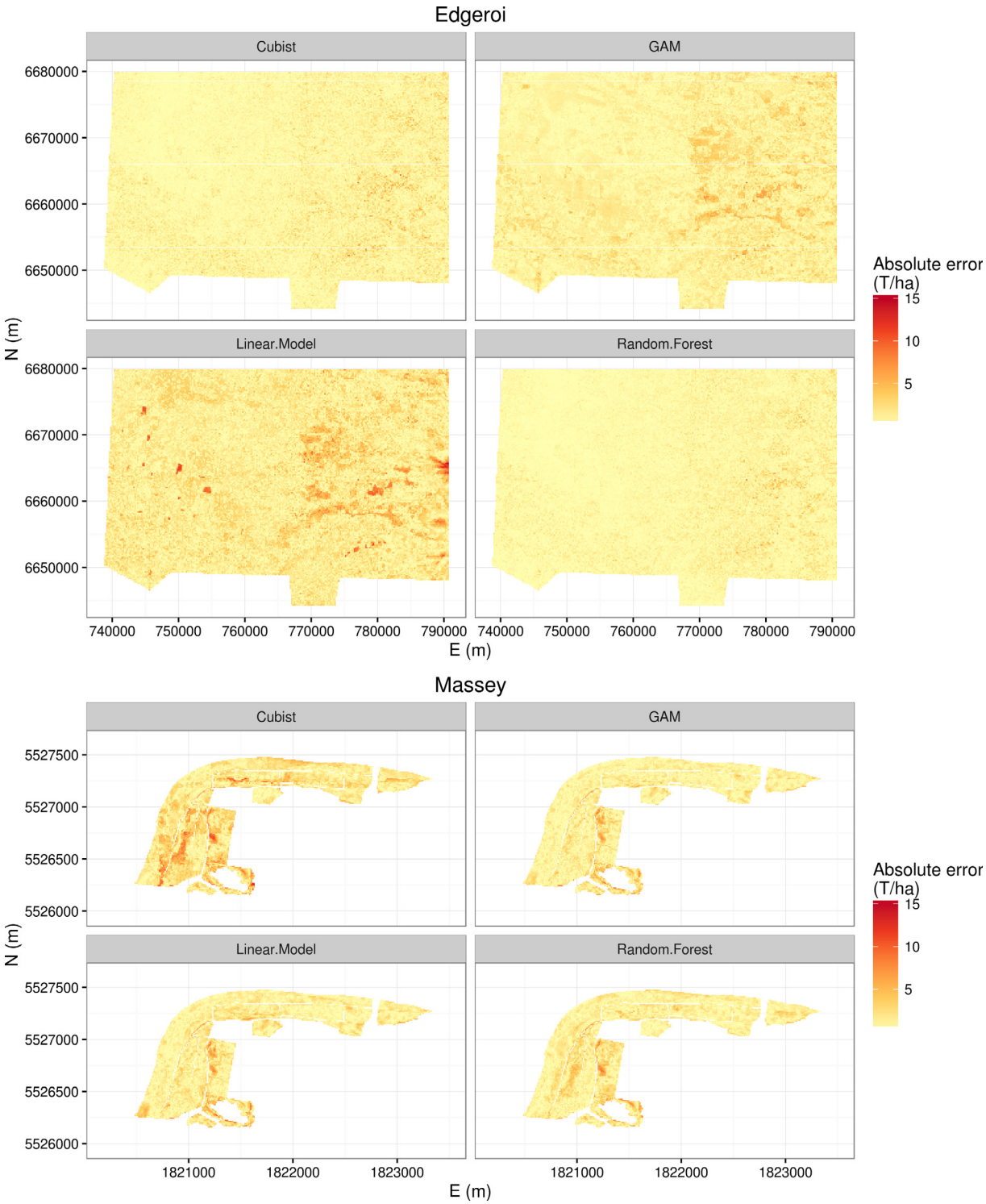

Fig. 7. Maps of the absolute errors between the reference map and the downscaled maps.

opposed to the other methods. From a general standpoint, errors seems to be comparatively larger on the West part of the map, which corresponds to the foothills. The LM model is also affected by water bodies that are present on the East part of the map. At the Massey site, results show that the downscaling error is more important on the river terrace located away from the river. The maps also show that the CU model is missing a lot of the fine scale variations contained in the reference map.

