Geoderma 160 (2011) 614–626

Contents lists available at ScienceDirect

Geoderma

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

Empirical estimates of uncertainty for mapping continuous depth functions of soil attributes

B.P. Malone [a,*], A.B. McBratney [b,1], B. Minasny [c,2]

[a] Faculty of Agriculture, Food & Natural Resources, The University of Sydney, S207 John Woolley Building, NSW 2006, Australia
[b] Faculty of Agriculture, Food & Natural Resources, The University of Sydney, S224 John Woolley Building, NSW 2006, Australia
[c] Faculty of Agriculture, Food & Natural Resources, The University of Sydney, S222 John Woolley Building, NSW 2006, Australia

ARTICLE INFO

Article history:
Received 4 May 2010
Received in revised form 16 November 2010
Accepted 24 November 2010
Available online 18 December 2010

Keywords:
Digital soil mapping
Model uncertainty
Soil depth functions
Fuzzy k-means with extragrades

ABSTRACT

We use an empirical method where model output uncertainties are expressed as a prediction interval (PI) of the underlying distribution of prediction errors. This method obviates the need to identify and determine the contribution of each source of uncertainty to the overall prediction uncertainty. Conceptually, in the context of digital soil mapping, rather than a single point estimate at every prediction location, a PI, characterised by upper and lower prediction limits, encloses the prediction (which lies somewhere on the interval) and ideally the true but unknown value $100(1-\alpha)\%$ of times on average the target variable (typically 95%). The idea is to partition the environmental covariate feature space into clusters which share similar attributes using fuzzy k-means with extragrades. Model error for predicting a target variable is then estimated from which cluster PIs are constructed on the basis of the empirical distribution of errors associated with the observations belonging to each cluster. PIs for each non-calibration observation are then formulated on the basis of the grade of membership each has to each cluster.

We demonstrate how we can apply this method for mapping continuous soil depth functions. First, using soil depth functions and digital soil mapping (DSM) methods, we map the continuous vertical and lateral distribution of organic carbon (OC) and available water capacity (AWC) across the Edgeroi district in north-western NSW, Australia. From those predictions we define a continuous PI for each prediction node, generating upper and lower prediction limits of both attributes. From an external validation dataset, preliminary results are encouraging where 91% and 93% of the OC and AWC observations respectively fall within the bounds of their 95% PIs. Ideally, 95% of instances should fall within these bounds.

1. Introduction

Soil scientists are acutely aware of the current issues concerning the natural environment because our expertise is intimately aligned with their understanding and alleviation. We know that sustainable soil management alleviates soil degradation, improves soil quality and will ultimately ensure food security. Critical to better soil management is information detailing the soil resource, its processes and its variation across landscapes. Consequently, under the broad umbrella of 'environmental monitoring,' there has been a growing need to acquire quantitative soil information (Grimm and Behrens, 2010; McBratney et al., 2003). The concerns of soil-related issues in reference to environmental management were raised by McBratney (1992) when stating that it is our duty as soil scientists, to ensure that the information we provide to the users of soil information is both accurate and precise, or at least of known accuracy and precision.

However, a difficulty we face is that soil can vary, seemingly erratically in the context of space and time (Webster, 2000). Thus the conundrum in model-based predictions of soil phenomena is that models are not 'error free.' The unpredictability of soil variation combined with simplistic representations of complex soil processes inevitably leads to errors in model outputs.

We do not know the true character and processes of soils and our models are merely abstractions of these real processes. We know this; or in other words, in the absence of such confidence, we know we are uncertain about the true properties and processes that characterise soils (Brown and Heuvelink, 2005). The key is therefore to determine to what extent our uncertainties are propagated through a model of which effect the final predictions of a real-world process.

In modelling exercises, uncertainty of the model output is the summation of the three main sources generally described as: model structure uncertainty, model parameter uncertainty and model input uncertainty (Brown and Heuvelink, 2005; Minasny and McBratney,

* Corresponding author. Tel.: +61 290 365 278.
E-mail addresses: brendan.malone@sydney.edu.au (B.P. Malone), Alex.McBratney@sydney.edu.au (A.B. McBratney), budiman.minasny@sydney.edu.au (B. Minasny).
[1] Tel.: +61 293 513 214.
[2] Tel.: +61 290 36 9 043.

0016-7061/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.geoderma.2010.11.013

2002b). The general procedure is to determine independently the contribution of each source to the overall uncertainty. One obvious issue of this is that generating estimates of uncertainty for each of the sources could become a prohibitive exercise in terms of time and cost.

With this in mind, there are a number of approaches to estimate the uncertainty of model outputs. One of these is an empirical approach in which the residuals between modelled outputs and corresponding observed data are used to formulate a prediction interval (PI). Such an approach was proposed by Shrestha and Solomatine (2006) where uncertainty is expressed in the form of two quantiles of the underlying distribution of model error (residuals). It is stated that the PI explicitly takes into account all sources of uncertainty and circumvents attempts to separate out the contribution of each source (Shrestha and Solomatine, 2006; Solomatine and Shrestha, 2009). Their idea is to partition a feature space into clusters (with a fuzzy k-means routine) which share similar model errors. A PI is constructed for each cluster on the basis of the empirical distribution of residual observations that belong to each cluster. A PI is then formulated for each observation in the feature space according to the grade of their memberships to each cluster. They applied this methodology to artificial and real hydrological data sets and it was found to be superior to other methods which estimate a PI. The Shrestha and Solomatine (2006) approach computes the PI independently and while free of the prediction model structure, it requires only the model or prediction outputs. Tranter et al. (2010) extended this approach to deal with observations that are outside of the training domain.

Application of the Shrestha and Solomatine (2006) approach for estimating model output uncertainty has not previously been attempted in a digital soil mapping (DSM) framework where they are often infrequently reported (Grunwald, 2009). Such was the case in Malone et al. (2009) where a methodology for mapping continuous depth functions of soil attributes was introduced. Other than to identify the sources, no attempt was made to address the contribution of each source to the overall uncertainty. Given the modest results in terms of accuracy, particularly with increasing soil depth it was assumed that significant model uncertainties existed. It is thought that given the complexity of the modelling component to generate predictions both in a vertical and lateral space, that traditional forms of uncertainty analysis would be prohibitive and time consuming. A pragmatic approach to this dilemma is to use an empirical methodology in a similar fashion to that presented by Shrestha and Solomatine (2006) and Tranter et al. (2010) who view the model residuals as the best quantitative measure of the discrepancy between a model and the modelled real-world process.

The method we present here modifies slightly the Shrestha and Solomatine (2006) and Tranter et al. (2010) approach to suit it for a DSM framework. It extends the idea of model uncertainty by extrapolating the uncertainty parameters across the extent of a defined area so that mapping the continuous depth functions (Malone et al., 2009) involves both the mapping of predictions and their uncertainties. In addition to presenting this modified approach to uncertainty analysis, we also perform an external validation in order to gauge how successful this method works for this particular application of DSM.

2. Theory and scope of work

2.1. The prediction interval as a measure of uncertainty

The characteristics of a PI include both upper and lower prediction limits. The interval between the prediction limits constitutes the PI (Fig. 1). Given a prescribed probability such as a 95% confidence level, a future unknown value is expected to lie somewhere along this interval. There is a clear distinction between a PI and a confidence interval (CI), however. A CI tells us how well or how accurate is the estimate of a true regression to predict one variable from another. Conversely, the PI deals with the accuracy of the prediction with respect to the corresponding observed value. Thus a PI is always wider than a CI because it includes both the uncertainty in knowing the value of the population mean, and the uncertainty of the new measurement (Altman and Gardner, 1988).

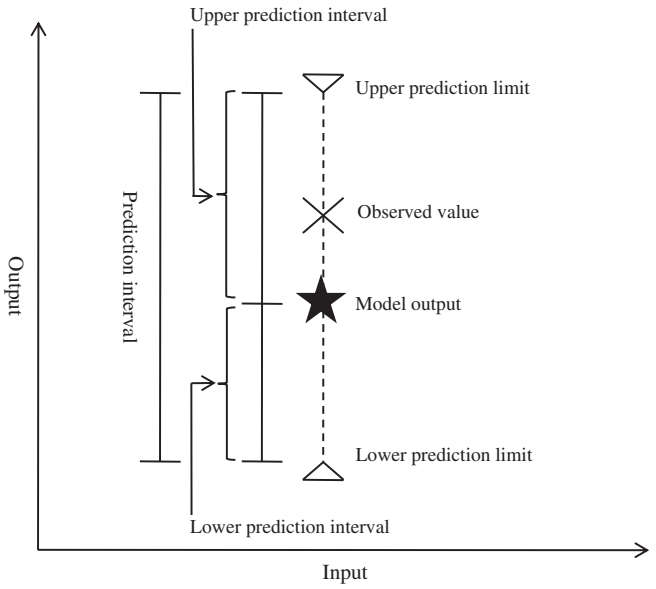

Fig. 1. The prediction interval, characteristic features and general descriptive terminology.
Adapted from Shrestha and Solomatine (2006).

