Geoderma 262 (2016) 243–253

Contents lists available at ScienceDirect

Geoderma

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

Comparing regression-based digital soil mapping and multiple-point geostatistics for the spatial extrapolation of soil data

Brendan P. Malone [a,*], Sanjeev K. Jha [b], Budiman Minasny [a], Alex B. McBratney [a]

[a] Department of Environmental Sciences, Faculty of Agriculture and Environment, C81 Biomedical Building, The University of Sydney, New South Wales 2006, Australia
[b] School of Civil and Environmental Engineering, University of New South Wales, Sydney, New South Wales 2052, Australia

ARTICLE INFO

Article history:
Received 12 April 2015
Received in revised form 24 August 2015
Accepted 25 August 2015
Available online 8 September 2015

Keywords:
Digital soil mapping
Gamma radiometrics
Multiple-point geostatistics
Regional soil mapping

ABSTRACT

In this study, two approaches for spatial data extrapolation are investigated. The intention here is to predict at fine spatial resolution, total gamma radiometric counts across a large mapping extent (recipient site) on the basis of finely resolved information collected from a nearby donor site. The extrapolation methods used were a digital soil mapping (DSM) regression model approach and a multivariate multiple-point statistical (MPS) approach. Qualitative interpretation of the results from both extrapolation approaches across the recipient site in the Lower Hunter Valley, Australia (area ≈ 220 km^2) shows promise in terms of highlighting known geochemical and physical variations of soils in this area. The extrapolated map was evaluated in a small portion of the study area (area ≈ 4 km^2) where similar high-resolution gamma radiometric data were available. Results show comparable performance of both approaches where a root-mean-square error of 87 ppm was found. A concordance correlation coefficient value of 0.04 was found for the DSM approach, but higher for the MPS approach (0.16). Under the Homosoil framework, where soil point data and mapping are sparse, either method investigated in this study would be suitable as a 'first-cut' approach for developing a comprehensive soil information system in those areas.

1. Introduction

One of the issues in developing high-resolution global and national spatial soil information systems of consistent coverage is reconciling some of the disparity between those areas that have well developed soil information resources with those that are comparatively underdeveloped (Minasny and McBratney, 2010). To address this disparity, most soil scientists would advocate a rebirth of soil survey and mapping programmes to rival the efforts made internationally during the early to mid-20th century (Brevik and Hartemink, 2010) in the areas where information is currently sparse. While appealing, we need to permit ourselves to consider alternative and possibly less costly approaches; with one being model extrapolation, to which is the focus of this investigation.

The concept of Homosoil (Mallavan et al., 2010) has particular relevance in that regard, because it aims, through similarity assessment, the evaluation of which soils (unknown) are similar to other soils (known). For example, if one specified area has very detailed soil mapping (donor site), and has similar soil forming factors to another area that has little to no soil mapping, then it may be possible to extrapolate the information or model from the detailed area to the sparse area (recipient site). These ideas have been around for a while; for example, Lagacherie et al. (1995) implemented an extrapolation concept in France where soil pattern rules were acquired from a reference area or donor site and applied across a wider area where a lower intensity of survey had been achieved. The extrapolation of data is a general concept, and one that can be applied for other variables that are not exclusively soil attributes or classes. For example, proximal soil sensing instruments are able to collect very detailed information about the geochemical and geophysical properties of soils (with gamma radiometrics and electromagnetic induction as a few common examples).

Such proximally sensed information has been demonstrated to be invaluable for soil studies in terms of digital mapping and precision agriculture (Viscarra Rossel et al., 2010). However, their application is commonly restricted to farm and field spatial extents. Using them at regional and larger extents is rare because it is difficult and costly to maintain the same sampling frequency at these scales as for field and farm extents. This issue of practicality has prompted a few recent studies to use proximal soil sensing instruments for regional scale studies. For example, both Viscarra Rossel et al. (2014) and Stockmann et al. (2015) developed efficient methods of traversing a landscape that dually attempt to minimise the time spent in the field yet maximise the potential to capture the spatial soil variation at their scale of investigation. In a similar context, Podgorski et al. (2015) demonstrated the value of integrating proximal sensed geophysical data – that was collected at limited

* Corresponding author.
E-mail addresses: brendan.malone@sydney.edu.au (B.P. Malone), s.jha@unsw.edu.au (S.K. Jha), budiman.minasny@sydney.edu.au (B. Minasny), alex.mcbratney@sydney.edu.au (A.B. McBratney).

http://dx.doi.org/10.1016/j.geoderma.2015.08.037
0016-7061/

sites – with airborne sensed data for constraining and delivering a more detailed hydrological and geological model across a large spatial extent of Botswana (Okavango Delta).

In this study we approach the problem of delivering detailed mapping differently by investigating the efficacy of model extrapolation through the use and subsequent comparison of two contrasting (extrapolation) approaches. The first is using a digital soil mapping approach (McBratney et al., 2003) as suggested in Mallavan et al. (2010). The second is via multiple-point statistics, in particular the Direct Sampling algorithm as described in Mariethoz et al. (2010).

The first extrapolation method hereafter referred to as the DSM approach, entails the following steps. From the area with detailed information, first the target variable of interest is decided upon. Using existing point observations (for which there should be many), or sampling directly from an available raster of the property of interest, these data are then intersected with a portfolio of spatially exhaustive environmental covariate data. This information could be retrieved from an available digital elevation model, remote sensing data platform or some other similar source (Mulder et al., 2011). A DSM model is then constructed, which is essentially a numerical model that relates the information on the variable of interest to the environmental factors. The constructed model is then applied to the recipient site. Grinand et al. (2008) used a DSM approach in France for mapping soil types to investigate the extent to which a model yields a valid prediction. The accuracy of predictions made for the extrapolated area (recipient site) was found to be lower than that made in the training or donor area. Intuitively, this type of result is expected because of the complexity of spatial soil variation, and the impossibility of matching soil forming factors between donor and recipient sites. The results from Grinand et al. (2008) are encouraging from the perspective that such an extrapolation approach would be useful to fill the gaps in present soil map coverage and to increase efficiency of ongoing soil survey to target areas of greatest uncertainty.

Multiple-point statistics (MPS) (Guardiano and Srivastava, 1993) has not before been used in the context of Homosoil. In fact, there have only been a limited number of soil science studies that have explored MPS, with Meerschman et al. (2013a) and Meerschman et al. (2014) being a few examples. Originally developed in the field of geological reservoir modelling, MPS represents an alternative to two-point statistics such as that of variogram modelling and subsequent kriging, and even DSM modelling, with recent applications in hydrogeology (Chugunova and Hu, 2008; Jha et al., 2014), geophysics (Liu et al., 2004; Comunian et al., 2014), and remote sensing (Ge and Bai, 2010; Mariethoz et al., 2012). A stated advantage of MPS is its ability to capture complex patterns and connectivity in data, which is difficult to do with two-point statistics (Mariethoz et al., 2010). In statistical literature, Markov Random Fields serve as the statistical construct that underpins MPS, e.g. Besag (1986) and Emery and Lantuéjoul (2014). Central to MPS, is the training image, which is a conceptual image of the expected spatial structure of the variable to be predicted. The idea of training images is that there may exist another site – a soil analogue in this case (i.e. the training image) – where large amounts of information are available, and from which it is possible to learn spatial or textural information. This idea is very much in line with the concept of Homosoil, making MPS an interesting candidate technique in this context. Spatial patterns learnt from a training image were particularly relevant for Meerschman et al. (2014) in processing proximal soil sensor data given a repeating polygonal fossil ice-wedge soil pattern. Extending MPS to include multivariate training images (Jha et al., 2013a, 2013b, 2015) provides an opportunity to explore its broader application for digital soil mapping efforts, and consequently for Homosoil. The hypothesis here is that environmental covariates together with detailed (soil) mapping from the donor site can be used as training image to inform the spatial pattern of mapping at the recipient site.

The subsequent investigation is a scoping study and details the use of the above-described methods of extrapolation for mapping the total count gamma-ray emission from soils across the Lower Hunter Valley, NSW (recipient site), given some existing detailed survey from the same area (albeit at a much smaller spatial extent). We firstly describe the study area and data used in this study. Secondly the theoretical underpinnings of DSM and MPS are described, followed by description of the procedures for implementing each of the approaches. Lastly, subsequent results and outputs are presented together with a broader discussion of their significance.