Fig. 8 and Table 3 compare the two values of the downscaled maps against the values of the reference map. At both study sites, there is generally good agreement between the downscaled and the reference datasets, with CCC > 0.8 for all regression methods tested. Additionally, no regression methods produced any bias. However, some methods performed noteably better than others, but the ranking of the regression methods changed depending on the study site. At the Edgeroi site, CU and RF produced the best results, with R-squared > 0.9, CCC > 0.95, and a RMSE close to 1 T/ha. GAM followed, with a R-squared of 0.84, CCC of 0.91, and a RMSEd < 2 T/ha. LM gave the worst performance statistics, with a R-squared of 0.65, CCC of 0.8, and RMSE > 2.2 T/ha. At the Massey

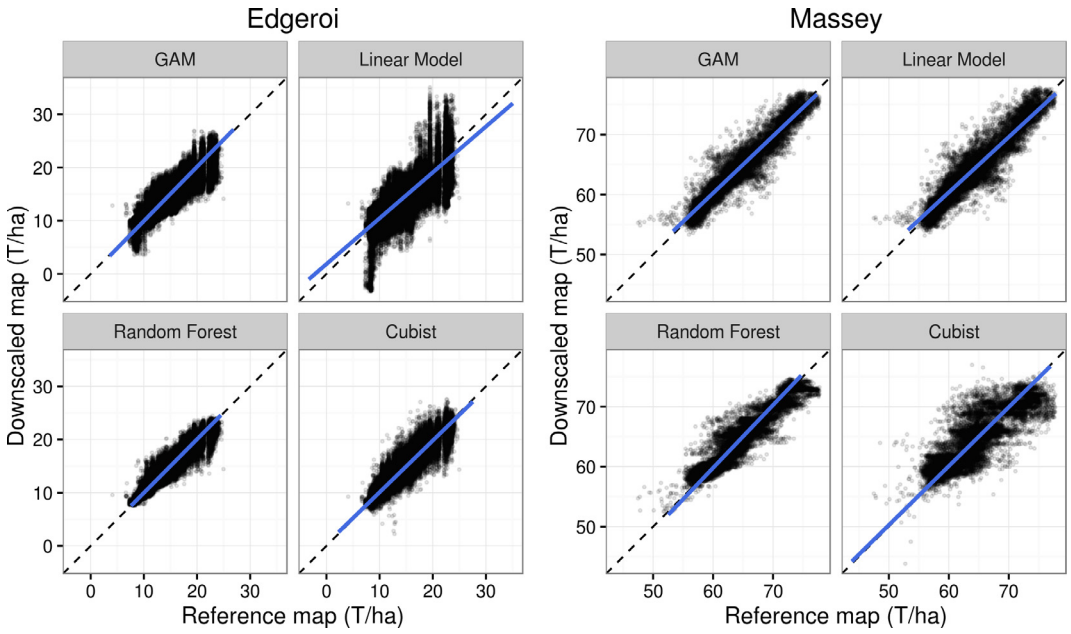

Fig. 8. Reference vs. downscaled values.

Table 3
Performance of the downscaling process using different regression methods.

Site	Model	RMSE (T/ha)	R-squared	CCC	Bias (T/ha)
Edgeroi	Cubist	1.07	0.91	0.96	0.00
Edgeroi	GAM	1.46	0.84	0.91	0.00
Edgeroi	Linear model	2.20	0.65	0.80	0.00
Edgeroi	Random forest	0.86	0.94	0.97	0.00
Massey	Cubist	2.88	0.71	0.83	−0.01
Massey	GAM	1.70	0.90	0.95	−0.01
Massey	Linear model	1.92	0.88	0.94	−0.01
Massey	Random forest	1.83	0.88	0.93	−0.01

site, LM, GAM, and RF performed similarly. They produced downscaled values with a R-squared around 0.9, CCC > 0.9, and RMSE < 2 T/ha. The CU model, as it has already be observed above, did not performed as well, with a R-squared of 0.71, CCC = 0.83, and RMSE = 2.88 T/ha.

Fig. 9 shows the cumulative probability distribution of the absolute error between the reference map and the downscaled maps. At the Edgeroi site, it shows a rather clear hierarchy in terms of performance, with RF giving the best results, followed by CU, GAM, and LM. This figure allows to express the results in terms of risk: the error threshold of 2 T/ha used by Malone et al. (2012) is observed by 96% and 93% of the downscaled locations using the RF and CU methods. This percentage falls to 85% and 68% for GAM and LM. At the Massey site, the figure shows that GAM, RF, and LM are very close in terms of prediction performance, with 82%, 77%, and 76% of the downscaled locations showing an error < 2 T/ha. For the CU model, this proportion falls down to 55%.

4. Discussion

The comparison of four different regression methods on two study cases showed that the best regression method to downscale information varied. The RF and CU algorithms showed better respect of mass-preserving constraint in both cases, but when comparing the downscaled maps with the reference maps, results from the Edegroi case study were the opposite of those from the Massey case study. In the first study case, the best performance was achieved using complex regression tree approaches (RF and CU), whilst in the second case study, simpler regression methods performed better (LM and GAM).

