Geoderma 232–234 (2014) 34–44

Contents lists available at ScienceDirect

Geoderma

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

Using model averaging to combine soil property rasters from legacy soil maps and from point data

Brendan P. Malone *, Budiman Minasny, Nathan P. Odgers, Alex B. McBratney

Department of Environmental Sciences, Faculty of Agriculture and Environment, C81 Biomedical Building, The University of Sydney, New South Wales 2006, Australia

ARTICLE INFO

Article history:
Received 4 December 2013
Received in revised form 17 April 2014
Accepted 25 April 2014
Available online 20 May 2014

Keywords:
Digital soil mapping
Ensemble models
Regional soil mapping
Legacy soil survey
Disaggregation
Scorpan

ABSTRACT

The objective of this study was to determine the efficacy of model averaging (ensemble modelling) as an approach for combining digital soil property maps derived from disaggregated legacy soil class maps and *scorpan* kriging (using soil point data). The study is based in the Dalrymple Shire, QLD and continues on the soil pH mapping work of Odgers et al. (2014a). Equal weights averaging (EW), Bates–Granger or variance weighted averaging (VW), Granger–Ramanathan averaging (GRA), and Bayesian model averaging (BMA) were compared in this study. Model averaged predictions were estimated to 2 m depth at regular depth intervals. 90% prediction intervals of the model averaged predictions were derived numerically. Neither the disaggregated soil map nor the *scorpan* kriging map was particularly accurate. Predictions from model averaging however did improve upon the accuracy, where at all depths, the combined predictions were an improvement on using either of the contributing soil maps alone. We recommend the use of GRA for digital soil mapping applications because its performance is equal to or better than the generally preferred BMA approach, yet far simpler to implement, and is computationally efficient. For regional soil studies where polygon mapping and soil point data are available, ensemble modelling is a useful combinatorial approach.

1. Introduction

Odgers et al. (in preparation) described a number of possible pedometric methods that could be implemented for creating digital soil property maps. One of these, and the focus of their investigations, was the generation of continuous soil property maps at regular depth intervals from disaggregated legacy soil class maps. In their work, they demonstrated the approach using the DSMART and PROPR algorithms to map soil pH (to 2 m depth) across the 68,000 km^2 area of the Dalrymple Shire QLD at a spatial resolution of 30 m.

This approach for digital soil property mapping is ideal when there are detailed soil maps with legends available, and when point data is scarce or even non-existent. But what do we do when there exist legacy soil maps and a reasonable coverage of soil point data? In terms of reasonable coverage, experience suggests that a density of between 1 and 10 observations per 1000 km^2 is required for making point predictions via a *scorpan* kriging digital soil mapping approach (GlobalSoilMap Science Committee, 2013). Subsequently, in the Dalrymple Shire in central Queensland, Australia, as investigated by Odgers et al. (in preparation), the sampling density is approximately 15 sites per 1000 km^2 — warranting a *scorpan* kriging digital soil mapping approach with the available points. With this possibility, there is the luxury of having two or more (if we use further yet different predictive approaches) realisations of the same target variable across the same study area, which could be considered as uncommon in some regions of the world. Naturally however; one will want to know which map is more accurate — the disaggregated conventional soil map or digital map derived from *scorpan* kriging. Indeed, enquiries of this nature have been investigated in other parts of the world by Bregt et al. (1987) and Kempen et al. (2012) as a few examples. Another question is: what if we combine both maps together, yielding a single new map? A single map is more useful than two or more independent realisations of the same target variable; and it is a tantalising prospect if the combined map is more accurate than each independent map alone (e.g. Heuvelink and Bierkens, 1992). As such, this study is concerned with investigating a number of different approaches for combining digital soil maps with the intention of yielding a single and more accurate digital soil map of soil pH across the aforementioned study area of the Dalrymple Shire, QLD.

The *scorpan* model allows incorporation of existing soil information as a covariate via the *s* (soil) factor (McBratney et al., 2003). Henderson et al. (2005) exemplified this by using existing legacy soil class mapping for predicting a number of soil properties across the Australian continent without kriging the residuals. Subsequently, using an existing soil map as a *scorpan* model input could be considered as one way to combine traditional soil map information with soil point data (Minsany and McBratney, 2010). Another is to treat the outputs from the

* Corresponding author.
E-mail addresses: brendan.malone@sydney.edu.au (B.P. Malone), budiman.minasny@sydney.edu.au (B. Minasny), nathan.odgers@sydney.edu.au (N.P. Odgers), alex.mcbratney@sydney.edu.au (A.B. McBratney).

http://dx.doi.org/10.1016/j.geoderma.2014.04.033
0016-7061/

disaggregated soil map and from *scorpan* kriging as outcomes or realisations of two different processes. They could be considered as an ensemble of outcomes (here, two) that one wishes to combine into a single outcome. This type of situation is common in atmospheric and hydrologic research fields (Wagener and Gupta, 2005) where multiple forecasts of a given process are derived from a number of competing predictive models. Each contributor model will have its own strengths and weaknesses. Rather than selecting the single best-performing model for a given situation or scenario (which is a traditional pursuit), combining model outcomes is a natural generalisation to this (Diks and Vrugt, 2010). Ideally, the new combined outcome is at least as good as any of the individual outcomes.

Combining different model outcomes is termed model ensemble or averaging (Rojas et al., 2008). Diks and Vrugt (2010) thoroughly described, applied, and compared to a number of different model averaging approaches with reference to point forecasting for hydrologic modelling applications. The fundamental basis of these approaches can be described with the following simple model:

$$Y_i = \sum_{k=1}^{Kk} W_k X_{ik} \quad (1)$$

where Y_i is the combined outcome at point i from K number of contributor models. X_{ik} is the realisation from the *kth* contributor model and W_k is the weighting attributed to that model. For most model averaging approaches, the weights from all the competing contributor outcomes sum to one. Given this, the variation between the different model averaging methods comes down to how W_k is estimated. The simplest option is to presume equal weighting across the different contributor outcomes. This is generally undesirable because it does not take into account the relative accuracy of the contributor outcomes. A better choice in that regard is that proposed by Bates and Granger (1969) which is to weight each contributor outcome by its associated variance. Predictions that have a higher prediction variance are given a lower weight than those with a lower prediction variance. It was this approach that was used in the study by Heuvelink and Bierkens (1992) for combining soil map predictions (from a legacy soil polygon map) with interpolated point predictions. Other model averaging approaches include information criterion averaging (Buckland et al., 1997), which is a relatively straightforward approach compared to the more sophisticated and computationally demanding Bayesian (Hoeting et al., 1999) and Mallows model averaging (Hjort and Claeskens, 2003) methods. Interestingly, a far simpler, but equally, if not better performing model averaging approach, used in Diks and Vrugt (2010) is Granger–Ramanathan averaging (Granger and Ramanathan, 1984). In this approach, the constraint of ensuring that the weights add to unity is relaxed, and to accommodate this, a constant term is added. The Granger and Ramanathan model averaging constant and weighting parameters are solved using ordinary-least-squares (OLS) regression, where the predictor variables are the different realisations from each competing model, and the target variable is the associated actual observations. This approach exploits the covariance structure that may be present between the errors of the competing model outcomes i.e. using OLS estimators within the linear regression model (Diks and Vrugt, 2010). Granger and Ramanathan model averaging is advantageous because the OLS regression optimises the fitting of the parameters to ensure the error between predictions and observations is minimised, which also results in an unbiased combined prediction; even if the contributing model outcomes are biased (Granger and Ramanathan, 1984).

For digital soil mapping, any savings in time and computational load is an advantage given the large mapping extent and/or high resolution at which some digital soil maps are produced, as in this study where soil pH is to be mapped across the 68,000 km^2 study area at a 30 m grid cell resolution. It is therefore worth comparing several model averaging methods in order to assess those that are most attractive for digital soil mapping with regard to computational efficiency and ability to return a more accurate map.

In this study, our aim is to investigate and compare some of the aforementioned model averaging methods. More specifically, we will investigate the use of:

i) Equal weights averaging,
ii) Bates–Granger averaging (variance weighted averaging),
iii) Bayesian model averaging (with finite mixture model),
iv) and Granger–Ramanathan averaging.