2. Materials and methods

2.1. Study area

The study area is located in the Lower Hunter Valley, NSW, Australia (32.83°S 151.35°E), approximately 140 km north of Sydney, NSW, Australia, and covers an area of approximately 220 km^2 (Fig. 1). This area is referred to as the Hunter Wine Country Private Irrigation District (HWCPID). This area is situated in a temperate climatic zone, and experiences warm humid summers, and relatively cooler yet also humid winters. Rainfall is mostly uniformly distributed throughout the year. The area receives on average just over 750 mm of rainfall annually (Australian Government Bureau of Meteorology, 2014). Topographically, this area consists mostly of undulating hills that ascend to low mountains to the south-west. The underlying geology includes predominantly Early Permian siltstones, marl, and some minor sandstone (Hawley et al., 1995). Other parent materials include Late Permian siltstones, and Middle Permian conglomerates, sandstones and siltstones. Soils are quite variable, but in general terms are weathered mixed kaolinitic–smectitic type soils.

2.2. The data

The recipient site for this study is the entire HWCPID. In 2013 an area of 15 km^2 was surveyed using a ground-based gamma-ray detector (Stockmann et al., 2015) to produce raster maps of the radiometric ROIs (regions of interest) with a raster cell size of 25 by 25 m (shown in yellow in Fig. 1). Specifically, that work entailed driving across the landscape following a network of pre-determined transects. A gamma-ray spectrometer was attached to the vehicle which recorded on-the-go radiometric signals being emitted from the soil surface. On average, the 'sampling' density of the on-the-go proximal sensing was 45 points per hectare. For the work of Stockmann et al. (2015), the data was collected for total gamma-ray count and the ROIs that corresponded to Potassium, Thorium, and Uranium. All data were mapped in the units of counts-per-second (cps). The mapped outputs from Stockmann et al. (2015) represent the donor site in this study — they are detailed data that need to be extrapolated to the entire HWCPID. It is possible that this extrapolated information could be used in the future for updating existing soil mapping, and more generally for digital soil mapping studies in this region such as the refinement of soil and landscape regions or terrons as described in Malone et al. (2014a). This study focuses specifically on the mapping of the total gamma-ray counts rather than each of the individual ROIs.

Both extrapolation methods (DSM and MPS) make use of spatially exhaustive covariate information derived principally from a digital elevation model (25 m × 25 m spatial resolution). In total 7 environmental covariates were used in this study: elevation, altitude above channel network, incoming solar insolation, mid-slope position, multi-resolution valley bottom flatness, terrain wetness index, and slope. The processing of the digital elevation model (DEM) to derive these additional terrain-based variables was performed using SAGA-GIS (System for Automated Geoscientific Analyses, http://www.saga-gis.org). Maps of each of the covariates are shown in the supplementary material associated with this manuscript.

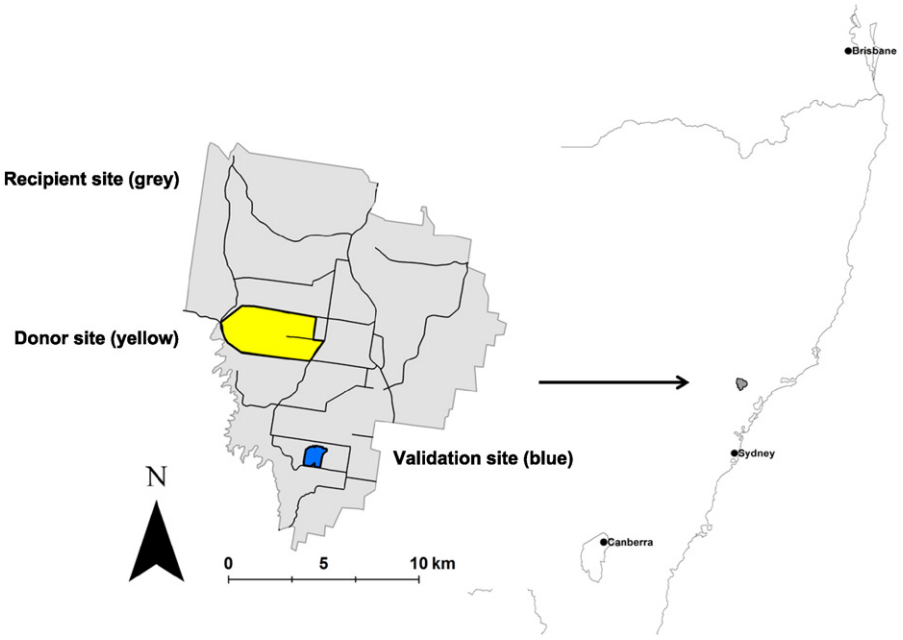

Fig. 1. Grey coloured map showing the boundary extent and road network of the Hunter Wine Country Private Irrigation District (HWCPID) situated in the Lower Hunter Valley. Geographical situation of HWCPID is displayed in relation to the major Australian cities of Brisbane, Sydney and Canberra. On the map, areas shown in grey, yellow and blue colour indicate the recipient, donor, and validation sites respectively.

2.3. How similar are the donor and recipient sites?

A key component of Homosoil is to evaluate via a taxonomic distance measure, the similarity of environment between the recipient site and potential donor sites. The motive in Homosoil is that the donor site selected has the lowest possible taxonomic separation to the recipient site compared to all other candidate donor sites. In the case of this study, the donor site has already been determined. Therefore it is necessary to evaluate the question of how similar the donor site is to the recipient site. In a normal situation, predictions would only be generated where the similarity passes some pre-determined threshold criteria.

In this study, taxonomic distance is quantified in terms of the Mahalanobis distance (Mahalanobis, 1936), where each pixel location of the recipient area (which will have a vector of values that correspond to each of the environmental covariates) is compared to each pixel (vector of environmental covariate values) across the donor site. As was described above, 7 covariates (all derived from a digital elevation model) were able to be sourced for this study. In mathematical terms, the matrix of environmental covariates for the recipient area can be defined as **R** which in this study is a 335,838 × 7 matrix, where each row is a pixel location and each column holds corresponding values to each given environmental covariate. Similarly **D** is defined as the 17,853 × 7 environmental covariate matrix for the donor site. At each pixel of the recipient area, **e** is created which is the vector of squared Mahalanobis distances a single pixel in the recipient area has to each pixel of the donor site **D**. The Mahalanobis distance requires a covariance matrix of the input variables (environmental covariates) which was estimated as the covariance matrix of **R**. For simplicity, a single taxonomic distance estimate is calculated at each pixel as the mean of the nearest 500 distance calculations of **e**. A threshold distance of 6.5 was chosen as the cut-off between whether a pixel was similar in terms of its environment covariate to the donor site. This value was determined on the basis of the distance calculations within the donor area, where 6.5 was the 97.5% percentile of taxonomic distance measurements across this site. Therefore, a low value (i.e. less than 6.5) indicates that donor and recipient sites are relatively similar.

2.4. Theory and implementation

2.4.1. Extrapolation approach based on digital soil mapping

For some background, digital soil mapping (DSM) is: "the creation and population of spatial soil information systems by numerical models inferring the spatial and temporal variations of soil types and soil properties from soil observation and knowledge from related environmental variables" (Lagacherie and McBratney, 2007). Formalised by McBratney et al. (2003), DSM uses the *clorpt* formulation of Jenny (1941) to describe the factors of soil formation. This is not for explanation, but for empirical quantitative descriptions of relationships between soils and spatially referenced environmental data, with a view of using these as soil spatial prediction functions. This is called the "*scorpan*" model, and is expressed as:

$$S_c[x,y,\sim t] \text{ or } S_p[x,y\sim t] = f(s[x,y,\sim t],\ c[x,y,\sim t],\ o[x,y,\sim t], r[x,y,\sim t], p[x,y,\sim t],\ a[x,y,\sim t],\ n) \quad (1)$$

where:

S_c	soil class
S_p	soil property
s	soils, other attributes of the soil at a point
c	climate, climatic properties of the environment at a point
o	organisms, vegetation, or fauna, or human activity
r	topography, landscape attributes
p	parent material, lithology
a	age, the time factor
n	space, spatial position
t	time (where t is defined as an approximate time)
x,y	the explicit spatial coordinates
f	function or soil spatial prediction function (SSPF).

In this study a rule-based model called Cubist (Quinlan, 1992) was used to regress the target variation (total gamma count) with the sourced environmental covariates (which were principally derived

from a digital elevation model only). The Cubist model is similar to a regression tree model in the sense that data are partitioned into smaller subsets based on the target variable and its relationship with the environmental covariates. However, the terminal nodes are multiple linear regression equations rather than predictions. A sensitivity analysis was performed using different sample sizes of the total gamma ray count map to establish the regression model. Sampling 25% of map pixels was found to result in similar model parameters to those models fitted using more or all the available pixels. A sample of less than 20% of the map pixels resulted into higher occurrences of dissimilar models. The Cubist model was then applied across the whole HWCPID or recipient site.

