HYDROLOGICAL PROCESSES
Hydrol. Process. (2014)
Published online in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/hyp.10383

Soil property variation mapping through data mining of soil category maps

Fei Du,[1] A-Xing Zhu,[2,3,4,1]* Lawrence Band[5] and Jing Liu[1]

[1] *Department of Geography, University of Wisconsin-Madison, Madison, WI 53706, USA*
[2] *School of Geographical Science, Nanjing Normal University, Nanjing, Jiangsu 210023, China*
[3] *Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing, Jiangsu 210023, China*
[4] *State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China*
[5] *Department of Geography and Institute for the Environment, University of North Carolina, Chapel Hill, NC 27599, USA*

Abstract:

Spatial information on soil properties is an important input to hydrological models. In current hydrological modelling practices, soil property information is often derived from soil category maps by the linking method in which a representative soil property value is linked to each soil polygon. Limited by the area-class nature of soil category maps, the derived soil property variation is discontinuous and less detailed than high resolution digital terrain or remote sensing data. This research proposed *dmSoil*, a data-mining-based approach to derive continuous and spatially detailed soil property information from soil category maps. First, the soil–environment relationships are extracted through data mining of a soil map. The similarity of the soil at each location to different soil types in the soil map is then estimated using the mined relationships. Prediction of soil property values at each location is made by combining the similarities of the soil at that location to different soil types and the representative soil property values of these soil types. The new approach was applied in the Raffelson Watershed and Pleasant Valley in the Driftless Area of Wisconsin, United States to map soil *A* horizon texture (in both areas) and depth to soil *C* horizon (in Pleasant Valley). The property maps from the *dmSoil* approach capture the spatial gradation and details of soil properties better than those from the linking method. The new approach also shows consistent accuracy improvement at validation points. In addition to the improved performances, the inputs for the *dmSoil* approach are easy to prepare, and the approach itself is simple to deploy. It provides an effective way to derive better soil property information from soil category maps for hydrological modelling.

KEY WORDS soil property; continuous variation; hydrological modelling; data mining; SoLIM; soil similarity model

Received 20 July 2014; Accepted 14 October 2014

INTRODUCTION

Information on spatial variation of soil properties is one of the key inputs for many hydrological models. Detailed spatial data on soil properties can better characterize soil conditions and have the potential to improve simulation accuracy (Zhu and Mackay, 2001; Bosch *et al*., 2004; Di Luzio *et al*., 2004; Chaplot, 2005; Anderson *et al*., 2006; Li *et al*., 2012). However, soil property data used in hydrological models is usually less detailed than other inputs (e.g. digital terrain data and remotely sensed vegetation data) as it often cannot provide pixel-level (continuous) spatial variation information. The mismatch in spatial details causes incompatibility between the soil information and other information (Band and Moore, 1995; Zhu, 1997). This incompatibility issue might be more evident in distributed hydrological modelling where spatially detailed simulation needs to be performed (Zhu and Mackay, 2001).

The detailed information on soil spatial variation can be derived through spatial prediction using field samples. Kriging techniques (Isaaks and Srivastava, 1989; Goovaerts, 1999) and machine learning/regression techniques (Moore *et al*., 1993; Bui *et al*., 2006) are commonly used for this purpose. Most of these methods, however, might require a large amount of field soil samples in order to build a meaningful predictive model. It is hard to obtain such sample sets for many hydrological modelling applications. Instead of using field samples, Zhu *et al.* (1996), Zhu (1999) and Zhu *et al*. (2001) developed the SoLIM approach to derive detailed soil spatial information based on fuzzy logic and expert knowledge. This approach, however, requires a knowledge acquisition process in which soil experts' knowledge on soil landscape relationship needs to be

*Correspondence to: A-Xing Zhu, Department of Geography, University of Wisconsin, 550 North Park Street, Madison, WI 53706, USA.
E-mail: azhu@wisc.edu

solicited. This is usually not a feasible solution for hydrologists. In addition to predictive methods, different sensing technologies are increasingly used to obtain soil spatial information directly. Remote sensing uses soil spectral reflectance to differentiate soil properties (Barnes *et al*., 2003; Mulder *et al*., 2011). Though it can obtain spatial information at large scale, there are still many challenges. For example, vegetation cover may obscure the soil response. In recent years, proximal soil sensing is becoming a promising technique in soil mapping. Proximal soil sensing involves the use of optical, geophysical and/or electromagnetic methods to directly measure soil properties (Rossel *et al*., 2010; Rossel *et al*., 2011). It has already gained great success in deriving high resolution soil property information for precision agriculture and some other fields. However scaling the technology to larger areas (e.g. watershed scale) is still a challenge.