The workflow of this investigation is as follows. The data that we will be using is first introduced. The steps for producing digital soil property maps using *scorpan* kriging will be detailed, along with the approach for quantifying the associated prediction uncertainties. We then discuss the characteristics of each of the model averaging techniques used and how they are implemented for combining the soil maps from *scorpan* kriging with those produced in Odgers et al. (in preparation). We then detail a numerical approach for estimating the associated uncertainties of the combined soil pH map. Here uncertainty for both *scorpan* kriging and model averaging is mapped continuously across the study area and is expressed as a 90% prediction interval. The results and broader discussion of this work are then presented.

2. Materials and methods

2.1. Study area

The study area comprises most of the former Dalrymple Shire in central Queensland, Australia (Fig. 1). It has an area of about 68,000 km^2 and is approximately 1000 km north of Brisbane (capital city of Queensland, Australia). The area comprises a large part of the northern Burdekin River catchment and is bounded on the east by the Seaview and Leichhardt Ranges, the Great Dividing Range in the west, and the Suttor and Belyando Rivers in the south-east. Most of the area is flat to gently undulating and elevation generally decreases towards the south-east. It is drained by the Burdekin River and its tributaries (Rogers et al., 1999). The Dalrymple Shire lies within the seasonally wet–dry tropics and has a warm, subhumid climate with a distinct hot-wet summer and a warm-dry winter. Average annual rainfall ranges from approximately 500 mm in the south-west of the area, to 1600 mm in the north-east (Rogers et al., 1999). The geology of the Dalrymple Shire is varied and complex, resulting in the development of a large number of soil types. A comprehensive description of the geology can be found in a series of 1:250,000 maps and explanatory notes by the Geological Survey of Queensland (Olgers, 1970). Geological landscapes in the area include Alluvial, Basalt, Cainozoic (includes Tertiary landscapes), Granodiorite, Igneous (other than Granodiorite and Basalt), Metamorphic, and Sedimentary.

2.2. The data

In this particular study we used a soil dataset containing 1080 soil profile observations of which we were specifically interested in the observed measurements of soil pH (1:5 soil–water solution). Odgers et al. (in preparation) had previously fitted mass-preserving depth splines to these soil profile data to harmonise pseudo-measurements of soil pH for the standardised depth intervals of 0–5 cm, 5–15 cm, 15–30 cm, 30–60 cm, 60–100 cm, and 100–200 cm. After removal of some spurious soil profiles (missing spatial coordinates, depth observation, etc.), 1048 remained. 300 of these (the same that were used in Odgers et al. (in preparation)) were kept aside for the purpose of externally validating predictions derived from *scorpan* kriging and those from model averaging. The map in Fig. 1 illustrates the locations of data used for fitting *scorpan* kriging models (calibration data) and for validation in this study.

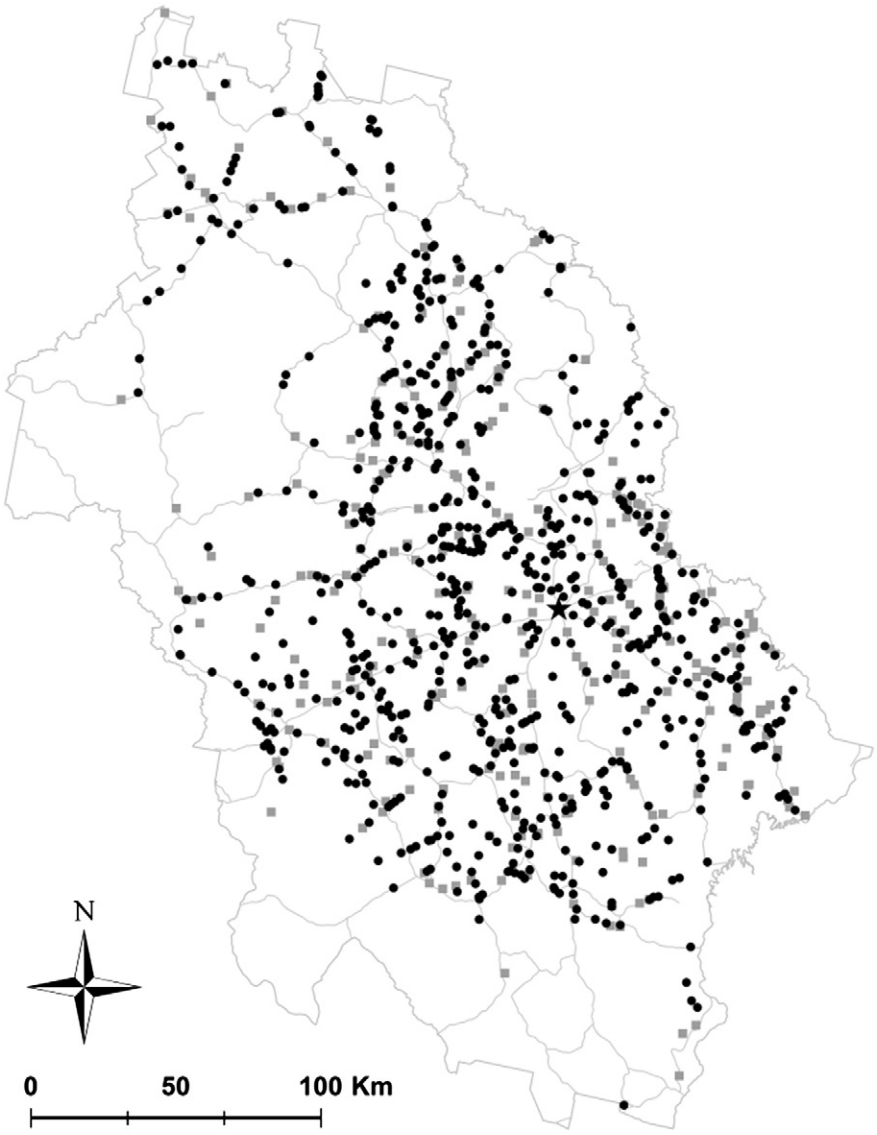

Fig. 1. The Dalrymple Shire, QLD, with road network and soil point data. Rounded points are soil data used for calibration of *scorpan* kriging models. Square points are the 300 validation point data. Star point is the township of Charters Towers [20.1° S, 146.3° E].

A number of environmental covariates were collated in order to serve as covariates in the *scorpan* kriging process. These covariates were also used in Odgers et al. (in preparation) and are predominantly derived from a digital elevation model, an air-borne gamma radiometric survey and from the Landsat 5 Thematic Mapper instrument. Wilford's (2012) weathering intensity index, derived from the interaction of topography and gamma radiometrics, was also used for this study. This weathering index provides a quantitative estimate of the degree to which the regolith is weathered. The covariates used in this study are listed in Table 1. All the topographic covariates have a 30 m grid resolution and are co-registered to the same raster grid used by Odgers et al. (in preparation). The radiometric data are provided as raster grids at 100 m grid resolution. Fine-gridding was used to coerce this information to the 30 m grid using the B-Spline interpolation algorithm from SAGA GIS. In this study the *scorpan* kriging and model averaging output was predicted onto the same 30 m raster grid.

2.3. Scorpan kriging: the prediction model

The *scorpan* kriging prediction is the sum of a deterministic component and a stochastic component (Odeh et al., 1995). The deterministic component requires fitting a predictive model between known values of the target variable (soil pH) and the values of the environmental

Table 1
Environmental covariate data that were under consideration for *scorpan* kriging models.

Covariate data source	Attribute
Digital elevation model	Elevation (E), hillshading (HS), mid-slope position (MSP), multi-resolution ridge top flatness (MRRTF), multi-resolution valley bottom flatness (MRVBF), terrain wetness index (TWI), slope gradient (S), slope height (SH), incoming solar radiation (SR), standardised height (H), valley depth (VD)
Air-borne gamma radiometrics	Potassium (K), thorium (TH), uranium (U), thorium–potassium ratio (TKr), thorium–uranium ratio (TUr), uranium–potassium ratio (UKr), weathering index (WI)
Landsat 5	Normalised difference vegetation index (NDVI)

covariates at the known points. The stochastic component requires interpolation of the residuals from the deterministic model.

In our case we chose to use the Cubist model as the deterministic component of the *scorpan* kriging procedure. The Cubist model is a data mining algorithm which allows one to explore non-linear relationships in observed data. It is similar to a typical regression tree model in terms of it being a data partitioning algorithm. The Cubist model is based on the M5 algorithm of Quinlan (1992). The Cubist model recursively partitions the data into subsets which are more internally homogeneous with respect to the target variable and covariates than the dataset as a whole. A series of rules defines the partitions, and these rules are arranged in a hierarchy. Each rule takes the form:

if [condition is true]

then [regress]
else [apply next rule].