It is worth noting that the two study sites exhibited significant differences. First, the change in resolution associated with the downscaling process is more important for the Massey site (from 250 m to 10 m) than for the Edgeroi site (from 1000 m to 90 m). Also, despite the base map at the Edgeroi site having a much larger number of pixels than the Massey site base map (1689 vs. 38 pixels), it has a smaller inter-quartile range (4.35 vs. 6.49 T/ha) and a smaller standard deviation (3.38 vs. 4.6 T/ha). With a small number of values to calibrate the model at the Massey site, it seems that simple models such as LM or GAM outperform more complex approaches such as CU and RF to capture the important variations observed on the small farm. At the Edgeroi site, with more data to calibrate regression models, CU and RF outperformed LM and GAM.

As the case studies presented above demonstrate, having the ability to easily test different regression methods provides the opportunity to find the best fit for purpose. While the original algorithm relied on GAM to model the relationship between the coarse resolution target variable and the fine resolution environmental covariates, the R implementation of `dissever` is more flexible in that it allows the user to test and use a very wide collection of predictive techniques. To do so, the `caret` package was used as a wrapper to access a very important range of regression techniques. To date, 192 regression methods can be tested, ranging from the simple multivariate linear model, to more cutting edge algorithms from the machine learning literature. At present, these regression methods can be compared, but a possible enhancement to the

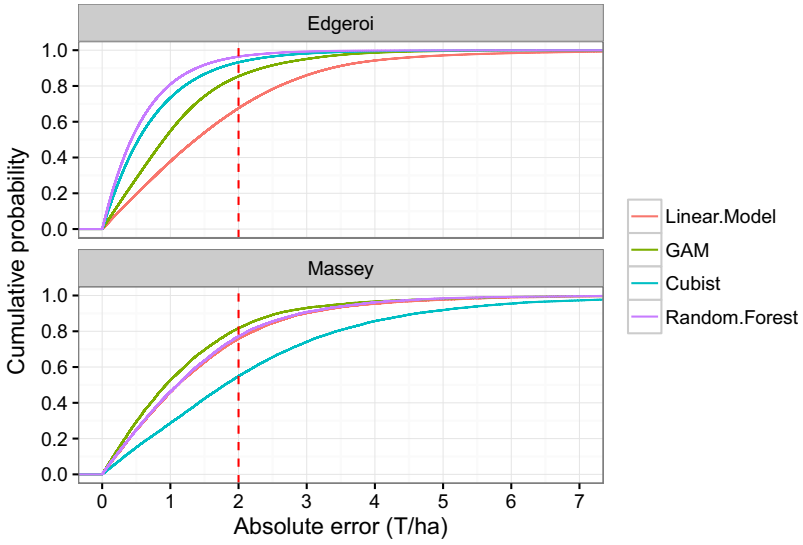

Fig. 9. Cumulative distribution of the absolute errors for the Edgeroi and the Massey sites.

method would be to have them to collaborate, using ensemble modelling. Ensemble modelling makes use of multiple prediction techniques collaboratively in order to obtain better predictions than using these techniques individually, since different prediction models can capture different aspects of the data.

Another point on which further development needs to focus is the ability to compare downscaled maps without reference maps. While for the purpose of this study, reference maps were available, in most applications this won't be the case, and only the mass-preservation could be assessed. However, performance assessement methods used for pan-sharpening in the remote sensing literature, such as indicators based on the entropy of the restored image (Leung et al., 2001), could be trialled.

5. Conclusions

The original `dissever` method for downscaling spatial information uses a GAM model to describe the relationship between a coarse resolution variable and a suite of fine resolution covariates. It has been extended so that the user can choose to model this relationship using a variety of regression techniques. To illustrate this, two case studies have been considered, in Australia and in New Zealand. Care must be taken when picking the more suitable regression technique to succesfully downscale a given map. While the more complex data mining approaches (Cubist, Random Forest) produced the best results for the larger Australian dataset (1689 pixels), on the smaller NZ site (38 pixels), simpler approaches such as linear model and GAM provided the best option for downscaling the coarse scale SOCS map down to farm management scale. Moreover, the fine-scale covariates used for the downscaling also need to be carefully selected, and explain the variations in the variable that is downscaled.

The availibility of a downscaling strategy based on covariates is an opportunity for adding value to national models, developed for national inventory exercises during the Kyoto Protocol reporting years. Using the increasing amount of high resolution environmental data recorded at the farm management scale, the disaggregation of these national models to farm scale will provide an initial framework for land owners to audit changes in soil organic carbon stocks through time, which is required by emerging carbon trading schemes that aim to audit and reward management strategies that maintain or sequester organic carbon into global soil resources.