Prediction uncertainties were defined empirically from the data used for fitting the extrapolation model. Uncertainty is expressed in the form of two quantiles of the underlying distribution of model error (residuals) which has previously been applied in hydrological (Shrestha and Solomatine, 2006) and soil (Malone et al., 2011) studies. The underlying distribution of errors was evaluated through leave-one-out cross validation (LOCV). Because a Cubist rule-based model was used in this study, the distributions of residuals were defined for each ruleset, following a partitioning of the data (according to the ruleset each data point belonged to). Within each ruleset, LOCV was performed such that n number of Cubist models (n being the number of contributing data to the ruleset) was fitted, with the contributing model set being composed of $n - 1$ data. With each fitted model, a different observation is removed each time. The model residual for the removed data however is evaluated by making a prediction for that observation using the fitted model, then calculating the subsequent residual (observed value − predicted value). For each ruleset, the uncertainty is expressed as a 90% prediction interval; which means that the lower 5% and upper 95% quantiles of the empirical model residual distribution are recorded. This empirical method of uncertainty quantification is described in Malone et al. (2014b). Upon extrapolation of the *scorpan* model to the recipient site, each pixel was interrogated to determine which ruleset it belonged to, based on the vector of covariate information at that pixel, and the partitioning criteria of the cubist model. With this defined, the associated rule prediction limits were added to the *scorpan* model prediction, resulting in a 90% prediction interval at each pixel in the recipient area.

2.4.2. Extrapolation approach based on MPS

The MPS methodology adopted for digital soil mapping extrapolation is based on the Direct Sampling (DS) geostatistical approach. The description of DS is presented in Mariethoz et al. (2010) and its recent application in hydrological application with multivariate training images and fusing dense and scarce data can be found in Mariethoz et al. (2012) and Jha et al. (2013b). Here we briefly present the main components of the approach.

The DS algorithm uses a training image, conditioning data, and simulation grid. The nodes of the simulation grid are visited according to a random path and a pattern is defined by its neighbouring values. When conditioning data are available, they are incorporated into the simulation grid by appending the value to the grid node it is spatially closest to. Subsequent spatial patterns (neighbourhood) from the DS have to be coherent with the conditioning values. A pattern with similar neighbourhood is searched in the training image, and a distance representing the mismatch is calculated between the patterns in the simulation grid and in the training image. If the distance is below a given threshold *dth*, the pattern from the training image is pasted in the simulation grid. The newly simulated value is then added to the available conditioning dataset and used for subsequent simulations. Sometimes there may not be any initial conditioning data, in which case results in an unconditional simulated value being made. As the simulation grid fills out with values, the number of conditional values for future simulation increases accordingly.

The algorithm used to search the pattern in the training image is as follows: let $\mathbf{U}$ denote a vector of coordinates for a pixel in the simulation grid and $\mathbf{V}$ coordinates of a value in the training image. $Z(\mathbf{U})$ is the variable to be simulated. $\boldsymbol{N}_{\mathbf{U}}$ is the ensemble of the n closest known pixel values of $\mathbf{U}$ either conditioning data or previously simulated values. For the case of a single variable the local neighbourhood of $\mathbf{U}$ is defined as $\boldsymbol{N}_{\mathbf{U}} = [Z(\mathbf{U}+\boldsymbol{h}_1), Z(\mathbf{U}+\boldsymbol{h}_2), \ldots Z(\mathbf{U}+\boldsymbol{h}_n)]$, where $\boldsymbol{h}$ is the lag vector between $\mathbf{U}$ and its neighbours. The idea of this process is to find a location in the training image that has a neighbourhood $\boldsymbol{N}_{\mathbf{V}} = [Z(\mathbf{V}+\boldsymbol{h}_1), Z(\mathbf{V}+\boldsymbol{h}_2), \ldots Z(\mathbf{V}+\boldsymbol{h}_n)]$ similar to $\boldsymbol{N}_{\mathbf{U}}$. Both neighbourhoods have the same lag vectors. Any mismatch between $\boldsymbol{N}_{\mathbf{U}}$ and $\boldsymbol{N}_{\mathbf{V}}$ is quantified by a distance measure $d[\boldsymbol{N}_{\mathbf{U}}, \boldsymbol{N}_{\mathbf{V}}]$. As soon as a mismatch value below the threshold of *dth* is found, the value $Z(\mathbf{V})$ in the training image is posted in the simulation at location $\mathbf{U}$ and the simulation proceeds to the next unknown pixel value. For continuous variables, a normalised Manhattan distance is used to compute the mismatch between neighbourhoods:

$$d[\boldsymbol{N}_{\mathbf{U}}, \boldsymbol{N}_{\mathbf{V}}] = \frac{1}{\boldsymbol{n}} \sum_{\boldsymbol{i}=1}^{\boldsymbol{n}} \frac{|Z(\mathbf{U}_i) - Z(\mathbf{V}_i)|}{\max_{\mathbf{V} \epsilon TI} Z(\mathbf{V}) - \min_{\mathbf{V} \epsilon TI} Z(\mathbf{V})} \epsilon [0, 1]. \quad (2)$$

Here n is the number of nodes in the neighbourhood being compared and *TI* is the training image.

In the case of a multivariate situation with m variables, the distance between the multiple variables is defined in order to find the pixel value matching the neighbourhoods considering all variables together. The result is that the sampled values have the same cross dependencies as the multivariate training image. The number of neighbouring nodes n_k may vary for each variable k, where $k = 1, \ldots m$. Thus for each variable k the individual neighbourhood will be given as: $\boldsymbol{N}_{\mathbf{U}}{}^{k} = [Z_k(\mathbf{U}+\boldsymbol{h}_i{}^{k}), \ldots Z_m(\mathbf{U}+\boldsymbol{h}_{nk}{}^{k})]$. The multivariate neighbourhood is the concentration of all m individual neighbours: $\boldsymbol{N}_{\mathbf{U}} = [\boldsymbol{N}_{\mathbf{U}}{}^{1}, \ldots \boldsymbol{N}_{\mathbf{U}}{}^{m}]$. Mismatch between such multivariate neighbourhoods is obtained by a weighted linear combination of individual distances between univariate neighbourhoods as given below, where the sum of w_k is 1:

$$d[\boldsymbol{N}_{\mathbf{U}}, \boldsymbol{N}_{\mathbf{V}}] = \sum_{\boldsymbol{i}=1}^{\boldsymbol{m}} w_k \; d\left[\boldsymbol{N}_{\mathbf{U}}{}^{k}, \boldsymbol{N}_{\mathbf{V}}{}^{k}\right]. \quad (3)$$

In this study, the training image(s) are derived from the donor site. It consists of the 7 environmental covariate data sources detailed previously, together with the raster of the gamma total count, as shown in Fig. 2. The simulation is performed upon each 25 × 25 m grid node of the recipient site, and includes both the donor and validation sites. The spatial resolution and extent of the simulation grid are identical to that of the environmental covariates that were arranged for the recipient site. The conditioning data were the 7 environmental covariates that have the full spatial coverage of the recipient site. These are shown in the supplementary material of this research. Outside the extent of the donor site, the total count gamma is unknown and needs to be simulated using MPS using both training and conditioning data. This DS simulation in this case is the operative procedure for extrapolation of total count gamma using multivariate MPS.

For the DS, we used a neighbourhood of 20 pixels. The distance threshold *dth* was assigned a value of 0.1. In the distance calculation, an equal weight of 0.125 was assigned for all of the training images. For comprehensive discussion on how to select these parameter values, readers are referred to Meerschman et al. (2013a, 2013b). 100 conditional realisations were obtained with this setting from which the mean at each pixel was estimated in order to obtain a single estimation. The uncertainty of the predictions was expressed as a 90% prediction