In current hydrological modelling practices, soil category maps are the major sources to derive soil spatial information. A commonly used approach for obtaining soil property information from soil category maps is the linking method in which a representative soil property value is assigned to each soil polygon (Grossman *et al.*, 1992). The linking method has the advantage of simplicity, but the derived soil property information suffers a major problem: the soil property is implied to be uniform within each of the soil polygons and changes only occur at polygon boundaries. The derived spatial variation of soil properties is thus discontinuous and lacking spatial details. This drawback is mainly caused by the area-class nature of soil category maps. Many efforts have been made to generate more detailed soil category maps from existing ones (Moran and Bui, 2002; Qi and Zhu, 2003; Li *et al*., 2004; Kempen *et al*., 2009; Yang *et al*., 2011). However, the outputs are still area-class maps, so deriving soil property information from those maps still suffers the discontinuity issue. Research has also been done to generate soil property information from soil category maps using ensemble approaches. Instead of a single soil property value, an empirical or theoretical probability distribution is linked with a soil type. Based on the probability distributions, a statistical simulation is conducted to derive many realizations of soil property maps (Heuvelink and Webster, 2001; Webb and Lilburne, 2005; Loosvelt *et al*., 2011). A nice feature of this approach is it can provide uncertain information about the derived soil information. However, the derived soil property maps may look 'unnatural' as they are purely from statistical simulation. The spatial nature of the variation in a soil polygon is not explicit stated, and its relationships to environmental covariate are not captured.

This paper presents a new approach to characterize the spatial variation of soil properties from soil category maps. The approach, referred to as *dmSoil,* employs the similarity model (Zhu, 1997; Zhu *et al*., 2001) to represent the spatial gradation of soils in an area. The similarity model is populated by coupling soil environmental covariates with soil–environment relationships extracted through data mining of the soil category maps. The information represented in the similarity model is then combined with representative property values of the soil types in the area to map the spatial gradation of soil property (Zhu *et al*., 2010).

The major contribution of this paper is it provides a new approach to obtain spatially continuous soil property information from existing soil category maps. The required inputs for the new approach are readily available, and the procedure is easy to deploy for hydrologists. The derived soil property information is compatible with other inputs (e.g. digital terrain data, vegetation data) in hydrological modelling.

METHODOLOGY

Basic idea

Instead of assuming uniform soil properties within each soil polygon (the current assumption conveyed by use of the linking method commonly adopted in hydrological modelling), a raster data model is used to represent soil property variation at pixel level. The spatial resolution can be set to be the same as other inputs data (e.g. Digital Elevation Model) of hydrological models. Following the similarity model proposed by Zhu (1997), each location (pixel) of a study area holds a similarity vector with its elements representing a 0–1 similarity measure to each soil type present in a soil category map. The similarity of the soil at a location to a soil type is approximated by the degree to which the environmental condition of the location is optimal for the soil type (Zhu *et al*., 1996). The optimality of environmental conditions for a given soil type can be approximated by the environmental frequency curves constructed by overlaying the soil category map with environmental covariates (environmental variables that co-vary with soil), as shown in Figure 1. In other words different environmental frequencies can be taken as indication of the degree to which the environmental conditions are optimal for the soil type. The highest frequency corresponds to optimality value of 1. For a location, their optimality values at individual environmental variable level are determined based on the frequency curves. Those optimality values can then be integrated to get the overall environmental condition optimality. The overall optimality is used as the final similarity measure between the soil at the location and the soil type.

Once the similarity of the soil at a location to a set of prescribed soil types is determined, the value of a given soil property at the location can be computed using the

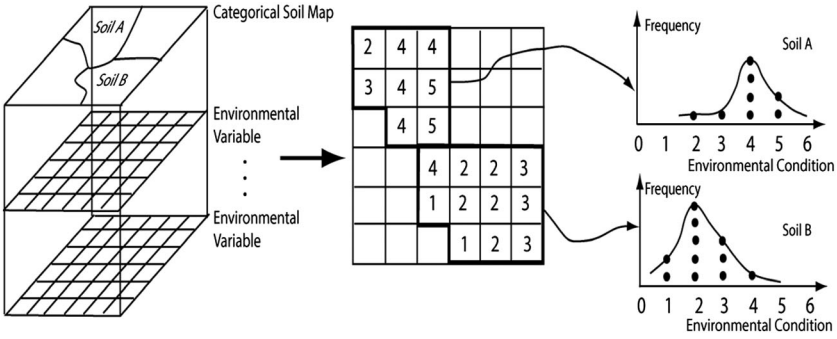

Figure 1. Data mining on soil category maps to extract the optimality of environmental conditions for each soil type

similarity weighted approach (Zhu *et al*., 2010). The combination of similarity representation of soils and the similarity weighted approach will provide the mechanism for a better approximation of soil spatial variation.