Calculating the PI for a given observation following the method of Shrestha and Solomatine (2006) is performed independently of the model building or calibration process. Solomatine and Shrestha (2009) refer to this as "uncertainty estimation based on local errors and clustering" (UNEEC). The purpose of the UNEEC is to derive the upper and lower prediction limits based on the model error, and since it is estimated through empirical distribution, it is not necessary to make any assumption about residuals (Solomatine and Shrestha, 2009). First a user-defined class of regression is performed to estimate the target variable from one or a suite of predictor variables or covariates. The prediction outputs are compared to their observed corresponding values; the residuals are recorded. Using a clustering routine such as fuzzy k-means (Bezdek, 1981), the calibration dataset is partitioned into c clusters corresponding to different values or distributions of the residuals. It is assumed that the region in the feature space associated with any particular cluster has similar residuals or residuals with similar distributions. Once the clusters have been identified, the PIs for each cluster are computed from empirical distributions of the corresponding residuals. To construct a $100(1-\alpha)\%$ prediction, the $(\alpha/2)\times 100$ and $(1-\alpha/2)\times 100$ percentile values are taken from the empirical distribution of residuals for the lower and upper prediction limits respectively. The computation of the PI for each calibration observation is straightforward if it belongs entirely to one cluster as would be the case where the input space is divided into crisp clusters e.g. hard clustering. However in the case of fuzzy clustering, where each observation belongs to all available clusters with respect to a membership grade, a "fuzzy committee" approach is used where the PI is computed using the weighted mean of the PI of each cluster (Shrestha and Solomatine, 2006). This can be defined mathematically as:

$$PI_j^L = \sum_{i=1}^{c} m_{ij} PIC_j^L$$
$$PI_j^U = \sum_{i=1}^{c} m_{ij} PIC_j^U \quad (1)$$

where PI_i^L and PI_i^U correspond to the weighted lower and upper PI for the ith observation. PIC_i^L and PIC_i^U are the lower and upper PIs for each cluster j, and m_{ij} is the membership grade of ith observation to cluster j. Finally, the lower and upper prediction limits (PL_i^L and PL_i^U respectively) are derived for each calibration observation by adding the prediction (from the prediction model) to PI_i^L and PI_i^U.

2.2. Validating the prediction interval

Validation of the PI is performed externally with a dataset separate from the calibration dataset. In the context of DSM, data splitting or collecting additional samples using some sort of probability sampling are the most common methods for which validations are then based upon; see Grinand et al. (2008) and Kempen et al. (2009) for recent examples of each. The procedure for validating (the PI) follows closely with Shrestha and Solomatine (2006) in which the idea is to simply determine whether each of the validating observed values is inside their respective prediction limits. By definition, the prediction limits enclose the true but unknown value $(1-\alpha)\%$ of times on average (typically 95%). The performance of the model is therefore evaluated by means of a prediction interval coverage probability (PICP) (Shrestha and Solomatine, 2006) whereby the PICP is the probability that all observed values fit within their prediction limits and is estimated by:

$$\mathrm{PICP} = \frac{1}{V}\,\mathrm{count}\ i \quad i: PL_i^L \le obs_i \le PL_i^U \tag{2}$$

V is the number of observations in the validation dataset. The clustering technique and uncertainty model is said to be optimal when the PICP value is close to the $100(1-\alpha)\%$.

2.3. Fuzzy clustering

Particularly important in this study is what methodology of clustering we use, especially in the context of soil variability and identifying regions in a study area where predictions are more certain in some areas in comparison to other areas. In general terms, clustering is the unsupervised partitioning of a feature space into natural groups or clusters which share some measure of similarity. Many clustering techniques are in existence for which Jain et al. (1999) comprehensively reviews. In the domain of soil science, the most widespread clustering algorithms are non-hierarchical or in other words, have a partitional basis (McBratney and Odeh, 1997). The k-means algorithm is the simplest partitional clustering method and aims to minimise the within-class sum of square distances between the input space observations and the corresponding cluster centroids (McQueen, 1967). An extension of the k-means algorithm is fuzzy k-means (FKM) which allows each observation a degree of membership to j clusters (Bezdek et al., 1984). Because soil is both spatially and temporally continuous, the FKM approach to classification is more commonly used. The FKM algorithm minimises the objective function:

$$J(C,M) = \sum_{i=1}^{n}\sum_{j=1}^{c} m_{ij}^{\varphi} d_{ij}^2 \quad i = 1,\ldots,n; j = 1,\ldots,c \tag{3}$$

where C is the $c\times p$ matrix of class centres, M is the $n\times c$ matrix of partial memberships, $m_{ij}\in[0,1]$ is the partial membership of the ith observation to the jth cluster, $\varphi\ge 1$ is the fuzziness exponent. Increasing φ results in a fuzzier partition between clusters. The square distance between the ith observation and jth cluster centre is d_{ij}^2. A more detailed explanation of the FKM algorithm can be found in Bezdek (1981). Essentially FKM gives the number of clusters; it defines class centroids based on each variable and calculates optimally the memberships of each observation to each defined cluster.

McBratney and de Gruijter (1992) recognised a limitation of the FKM algorithm in that it had the inability to distinguish between observations very far from the cluster centroids and those at the centre of the centroid configuration. The observations were termed extragrades as opposed to intragrades, which are the observations that lie between the main clusters. The extragrades are considered the outliers of the data set and have a distorting influence on the configuration of the main clusters (Lagacherie et al., 1997). McBratney and de Gruijter (1992) developed an adaptation to the FKM algorithm which distinguishes observations that should belong to an extragrade class. The FKM with extragrades algorithm minimises the objective function:

$$J_e(C,M) = \alpha\sum_{i=1}^{n}\sum_{j=1}^{c} m_{ij}^{\varphi} d_{ij}^2 + (1-\alpha)\sum_{i=1}^{n} m_{i^*}^{\varphi}\sum_{j=1}^{c} d_{ij}^{-2} \tag{4}$$

The notation is similar to the FKM algorithm except where m_{i^*} denotes the membership to the extragrade class. This function also requires the parameter *alpha* (α) to be defined which determines the degree of importance attributed to the extragrade class. Details of FKM with extragrades are comprehensively discussed in McBratney and de Gruijter (1992) and Odeh et al. (1992).

Shrestha and Solomatine (2006) used a FKM algorithm for their clustering routine which also implemented the Euclidean distance measure where equal weight is given to all the variables in the feature space. For this study we implement the FKM with extragrades algorithm for the reasons described above. Tranter et al. (2010) also point out that extragrade instances exist spatially in regions of low density calibration data. As a consequence, this fact also confers a low reliability of prediction in these areas which is an important consideration in the context of DSM. A Mahalanobis distance measure is also used in our procedure and for reasons discussed in more detail later is on the basis that we cluster the feature space based on a suite of available soil prediction covariates rather than the prediction errors themselves. The Mahalanobis distance takes into account the correlation between variables in the feature space.

2.4. Adaptation of the of the UNEEC approach for digital soil mapping of continuous depth functions

In order to use the UNEEC procedure of Shrestha and Solomatine (2006) and Solomatine and Shrestha (2009) within a DSM framework, some critical modifications and assumptions need to be made. The first involves the feature space clustering such that the model errors are calculated on the basis of the available soil state factors. The idea is to perform the clustering routine prior to estimating model errors after which the cluster PIs are then formulated. The key assumption of this paper therefore is that particular areas within a landscape will have similar residuals or distribution of residuals and ultimately share a similar range of uncertainty.

The second modification or moulding to a DSM framework is the question of how we extend the PI to prediction nodes that have not been visited. Like in any DSM project, training rules are constructed on calibration data which are then extrapolated across a study area where only the prediction covariates are known. This approach is maintained for estimating the uncertainties at these sites whereby the cluster centroids derived from the calibration procedure are used to determine the membership grade of each prediction node across the study area to each cluster. Once these are known, PI_i^L and PI_i^U can be calculated after which PL_i^L and PL_i^U are derived once the model prediction is made.

In order to derive a continuous depth estimation of the upper and lower prediction limits for each prediction node, we follow the same routine as for mapping continuous functions of soil attributes as

presented in Malone et al. (2009). Once the standard depths of prediction are determined, an uncertainty model is used to estimate the PI at each depth increment. Using the PIs at each depth as parameters, we can then perform a mass-preserving spline reconstruction to generate continuous representations of the upper and lower prediction limits to a prescribed maximum depth.

2.5. Procedure

There are thus three components that need to be adhered to replicate this approach in a DSM framework. The first of which is the prediction model which essentially recreates the method of Malone et al. (2009) for using splines and regression kriging modelling for the prediction of soil attributes at standardised depths. The second component involves the training of the empirical uncertainty model, from which cluster PIs can be derived. For both components, validation is performed using an independent dataset. In this study, both calibration and validation sets are the same for these two components. The third component involves the mapping of the predictions and associated PIs in the lateral and vertical dimensions. A summary of the stepwise procedure for achieving these outcomes is as follows and illustrated in Fig. 2:

- Soil data is pre-processed then arranged for analysis. The framework we present here applies to available soil datasets where soil observations within a profile have been made at horizon and/or regular depth increments. Often the depths of observation between profiles are not the same;
- Mass preserving spline functions (Bishop et al., 1999) are used to standardise the depth increments of prediction for all available site locations;
- The dataset is then spatially intersected with a suite of available environmental covariates;
- The dataset is then randomly split into two; we use 85% of total site observations for calibration and 15% for validation.

The prediction model

- The calibration dataset is used to train a prediction model via a regression technique in which the prediction variables are the environmental covariates;
- For validation of the prediction model, the training rules are extended to the validation data. Then evaluate statistically the accuracy of the predictions for each of the standard depths for example the root mean square error (RMSE). It needs to be kept in mind that the validation data in this case are not the actual observations or 'hard data,' but 'soft data' that in addition to sampling and measurement errors are prone also to interpolation errors due to the fitting of the splines. While not optimal, it is necessary to perform validation in this way (at the standardised depths) because the 'hard data' is rarely observed systematically at set depths from one profile to the next;
- The training rules are further saved for the mapping component.