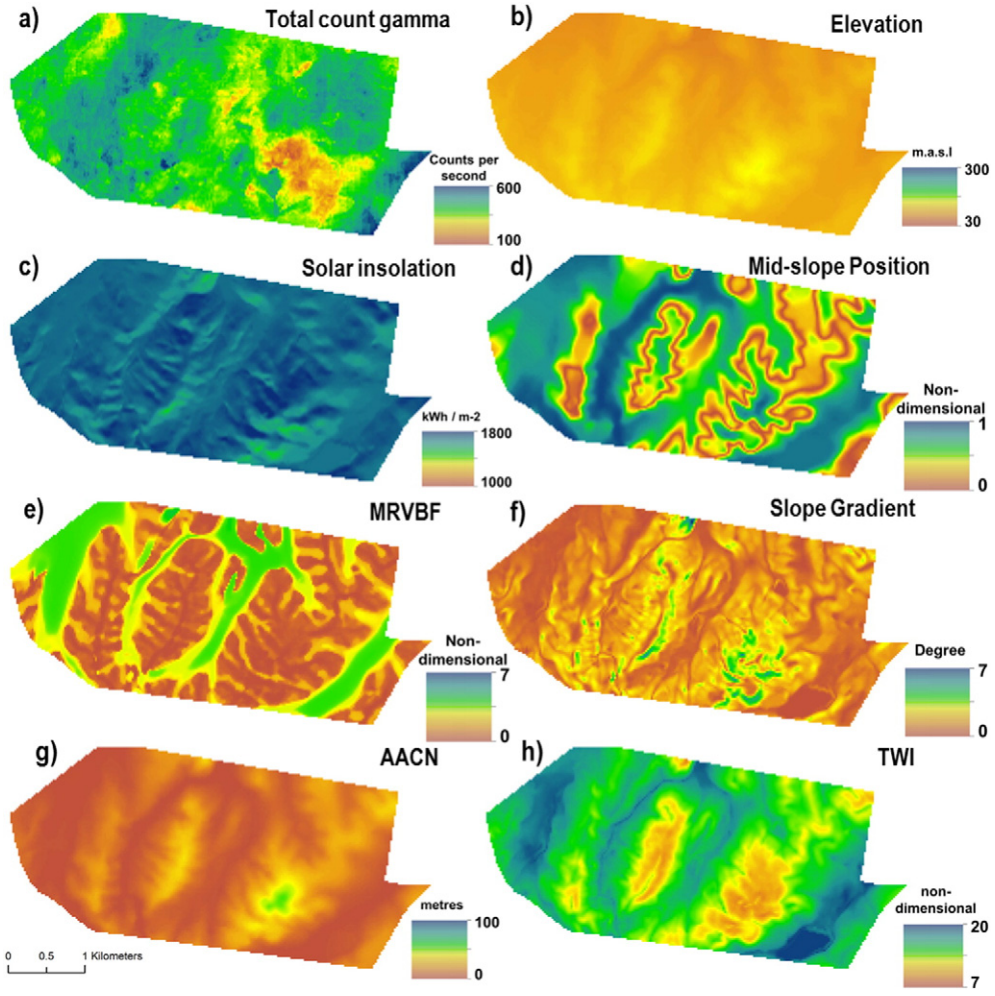

Fig. 2. Training images used for MPS. Training images correspond to spatial information from the donor site to be used for the extrapolation of total count gamma radiometric data. Training images include: a) raster map of surveyed gamma radiometric total count, b) elevation, c) solar insolation, d) mid-slope position, e) multi-resolution valley bottom flatness (MRVBF), f) slope gradient, g) altitude above channel network (AACN), and h) terrain wetness index (TWI).

interval and was estimated at each pixel on the basis of the lower 5% and upper 95% quantiles from the 100 simulations.

2.5. Validation

It is difficult to perform validation in the context of Homosoil because by definition, the recipient site has little or no data against which to check the quality of the predictions. In situations where there is no data available, soil expert qualitative assessments would only be feasible. However, for validation in this study there occurs an additional landholding that has detailed gamma radiometric data that was collected and mapped (25 m × 25 m grid resolution) from a previous survey effort (Fig. 1, shown in green). The mapped total count gamma data at this site (validation site) was compared to the corresponding predictions and associated quantifications of uncertainty from both extrapolation methods. The root mean square error (RMSE) and concordance correlation coefficient were used as goodness of fit criteria to assess the quality of the predictions, while the prediction interval coverage probability (PICP) was used to determine the efficacy of the uncertainty estimates. The PICP is simply the proportion of observations that are encapsulated by the corresponding prediction interval. If the uncertainty estimates have been reasonably defined, the PICP should result in an estimate of 90% for a 90% prediction interval.

3. Results and discussion

3.1. How similar are the donor and recipient sites?

Fig. 3 illustrates that approximately 47% of the area was estimated to be similar to the donor site. This result indicates that there is limited extrapolation ability of the donor site, to which has implications about the certainty of the subsequent predictions, which is discussed further on.

3.2. Donor site

The gamma radiometric total count map of the donor site is shown in Fig. 4a. In this area, the low values correspond to a widespread area of marl parent materials — earthy deposits (indurated marine deposits from the Permian) consisting chiefly of an intimate mixture of clay and calcium carbonate (Stockmann et al., 2015). It is common that carbonate-rich parent materials and soils formed from them are expected to have low radiometric responses (Dickson and Scott, 1997). High values are related to a sedimentary parent rock of mudstones which ultimately weather to fine grained soils. Soils with higher clay content generally have a corresponding high response in total radioelement content relative to other soil with low clay contents (Dickson and Scott, 1997).

For the DSM extrapolation, a Cubist model entailing three rule-sets was defined. The rulesets were partitioned on the basis of threshold

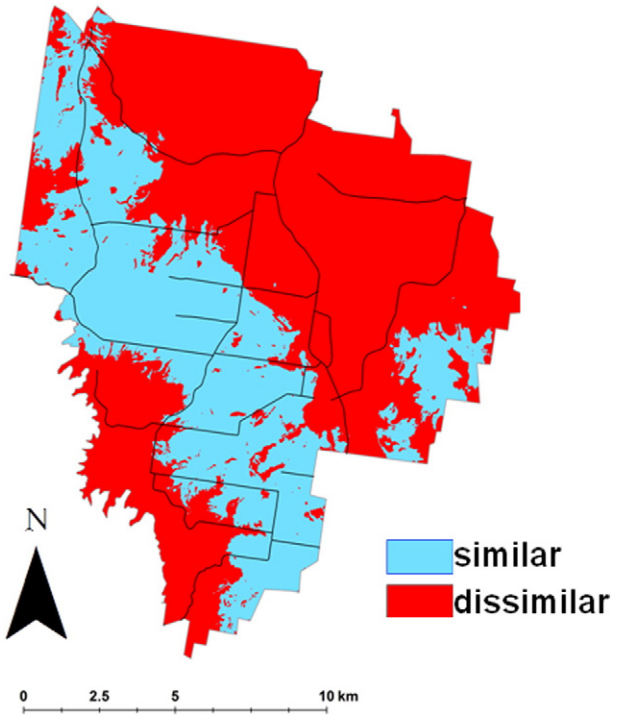

Fig. 3. Map of Hunter Wine Country Private Irrigation District (HWCPID) indicating similarity and dissimilarity to donor site on the basis of environmental covariates.

values of altitude above channel network and mid-slope position. All of the seven environmental covariates were used in each of the three rulesets. The fitted cubist model was able to explain 65% of the variation in the target variable within the donor site. Fig. 4b shows the resulting map, and Fig. 4c is a scatter plot between the detailed (donor) mapping and Cubist model predictions. The scatter plot shows some dispersion around the 1:1 line for low and high values of the target variable and their associated predictions (concordance = 0.43). As can be seen from the map (Fig. 4b) it retains the same general spatial pattern shown in Fig. 4a. Fig. 4d on the other hand highlights a more superior pattern-matching than what was achieved from the DSM approach. This map illustrates the mean of 100 simulations of the MPS algorithm. Fig. 4e is the scatter plot between the detailed mapping and simulation predictions where a concordance of 0.71 was observed. Fig. 5a shows omnidirectional semi-variograms derived for each map from Fig. 4. The shape of the semi-variogram of the predictions from both extrapolations corresponds well with that of the observed data. The semi-variance of the observed data with increasing short-range separation reflects the short-range variability of gamma radiometric data in general. The extrapolations on the other hand appear much smoother and are indicated by the relatively lower slope of the semi-variograms about the origin.

3.3. Recipient site

Both model extrapolations upon the recipient site resulted in visually similar prediction maps. Fig. 6a–b shows the final prediction and prediction interval range respectively for the DSM extrapolation. Fig. 6c–d shows the corresponding maps from MPS. A correlation coefficient of 0.67 was calculated between both maps (Fig. 6a and c). At this large spatial extent, total count values correspond broadly with soil texture mapping across this area (Malone et al., 2014a). High values generally correspond with soils that have high clay content and vice versa. Yet both maps also reflect the differences in soil geochemistry and topography that is independent of soil texture. For example, both maps show low radiometric counts to the south-west of the study area where there is a significant region of marl presence (Malone et al., 2014a), together also with thin and skeletal soils that occur here as well. Soils containing the marl appear to be better delineated by the MPS predictions, while the skeletal soils are better accentuated by the DSM predictions. The south-western region of the study area is bounded by a small mountain range (Brokenback Range), to which contributes to the occurrence of the young and skeletal soils in this area. For other parts of the recipient area, low radioelement values for both maps also correspond broadly with marl occurrence and with topographically features that include ridges and crests where shallow and skeletal soils would generally be found. Low and intermediate total count values also correspond generally with drainage lines where soils have a mixed pedogeneisis from both alluvial and colluvial processes.