Acknowledgements

This research was funded by the New Zealand Government to support the objectives of the Livestock Research Group of the Global Research Alliance on Agricultural Greenhouse Gases. Any view or opinion expressed does not necessarily represent the view of the Global Research Alliance. The New Zealand based authors wish to thank Massey University for allowing access to their farms, Horizon Regional Council for provision of LiDAR data, and John Dando and Paul Peterson for technical support. Australian based authors were supported from funding from the Australian Department of Agriculture, Round 2—Filling the Research Gap Program (1194105-66) "Farm scale assessment of SOC from disaggregated national/regional scale models".

References

Arrouays, D., Grundy, M.G., Hartemink, A.E., Hempel, J.W., Heuvelink, G.B., Hong, S. Y., Lagacherie, P., Lelyk, G., McBratney, A.B., McKenzie, N.J., et al., 2014. GlobalSoilMap: toward a fine-resolution global grid of soil properties. Adv. Agronomy 125, 93–134.

Behrens, T., Zhu, A.-X., Schmidt, K., Scholten, T., 2010. Multi-scale digital terrain analysis and feature selection for digital soil mapping. Geoderma 155 (3), 175–185.

de Gruijter, J., McBratney, A., Minasny, B., Wheeler, I., Malone, B., Stockmann, U., 2016. Farm-scale soil carbon auditing. Geoderma 265, 120–130.

DeGloria, S.D., Beaudette, D.E., Irons, J.R., Libohova, Z., O'Neill, P.E., Owens, P.R., Schoeneberger, P.J., West, L.T., Wysocki, D.A., 2014. Emergent imaging and geospatial technologies for soil investigations. Photogram. Eng. Remote Sensing 80 (4), 289–294.

Fink, C.M., Drohan, P.J., 2016. High resolution hydric soil mapping using lidar digital terrain modeling. Soil Sci. Soc. Am. J. 80 (2), 355–363.

Hastie, T., 2015. gam: Generalized Additive Models. R package version 1.12. <https://CRAN.R-project.org/package=gam>.

Heung, B., Ho, H.C., Zhang, J., Knudby, A., Bulmer, C.E., Schmidt, M.G., 2016. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma 265, 62–77.

Kuhn, M., 2008. Building predictive models in R using the `caret` package. J. Stat. Softw. 28 (5), 1–26.

Kuhn, M., Weston, S., Keefer, C., code for Cubist by Ross Quinlan, N.C.C., 2014. Cubist: Rule- and Instance-Based Regression Modeling. R package version 0.0.18. <https://CRAN.R-project.org/package=Cubist>.

Leung, L.W., King, B., Vohora, V., 2001. Comparison of image data fusion techniques using entropy and ini. In: Proceedings of the 22nd Asian Conference on Remote Sensing, Singapore, vol. 5, p. 9.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News 2 (3), 18–22. <http://CRAN.R-project.org/doc/Rnews/>.

Lin, L.I., 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics 45 (1), 255–268.

Malone, B., McBratney, A., Minasny, B., Wheeler, I., 2012. A general method for downscaling earth resource information. Comput. Geosci. 41, 119–125.

Malone, B.P., McBratney, A.B., Minasny, B., 2013. Spatial scaling for digital soil mapping. Soil Sci. Soc. Am. J. 77 (3), 890–902.
McBratney, A.B., Santos, M.M., Minasny, B., 2003. On digital soil mapping. Geoderma 117 (1), 3–52.
Minasny, B., McBratney, A.B., 2005. The Matérn function as a general model for soil variograms. Geoderma 128 (3–4), 192–207.
Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. Comput. Geosci. 30, 683–691.
R Core Team, 2015. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.
Roudier, P., Ritchie, A., Hedley, C., Medyckyj-Scott, D., 2015. The rise of information science: a changing landscape for soil science 25 (1), 012023.
Smith, M.P., Zhu, A.-X., Burt, J.E., Stiles, C., 2006. The effects of DEM resolution and neighborhood size on digital soil survey. Geoderma 137 (1), 58–69.
Stockmann, U., Malone, B., McBratney, A., Minasny, B., 2015. Landscape-scale exploratory radiometric mapping using proximal soil sensing. Geoderma 239, 115–129.
Taylor, J., Jacob, F., Galleguillos, M., Prévot, L., Guix, N., Lagacherie, P., 2013. The utility of remotely-sensed vegetative and terrain covariates at different spatial resolutions in modelling soil and watertable depth (for digital soil mapping). Geoderma 193, 83–93.
Viscarra Rossel, R., Adamchuk, V., Sudduth, K., McKenzie, N., Lobsey, C., 2011. Proximal soil sensing: an effective approach for soil measurements in space and time. Adv. Agronomy 113, 237–282.
Viscarra Rossel, R., Chen, C., Grundy, M., Searle, R., Clifford, D., Campbell, P., 2015. The Australian three-dimensional soil grid: Australia's contribution to the GlobalSoilMap project. Soil Res. 53 (8), 845–864.