Each condition is based on a threshold for one or more covariates. For example, if elevation is greater than 300 and radiometric potassium is greater than 1% is one such situation of thresholding where more than one covariate is used. If the condition returns true then the next step is the prediction of the target variable by OLS regression on the covariates within that partition. If the condition returns false, then the rule identifies the next node in the tree to move to, and the sequence of if-then-else is repeated. The result is that a separate regression equation is fit within each partition and the errors are smaller than they would be if a single regression was fit to the entire dataset (Quinlan, 1992).

The kriging component of the *scorpan* kriging model involved modelling the spatial structure of the Cubist model residuals (difference between observations and associated modelled predictions). For each depth increment, the spatial structure of the residuals was modelled with a global spherical variogram, the parameters of which were used to estimate the residuals via kriging across the extent of the study area.

Mapping soil pH at each depth increment involved recalling the associated fitted Cubist model and spatial residual model particular to that depth and, applying them together, with the two independent outputs being summed to generate final regression kriging predictions. Independent validation of the regression kriging models were evaluated by applying them to the 300 withheld site data points. Validation criteria used in this study were the co-efficient of determination (R^2) and the root mean square error of prediction (RMSE). These criteria are popularly used in digital soil mapping for assessing the agreement between observations and corresponding model predictions.

2.4. Scorpan kriging: estimates of uncertainty

Quantification of uncertainties expected from the *scorpan* kriging model (at each depth interval), was expressed as 90% prediction intervals. In this study we used a modified version of the method from Malone et al. (2011b), where prediction uncertainty is estimated empirically on the basis of the underlying regression kriging residuals. The modification of the approach regards how the geographical space is partitioned. Because the Cubist model divides the input data into a series of rule sets, it is probably more appropriate to examine the empirical distribution of regression kriging residuals within each of these rule sets, rather than perform an unsupervised classification of the covariate data space as demonstrated in Malone et al. (2011b). With this slight modification the underlying assumption is that, not only do the contributing environmental covariates that parameterise a given rule determine the spatial distribution of a given target soil property, but also they determine the magnitude of the prediction uncertainties too.

For each depth interval, we examined the rule sets that contributed to each model. We were mindful of the fact that to get a meaningful distribution of model residuals in each rule, intuitively, it was determined that 30 or more observations would be needed contribute to that particular rule. For this study, we did not encounter this issue, but if we had, we could have managed it by limiting the number of rules that could be realised by the given Cubist model. For each rule, the contributing n number of observations was subsetted for performing leave-one-out cross validation analysis (LOCV). Using only the covariates that parameterised the regression model in the rule, LOCV was performed. LOCV entailed, for each iteration, regression kriging with $n - 1$ observations using a single-rule Cubist model (i.e. data was not partitioned into smaller subsets and essentially equates to a multiple linear regression model) for modelling using the selected covariates, followed by fitting a global variogram model to the associated cubist model residuals. This regression kriging model was then applied to the 'left out' observation, from which a residual was estimated — the deviation between observed valued and regression kriging prediction (summation of cubist prediction and interpolated residual). At the end of each LOCV, for each rule set, the empirical distribution of residuals were formed. For forming 90% prediction intervals to each rule, the lower 5% and upper 95% percentile values of the empirical distributions were taken.

To map estimates of uncertainty across the entire study area, we determined which rule was applied at every cell of the prediction grid. We then added the corresponding upper (95th percentile) and lower (5th percentile) values for that rule to the soil pH predictions that were derived from the *scorpan* kriging. This generated two maps for each depth increment, with the first indicating the lower prediction limit; the second, the upper prediction limit.

Validation of the quantifications of uncertainty involved estimating the prediction interval coverage probability (PICP). From Malone et al. (2011b), the PICP is the probability that all observed values fit within their estimated prediction interval. This probability was estimated for the 300 validation points, and because a 90% prediction interval has been defined for each observation, we should expect 90% of all the observations to fit within their given prediction limits. The uncertainty model is said to be optimal when this occurs.

2.5. Soil property predictions from disaggregated soil maps

Odgers et al. (2014) used the DSMART algorithm to disaggregate the map units of a 1:250,000-scale legacy soil polygon map covering the study area. The result of the spatial disaggregation was a map of the estimated probability of occurrence for each of the legacy map's 72 soil classes. Given reference soil property information for the Dalrymple Shire soil classes, Odgers et al. (in preparation) introduced the PROPR algorithm in which they used the probability rasters (all 72 of them) from DSMART as weights to calculate the weighted mean of soil pH across the study area at the same standardised depth increments used here. The probability rasters were also used in a procedure to estimate the 90% prediction interval for the weighted mean pH.

2.6. Model averaging

With the digital soil maps from *scorpan* kriging and those derived from Odgers et al. (in preparation), we effectively have an ensemble of model realisations. We will refer to these as source maps, where source map 1 is the one derived from Odgers et al. (in preparation), while source map 2 is that derived from *scorpan* kriging. The purpose of this work is to try to combine both source maps using model averaging methods. The idea behind model averaging is that we can combine predictions from two or more methods by enhancing the strength and reducing the weakness of each source map. The model averaging approaches under consideration in the study are: equal weights (EW), Bates–Granger or variance weighted (VW), Bayesian model averaging (with finite mixture model; BMA), and Granger–Ramanathan averaging (GRA). The performance of each of these approaches is assessed using the withheld 300 observation points. As for *scorpan* kriging, the RMSE and R^2 statistics are the quantitative indices for performance evaluation of each model averaging method. In the case of BMA and GRA

(described below), the 300 points are actually used to define the weights attributed to each different model realisation. We do not map the outputs from all the model averaging methods; rather we do so only for the one that performs the best overall for all depth intervals.

The general model for model averaging is:

$$Y_i = \sum_{k=1}^{Kk} W_k X_{ik} \quad (2)$$

where Y_i is the combined outcome at point i from K number of contributor models. X_{ik} is the realisation from the kth contributor model and W_k is the weighting attributed to that model. In this study $K = 2$. The problem is finding the optimal weight parameter W. Now following is a brief description of each approach used in this study. Consult Diks and Vrugt (2010) for detailed information on their theoretical underpinnings.

2.7. Model averaging: equal weights

Under EW, the combined map is simply obtained by giving each source map equal weight. In this case $W = 0.5$ for both maps.

2.8. Model averaging: Bates–Granger or variance weighted

Under VW, each source map is weighted by $\frac{1}{\sigma_i^2}$, where σ_i^2 is the attributed variance associated with a prediction i. From the 90% prediction intervals of both source maps, and assuming a normal distribution, we may approximate the prediction variance as:

$$\sigma_i^2 = \left(\frac{\mathrm{UPL}_i - \mathrm{LPL}_i}{2 \times z}\right)^2. \quad (3)$$

Here UPL_i is the upper 90% prediction limit at point i. Similarly LPL_i is the lower prediction limit; and z is the z-value for a given confidence interval, which in our case we attribute a 90% confidence to our predictions, meaning that z equals approximately 1.64. If we want to predict W_i for the *scorpan* kriging map ($W_{i\ (SK)}$) predictions, it is estimated as:

$$W_{i\ (SK)} = \frac{\frac{1}{\sigma_{i\ (SK)}^2}}{\frac{1}{\sigma_{i\ (SK)}^2} + \frac{1}{\sigma_{i\ (DS)}^2}} \quad (4)$$

where $\sigma_{i\ (SK)}^2$ and $\sigma_{i\ (DS)}^2$ are the predicted variances from both *scorpan* kriging (*SK*) and disaggregated (Odgers et al., in preparation) (*DS*) source maps at point i. Naturally $W_{i\ (DS)}$ will equal $1 - W_{i\ (SK)}$. Eq. (4) assumes that the variances from each source map are uncorrelated, and this is a reasonable assumption given that both maps were produced from two very different approaches.

2.9. Model averaging: Bayesian model (in the finite mixture model)

BMA is similar to the general model averaging method (Eq. (2)); however, rather than having a single estimate of the weighting factors, each source map prediction is associated with a conditional probability density function (PDF). It is considered as a combined forecast density (Diks and Vrugt, 2010):

$$g(y) = \sum_{i=1}^{k} W_i f_i(y). \quad (5)$$

The finite mixture model assumes that the weights W are non-negative and sum to unity. They are estimated conditionally based on the observed values and the corresponding predictions from both source maps. The aim is to estimate the weight parameter density W and its variance s^2. This is achieved using an enhanced Markov Chain Monte Carlo simulation method called DREAM (DiffeRential Evolution Adaptive Metropolis) developed by Vrugt et al. (2009). For more detail on the theory and algorithm of BMA, we refer the reader to Vrugt et al. (2008) and Diks and Vrugt (2010).