A divergent feature between the two extrapolation methods is the magnitude of uncertainty as indicated by the prediction limit ranges (Fig. 6b and d), where they are generally larger for the DSM extrapolation than they are for the MPS extrapolation. One similarity however is that where there is a high prediction range for MPS extrapolation; it is equally as high for DSM extrapolation. This is an interesting observation because for the DSM method, the uncertainties are related to the quality of the fitted Cubist model. The spatial pattern of the prediction limit range for the DSM extrapolation reflects the differences in the magnitude of uncertainty attributed to each ruleset (3 were defined in this study) of the Cubist model. The uncertainty thus appears as discrete areas of relatively high, medium and low prediction ranges. Across the recipient area this has resulted in relatively lower uncertainties in the depression areas of the landscape, compared to areas that are topographically positioned higher. On the other hand for MPS, the uncertainties are an expression of the prediction variance attributed to running multiple simulations of the MPS algorithm — they are a simulation-derived measure of uncertainty as opposed to an empirical-based uncertainty. Where there is a high prediction limit range for the MPS extrapolation, it is a reflection of the fact that there is a significant dissimilarity between the training image and the areas where the extrapolation is made. This observation together with corresponding high prediction uncertainties from DSM extrapolation corroborates with the map in Fig. 3 such that those areas that are dissimilar in terms of their environmental similarity to the donor site.

Efforts to minimise the uncertainty for either extrapolation method may be facilitated using an identical approach, which is sourcing additional and/or alternative environmental covariates. As an example, in the future there may be a detailed geology map that is developed for the area, where currently the available mapping is too general and does not add to the predictive power of DSM models. For the DSM extrapolation, sourcing of additional covariates will aid in efforts to derive a more accurate spatial model of the target variable that will by association also reduce the magnitude of uncertainties. For MPS, considering alternative training images will provide a diverse ensemble of realisations and thus the uncertainty might be decreased. Emery and Lantuéjoul (2014) indicate that the training image should be at least of the same size as the simulation domain so that patterns and ranges of values of the training image can be used in the simulation. Intuitively however, if the training image has enough patterns or variability and the recipient area is actually very similar to the donor area, then the size of the training image may be irrelevant.

Another point for consideration in regard to MPS is that the uncertainties will vary by modifying the threshold *dth* on the Manhattan distance used for accepting the image pattern for a given prediction. At the moment, there is no unified way of determining the value of *dth* and other parameters used in MPS simulation. One way is with a sensitivity analysis, which will determine the influence of each parameter on the uncertainty range. Since the focus of this manuscript is to explore the possibility of using MPS in the context of Homosoil, we did not find it necessary to carry out a complete sensitivity analysis of each parameter. We relied on the guidelines provided in Meerschman et al. (2013a, 2013b) and the range of values used in previous studies e.g., Jha et al.

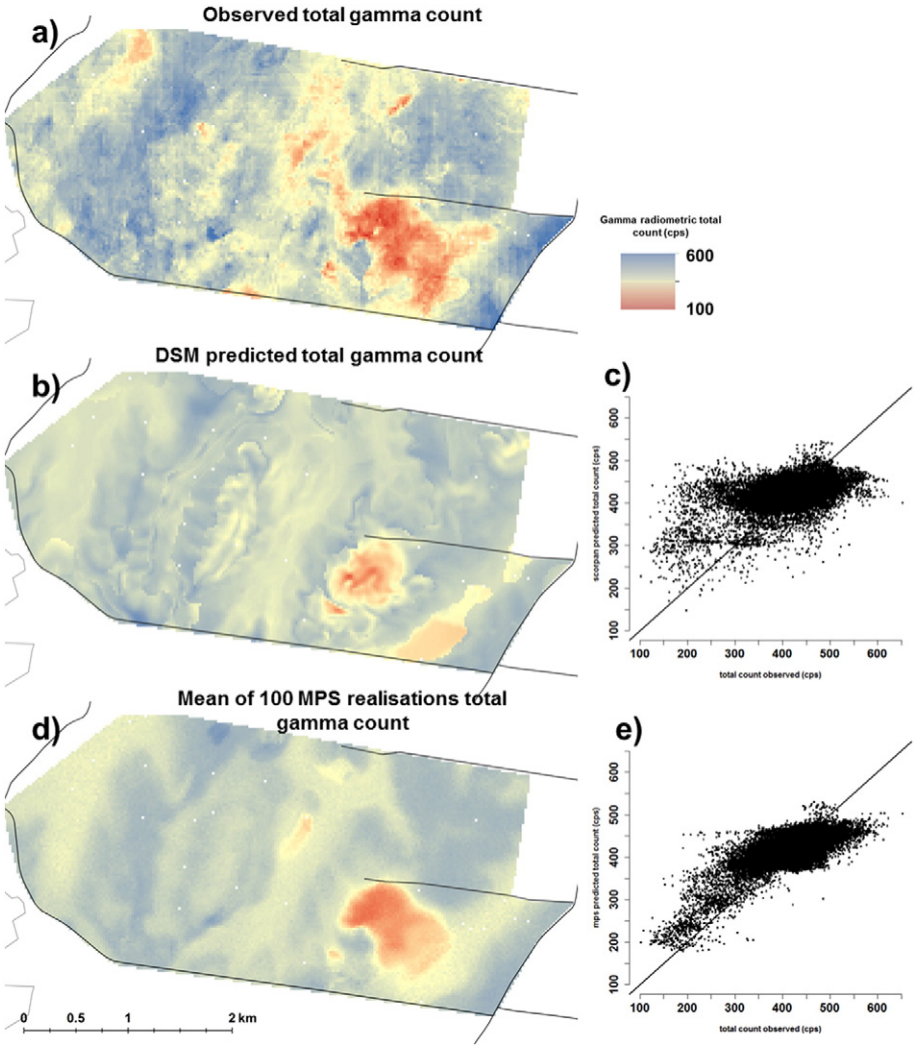

Fig. 4. Donor site maps of gamma radiometric total count (a) observed, (b) predicted using DSM model, (c) scatter plot of comparison between observed and DSM predicted total count, (d) predicted using MPS, and (e) scatter plot of comparison between observed and MPS predicted total count.

(2013a), Jha et al. (2013b) and Jha et al. (2015). We agree that it may be possible to tune the values of parameters but we do not believe that it would change the main findings of this paper.

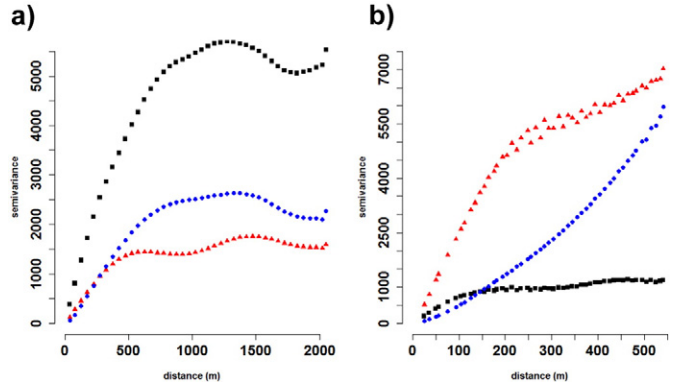

Fig. 5. Omni-directional semi-variograms of the total count radiometric maps for observed data (black squares), MPS predictions (blue diamonds), and DSM predictions (red triangles) at a) the donor site, and b) the validation site.