The empirical uncertainty model

- Clustering is performed where the feature space is the suite of environmental covariates observed at each calibration site. Once the optimal number of clusters is determined, cluster centroids are saved;
- Estimation of the model error is made. We would opt to use the same class of model as used in the prediction model component but generate predictions via a leave-one-cross-validation method. The idea is that we save only the residual between the predicted (model output) and observed value at each standardised depth;

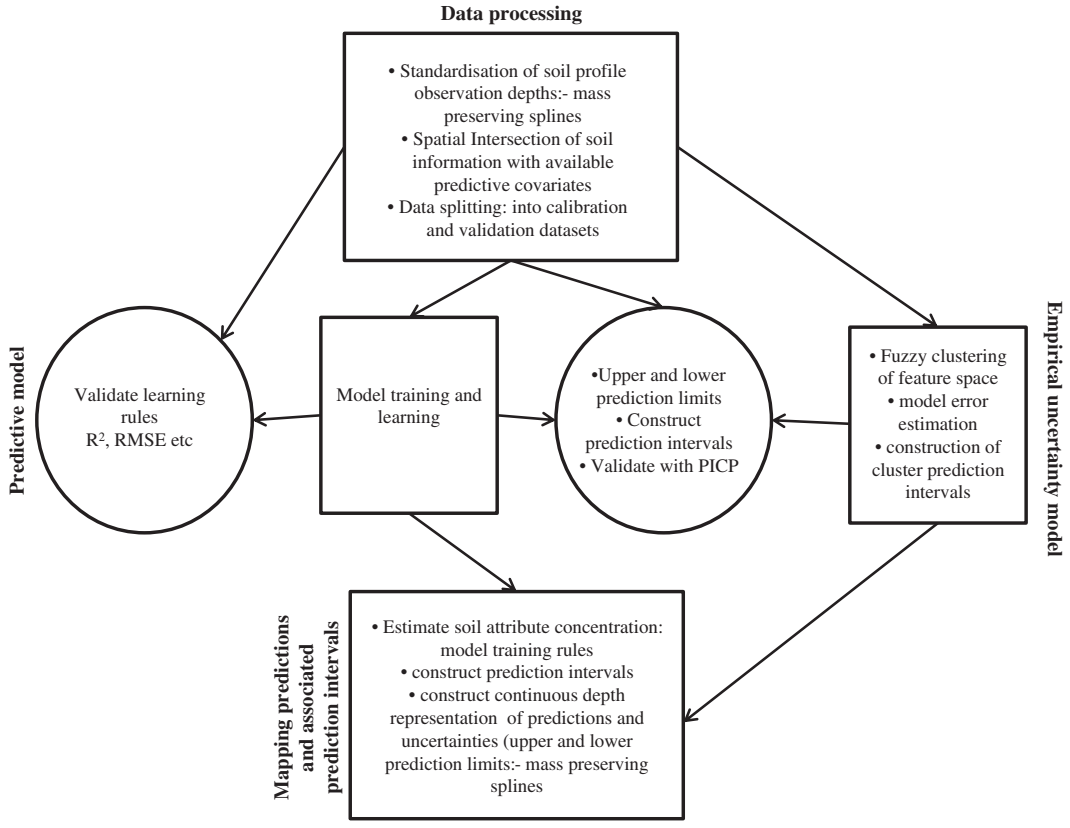

Fig. 2. Flow diagram of the general procedure for achieving the outcome of mapping predictions and their uncertainties (upper and lower prediction limits) within a digital soil mapping framework. The 3 components for achieving this outcome are the prediction model, the empirical uncertainty model and the mapping component.

- The highest membership value to a cluster determines which cluster each observation is assigned to. Cluster PIs (PIC_i^L and PIC_i^U) are found for each cluster on the basis of the distribution of residuals within each cluster. To construct a 95% PIs (for each cluster) we take the lower 2.5% and upper 97.5% percentile values from the empirical distribution of residuals in each cluster;
- Validation (using the validation dataset) of the uncertainty procedure involves construction of the PI for each observation i.e. PI_i^L and PI_i^U (Eq. (1)) where the requirements are the cluster PIs and cluster centroids, from which the cluster membership values can be derived. Adding the prediction of the soil attribute (from prediction model) to PI_i^L and PI_i^U yields the upper and lower prediction limits (PL_i^L and PL_i^U) for each validation observation. Calculating the PICP determines (for a given confidence level) the proportion of observed values which fit within their respective prediction limits (Eq. (2)).

Mapping of the predictions and associated PI

- Training rules from the calibration procedure are extended to all prediction nodes across a study area where only covariate information is available;
- Cluster centroids from the uncertainty model are used to determine the membership grade each prediction node has to each cluster, from which PI_i^L and PI_i^U are formulated;
- Adding the model prediction to PI_i^L and PI_i^U yields the upper and lower prediction limits at each prediction node;
- Continuous depth representations of the predictions and upper and lower prediction limits can be generated by using the mass-preserving spline reconstruction method where the observations at the standardised depths are the only required parameters.

3. Materials and methods

3.1. The data

We test the approach described above using actual soil data where our target properties are organic carbon (kg m^{-3}) (OC) and available water capacity (m m^{-1}) (AWC). The area from which the data has been collected is the Edgeroi district located near Narrabri (30.32S 149.78E) in north-western New South Wales, Australia. Details of this predominantly agricultural district can be found in McGarry et al. (1989) and Ward (1999). The soil dataset consists of 341 soil profiles, 210 of which were sampled on a systematic, equilateral triangular grid. At most profile sites observation of OC and AWC were made at the depth increments of 0–0.1, 0.1–0.2, 0.3–0.4, 0.7–0.8, 1.2–1.3 and 2.5–2.6 m (McGarry et al., 1989).

Environmental covariates were compiled for the whole Edgeroi study area ($\approx$1500 km^2) on a grid with spatial resolution of 90 m × 90 m. These included a digital elevation model and its derivatives, Landsat 7 ETM images from 2003 and gamma radiometric data from airborne survey (Geosciences Australia, 2008). Specifically, the environmental covariates used for analysis in this study were elevation, slope, altitude above channel network (AOCN), flow path length (FPL), multi-resolution index of valley bottom flatness (MRVBF) (Gallant and Dowling, 2003) and SAGA wetness index which is similar to the topographic wetness index (TWI) (Boehner et al., 2002). Bands 1–5 and 7 of the Landsat ETM were used in addition to the Normalised Difference Vegetation Index (NDVI) and soil enhancement ratios of b3/b2, b3/b7 and b5/b7 (Saunders and Boettinger, 2007). For gamma radiometric data, the percentage measure of radiometric K was used in addition to the ppm measures of both radiometric U and Th.

For all soil profiles (341), the procedure as described in Malone et al. (2009) was used to fit splines to the observed values down to 1 m. This generated continuous profile descriptions to 1 m for both soil attributes. From the fitted splines of the observed data, the mean value of each soil attribute was derived at the specified depth increments of: 0–10, 10–20, 20–30, 30–40, 40–50, 50–70, 70–80, 80–100 cm. These 8 surfaces became the target inputs to be modelled against the suite of environmental covariates which were then intersected to the data based on the spatial location of each soil profile description. Lastly, the dataset was then randomly divided into two sets: 291 profiles for calibration and 50 profiles for validation.

3.2. The prediction model

3.2.1. Prediction model calibration

In terms of the modelling process, the systematic approach for model calibration from Malone et al. (2009) was used to generate rules or formulae based on the relationship between calibration data at the eight specified depth increments and the suite of environmental covariates. Like Malone et al. (2009) this study uses a regression kriging approach for prediction where neural networks model the deterministic component of variation and residual kriging to model the stochastic component of the variation. Once an appropriate model was constructed for each soil attribute, residuals were calculated and then saved for use in the following validation and mapping procedures.

3.2.2. Prediction model validation

For validation, model formulae generated in the calibration procedure were used to derive the initial predictions of OC and AWC from the suite of environmental covariates that existed at each validating site. Kriging was used to interpolate the residuals at each validation observation based on the localised exponential variogram model of the 100 nearest residuals found for the calibration procedure. A final prediction was derived from the summation of the model prediction and the interpolated residual. To determine the accuracy of the final predictions with their corresponding observed values we used the root mean square error (RMSE) and Lin's Concordance Correlation Coefficient (CCC) (Lin, 1989).

3.3. The empirical uncertainty model

3.3.1. Uncertainty model calibration: fuzzy clustering and formulation of cluster prediction intervals

To establish the optimal cluster size and φ value of the calibration data, fuzzy classification was performed with the *FuzME* software (Minasny and McBratney, 2002a). As discussed previously, in this study we used the FKM with extragrades function. The environmental covariates of the calibration data were arranged in a matrix of N observations × M covariates (291 × 19). Iteratively, using cluster sizes of 2 through 15, the FKM with extragrades function ran using successive φ values ranging from 1 through 2 with step length of 0.01. For this study along with an intuitive guide, we adopted an internal criterion approach to determine the optimal cluster size and φ value using both the modified partition entropy (MPE) and the derivative of the objective function with respect to the fuzzy exponent (φ), $-(\delta J/\delta\varphi)c^{0.5}$ (McBratney and Moore, 1985). Such indices have been used previously by Bragato (2004) and Odeh et al. (1992) where more detailed discussion is made about them. The MPE establishes the degree of fuzziness created by a specified number of classes for a defined φ value. The notion is that the smaller the MPE, the more suitable is the corresponding number of classes at the given φ value. The derivative of $J_e(C,M)$ with respect to φ is used to simultaneously establish the optimal φ and cluster size. The optimal φ will maximise $-(\delta J/\delta\varphi)c^{0.5}$ and the most suitable cluster size will produce the curve with the lowest maximum. In this study these indices are used as a general guide; once the range of possible combinations has been narrowed, we then intuitively decide on the most suitable cluster size and φ value on other non-clustering related criteria such as the number of observations within each cluster and the associated distribution of errors

assigned to each cluster i.e. as close to being normally distributed as manageable.