Procedure

The proposed *dmSoil* approach consists of the following major steps:

Preparation of environmental covariates. Based on Jenny's soil formation factors (Jenny, 1941) and the *scorpan* concept (McBratney *et al*., 2003), environment covariates include climate, organism, topography, parent material etc. The specific set of environment covariates used depends on the nature of the study area and the soil property to be predicted. For example, in study areas where topography plays a major role in soil formation, a Digital Elevation Model (DEM) and the derived terrain attributes from the DEM, such as slope, planform curvature, profile curvature and wetness index, can be used to characterize topographic conditions.

Data mining of soil category maps. As described in 'Basic idea', the goal for data mining of soil category maps is to model frequency distributions as optimality. Specifically, for a nominal (categorical) environmental variable, the dominant value (the category with the highest frequency) is regarded as optimal environmental value. For example, if the polygons of a soil type overlap with different parent material types, the modal parent material that dominates the polygons is regarded as the optimal parent material for the soil type.

For a continuous environmental variable, such as elevation, a curve can be used to mathematically model the frequency distribution. Both parametric method and non-parametric method could be used for this purpose:

Parametric method to fit frequency distribution for continuous variable: In the parametric method, the frequency distribution is usually assumed to follow a theoretical model, and the parameters are calibrated for that model. Many frequency distributions can be fitted using a Gaussian model, in which the only parameters needed are the mean and standard deviation. The Gaussian model is symmetric, which often does not approximate natural frequency distributions very well. To model asymmetric distributions, first the mode of the frequency distribution is used to separate the environmental values into two parts. Then the standard deviations from the mode are calculated for values below and above the mode. With the mode and the two standard deviations obtained, the frequency distribution on the left-hand side and that on the right-hand side of the mode can be approximated by the following equation:

$$\begin{cases} f(x) = \exp\left[\left(\dfrac{x-b}{\sqrt{2*\log(2)}*\sigma_l}\right)^2 * \log(0.5)\right] & (x \le b) \\ f(x) = \exp\left[\left(\dfrac{x-b}{\sqrt{2*\log(2)}*\sigma_r}\right)^2 * \log(0.5)\right] & (x \ge b) \end{cases} \quad (1)$$

where b is the modal value of the frequency distribution, σ_l is the standard deviation for the values below the mode and σ_r is the standard deviation for the values above the mode. Outputs from this function are within the range [0, 1]. The modal value (highest frequency) corresponds to value 1 which means it is assumed as the optimal environmental value.

Non-parametric method to fit frequency distribution for continuous variable: The frequency distribution of environmental variable from soil maps may not comply with the predefined shapes as assumed in parametric methods. The non-parametric method, such as Kernel Density Estimation (KDE), can be used to model the frequency distribution without assuming any curve shape in advance. The following equation shows the KDE method for frequency distribution fitting:

$$f(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) \quad (2)$$

where K is a kernel density function and h is bandwidth. Gaussian kernel is usually used as the kernel function and the 'rule of thumb' algorithm (Scott, 1992) can be used to determine h. The derived frequency values can be easily normalized to range [0, 1]. After the normalization, the environmental value with the highest frequency will get a value 1, which means it is the most optimal value.

Similarity calculation. With the data mining result, the similarity of the soil at a location to a soil type in the soil category map can be quantified. To calculate this similarity, the optimality of each environmental variable at the location to the soil type is determined first.

For a nominal (categorical) environmental variable, if the environmental value at a location matches the optimal environmental value of the soil type, the optimality is 1. Otherwise, the optimality is 0.

For a continuous environmental variable, the fitted frequency distribution curve from either parametric or non-parametric method is used. The optimality value for a location is the normalized frequency value (retrieved from the fitted curve) corresponding to the environmental value of the variable at the location.