3.4. Validation site

The validation site provided a situation to independently assess the quality of both extrapolation methods. Bearing in mind however, that this validation is opportunistic (because some existing gamma radiometric is present here), and does not necessarily reflect the quality of the mapping across the entire recipient area. For an independent validation of the mapping, a probability sample of the mapping domain would be necessary. Fig. 7a shows the observed total count radiometric map for this small site. Fig. 7b indicates on the map the areas which are similar and dissimilar to the donor site. Approximately 50% of this site is similar to the donor site. Fig. 7c–e shows the corresponding predictions from the DSM extrapolation, the PICP map — green indicates the areas of the predicted mapping where the 90% prediction interval encapsulates the corresponding observation, and a scatter plot of the DSM predictions compared to the observations. The scatter plot marks are coloured according to the similarity assessment with blue indicating similarity to the donor site and red indicating dissimilarity. Fig. 7f–h shows the corresponding figures for the MPS extrapolation. For the DSM extrapolation, the spatial pattern of the predictions roughly corresponds to that for the observations. The associated scatter plot shows that there is quite some dispersion around the 1:1 line. Comparing observations with the DSM predictions a RMSE of 87 cps and concordance of 0.04 was found. When considering the similarity to the donor site, it was

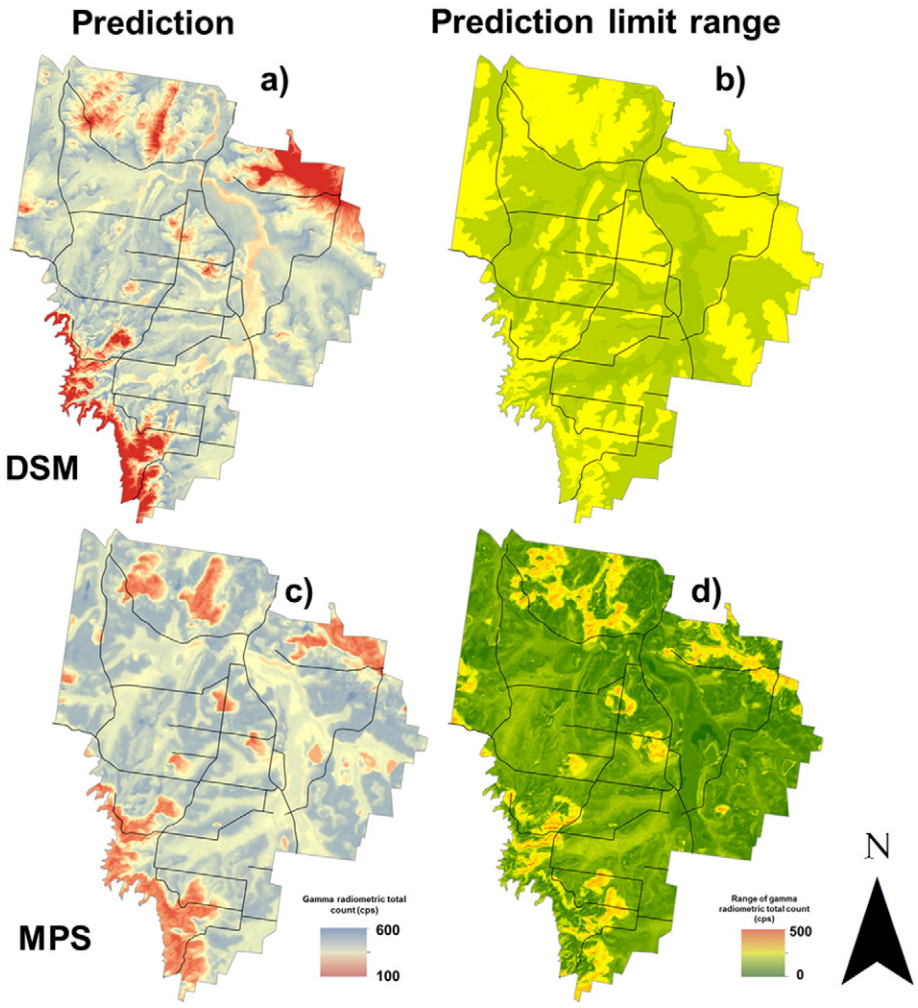

Fig. 6. Recipient site maps of gamma radiometric total count using DSM extrapolation (a and b), and MPS extrapolation (c and d). Maps correspond to prediction and prediction interval range for each extrapolation method. For MPS the prediction corresponds to the mean of 100 MPS realisations.

found that the RMSE was 73 cps and 101 cps for similar and dissimilar locations respectively. This is a subtle indication of the fact that extrapolation is more accurate in areas where there is an associated environmental similarity to the donor site. While the DSM prediction interval range across the recipient site roughly corresponds to areas of similarity and dissimilarity; at the small spatial extent of the validation area, the range is relatively homogenous regardless of the similarity assessment. The averaged prediction interval range from DSM extrapolation was 217 cps. For the MPS extrapolation, there is a subtle difference in prediction interval range where the average range in areas defined as similar was 175 cps, while for the dissimilar areas the average was 191 cps. The estimated RMSE between observations and associated MPS predictions was also 87 cps; a concordance of 0.16 was also found. Breaking this down according to the similarity, the RMSE was 62 cps and 106 cps for similar and dissimilar areas respectively. Overall, the MPS map appears smoother than the DSM map. This is because it is the outcome of calculating the mean of the 100 realisations. Fig. 8a–c shows maps from three randomly selected simulations to provide an example that the outputs from each simulation can be quite different, and can locally appear to be quite noisy. Calculating the concordance between the observations and each of the 100 simulations, it was found to range between 0 and 0.19. Fig. 5b shows the omnidirectional semivariograms of the radiometric mapping in Fig. 7, and provides an indication of the fidelity of spatial structure between observations and subsequent extrapolation methods. Interestingly, the semi-variance of the observations with increasing distance does not increase to the magnitude to what is found for the extrapolation methods. Essentially, the spatial variation of total count in this area is relatively small and its structure is not strongly related to the environmental covariates, which explains the lesser performance of the DSM approach.

For the DSM predictions it was found that the PICP was 76%; meaning that for 76% of locations, the observation is encapsulated by the associated prediction intervals. While a reasonable result, this outcome implies that the quantification of uncertainties is underpredicted in this case. Despite the range of the prediction interval being relatively homogenous across the validation site, areas deemed to be dissimilar to the donor site (Fig. 3) show general agreement to areas where the prediction interval was unspecified. This is also the case for the MPS extrapolation. In terms of the PICP for the MPS extrapolation, 72% of observations were encapsulated by their prediction interval. However, as established above, the prediction intervals are in general narrower for MPS than for DSM. This is an encouraging result — there is reasonable confidence in the performance of MPS for extrapolation studies as it performs more-or-less similarly to the DSM approach. A distinguishing advantage of MPS however is the computational efficiency in being able to generate multiple realisations.

4. General discussion

The research has been a scoping study of the efficacy of different extrapolation methods that could be used in a Homosoil or similar other framework where knowledge is transferred from a donor site

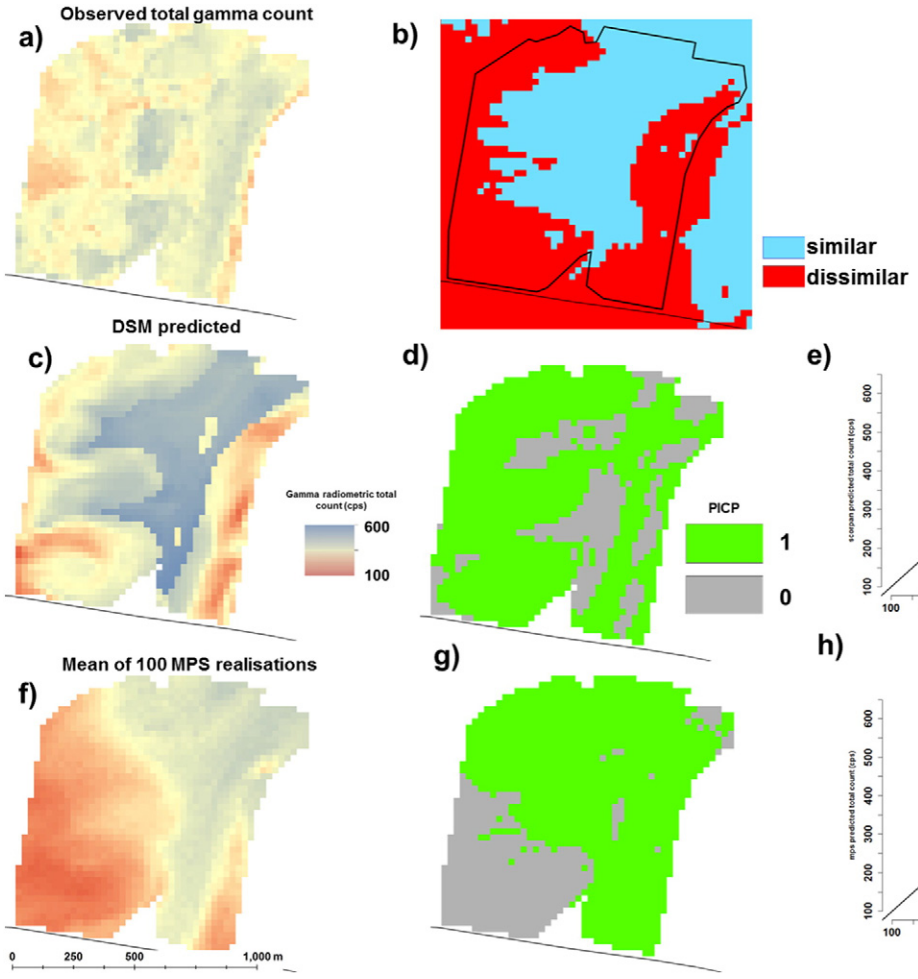

Fig. 7. Validation site maps of gamma radiometric total count. (a) Observed data, (b) similarity of validation site to donor site, (c)–(e) DSM extrapolation, and (f)–(h) MPS. Prediction maps are the final prediction, prediction interval coverage probability (PICP; green colour = 1, observation fits within interval), and scatter plot of observation compared to prediction. Blue and red markings on the scatter plot correspond to locations similar and dissimilar to donor site respectively.