To determine model error we employed a leave-one-out-cross-validation procedure (Hastie et al., 2009). With this form of cross-validation, there were $N=291$ sets of computations. Using neural networks, a prediction is made for each successive calibration profile (at each depth increment) based on the learning rules of the remaining $N-1$ calibration profiles. Furthermore, with each computation, kriging was used to interpolate the residual at each depth increment based on the spatial auto-correlation of residuals of the $N-1$ calibration profiles. A final prediction resulted from the summation of the prediction and interpolated residual. Thus model error in this case was determined to be the difference between the observed value at a specific depth increment and its corresponding final prediction.

To calculate the cluster PIs (PIC_i^L and PIC_i^U) at each depth increment, we first arranged the observations into their respective clusters on the basis of their highest cluster membership grade. For a 95% PI, we took the upper 97.5% and lower 2.5% quantiles of each distribution for every cluster. In terms of handling the extragrade cluster error distributions, we follow the procedure of Tranter et al. (2010) whom suggested a penalisation calculation to extragrade areas where there is very low model prediction confidence. The PI for the extragrade class can be evaluated by:

$$\begin{aligned} PIC_{ej}^L &= 2 \times q_{2.5} \\ PIC_{ej}^U &= 2 \times q_{97.5} \end{aligned} \quad (5)$$

where PIC_{ej}^L and PIC_{ej}^U are the lower and upper PIs of the extragrade class, and q is the quantile value of the extragrade cluster error distribution at each depth increment.

3.3.2. Uncertainty model validation

Validation of the uncertainty model follows the description as detailed in Section 2.5. where PI_i^L and PI_i^U are formulated. The PICP is estimated accordingly for a 95% PI based on the count of observed values that lie within the PI for each site at each depth. As such the PICP considers all observations (site and depth increments) as independent observations. Thus the PICP is the proportion at all depths across all observations which lie within the 95% PI. To assess the sensitivity of the model by means of reducing the confidence limit sequentially, we constructed PIs for various confidence levels ranging from 5% to 99%. As for a 95% prediction level, by definition you would expect the PICP value to be close to the corresponding confidence or $100(1-\alpha)\%$ level.

3.4. Mapping of predictions and their uncertainties

For mapping of the predictions and associated prediction PIs follows precisely the steps as outlined in Section 2.5. Along with generating maps of the predictions and their uncertainties (displayed as upper and lower prediction limits) at the standard depths we also demonstrate the functionality of the splines in a DSM framework which was also demonstrated in Malone et al. (2009). In this study we determine to total predicted AWC and OC across the study area to a depth of 1 m. These predictions are also accompanied by upper and lower prediction limits.

4. Results and discussion

4.1. The prediction model

4.1.1. Prediction model calibration

For model calibration of OC, we found that a neural network model with 4 hidden nodes was appropriate in terms of predictive power without over-fitting the data. For AWC, 3 nodes was found to be the most appropriate model configuration. The coefficients of determination (R^2) were reasonable at 54% for OC and 48% for AWC.

4.1.2. Prediction model validation

Generally predictions of OC were strongest at the surface and poorest at the bottom of the soil profile. As shown in Fig. 3a–c, CCC ranged from 0.28 (RMSE = 0.3) in the 0–10 cm depth increment, to 0.11 (RMSE = 0.54) at 30–40 cm through to −0.05 (RMSE = 0.90) for the 80–100 cm depth increment.

It was apparent that there was a low spatial auto-correlation between residuals at all 8 depth increments (data not shown), which was also observed for OC in Malone et al. (2009). Nevertheless, adding the model prediction to the interpolated validation residuals resulted in modest improvements to the prediction of OC. In Fig. 3d–f, CCC increased to 0.38 and 0.14 at the 0–10 cm and 30–40 cm depth increments respectively, but no change was observed at 80–100 cm. In each of the cases there was no improvement in the RMSE values.

For AWC, the validation results were also modest where predictions were strongest in the 80–100 cm depth increment (CCC = 0.1) (data not shown). In terms of all 8 depth increments CCC ranged from 0.04 to 0.1(RMSE = 0.015–0.027). As for OC, there was not a well defined function for the spatial distribution of residuals. Additionally, any spatial auto-correlation that we were able to define was independent for each depth increment. Overall, adding the residuals to the predictions of AWC resulted in little to no improvements.

The average CCC of the final validation predictions at the 8 depth increments for AWC and OC found for Malone et al. (2009) was 0.44 and 0.38 respectively. Essentially the validation results in this paper are weaker. An explanation for this is that we did not perform independent multivariate analyses for both attributes to determine what the most correlated covariates were before construction of the neural networks. This is because, as discussed later, we performed the clustering process of the environmental covariates once only. Thus the defined clusters could be used simultaneously for both soil attributes. Given that the final predictions of both OC and AWC are modest in this paper, future studies of this type will need to include an independent multivariate analysis prior to modelling. This would also mean that the clustering process would become independent for each predicted soil attribute and that the error determination through the leave-one-out-cross-validation would only include those environmental covariates that are significant for each soil attribute.

4.2. The empirical uncertainty model

4.2.1. Uncertainty model calibration: fuzzy clustering and formulation of cluster prediction intervals

The optimal cluster size for the given environmental covariates using the FKM with extragrades algorithm was found to be 6, including the extragrade class. Clustering resulted in 61 observations belonging to cluster A, while 38, 43, 54 and 61 observations belonged to cluster B, cluster C, cluster D, cluster E and the extragrade cluster respectively.

Box plots of the empirical distribution are shown for AWC (Fig. 4a–c) and OC (Fig. 4d–f) at the depth increments of 0–10 cm, 40–50 cm, and 80–100. At all depth increments the distribution of model errors are different for each cluster. For both soil attributes there was a decreasing distribution of model errors with increasing depth down the profile, which was proportional to the observed values at each depth increment. For example the proportion of the error and observed value (error/observed value) for OC and AWC regardless of cluster at 0–10 cm and 80–100 cm were both found to be very close to 1. For AWC the distribution of model errors was largest for the extragrade cluster at all depth increments. This upholds the notion that reliability of the prediction in this part of the feature space is low in comparison to the other clusters. This relationship was