Once the optimality values at individual environmental variable level are determined for a location, they are integrated to get the overall environmental optimality. As suggested by Zhu *et al*. (1996) and Zhu *et al*. (2001), the fuzzy MIN operator can be used here for the integration based on the limiting factor principle in ecology. That is, the minimum optimality among all environmental variables at the location is used as the integrated optimality of the environmental condition at the location to the soil type. The integrated optimality is used as the final similarity of the location to the soil type. The similarities of the soil at a location to every soil type can be calculated using this procedure.

Derivation of representative soil property value for each soil type. The representative soil property value for each soil type can be obtained from the associated description of the soil category maps. If they are not recorded in the description, regional to national soil databases are other sources to obtain those values. Furthermore, property values from typical field soil profiles can also be used if they are available. If the representative soil property is given as a range, the mid-point value can be used as the representative soil property value.

Soil property prediction. After the representative soil property values for different soil types are obtained, a similarity based weighting procedure is used to determine the soil property value at each location with the following equation (Zhu *et al*., 2010):

$$v_{ij} = \frac{\sum_{k=1}^{n} s_{ij}^{k} v^{k}}{\sum_{k=1}^{n} s_{ij}^{k}} \tag{3}$$

where v_{ij} is the property at location (i, j), v^k is the representative soil property value of soil type k, s_{ij}^k is the similarity to soil type k at location (i, j) and n is the total number of soil types in the soil category map.

CASE STUDIES

Study areas

Two study areas were selected to evaluate the ability of the *dmSoil* approach to map spatial variation of different soil properties. The location and topography of the two sites are shown in Figure 2. The first area is the Raffelson Watershed in La Crosse County, Wisconsin, United States. It is 1.6 km in the north/south direction and 2.2 km in the east/west direction. The second area, Pleasant Valley, is in Dane County, Wisconsin, United States. It is 3.2 km in the north/south direction and 3.7 km in the east/west direction. Both study areas are in the Driftless Area of Wisconsin, which has remained free of direct impact from late Pleistocene era continental glaciers. As a result, the two areas have a relatively mature topography and are typical ridge and valley terrain. The main reason for using two study areas is that the soil maps were produced by different soil surveyors. The comparison of the application of *dmSoil* between the two areas would provide information of the applicability of this approach. However, a more comprehensive analysis of the behavior of this approach across different landscape types is beyond the scope of this research.

The soil category maps of the two study areas were obtained from the NRCS Soil Survey Geographic database (SSURGO). SSURGO is the most detailed digital soil database in the United States. Data are stored in over 60 tables that provide information at three basic levels: map units, soil components and soil horizons. Each map unit has one or more components (soil series), and each component is associated with a few soil horizons. Map unit polygons provide the basic spatial units of the SSURGO dataset. In this research, the major component (soil series) of the map unit is used as the soil type of the map unit polygon. The soil category maps of the two study areas are shown in Figure 3.