(where there is lots of information) to a recipient site (where there is scarce information), under the assumption that both sites have similar environmental conditions, or soil forming factors if studies are concerned with mapping soil property information. Across the whole recipient site there are subtle differences between the predictions from both extrapolation methods. Despite the limited extrapolation potential of the donor site – on the basis of limited similarity to the recipient site — the spatial pattern for both predictions is coherent in the sense that areas of high, intermediate, and low radiometric total counts correspond to known physical and geochemical variations of soil in this area. The quantifications of uncertainty between both extrapolation methods however are approached differently, and consequently their magnitude differs markedly. Despite being evaluated separately and therefore independent of each other, there is a subtle correlation between similarity to the donor site and the associated extrapolation uncertainties for either approach. The validation example highlighted the potential dangers of extrapolation where results were generally unsatisfactory for both approaches in this small area. The performance of the uncertainty quantifications was however reasonable from the viewpoint that they correctly covered at least 70% of the validation area. Consequently, in terms of the prediction interval, once one has been able to quantify the magnitude of uncertainties, objective strategies can be emplaced to bring about their minimisation or narrowing. Some strategies include sourcing new data by way of environmental covariates in order to improve the modelling, or by discovery, through the implementation of field sampling and survey. With such additional information, a clearer distinction between donor site similarity and dissimilarity in terms of extrapolation uncertainty may also be realised.

The Homosoil approach or its implementation is meant for situations where there is very little available information with which to generate digital soil mapping products. It is encapsulated within a more general framework for global soil mapping (Minasny and McBratney, 2010) where decisions on what approach to use are determined on the basis of available data. Most digital soil mapping studies are concerned with the use of soil point data (McBratney et al., 2003). However there have only been few digital soil mapping studies concerned with the use of detailed polygon soil maps and soil point data (Malone et al., 2014b) or detailed soil type polygon maps only (Odgers et al., 2015). Nonetheless, there is a legitimate need for consistent global soil mapping (Sanchez et al., 2009). Consequently, the Homosoil approach has particular relevance for meeting these needs as situations of poor soil data coverage are a widespread problem across the globe. As can be seen from Batjes (2009), poor soil data coverage is not just constrained to developing countries. It is not expected that extrapolation approaches are final products to the ongoing construction and maintenance of digital soil information resources. Rather, such approaches would be used as a first cut or version to the ongoing development of such digital resources with the long term aim of continual improvement and revitalisation.

A consideration of this study is to weigh up the comparative operational advantages and disadvantages of each extrapolation method. At the scale or spatial extent of the recipient area, there are subtle

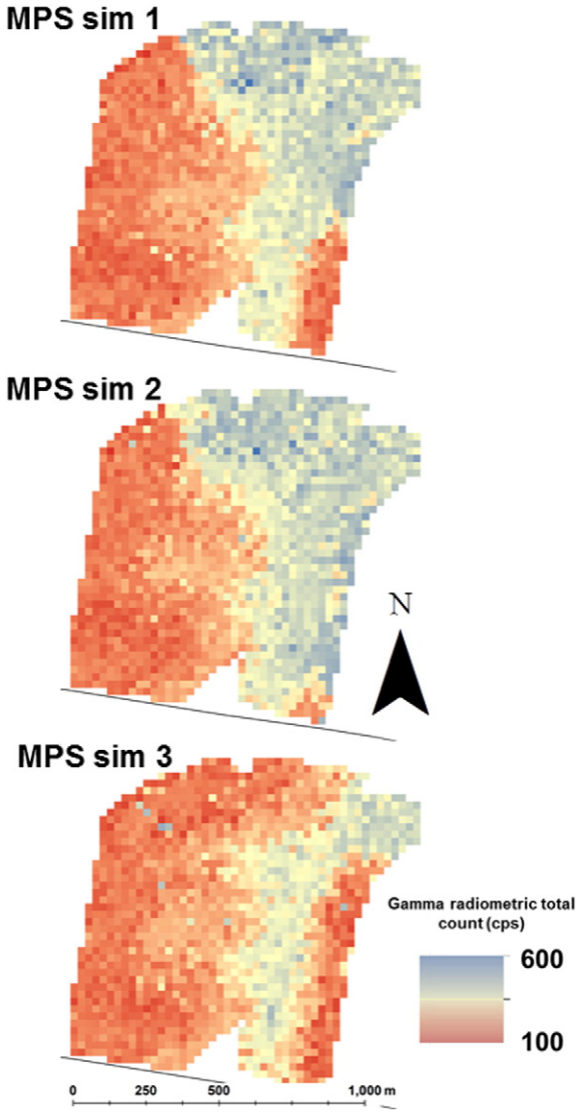

Fig. 8. Three randomly selected realisations of gamma radiometric total count from the MPS extrapolation for the validation site.

differences to the approaches as already detailed. We have already discussed the differences observed between both approaches in terms of the validation area and the quantifications of uncertainty. Operationally, both methods are computationally the same in terms of execution when considering the running time to generate the simulations from the MPS and model fitting, subsequent extrapolation and quantification of uncertainties for the DSM approach. The DSM approach is favourable because it is model derived — the extrapolation predictions are generalised as a predictive function of the available covariates from the donor area. Ultimately, the accuracy of the predictions will be improved if there is a high fidelity between donor and recipient sites in terms of the covariates. The success of the MPS approach is equally reliant on having this high fidelity, however an operational issue with this approach is that there are some parameters to tune in order to generate the predictions, although Meerschman et al. (2013b) provide guidance on how to do this. Intuitively, the MPS approach may be a more acceptable alternative in situations where the donor area contains repeating landscape characteristics, an example being the spatial pattern of a drumlin landscape. Well-defined and repeating features seem to be a necessary input for MPS (Mariethoz et al., 2010; Meerschman et al., 2013a, 2013b), which may be comparatively overlooked if a point-based DSM approach were used for extrapolation. Further investigation will be necessary to determine whether this idea holds true however.

While Homosoil is framed in the context of global soil mapping efforts, it has been demonstrated in this study that it can be applied to a regional context. The study has been a scoping study about differing extrapolation methods. It has possibly been limited by the fact that a small donor area was used, meaning some rather large and possibly insurmountable assumptions about the donor site being 'representative' of the recipient site are being proposed. Nonetheless, in the context of this study, the Homosoil framework is but one line of research enquiry to ongoing investigations to derive detailed gamma radiometric mapping across the recipient site. Another is detailed in Stockmann et al. (2015). Future investigations in the Homosoil framework regarding the data infrastructure development in the region are to apply it using a patchwork (rather than one) of donor sites distributed across the area. This is likened to a gap filling exercise which could also have applications for studies at national and global extents.

5. Conclusions

The main outcomes of this research were:

1. We investigated the concept of Homosoil for spatial soil mapping across a data scarce area. Both DSM and MPS were used as different extrapolation methods to fulfil high resolution mapping in an otherwise data scarce area, using detailed information from a relatively small donor area.
2. We demonstrated an approach based on the Mahalanobis distance for assessing the similarity between donor and recipient sites. Once a threshold has been established to distinguish between similar and dissimilar, it may be used to constrain the extent of extrapolation. Nevertheless, while the similarity assessment and uncertainties of the extrapolation approaches were evaluated independently, our study demonstrated that areas of dissimilarity to the donor site have a relatively larger uncertainty compared to those areas that are similar.
3. From a limited validation area, it was demonstrated that Homosoil approaches for spatial mapping should not be used as a final deliverable, but as a 'first cut' to realising high information content digital soil mapping systems in otherwise data scarce areas.
4. Both the DSM and MPS approaches were comparable in terms of mapping the spatial pattern of gamma radiometric total count across the recipient site. The approaches are also comparable in terms of computational efficiency, taking into account that uncertainties of the predictions were also quantified — albeit for the DSM approach they are derived from a model, and for the MPS they are the quantified simulation response of varying the direct sampling of a training image/s. For the MPS approach however, there are a number of parameters to tune the algorithm. These may be optimised via a sensitivity analysis, of which was not carried out in this study, but may contribute to additional computation time. MPS could possibly be a good alternative compared to DSM for homosoil applications if there is a well-defined and repeating landscape feature observed in the donor site that is also present across a recipient site.