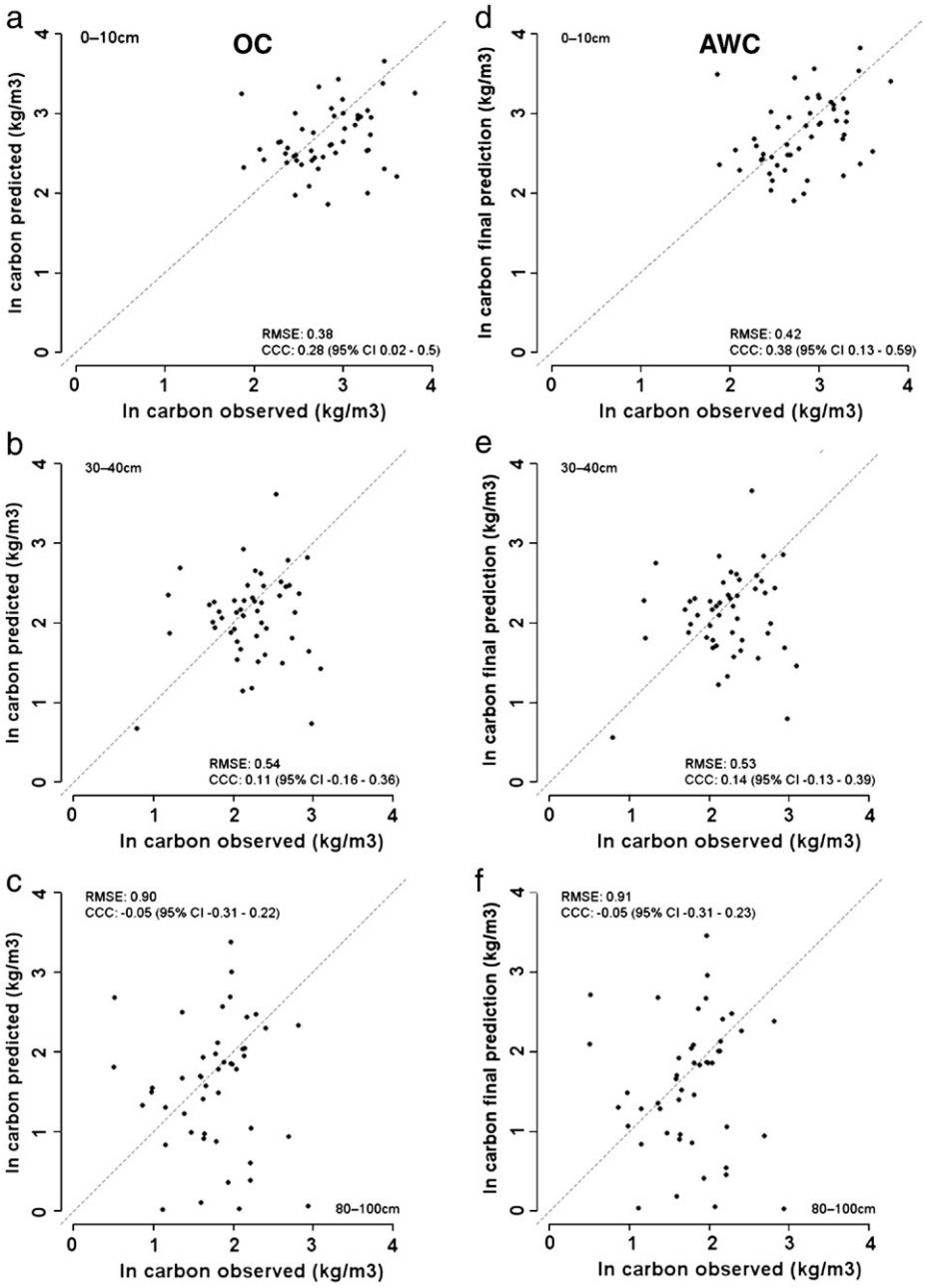

Fig. 3. Model validations of OC. Observed vs. fitted plots at 0–10 cm, 30–40 cm and 80–100 cm before adding residuals (a–c) and after adding residuals (d–f).

mostly observed also for OC, but there some exceptions (for example Fig. 4d). Applying the penalisation calculation for the extragrade class (Tranter et al., 2010) ensured that PIC_{ej}^{L} and PIC_{ej}^{U} were larger than those of the other clusters at each depth increment.

4.2.2. Uncertainty model validation

The final prediction and corresponding upper and lower predictions at the 8 specified depth intervals were used as parameters to construct estimated splines. This process created 3 continuous profiles for each observation: a final prediction profile and lower and upper prediction limit profiles. We randomly selected 5 validation sites to illustrate how these splines compare to the measured values of OC and AWC at each of these sites. As can be seen in Fig. 5 that the dotted lines constitute the final prediction, while the solid lines equate to the lower and upper prediction limits, conferring the 95% PI. The bars represent the measured values of AWC (Fig. 5a–e) and OC (Fig. 5f–j). As can be seen, a measured value fits within a PI if the right vertical side of the bar fits completely within the confines of the solid lines.

Due to the proportionality of the observed values with the prediction errors, the PIs are mostly widest at the soil surface. Generally for AWC, the PI then narrows gradually from about 40 cm to a roughly equally spaced interval to 1 m. In the example of (Fig. 5b), however, the interval is narrowest at 1 m. A similar situation is evident for OC (Fig. 5f–j) where uncertainty is greatest towards the soil surface. Additionally for OC there are a couple of instances where the PI does not completely confine the measured value (Fig. 5i and j). By deriving the relative proportion of the PI (range of the upper and lower prediction limits) and the corresponding prediction of all validation observations at each depth, it was found that the proportions slightly increased with depth. This indicates an increasing uncertainty with increasing soil depth, which was also the finding when validating the prediction model. For AWC again the PI was

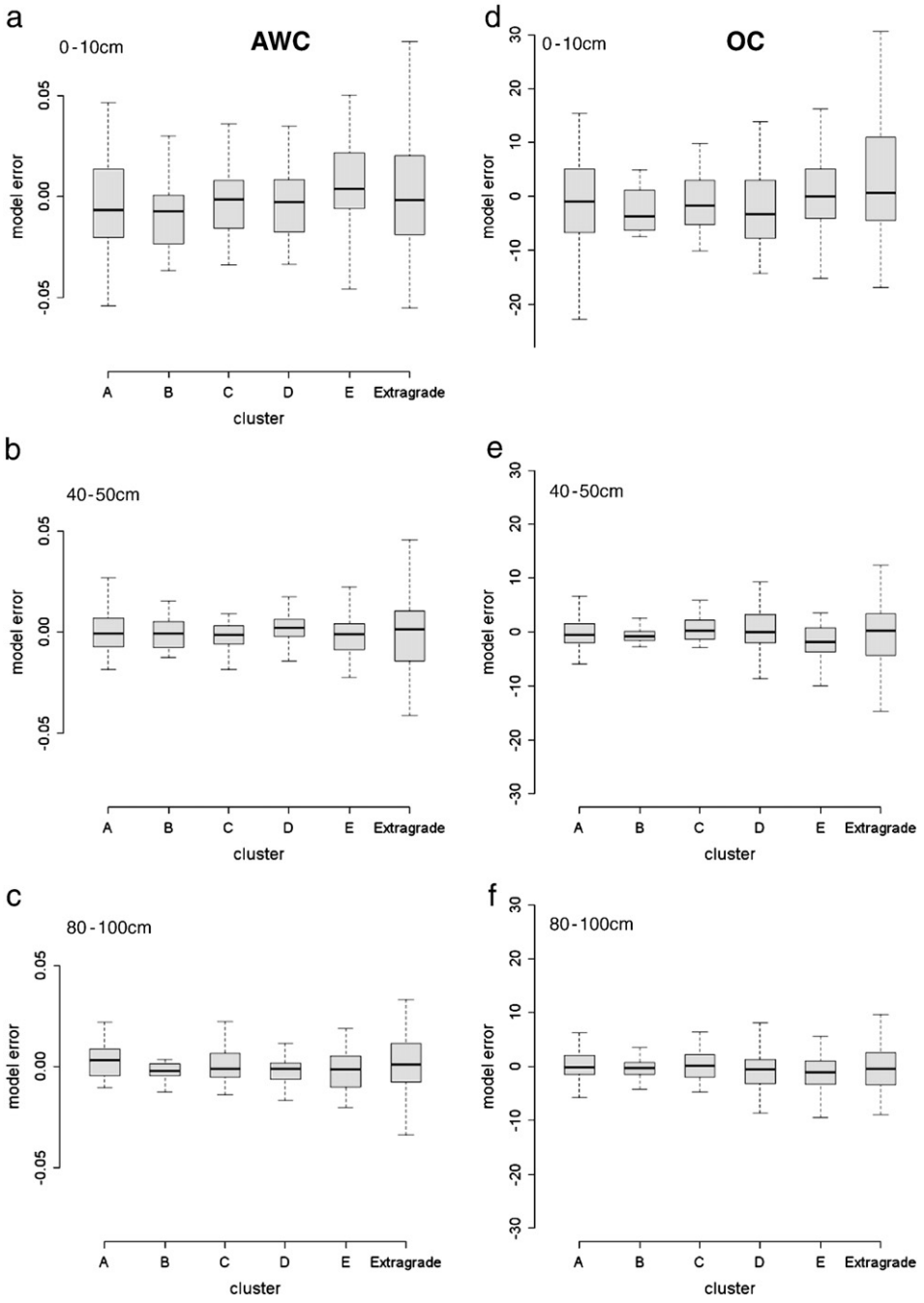

Fig. 4. Box plots of the empirical distributions of model error as derived from the empirical uncertainty model at the selected depths of 0–10 cm, 40–50 cm and 80–100 cm for AWC (a–c) and OC (d–f).

found generally to be proportional to the prediction at a given depth; however the predictions were strongest at the sub-soil depth increments. This was also reflected when validating the prediction model for AWC and counters that found in Malone et al. (2009) which could be put down to the difference in covariates and modelling steps used between the two studies.

The results of the PICP analysis indicate that at the desired confidence level of 95%, 91 and 93% of all observations fitted with their given PIs for OC and AWC respectively, indicating with this type of validation, that the empirical uncertainty model is optimal for both soil attributes (Fig. 6a–b) . Furthermore, with each successive decrease in the confidence levels a near corresponding decrease in the PICP is observed for both attributes indicating a required outcome in terms of sensitivity of the PI to changing confidence levels.

4.3. Mapping of predictions and their uncertainties

The maps in Fig. 7 illustrate the spatial variability of the degree of membership each prediction node has to each class, based on the cluster centroids derived from the empirical uncertainty model procedure. This gives a good representation of which areas in the extent of the study area share a similarity based on the given suite of environmental covariates.