Effectively for this BMA approach, conditionally based on the 300 validation observations and their corresponding predictions from both source maps, the DREAM algorithm was set to generate 10,000 estimates of the W parameters. Of these, 8000 were discarded as burn-in samples. The final 2000 samples were used to generate the distribution of W for each source map. For each realisation, and associated W parameter set, the general model averaging formula was applied. Calculation of the RMSE and R^2 statistics was performed for each realisation.

2.10. Model averaging: Granger–Ramanathan

Granger and Ramanathan (1984) proposed that the problem of combining model outcomes could be approached by using traditional OLS methods. Whatever covariance structure that may be present in the prediction errors, OLS estimation is able to exploit this, which for our purposes, is to derive optimal W estimates for each source map. Essentially GRA involves fitting a multiple linear regression model where observed values are regressed against the corresponding predictions derived from the different source maps. Such that with the 300 validation points, we fit the model:

$$Y = W_0 + (W_{SK} \cdot X_{SK}) + (W_{DS} \cdot X_{DS}). \quad (6)$$

Here Y is the vector of observed values (soil pH) and X_{SK} and X_{DS} are their corresponding predictions from both source maps. OLS is used to solve for the parameters: W_0, W_{SK}, and W_{DS}. Here the weights W_{SK} and W_{DS} do not necessarily have to sum to one. W_0, the intercept term, is an 'in-built' bias correction term between the observed values and the individual model source map predictions. Once this model is fitted, it is used as a global model to derive the combined digital soil map.

2.11. Model averaging: uncertainty estimation

We use a numerical approach to quantify estimates of uncertainty associated with model averaging. The approach is easiest explained by considering Fig. 2. Here the plot represents what would be expected when we compare the predictions and associated uncertainties of both source maps at a single location or grid cell. We assume the predictions represent the mean, and we can estimate the variance (and standard deviation) from the prediction intervals as described before. Therefore, with these parameters, we can derive the full PDF at this

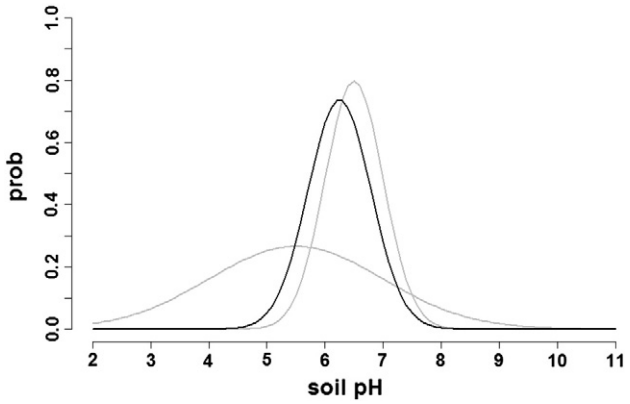

Fig. 2. A hypothetical situation for combining two PDFs, in this case for soil pH. Grey lines indicate PDFs at the same location from two source maps. Source map 1 (mean = 5.5, standard deviation 1.5); source map 2 (mean = 6.5, standard deviation = 0.5). The black line is the result of combining or mixing both contributing PDFs together. The mean is somewhere between both contributing means while the spread of the distribution is a compromise between the two contributing PDFs.

point for both source maps. In this example, a PDF for a single point from source map 1 has a mean of 5.5 and standard deviation of 1.5. The PDF associated with soil pH from source map 2 at the same point has a mean of 6.5 and standard deviation 0.5. To estimate the uncertainty associated with the combined prediction (which is represented by the black line on the plot), we effectively need to combine or mix together the two source map PDFs. We do this numerically such that each distribution is randomly sampled x number of times. On each sample x, the model average is calculated — in this case we used the variance weighted approach for this example ($W = 0.25$ for source map 1 and $W = 0.75$ for source map 2). This will ultimately generate a new PDF, albeit mixed, but where the mean is somewhere between the predictions of both source maps, and the PDF is a compromise between the two contributing PDFs. A condition to this approach is that for each of the model average methods, the weighting remains fixed i.e. the weightings are not re-computed for each sampling iteration. In the case of BMA where there is also full distribution of W parameters to sample from, things become quite complicated in terms of time and required computations. Here not only do we need to sample the PDFs at each point, we also have to apply these samples to each realisation (2000) of the BMA W parameters.

In order to ensure that we adequately sample the full PDFs at each point, we could run the random sampling for many thousands of iterations. However, from a computational perspective this is not optimal because the process has to be repeated at every point or grid cell on the digital soil map (for the study area there are $\approx 1.3 \times 10^8$ raster cells to process). A more efficient alternative to simple random sampling is Latin hypercube sampling (LHS) (McKay et al., 1979; Pebesma and Heuvelink, 1999). LHS is a stratified random sampling technique that ensures full coverage of the range of each variable to be sampled by maximally stratifying the marginal distribution. A sample is maximally stratified when the number of strata equals the sample size n. For independent variables, the cumulative distribution for each variable is divided into n number of equi-probable intervals. A value is then selected randomly from each interval. The n values obtained from each distribution are then matched randomly with those of the other distribution. In this study we set the sample size to 50 after we found that based on the validation data, there was little difference in outcomes if the PDFs were sampled 50, 100, or 1000 times. Samples sizes of less than 50, generally resulted in some fluctuating results.

After the LHS was taken and the model average was calculated, the 90% prediction interval was taken by retrieving the 95th and 5th percentiles of the resulting distribution. This was determined for all model averaging techniques using the validation data. Mapping the uncertainties for the combined digital soil map of soil pH was only carried out for the model averaging method that performed best overall, in consideration of the outcomes of each soil depth interval. Validation of the resulting prediction uncertainties involved estimation of the PICP as described in the methods for the *scorpan* kriging. The difference between BMA and the other model averaging methods is that the PICP was estimated for each realisation of the BMA parameters.

2.12. Implementation of methods

For the most part, the R statistical open-source software (R Core Team 2013) was used for running the statistics, modelling, and mapping procedures in this study. Besides the base R functionality, the R packages used in this study included "Cubist" (Kuhn et al., 2013) for fitting cubist models; "gstat" (Pebesma, 2004) for variogram fitting; and "raster" (Hijmans and van Etten, 2013) for handling raster layers and

Table 2
Cubist rules and variogram parameters of the cubist models for each depth. The number in square braces in the rule conditions is the number of contributing observations making up that rule. Spherical models were used for estimating the nugget, sill and distance parameters of the residual variograms.

Depth	Cubist model	Residual variogram model (global)
0–-5 cm	*Rule 1 [454]: where E > 286* pH = 6.8 + 0.0053VD − 0.4MSP − 1.3NDVI + 0.063WI − 0.00064E + 0.085 K − 0.07U *Rule 2 [239]: where E <= 286* pH = 7.3 − 0.012TKr − 1.3NDVI	Nugget: 0.24 Sill: 0.45 Distance: 5600 m Nugget-to-Sill ratio: 0.53
5–15 cm	*Rule 1 [456]: where E > 286* pH = 6.7 + 0.0061VD + 0.191 K + 0.096WI − 0.42MSP − 0.00093E − 0.017TH − 1NDVI *Rule 2 [239]: where E <= 286* pH = 7.1 − 0.0096TKr − 0.00022E − 0.02U − 0.1NDVI	Nugget: 0.32 Sill: 0.47 Distance: 7070 m Nugget-to-Sill ratio: 0.68
15–30 cm	*Rule 1 [141]: where E > 300 and UKr > 2.09* pH = 6.9 − 21.3S − 2.4NDVI + 0.00156E − 0.49MSP + 0.0003VD *Rule 2 [289]: where E > 300 and UKr <= 2.09* pH = 7.5 − 0.00261E + 0.0056VD + 0.0013SH − 0.13U *Rule 3 [285]: where E <= 300* pH = 7.2 − 0.0146TKr + 0.128WI − 1.7NDVI	Nugget: 0.35 Sill: 0.56 Distance: 726 m Nugget-to-Sill ratio: 0.63
30–60 cm	*Rule 1 [447]: where E > 300* pH = 8.3 − 2.3NDVI − 0.0011E − 0.081WI + 0.0028VD − 0.09 K − 0.06U *Rule 2 [286]: where E <= 300* pH = 7.7 − 0.0166TKr + 0.184WI − 2.2NDVI	Nugget: 0.38 Sill: 0.81 Distance: 1290 m Nugget-to-Sill ratio: 0.46
60–100 cm	*Rule 1 [48]: where E > 482* pH = 10.5 − 3.9NDVI − 0.152WI − 0.0017E − 0.13 K − 0.15U − 0.012SH − 0.042MRRTF *Rule 2 [148]: where E <= 482 and UKr > 2.21* pH = 7.2 + 1.29 K + 0.372WI − 0.0042E − 0.04SH − 2.9NDVI − 0.005TWI *Rule 3 [365]: where E <= 482 and UKr <= 2.21* pH = −2.5 − 0.0049E − 0.089TWI + 0.0061VD − 2.4NDVI − 0.23 K + 1.9HS − 0.073MRRTF + 0.072SR − 0.013WI	Nugget: 1.09 Sill: 1.09 Distance: 5000 m Nugget-to-Sill ratio: 1.00
100–200 cm	*Rule 1 [269]:* pH = 11.4 − 0.354WI − 4.2NDVI − 0.074TH − 0.028SH	Nugget: 0.47 Sill: 1.31 Distance: 6480 m Nugget-to-Sill ratio: 0.36