Acknowledgements

The authors thank: 1) Dr. Uta Stockmann (University of Sydney) for providing gamma radiometric data from the donor and validation site areas of this study; 2) Dr. Gregoire Mariethoz (University of Lausanne) for reviewing and providing advice to drafts of this manuscript; and 3) Prof. Cristine Morgan (Texas A&M University) who helped with the conceptualisation of this work.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.geoderma.2015.08.037.

References

Australian Government Bureau of Meteorology (BOM), 2014. Climate Statistics for Australian locations. Retrieved from http://www.bom.gov.au/climate/averages/tables/cw_061260.shtml.

Batjes, N.H., 2009. Harmonized soil profile data for applications at global and continental scales: updates to the WISE database. Soil Use Manag. 25 (2), 124–127.

Besag, J., 1986. On the statistical analysis of dirty pictures. J. Roy. Stat. Soc. Ser. B Methodol. 48 (3), 259–302.

Brevik, E.C., Hartemink, A.E., 2010. Early soil knowledge and the birth and development of soil science. Catena 83 (1), 23–33.

Chugunova, T.L., Hu, L.Y., 2008. Multiple-point simulations constrained by continuous auxiliary data. Math. Geosci. 40 (2), 133–146.

Comunian, A., Jha, S.K., Giambastiani, B.M.S., Mariethoz, G., Kelly, B.F.J., 2014. Training images from process-imitating methods. Math. Geosci. 46 (2), 241–260.

Dickson, B.L., Scott, K.M., 1997. Interpretation of aerial gamma-ray surveys: adding the geochemical factors. AGSO J. Aust. Geol. Geophys. 17 (2), 187–200.

Emery, X., Lantuéjoul, C., 2014. Can a training image be a substitute for a random field model? Math. Geosci. 46 (2), 133–147.

Ge, Y., Bai, H., 2010. MPS-based information extraction method for remotely sensed imagery: a comparison of fusion methods. Can. J. Remote. Sens. 36 (6), 763–779.

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.

Guardiano, F., Srivastava, M., 1993. Multivariate geostatistics: beyond bivariate moments. In: Soares, A. (Ed.), Geostatistics-Troia. Kluwer Acad, Dordrecht, Netherlands, pp. 133–144.

Hawley, S.P., Glen, R.A., Baker, C.J., 1995. Newcastle Coalfield Regional Geology 1:100 000. 1st edition. Geological Survey of New South Wales, Sydney, Australia.

Jenny, H.A., 1941. Factors of Soil Formation: A System of Quantitative Pedology. McGraw-Hill, New York.

Jha, S.K., Comunian, A., Mariethoz, G., Kelly, B.F.J., 2014. Parameterization of training images for aquifer 3-D facies modeling integrating geological interpretations and statistical inference. Water Resour. Res. 50 (10), 7731–7749.

Jha, S.K., Mariethoz, G., Evans, J.P., McCabe, M.F., 2013a. Demonstration of a geostatistical approach to physically consistent downscaling of climate modeling simulations. Water Resour. Res. 49 (1), 245–259.

Jha, S.K., Mariethoz, G., Evans, J., McCabe, M.F., Sharma, A., 2015. A space and time scale-dependent nonlinear geostatistical approach for downscaling daily precipitation and temperature. Water Resour. Res. http://dx.doi.org/10.1002/2014WR016729.

Jha, S.K., Mariethoz, G., Kelly, B., 2013b. Bathymetry fusion using multiple-point geostatistics: novelty and challenges in representing non-stationary bedforms. Environ. Model. Softw. 50, 66–76.

Lagacherie, P., Legros, J.P., Burrough, P.A., 1995. A soil survey procedure using the knowledge of soil pattern established on a previously mapped reference area. Geoderma 65 (3-4), 283–301.

Lagacherie, P., McBratney, A.B., 2007. Spatial soil information systems and spatial soil inference systems. In: Lagacherie, P., McBratney, A.B., Voltz, M. (Eds.), Digital Soil Mapping — An Introductory Perspective. Elsevier, Amsterdam, pp. 301–326.

Liu, Y.H., Harding, A., Abriel, W., Strebelle, S., 2004. Multiple-point simulation integrating wells, three-dimensional seismic data, and geology. AAPG Bull. 88 (7), 905–921.

Mahalanobis, P.C., 1936. On the generalised distance in statistics. Proc. Natl. Inst. Sci. India 2 (1), 49–55.

Mallavan, B.P., Minasny, B., McBratney, A.B., 2010. Homosoil: a methodology for quantitative extrapolation of soil information across the globe. In: Boettinger, J.L., Howell, D.W., More, A.C., Hartemink, A.E., Kienast-Brown, S. (Eds.), Digital Soil Mapping: Bridging Research, Environmental Application, and Operation. Springer, London, pp. 137–149.

Malone, B.P., Hughes, P., McBratney, A.B., Minasny, B., 2014a. A model for the identification of terrons in the Lower Hunter Valley, Australia. Geoderma Reg. 1, 31–47.

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

Malone, B.P., Minasny, B., Odgers, N.P., McBratney, A.B., 2014b. Using model averaging to combine soil property rasters from legacy soil maps and from point data. Geoderma 232–234, 34–44.

Mariethoz, G., McCabe, M.F., Renard, P., 2012. Spatiotemporal reconstruction of gaps in multivariate fields using the direct sampling approach. Water Resour. Res. 48 (10). http://dx.doi.org/10.1029/2012WR012115.

Mariethoz, G., Renard, P., Straubhaar, J., 2010. The Direct Sampling method to perform multiple-point geostatistical simulations. Water Resour. Res. 46 (11), W11536.

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

Meerschman, E., Pirot, G., Mariethoz, G., Straubhaar, J., Van Meirvenne, M., Renard, P., 2013b. A practical guide to performing multiple-point statistical simulations with the Direct Sampling algorithm. Comput. Geosci. 52, 307–324.

Meerschman, E., Van Meirvenne, M., Mariethoz, G., Islam, M.M., De Smedt, P., Van De Vijver, E., Saey, T., 2014. Using bivariate multiple-point statistics and proximal soil sensor data to map fossil ice-wedge polygons. Geoderma 213(0), 571–577.

Meerschman, E., Van Meirvenne, M., Van De Vijver, E., De Smedt, P., Islam, M.M., Saey, T., 2013a. Mapping complex soil patterns with multiple-point geostatistics. Eur. J. Soil Sci. 64 (2), 183–191.

Minasny, B., McBratney, A.B., 2010. Methodologies for global soil mapping. In: Boettinger, J.L., Howell, D.W., More, A.C., Hartemink, A.E., Kienast-Brown, S. (Eds.), Digital Soil Mapping: Bridging Research, Environmental Application, and Operation. Springer, London, pp. 429–436.

Mulder, V.L., de Bruin, S., Schaepman, M.E., Mayr, T.R., 2011. The use of remote sensing in soil and terrain mapping — a review. Geoderma 162 (1–2), 1–19.

Odgers, N.P., McBratney, A.B., Minasny, B., 2015. Digital soil property mapping and uncertainty estimation using soil class probability rasters. Geoderma 237–238(0), 190–198.

Podgorski, J.E., Green, A.G., Kalscheuer, T., Kinzelbach, W.K.H., Horstmeyer, H., Maurer, H., Rabenstein, L., Doetsch, J., Auken, E., Ngwisanyi, T., Tshoso, G., Jaba, B.C., Ntibinyane, O., Laletsang, K., 2015. Integrated interpretation of helicopter and ground-based geophysical data recorded within the Okavango Delta, Botswana. J. Appl. Geophys. 114, 52–67.

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

Sanchez, P.A., Ahamed, S., Carré, F., Hartemink, A.E., Hempel, J., Huising, J., Lagacherie, P., McBratney, A.B., McKenzie, N.J., Mendonça-Santos, M.d.L., Minasny, B., Montanarella, L., Okoth, P., Palm, C.A., Sachs, J.D., Shepherd, K.D., Vågen, T.-G., Vanlauwe, B., Walsh, M.G., Winowiecki, L.A., Zhang, G.-L., 2009. Digital soil map of the world. Science 325 (5941), 680–681.

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

Stockmann, U., Malone, B.P., McBratney, A.B., Minasny, B., 2015. Landscape-scale exploratory radiometric mapping using proximal soil sensing. Geoderma 239–240(0), 115–129.

Viscarra Rossel, R.A., McBratney, A.B., Minasny, B. (Eds.), 2010. Proximal Soil Sensing. Springer Netherlands, Netherlands.

Viscarra Rossel, R.A., Webster, R., Kidd, D., 2014. Mapping gamma radiation and its uncertainty from weathering products in a Tasmanian landscape with a proximal sensor and random forest kriging. Earth Surf. Process. Landf. 39 (6), 735–748.