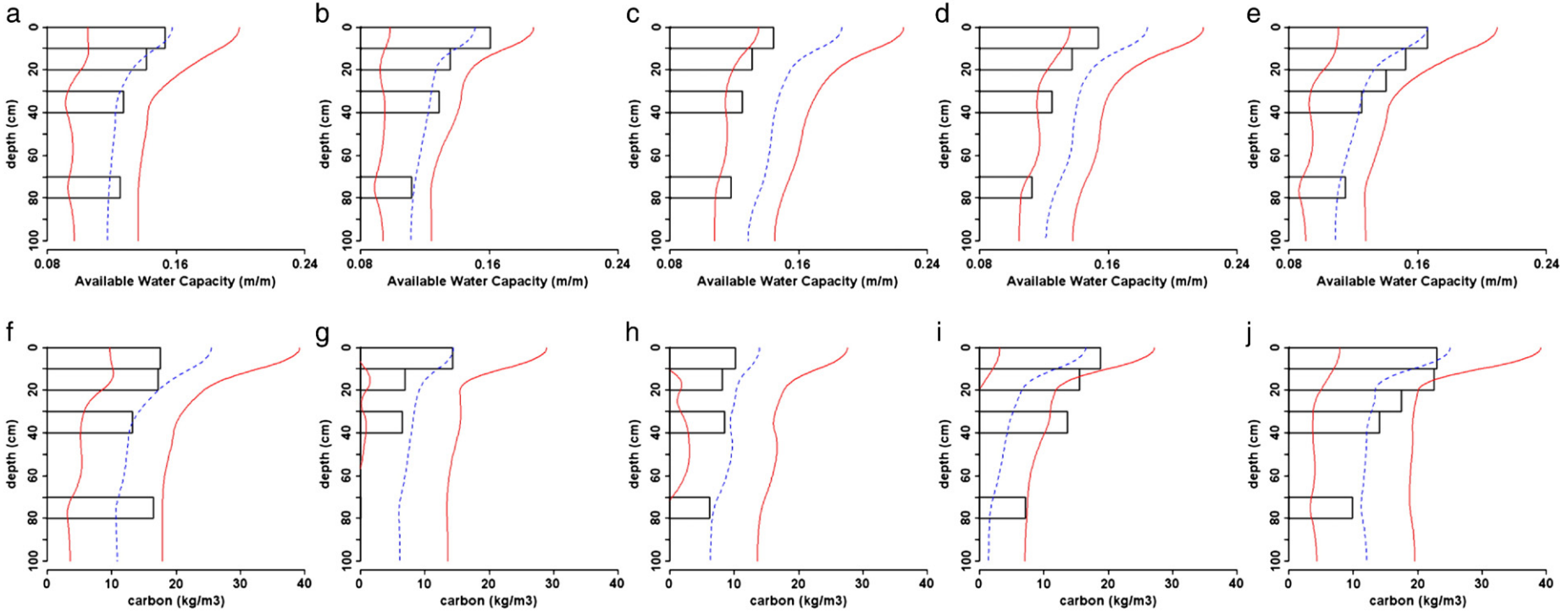

Fig. 5. Profile plots of AWC (a–e) and OC (f–j) at randomly selected validation sites. Bars represent actual observed values. Dotted lines represent final DSM predictions. Solid lines represent upper and lower prediction limits.

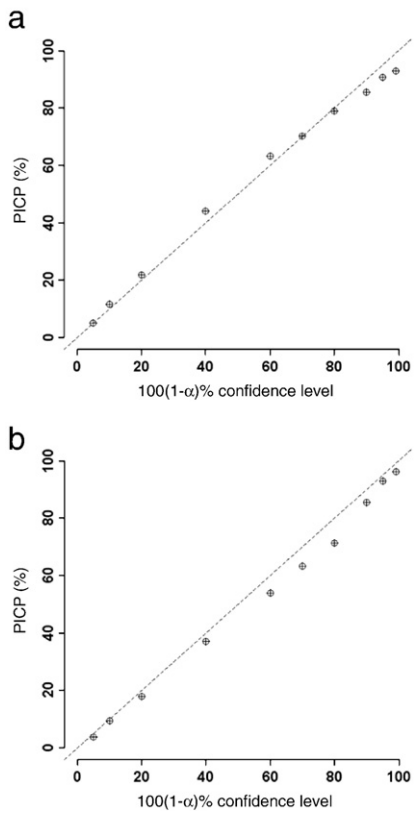

Fig. 6. Prediction interval coverage probability plots (PICPs) for OC (a) and AWC (b).

Areas with a high degree of membership to the extragrade class (Fig. 7a) appear to correspond to regions that are topographically diverse such as that to the east and south west where undulating slopes and hills are situated. These areas are also moderately to densely populated by native forest. While sample sites exist across these landscapes in a predominantly equilateral triangular grid design (McGarry et al., 1989), it is likely that given the combination of a diverse landscape (rolling and undulating hills) and forest that these areas have not been sufficiently defined from the few sites that were taken in the area in terms of our clustering procedure. Because we associate instances that have a high belongingness to the extragrade class as having a high prediction uncertainty, such areas could become the focus of future targeted sampling projects in order to generate new knowledge. In other instances, farm reservoirs (which if true should be eliminated from future analyses); as seen by symmetric shapes predominantly in the west of the study area also have a high extragrade membership.

Fig. 8a–c shows the variability of OC across the study area at 3 selected model depth increments of 0–10 cm, 30–40 cm and 80–100 cm. At each depth increment there is a lower prediction limit map (Fig. 8a1–c1), the final predicted map (Fig. 8a2–c2) and an upper prediction limit map (Fig. 8a3–c3). Similarly in Fig. 9a–c, the spatial variability of AWC at the same depth increments as a lower (Fig. 9a1–c1), final (Fig. 9a2–c2), and upper prediction (Fig. 9a3–c3).

Based on the lower and upper prediction limits at 0–10 cm across the study area, the average concentration of OC was predicted to range from 4 to 35 kg m^3. The average predicted OC concentration at this depth was 18 kg m^3. At 30–40 cm the predicted average OC was 10 kg m^3. We are 95% confident that the true average of OC at this depth is between 3 and 19 kg m^3.

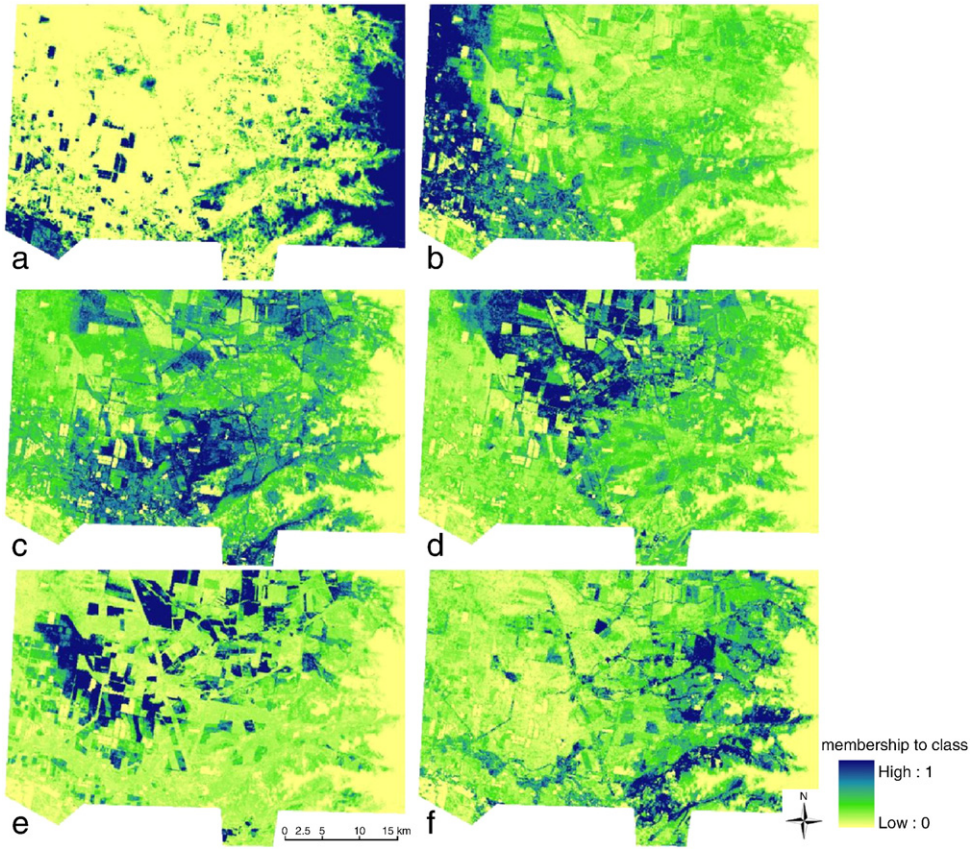

Fig. 7. Spatial variation of the degree of membership each instance has to each cluster including the extragrade class. Extragrade (a), cluster A (b), cluster B (c), cluster C (d), cluster D (f), and cluster E (g).