generating soil map predictions. The DREAM algorithm implemented for BMA (Vrugt et al., 2009) was run using script developed for Matlab (Mathworks 2012). All soil maps were produced in ArcGIS version 10 (ESRI 2012).

3. Results and discussion

3.1. Scorpan kriging and uncertainty estimation

A summary of the regression kriging models for each depth interval is presented in Table 2. For the Cubist modelling, two rules were defined for the 0–5 cm, 5–15 cm, and 30–60 cm depth intervals. Three rules were defined for 15–30 cm and 60–100 cm. One rule only was defined for the 100–200 cm increment where there was only 269 observations with which to calibrate the model. At the depths where there are two or more rules, the root of each rule is a conditional statement which directs the prediction to one of two regression equations depending on a threshold for the covariate identified in the conditional statement. Elevation (E) was used in all the conditional statements. At the 0–5 cm and 5–15 cm depth increments, the elevation threshold was 286 m; for the 15–30 cm and 30–60 cm depth increments the elevation threshold was 300 m; for the 60–100 cm depth increment the elevation threshold was 482 m. The UKr was a further conditional variable for the rules at the 15–30 cm and 60–100 cm depth intervals. Therefore, for this study area, and from the soil data collected from it, E and sometimes together with UKr (15–30 cm and 60–100 cm) capture the large scale pedogenic processes useful for mapping soil pH. E corresponds to one type of topographical characteristic of the *scorpan* digital soil mapping model (McBratney et al., 2003); while UKr is a quantitative proxy for parent materials, and the weathering processes they have undergone (Wilford, 2012).

Fig. 3 shows an example of the geographical nature to which cubist rules are applied when we extrapolate the rule thresholding criteria of a given model, in this case for the 15–30 cm depth increment where there were three rules defined. Here rule 1 indicates areas where the elevation is greater than 300 m and the radiometric uranium–potassium ratio is greater than 2.09; while rule 2 delineates areas where also the elevation is greater than 300 m but the radiometric uranium–potassium ratio is less than 2.09; and rule 3 simply delineates the area where the elevation is less than 300 m, which roughly corresponds to the basin area of the catchment.

Examining the environmental covariates used in the linear models of each rule, we may consider them as variables that capture the local variations of soil pH across the different landscapes and different depths. This is excepting of the 100–200 cm depth increment where one rule was defined, to which the covariates used — WI, NDVI, TH and SH — would be considered universal variables i.e. used for the entire study area. In all, the Cubist rule set contained 13 individual linear models. When we examine the covariates that contributed to these linear models, the most highly used in order of frequency were (see full names in Table 1): NDVI (13), WI (9), E (9), K (6), VD (6), U (5), TKr (4), SH (4), MSP (3), MRRTF (2), TWI (2), TH (2), S (1), HS (1), and SR

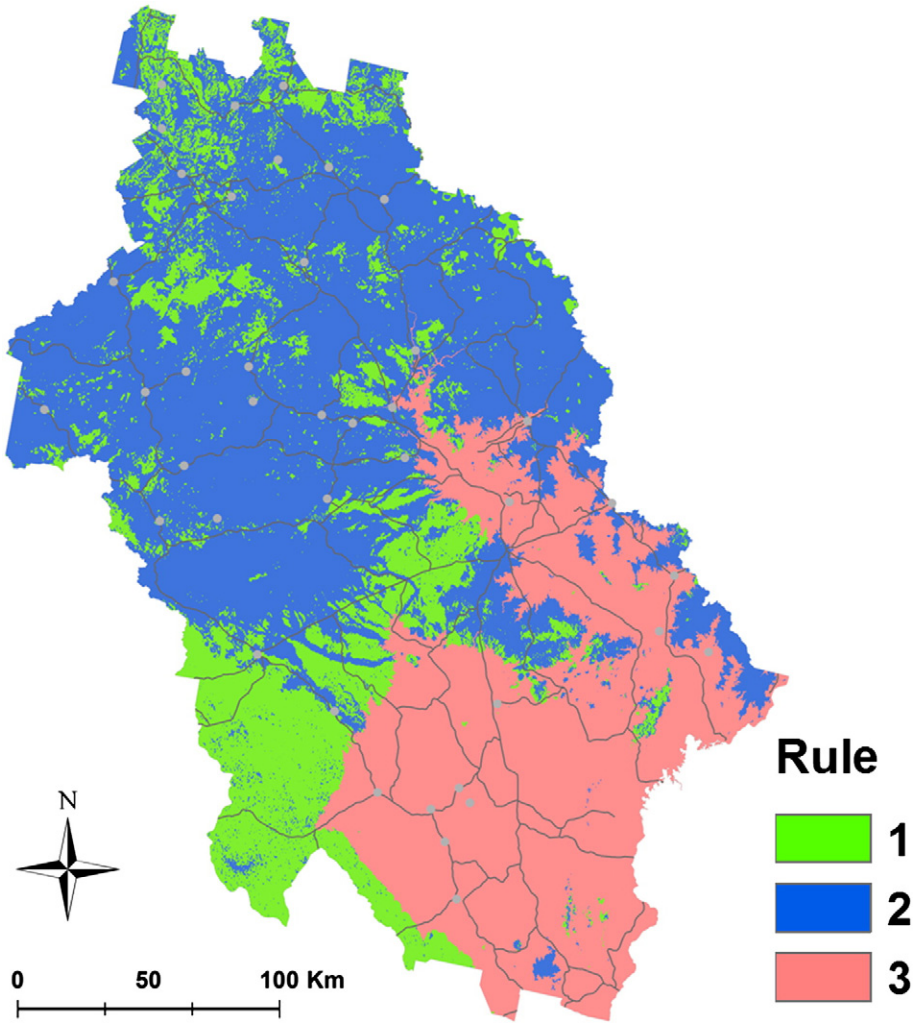

Fig. 3. Geographical partitioning of Cubist model rules for the 15–30 cm depth interval.

(1). NDVI corresponds to vegetation type or intensity, and fulfils the *o* (organism) criteria of the *scorpan* model. K, U, TKr, TH and WI are the gamma radiometric variables and fulfil the (p) parent material *scorpan* criteria. WI may also fulfil the (*a*) age criteria of the *scorpan* model as well (Wilford, 2012). The other variables E, VD, SH, MSP, MRRTF, TWI, S, HS, and SR all correspond to topographic variables that fulfil the (*r*) relief criteria.

A spherical model was fitted to the variogram of residuals resulting from cubist modelling at each depth interval. Other than the 60–100 cm depth increment, the model residuals displayed a moderate spatial autocorrelation, such that the highest observed nugget-to-sill ratio was 0.68, which corresponded to the 5–15 cm depth increment. For the 60–100 cm depth increment there was no spatial dependence i.e. pure nugget effect. The parameters of the fitted spherical models (nugget, sill, and range, nugget-to-sill ratio) are shown in Table 2. The nugget variation increased with depth (excluding the 60–100 cm depth interval), as did the variance of the residuals. Apart from the 15–30 cm depth, the residuals at the other depths displayed spatial autocorrelation to 5000 m or more.

Maps of the spatial distribution of soil pH at the 0–5 cm depth interval are shown in Fig. 4. In this figure, the centre map indicates the prediction resulting from *scorpan* kriging, while the map on the left shows the lower prediction limit and the map on the right shows the upper prediction limit. These prediction limits constitute a 90% prediction interval. Supplementary material shows the resulting maps for the other depth intervals.

