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
Available online 1 September 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.