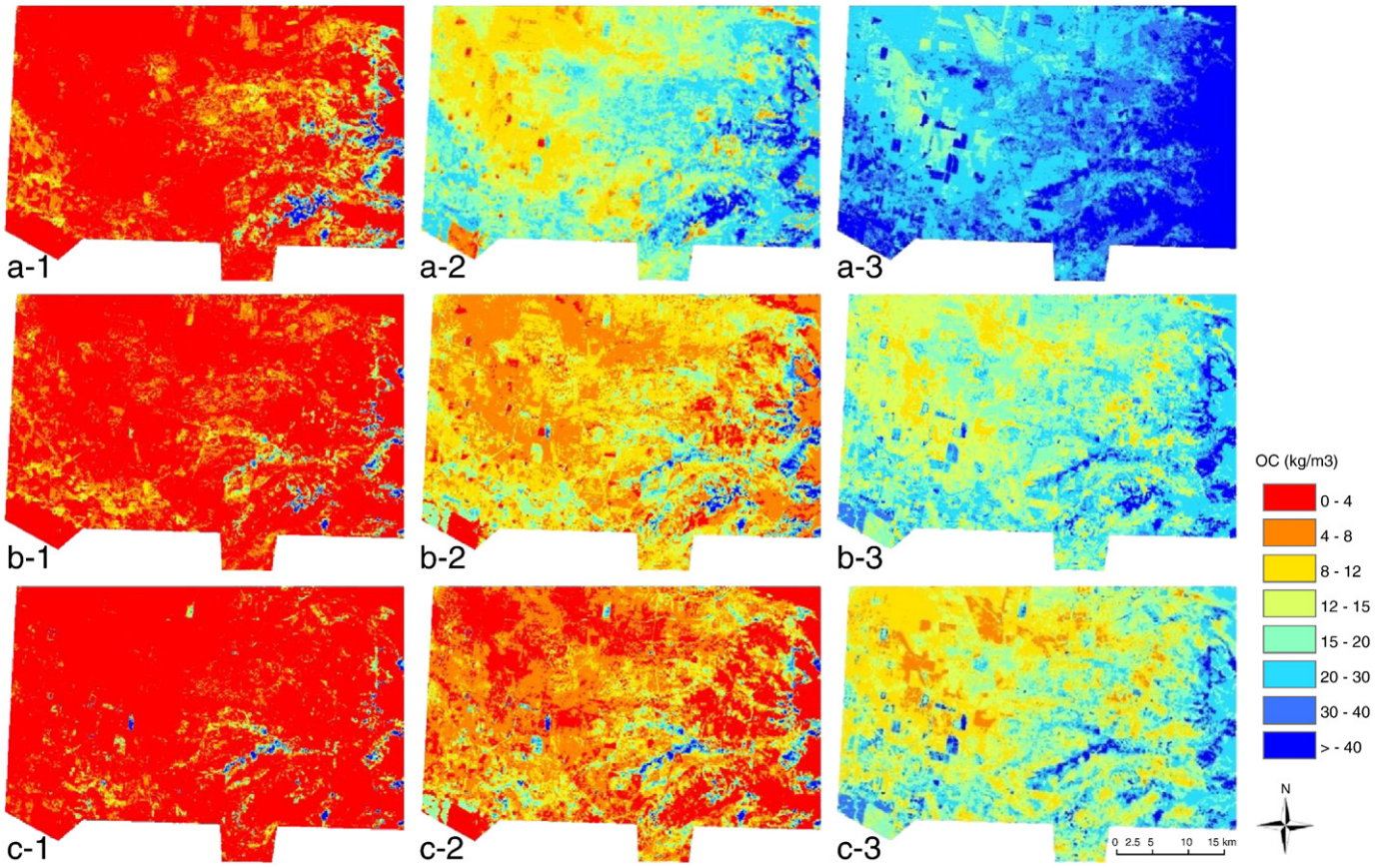

Fig. 8. Variability of OC at 0–10 cm, 30–40 cm and 80–100 cm across the Edgeroi study area. Lower prediction limit (1), DSM final prediction (2), and upper prediction limit (3).

While at the 80–100 cm depth increment, the predicted average of OC was 7 kg m^3, and we are 95% confident that the true average is between 2 and 16 kg m^3.

For AWC at 0–10 cm the average was predicted to be 0.16 m/m. With 95% confidence the true average is expected to be between 0.11 and 0.21 m/m at this depth. At the depth increments of 30–40 cm and 80–100 cm the average AWC was predicted to be 0.13 (0.08–0.16) m/m and 0.11 (0.08–0.14) m/m respectively.

Total AWC and OC maps to 1 m are shown in Fig. 10a and b respectively. Based on these maps we predicted that the average total water (m^2) to 1 m is 127 (88–158) mm. The total OC estimated across the extent of the study area is predicted to be 191 (50–385) Gg.

5. Conclusions

In this paper we have established a methodological framework for mapping uncertainties in the form of a PI for predicted soil attributes as they vary continuously with depth and space for a defined study area. This methodology complements the continuous prediction of soil attributes in a vertical and lateral space using splines and DSM methods. The methodology for deriving PIs is independent of the prediction model, requiring only the model outputs and the measure of error associated with those predictions.

The best available quantitative measure of the deviation between the modelled output and the modelled real-world process is the residual or error. Therefore, the empirical uncertainty model explicitly accounts for all sources of uncertainty without the requirement to separate out the contribution of each error source to the overall uncertainty. We have demonstrated that this method performed well for both OC and AWC where for a given confidence level, a near matching proportion of validation observations were within the confines of the PI. While an encouraging result, we accept that this methodology represents a pragmatic approach to estimating uncertainties both spatially and laterally in a DSM framework. It is likely their estimation may be a lot more complex than that formulated in this study. Future research will obviously need to investigate the extent of this perceived complexity and the scope of future research would initially involve comparison of the empirical approach we have presented with other approaches such as Monte Carlo simulations to construct PIs. During such a comparative exercise, one would then also need to consider in addition to the assessing the accuracy of the estimated uncertainties, the time and costs for implementing such alternative approaches.

In the course of this work however, a number of issues were presented that need to be addressed in order for improvements to be made to our approach. First, it begins with an independent model prediction framework where prior to modelling, a multivariate analysis should be performed to determine what the most closely correlated environmental covariates are for each soil attribute. This will ideally address some of the issues regarding the performance of the prediction models. Having a stronger prediction will naturally transfer to a reduced error. Ultimately, the distribution of errors for each class will be narrower, resulting in PIs that display more precision than that which we have just presented. By having an independent modelling process also means the requirement of an independent LOCV and clustering process. This inevitably generates more work. However, while better prediction outcomes are expected, the empirical uncertainty method is neither computationally demanding nor difficult to implement.

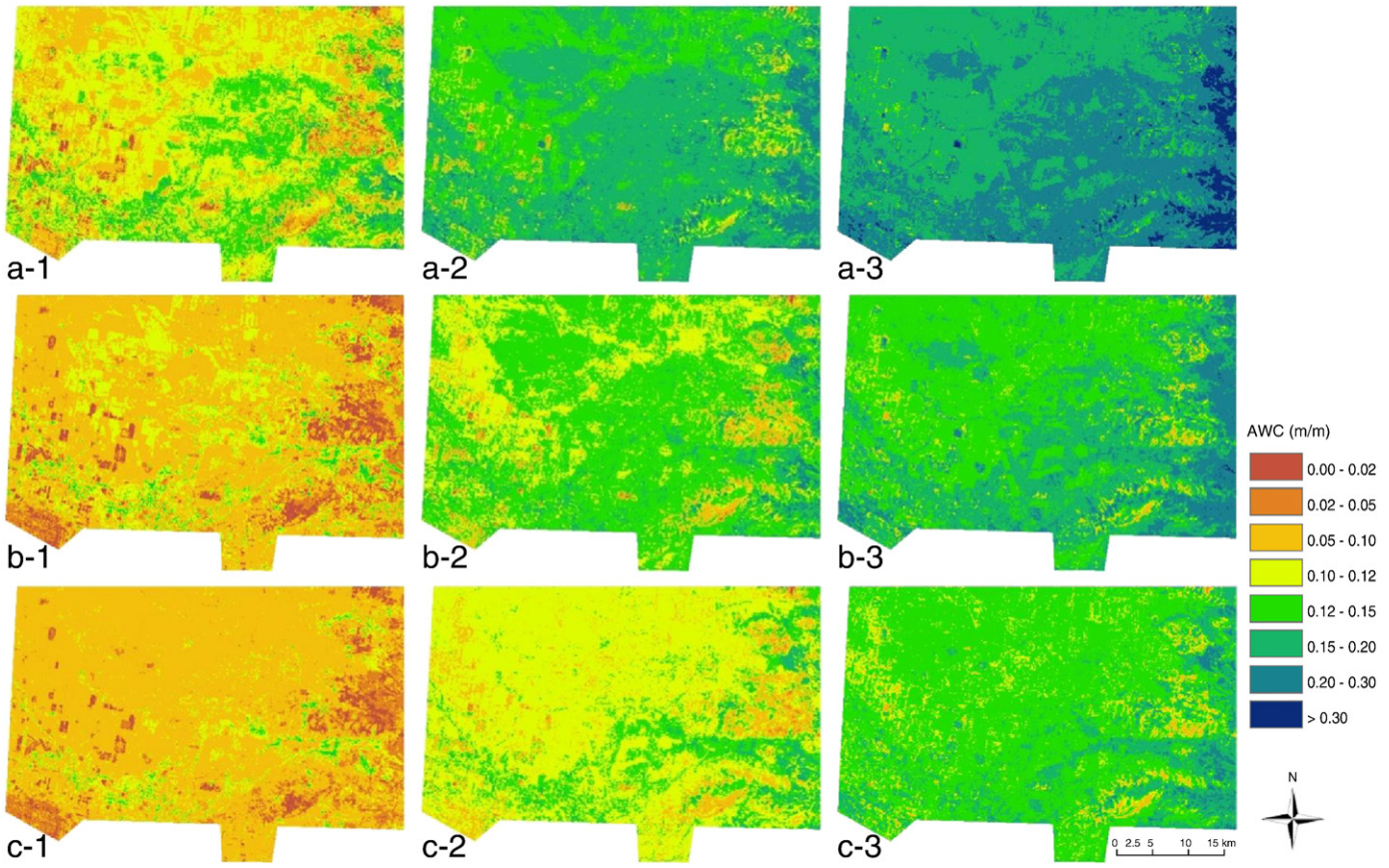

Fig. 9. Variability of AWC at 0–10 cm, 30–40 cm and 80–100 cm across the Edgeroi study area. Lower prediction limit (1), DSM final prediction (2), and upper prediction limit (3).

Beyond such pertinent modifications, future work could determine how well the methodology handles both different calibration sample sizes and types of soil attribute data. It would also be beneficial to test this methodology through field validation of the prediction and ultimately the PI. This would therefore require the implementation of a suitable sampling scheme. On the idea of sampling, it would be ideal to investigate whether sampling in areas that have a high membership to the extragrade class would facilitate narrowing the uncertainty where the uncertainty is believed to be greatest. Methods for determining the optimal class size will also need to be investigated.