Validation results from *scorpan* kriging are shown in Table 3 for all depth intervals. The number of points contributing to the validation is shown in brackets in the soil depth column. The number of available observations decreases down the profile. The RMSE increased from 0.69 pH units at the 0–5 cm depth interval to 1.13 pH units at the 100–200 cm depth interval. This result appears indicative of an increasing uncertainty in models with increasing soil depth, which is likely due to the paucity of covariates that adequately describe the subsoil spatial variations of soil properties including soil pH. This observation is also reflected in the R^2 values where we calculated values of 0.14 at 0–5 cm which then subsequently decreased to 0.08 at 60–100 cm, before jumping up again to 0.17 at 100–200 cm where the calculation was based upon 88 observations. The PICPs for each depth indicate that the prediction intervals are well defined, such that for the most part, 90% of the time, an observed value fitted within its estimated prediction interval.

Overall, the results of *scorpan* kriging in this study are consistent with previous studies examining the spatial variation patterns of soil pH, despite differences in model structure. For example Minasny and McBratney (2007) in comparing REML-EBLUP, universal kriging, and ordinary kriging spatial prediction methods for mapping soil pH (0–10 cm) across the Lower Hunter Valley, Australia found a RMSE of 0.674, 0.682, and 0.690 for each respective method from an independent validation at this soil depth increment. Similarly Malone et al. (2011a) in the same study area using a neural network approach combined with residual kriging found a similar prediction accuracy (RMSE = 0.6) for the 0–5 cm depth increment. Like in the current study, Malone et al. (2011a) made quantifications of the uncertainty (that were validated), and also found an increasing RSME with increasing depth, such that for the 60–100 cm depth increment the RMSE was found to be 1.6 from an independent validation. This trend of decreasing accuracy with increasing soil depth as assessed with the RMSE statistic has also been observed by Sulaeman et al. (2012) and Vaysse et al. (2014) in recent studies that investigated soil pH mapping. Adhikari et al. (2014) for national extent mapping of soil pH across Demark at multiple soil depths, despite plentiful data, recorded only a moderate improvement of the results found for this current study, where the most accurate prediction was for the 5–15 cm depth with an RMSE of 0.61. Adhikari et al. (2014) implemented a similar *scorpan* kriging approach to that used in this current study. Sulaeman et al. (2012) used both conceptual (expert-orientated models) and Cubist models. Vaysse et al. (2014) investigated the use of random forest modelling coupled with model residual kriging.

Comparing the *scorpan* kriging results with those predictions made by Odgers et al. (in preparation) at the same points, both are more-or-less comparable in the sense that there is a systematic increase in RMSE estimates with increasing soil depth (Table 3). R^2 values from the validation of predictions from Odgers et al. (in preparation) seldom breach 0.1 for any depth increment. Based on the RMSE and R^2 criteria, *scorpan* kriging performed marginally better than the disaggregated soil property maps for all depths except the 60–100 cm depth, where the performance of both models is equivalent.

3.2. Model averaging

The validation results for each of the model averaging methods are also shown in Table 3. The *W* parameters attributed to each source map are indicated as W_{SK} for the *scorpan* kriging predictions and W_{DS}

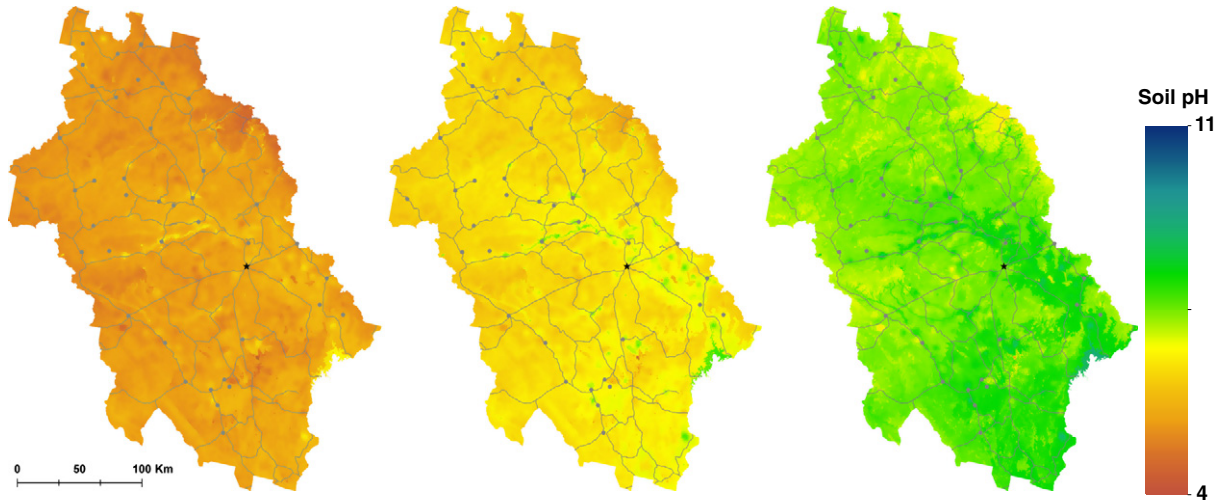

Fig. 4. Digital soil maps of soil pH for the 0–5 cm depth interval across the Dalrymple Shire. These maps were produced from *scorpan* kriging. The centre map represents *scorpan kriging prediction*, while the left and right maps respectively are the 90% lower and upper prediction limits. Supplementary material has further maps for the other subsequent depth intervals.

for the disaggregated soil map predictions. The GRA method also has a W_0 parameter, which indicates the intercept value of the multiple linear regression model. For the other model averaging methods this parameter is always zero. For EW, the W_{SK} and W_{DS} are always 0.5. The weighting parameters from VW are always locally determined i.e. they change from point to point — for comparative purposes we have shown the average of W_{SK} and W_{DS} resulting from VW based on the 300 validation points. In the case of the BMA approach we have reported the 2.5th and 97.5th quantiles of W_{SK} and W_{DS} found for each source map.

A general observation is that for the VW, GRA, and BMA approaches, the weights preference the *scorpan* kriging predictions over the disaggregated soil map predictions. This is particularly the case for the 0–5 cm, 5–15 cm, 15–30 cm and 100–200 cm depth intervals, but the weights are more even for the 30–60 cm and 60–100 cm depth intervals.

In terms of performance, all the model averaging approaches resulted in an improvement in prediction accuracy when compared to the predictions from each source map alone as based on the RSME and R^2 statistics from independent validation, and was apparent for all depth intervals. Differences between the different model averaging approaches were not really apparent in terms of the RMSE and R^2 statistics. Of course with the BMA method, it is possible to get some sense of what to expect in terms of gain in performance by implementing model averaging because we can retrieve the full distribution of model diagnostics by running each W parameter set and calculating the RMSE and R^2 statistics. The performance statistics for the other model averaging methods always fitted somewhere within the distribution of possible outcomes realised from BMA. In terms of validating the estimations of uncertainty regarding the combined predictions, there was a tendency for them to be under-predicted as indicated by the PICP results. For the GRA approach, PICPs were generally acceptable for the top three depth intervals, but with increasing depth PICPs went from 82%, 75%, to 83% for the last three depths. PICPs from the VW were generally lower compared to the other model averaging approaches at all depths. PICPs for EW were comparable to that of the GRA approach. The estimations of uncertainty relating to the model averaged predictions are a compromise of those from the input maps,

Table 3
Weight parameters and soil map quality statistics for *scorpan* kriging, disaggregated soil map predictions, and model averaging. SK (*scorpan* kriging), DS (disaggregated map), EW (equal weights), VW (variance weighted), GRA (Granger–Ramanathan averaging), BMA (Bayesian model averaging).

Soil depth	Method	W_0	W_{RK}	W_{DS}	RMSE	R^2	PICP
0–5 cm (300)	Individual models						
	RK	0	1	0	0.69	14	90
	DS	0	0	1	0.75	6	96
	Combined models						
	EW	0	0.5	0.5	0.69	15	87
	VW[a]	0	0.68	0.32	0.69	16	84
	GRA	−0.77	0.79	0.33	0.68	16	87
	BMA[b]	0	0.44–0.97	0.03–0.56	0.68–0.7	14–16	84–88
5–15 cm (299)	Individual models						
	RK	0	1	0	0.7	14	89
	DS	0	0	1	0.75	6	95
	Combined models						
	EW	0	0.5	0.5	0.69	16	84
	VW[a]	0	0.64	0.36	0.69	17	81
	GRA	−1.40	0.82	0.39	0.69	18	90
	BMA[b]	0	0.4–0.94	0.06–0.6	0.69–0.7	14–18	84–90
15–30 cm (299)	Individual models						
	RK	0	1	0	0.82	9	88
	DS	0	0	1	0.84	6	94
	Combined models						
	EW	0	0.5	0.5	0.8	13	85
	VW[a]	0	0.61	0.39	0.8	13	85
	GRA	−0.61	0.59	0.5	0.8	13	86
	BMA[b]	0	0.4–0.94	0.06–0.6	0.8–0.82	10–13	81–86
30–60 cm (290)	Individual models						
	RK	0	1	0	0.96	9	90
	DS	0	0	1	0.95	9	96
	Combined models						
	EW	0	0.5	0.5	0.92	13	82
	VW[a]	0	0.59	0.41	0.93	12	80
	GRA	0.25	0.45	0.51	0.92	13	82
	BMA[b]	0	0.17–0.84	0.16–0.83	0.92–0.94	10–13	82–85
60–100 cm (215)	Individual models						
	RK	0	1	0	1.13	8	87
	DS	0	0	1	1.13	7	96
	Combined models						
	EW	0	0.5	0.5	1.09	11	80
	VW[a]	0	0.6	0.4	1.09	11	77
	GRA	0.93	0.46	0.43	1.09	11	75
	BMA[b]	0	0.13–0.98	0.02–0.87	1.09–1.13	8–11	78–80
100–200 cm (88)	Individual models						
	RK	0	1	0	1.13	17	90
	DS	0	0	1	1.19	9	97
	Combined models						
	EW	0	0.5	0.5	1.11	18	83
	VW[a]	0	0.61	0.39	1.10	20	81
	GRA	0.14	0.67	0.32	1.10	19	83
	BMA[b]	0	0.17–0.98	0.02–0.83	1.10	19	81–84

[a] VW weighting parameters represent the average based on the 300 validation points.
[b] BMA weighting parameters and quality statistics are reported using the 2.5th and 97.5th percentiles of the distributions based on 2000 MCMC samples.

which is why they subsequently don't perform as well in terms of the PICP validation. While we find the uncertainty estimates acceptable from this study; given that in the worst case (60–100 cm), 75 times out of 100 the prediction intervals will cover the reality of what is observed in the field. Ideally, future studies should look at alternative methods of deriving uncertainty estimates for an ensemble of spatial soil models. This will be to assess whether the systematic underestimation of the quantifications of uncertainty is an outcome of the particular uncertainty method used, or indeed corroborates the finding that there is an additional source of uncertainty that has not been accounted for, to which further investigative efforts should seek to address.

To generate digital soil maps of combined model predictions and their uncertainties, we opted to use the GRA approach because it performed marginally better than the other methods based on the validation statistics of the predictions and their uncertainties on consideration of all depth intervals. Fig. 5 shows the predicted map and their associated lower and upper prediction limits of soil pH for the 0–5 cm depth interval. Supplementary material shows the maps for the remaining depth intervals. Given that the GRA weights for the 0–5 cm depth interval favoured the *scorpan* kriging map, the combined map is very similar to the *scorpan* kriging map. The subtle differences that do exist are the result of the contribution of the disaggregated soil map predictions.

4. General discussion

We have shown that model averaging of soil maps generated from two different processes — *scorpan* kriging and disaggregated legacy soil maps — results in a digital soil property map that is more accurate than either of the contributor maps. This is an expected outcome for this type of ensemble modelling approach (Diks and Vrugt, 2010). Despite the results, it is clear from this work that neither the source maps nor the model-averaged predictions are particularly accurate in terms of RMSE or R^2 statistics. The fairly poor prediction performance of both sets of source maps, especially in the subsoil, should not be interpreted as a failure in the method used to create them rather than as a limitation of (i) the available legacy data and (ii) the predictive ability of the suite of *scorpan* covariates on which prediction models were based. For example, Odgers et al. (2014) describes that the point data used in this study have a geo-locational error of up to 100 m. This source of uncertainty could potentially have severe ramifications to the model fitting process due to an inappropriate co-location of environmental covariates with the observed soil data. It is difficult to explicitly quantify this source of uncertainty, yet it is likely to have contributed to some degree to the overall prediction performance of *scorpan* kriging, irrespective of soil depth.

Nonetheless, what the model averaging is analogous to doing is leveraging the best aspects of each contributing model, and discarding the worst aspects. If both contributing models are poor, ultimately the quality of the combined outcome will also be relatively poor; however, one can at least expect the quality of the combined output to be comparable to or better than the best of the contributing models. Furthermore, the fact that we can quantify estimates of uncertainty of the predictions of soil pH to a depth of 2 m provides a strong impetus for further investment chiefly in the acquisition of new soil data, but also in investigating potential covariates that could improve upon the results we obtained.

Our results indicated that none of the model averaging approaches are a particular standout in terms of gains in accuracy. Despite what the results indicate, the EW approach will generally not be appropriate if quantifications of uncertainty have been derived; as one will essentially be neglecting to consider an important quality diagnostic of the data being used. Subsequently, model averaging approach that preferentially weights more accurate predictions is desirable, in which case the VW, GRA and BMA approaches are a worthwhile consideration for similar digital soil mapping applications.

BMA is advantageous from the point that one can explore what can be maximally achieved through model averaging i.e. we can generate a full distribution of expected model averaged outcomes. Its one disadvantage is the computational cost required to achieve that outcome. In that sense, VW or GRA are reasonable alternatives because the weights determined from these approaches generally resulted in the best outcome possible that could be achieved by combining the two source maps used in this study. The VW approach would work best given accurate estimates of uncertainty. However, comparing these two approaches in this study, GRA marginally performed better than VW.

When it comes to mapping the combined predictions and their uncertainties, due consideration of the computational costs associated with applying model averaging is necessary. We did not conduct a formal analysis to compare the mapping using either approach. But if one were to compare the VW with GRA, the computational cost of VW is that the weighting parameters need to be computed for all raster cells, after which the sampling of the PDFs is required for approximation of the uncertainties. The raster cell by raster cell approximation of the uncertainties is still required by GRA, however; one does not need to estimate the weighting at each raster cell; instead the fitted multiple linear regression model is applied. This computational difference

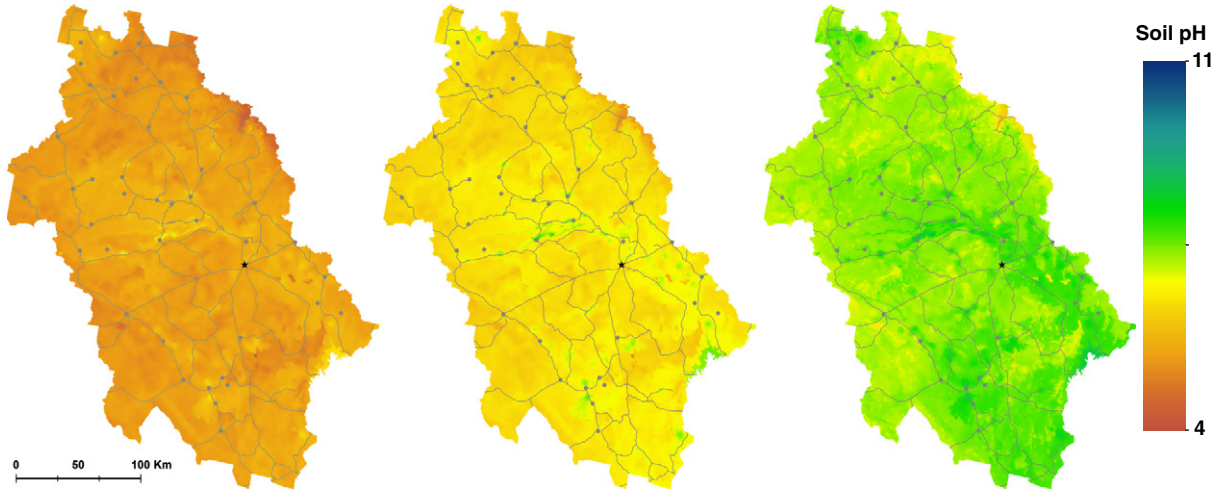

Fig. 5. Model average digital soil maps of soil pH for the 0–5 cm depth interval across the Dalrymple Shire. These maps were produced using GRA approach. The centre map represents the combined prediction, while the left and right maps respectively are the 90% lower and upper prediction limits. Supplementary material has further maps for the other depth intervals.

becomes poignant with particularly large mapping extents or with fine-resolution raster grids, as exemplified by the current study. On this basis, we would recommend the use of GRA for the reasons that it produces similar or better outcomes when compared to other model averaging approaches, and that it is also relatively easy and efficient to apply spatially.