Acknowledgement

The authors would like to thank Grant Tranter for assistance with some of the computer programming needed for this study.

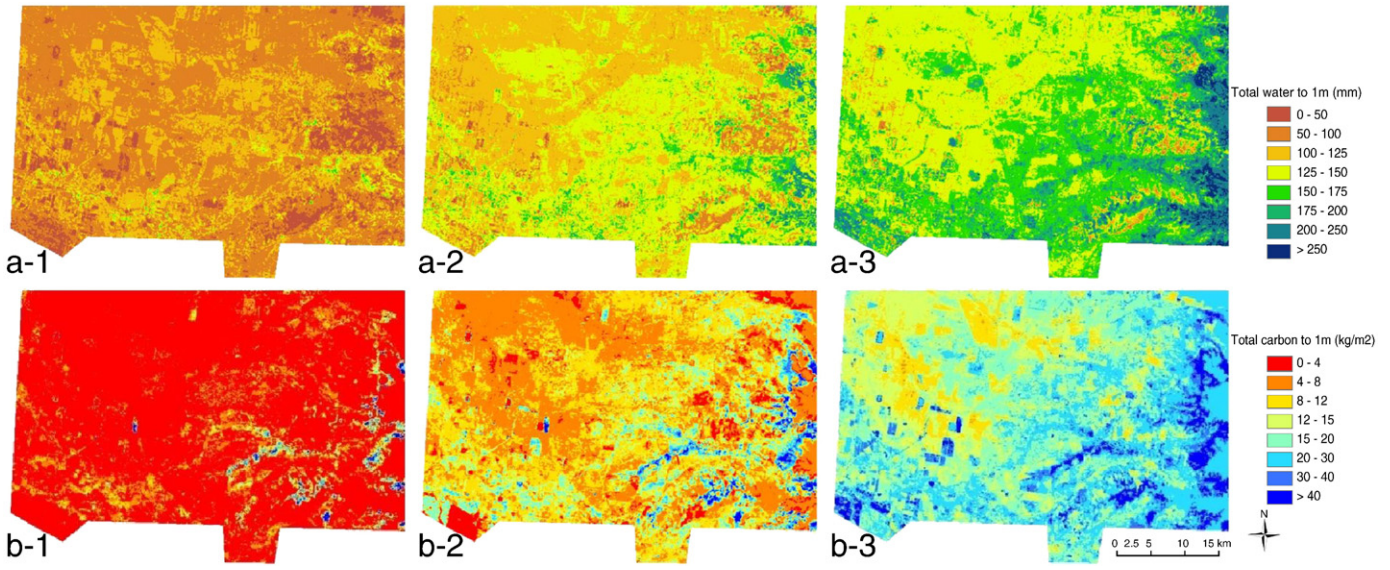

Fig. 10. Total water to 1 m (a) and total OC to 1 m (b) across the Edgeroi study area. Lower prediction limit (1), DSM final prediction (2), and upper prediction limit (3).

References

Altman, D.G., Gardner, M.J., 1988. Calculating confidence intervals for regression and correlation. British Medical Journal 296 (6631), 1238–1242.

Bezdek, J.C., 1981. Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York.

Bezdek, J.C., Ehrlich, R., Full, W., 1984. FCM — the fuzzy c-means clustering algorithm. Computers and Geosciences 10 (2–3), 191–203.

Bishop, T.F.A., McBratney, A.B., Laslett, G.M., 1999. Modelling soil attribute depth functions with equal-area quadratic smoothing splines. Geoderma 91 (1–2), 27–45.

Boehner, J., Koethe, R., Conrad, O., Gross, J., Ringeler, A., Selige, T., 2002. Soil regionalisation by means of terrain analysis and process paremeterisation. In: Micheli, E., Nachtergaele, F., Montanarella, L. (Eds.), Soil Classification 2001. European Soil Bureau, Research Report No. 7, EUR 20398 EN, Luxembourg, pp. 213–222.

Bragato, G., 2004. Fuzzy continuous classification and spatial interpolation in conventional soil survey for soil mapping of the lower Piave plain. Geoderma 118 (1–2), 1–16.

Brown, J.D., Heuvelink, G.B.M., 2005. Assessing uncertainty propagation through physically based models of soil water flow solute transport. In: Anderson, M. (Ed.), Encyclopedia of Hydrological Sciences. John Wiley and Sons, Chichester.

Gallant, J.C., Dowling, T.I., 2003. A multiresolution index of valley bottom flatness for mapping depositional areas. Water Resources Research 39 (12).

Geosciences Australia, 2008. Radiometric Data of the Narrabri, Moree, Inverell and Manilla 1:250 000 Topographic Map Sheets. Geophysical Archive Data Delivery System (GADDS) Website. date accessed 15/05/08 http://www.geoscience.gov.au/bin/mapserv36?map=/public/http/www/geoportal/gadds/gadds.

Grimm, R., Behrens, T., 2010. Uncertainty analysis of sample locations within digital soil mapping approaches. Geoderma 155 (3–4), 154–163.

Grinand, C., Arrouays, D., Laroche, B., Martin, M.P., 2008. Extrapolating regional soil landscapes from an existing soil map: sampling intensity, validation procedures, and integration of spatial context. Geoderma 143 (1–2), 180–190.

Grunwald, S., 2009. Multi-criteria characterization of recent digital soil mapping and modelling approaches. Geoderma 152 (3–4), 195–207.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer, New York, N.Y.

Jain, A.K., Murty, M.N., Flynn, P.J., 1999. Data clustering: a review. ACM Computing Surveys 31 (3), 264–323.

Kempen, B., Brus, D.J., Heuvelink, G.B.M., Stoorvogel, J.J., 2009. Updating the 1:50,000 Dutch soil map using legacy soil data: a multinomial logistic regression approach. Geoderma 151 (3–4), 311–326.

Lagacherie, P., Cazemier, D.R., vanGaans, P.F.M., Burrough, P.A., 1997. Fuzzy k-means clustering of fields in an elementary catchment and extrapolation to a larger area. Geoderma 77 (2–4), 197–216.

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

Malone, B.P., McBratney, A.B., Minasny, B., Laslett, G.M., 2009. Mapping continuous depth functions of soil carbon storage and available water capacity. Geoderma 154 (1–2), 138–152.

McBratney, A.B., 1992. On variation, uncertainty and informatics in environmental soil management. Australian Journal of Soil Research 30 (6), 913–935.

McBratney, A.B., de Gruijter, J.J., 1992. A continuum approach to soil classification by modified fuzzy k-means with extragrades. Journal of Soil Science 43 (1), 159–175.

McBratney, A.B., Moore, A.W., 1985. Application of fuzzy-sets to climatic classification. Agricultural and Forest Meteorology 35 (1–4), 165–185.

McBratney, A.B., Odeh, I.O.A., 1997. Application of fuzzy sets in soil science: fuzzy logic, fuzzy measurements and fuzzy decisions. Geoderma 77 (2–4), 85–113.

McBratney, A.B., Santos, M.L.M., Minasny, B., 2003. On digital soil mapping. Geoderma 117 (1–2), 3–52.

McGarry, D., Ward, W.T., McBratney, A.B., 1989. Soil studies in the Lower Namoi Valley: methods and data. The Edgeroi Dataset. CSIRO Division of Soils, Adelaide.

McQueen, J., 1967. Some Methods for Classifcation and Analysis of Multivariate Observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, Berkeley, CA, pp. 281–297.

Minasny, B., McBratney, A.B., 2002a. FuzME Version 3.0, Australian Centre for Precision Agriculture. The University of Sydney, Australia http://www.usyd.edu.au/agriculture/acpa/software/index.shtml.

Minasny, B., McBratney, A.B., 2002b. Uncertainty analysis for pedotransfer functions. European Journal of Soil Science 53 (3), 417–429.

Odeh, I.O.A., McBratney, A.B., Chittleborough, D.J., 1992. Soil pattern recognition with fuzzy-c-means: application to classification and soil-landform interrelationships. Soil Science Society of America Journal 56 (2), 506–516.

Saunders, A.M., Boettinger, J.L., 2007. Incorporating classifcation trees into a pedogenic understanding raster classifcation methodology, Green River Basin, Wyoming, USA. In: Lagacherie, P., McBratney, A.B., Voltz, M. (Eds.), Digital Soil Mapping: An Introductory Perspective. : Developments in Soil Science. Elsevier, Amsterdam, pp. 389–399.

Shrestha, D.L., Solomatine, D.P., 2006. Machine learning approaches for estimation of prediction interval for the model output. Neural Networks 19 (2), 225–235.

Solomatine, D.P., Shrestha, D.L., 2009. A novel method to estimate model uncertainty using machine learning techniques. Water Resources Research 45 Article Number: W00B11.

Tranter, G., Minasny, B., McBratney, A.B., 2010. Estimating pedotransfer function prediction limits using fuzzy k-means with extragrades. Soil Science Society of America Journal 74 (6), 1967–1975.

Ward, W.T., 1999. Soils and Landscapes Near Narrabri and Edgeroi, New South Wales, with Data Analysis Using Fuzzy k-means. CSIRO Division of Soils Divisional Report.

Webster, R., 2000. Is soil variation random? Geoderma 97 (3–4), 149–163.