5. Conclusions

- Traditional soil maps are the legacy of a skilled group of individuals who invested much expertise, time and cost to produce them. Digital soil mapping will be the richer by including these valuable soil information resources within contemporary soil mapping projects.
- We have demonstrated that one way to do this is with model averaging. Here, taking a disaggregated traditional soil map (Odgers et al., in preparation) and combining it with a digital soil map created from *scorpan* kriging (from point data), we generated a new soil map (at regular depth intervals) that were better than either contributing map alone.
- A number of model averaging techniques could potentially be used, but we would recommend Granger–Ramanathan averaging for digital soil mapping projects because it performs demonstrably as good as the more sophisticated methods available such as BMA, and is efficient to apply in large mapping extents, or for finely resolved mapping.
- In digital soil mapping where multiple potential models can be used for fitting spatial prediction functions, we would propose that model averaging is a good approach to combine such an ensemble of models so that one can keep the best, and discard the worst aspects of each.

Acknowledgements

The authors gratefully acknowledge the provision of the soil data for this work from the Queensland Department of Science, Information Technology, Innovation and the Arts.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.geoderma.2014.04.033.

References

Adhikari, K., Bou Kheir, R., Greve, M.B., Greve, M.H., Malone, B.P., Minasny, B., McBratney, A.B., Richer de Forges, A., 2014. Mapping soil pH and bulk density at multiple depths in Denmark. In: Arrouays, D., McKenzie, N., Hempel, J., McBratney, A.B. (Eds.), GlobalSoilMap: Basis of the Global Spatial Soil Information System. CRC Press, pp. 155–166.

Bates, J.M., Granger, C.W.J., 1969. Combination of forecasts. Oper. Res. Q. 20 (4) (451-&).

Bregt, A.K., Bouma, J., Jellinek, M., 1987. Comparison of thematic maps derived from a soil map and from kriging of point data. Geoderma 39 (4), 281–291.

Buckland, S.T., Burnham, K.P., Augustin, N.H., 1997. Model selection: an integral part of inference. Biometrics 53 (2), 603–618.

Diks, C.G.H., Vrugt, J.A., 2010. Comparison of point forecast accuracy of model averaging methods in hydrologic applications. Stoch. Env. Res. Risk A. 24 (6), 809–820.

GlobalSoilMap Science Committee, 2013. Specifications: Tiered GlobalSoilMap.net Products, Release 2.3. GlobalSoilMap Science Committee.

Granger, C.W.J., Ramanathan, R., 1984. Improved methods of combining forecasts. J. Forecast. 3 (2), 197–204.

Henderson, B.L., Bui, E.N., Moran, C.J., Simon, D.A.P., 2005. Australia-wide predictions of soil properties using, decision trees. Geoderma 124 (3–4), 383–398.

Heuvelink, G.B.M., Bierkens, M.F.P., 1992. Combining soil maps with interpolations from point observations to predict quantitative soil properties. Geoderma 55 (1–2), 1–15.

Hijmans, R.J., van Etten, J., 2013. Raster: geographic data analysis and modeling. R Package Version 2.1-37 (http://CRAN.R-project.org/package=raster).

Hjort, N.L., Claeskens, G., 2003. Frequentist model average estimators. J. Am. Stat. Assoc. 98 (464), 879–899.

Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian model averaging: a tutorial. Stat. Sci. 14 (4), 382–401.

Kempen, B., Brus, D.J., Stoorvogel, J.J., Heuvelink, G.B.M., de Vries, F., 2012. Efficiency comparison of conventional and digital soil mapping for updating soil maps. Soil Sci. Soc. Am. J. 76 (6), 2097–2115.

Kuhn, M., Weston, S., Keefer, C., Coulter, N., Quinlan, C code for Cubist by Ross Quinlan, 2013. Cubist: rule- and instance-based regression modelling. R Package Version 0.0.13 (http://CRAN.R-project.org/package=Cubist).

Malone, B.P., de Gruijter, J.J., McBratney, A.B., Minasny, B., Brus, D.J., 2011a. Using additional criteria for measuring the quality of predictions and their uncertainties in a digital soil mapping framework. Soil Sci. Soc. Am. J. 75 (3), 1032–1043.

Malone, B.P., McBratney, A.B., Minasny, B., 2011b. Empirical estimates of uncertainty for mapping continuous depth functions of soil attributes. Geoderma 160 (3–4), 614–626.

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

McKay, M.D., Beckman, R.J., Conover, W.J., 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21 (2), 239–245.

Minasny, B., McBratney, A.B., 2007. Spatial prediction of soil properties using EBLUP with the Matern covariance function. Geoderma 140 (4), 324–336.

Minasny, B., McBratney, A.B., 2010. Methodologies for global soil mapping. In: Boettinger, J.L., Howell, D.W., Moore, A.C., Hartemink, A.E., Kienast-Brown, S. (Eds.), Digital Soil Mapping: Bridging Research, Environmental Application, and Operation. Springer Science, New York.

Odeh, I.O.A., McBratney, A.B., Chittleborough, D.J., 1995. Further results on prediction of soil properties from terrain attributes: heterotopic co-kriging, and regression kriging. Geoderma 67 (3–4), 215–226.

Odgers, N.P., McBratney, A.B., Minasny, B., 2014n. Digital soil property mapping and uncertainty estimation using soil class probability rasters. Geoderma (in preparation).

Odgers, N.P., Sun, W., McBratney, A.B., Minasny, B., Clifford, D., 2014. Disaggregating and harmonising soil map units through resampled classification trees. Geoderma 214, 91–100.

Olgers, F., 1970. Buchanan, Queensland — 1:250,000 Geological Series — Explanatory Notes (SF/55-6). Bureau of Mineral Resources, Geology and Geophysics, Canberra.

Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. Comput. Geosci. 30, 683–691.

Pebesma, E.J., Heuvelink, G.B.M., 1999. Latin hypercube sampling of Gaussian random fields. Technometrics 41 (4), 303–312.

Quinlan, R., 1992. Learning with continuous classes. Proceedings of the 5th Australian Joint Conference on Artificial Intelligence, Hobart, Tasmania, pp. 343–348.

Rogers, L.G., Cannon, M.G., Barry, E.V., 1999. Land Resources of the Dalrymple Shire, 1. Land Resources Bulletin DNRQ980090Queensland Department of Natural Resources, Brisbane, Queensland.

Rojas, R., Feyen, L., Dassargues, A., 2008. Conceptual model uncertainty in groundwater modeling: combining generalized likelihood uncertainty estimation and Bayesian model averaging. Water Resour. Res. 44 (12).

Sulaeman, Y., Sarwani, M., Minasny, B., McBratney, A.B., Sutandi, A., Barus, B., 2012. Soil-landscape models to predict soil pH variation in the Subang region of West Java, Indonesia. In: Minasny, B., Malone, B.P., McBratney, A.B. (Eds.), Digital Soil Assessments and Beyond. CRC Press, The Netherlands, pp. 317–323.

Vaysse, K., Arrouays, D., McKenzie, N.J., Coste, S., Lagacherie, P., 2014. Estimation of GlobalSoilMap.net grids from legacy soil data at the regional scale in Southern France. In: Arrouays, D., McKenzie, N., Hempel, J., Richer de Forges, A., McBratney, A.B. (Eds.), GlobalSoilMap: Basis of the Global Spatial Soil Information System. CRC Press, pp. 133–138.

Vrugt, J.A., Diks, C.G.H., Clark, M.P., 2008. Ensemble Bayesian model averaging using Markov Chain Monte Carlo sampling. Environ. Fluid Mech. 8 (5–6), 579–595.

Vrugt, J.A., ter Braak, C.J.F., Diks, C.G.H., Robinson, B.A., Hyman, J.M., Higdon, D., 2009. Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling. Int. J. Nonlinear Sci. Numer. Simul. 10 (3), 273–290.

Wagener, T., Gupta, H.V., 2005. Model identification for hydrological forecasting under uncertainty. Stoch. Env. Res. Risk A. 19 (6), 378–387.

Wilford, J., 2012. A weathering intensity index for the Australian continent using airborne gamma-ray spectrometry and digital terrain analysis. Geoderma 183, 124–142.