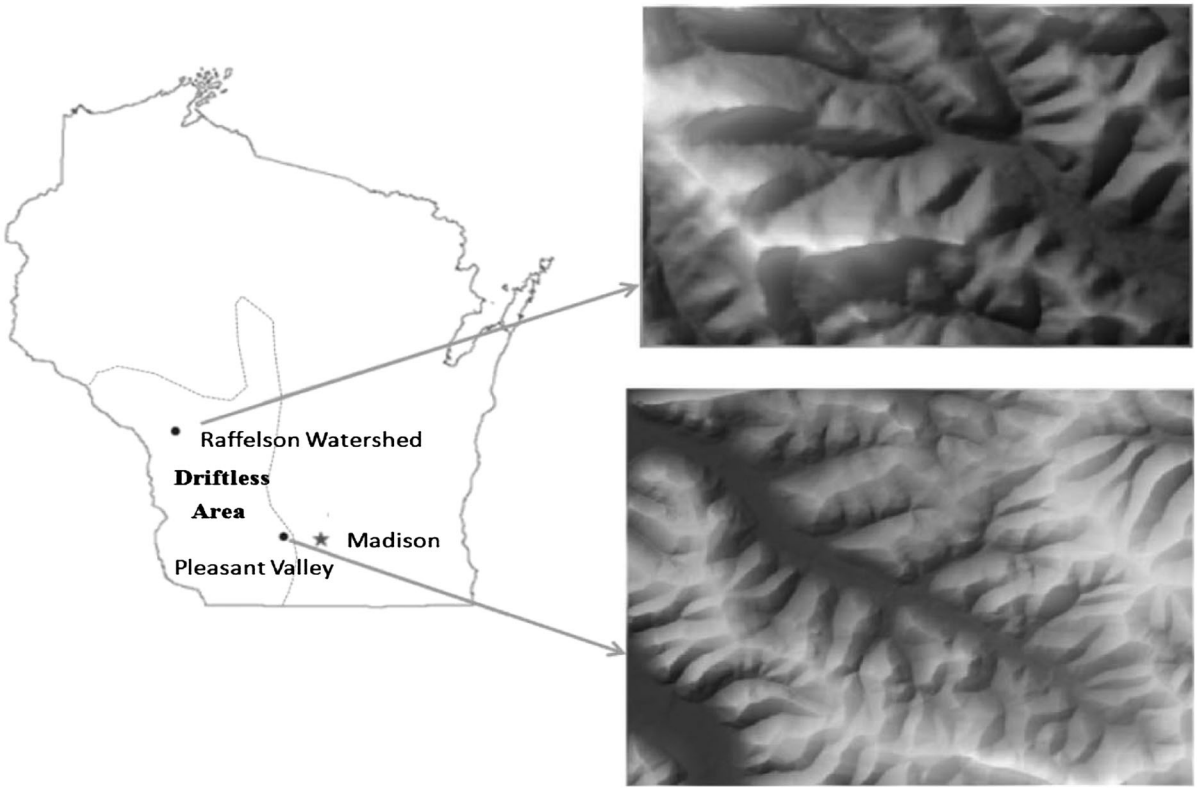

Figure 2. Locations and topography of two study areas

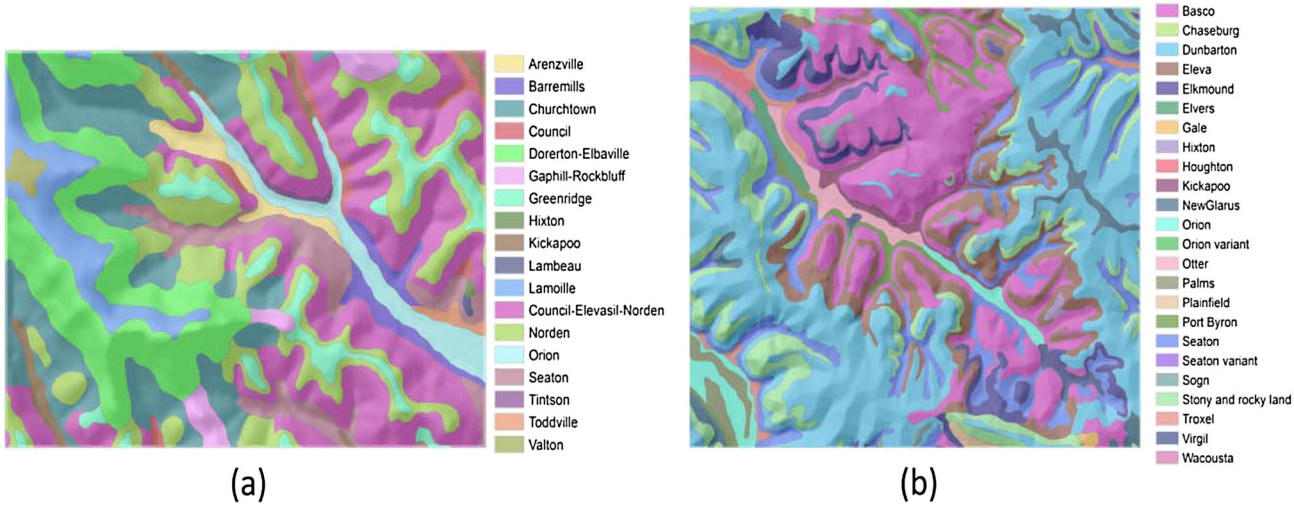

Figure 3. Soil category maps of the two study areas obtained from SSURGO database: (a) Raffelson Watershed; (b) Pleasant Valley

Experiments

Covariate selection. In both study areas, topography is the primary factor that determines the soil variation. This can be further verified by the SSURGO soil maps which show clear correlation between soil and terrain conditions. In the experiments 10-m resolution DEMs (obtained from Wisconsin Office of Natural Resource Conservation Service, United States Department of Agriculture) were used to generate slope gradient, planform curvature, profile curvature, topographic wetness index and slope aspect to characterize topography. Geological information (major bedrock geology breaks) was also included to depict soil

parent material conditions of the two study areas. This information can usually be obtained from geology maps and local soil survey reports. In this research, the geologic maps provided by the local soil scientist (Duane Simonson) in the Driftless Area were used.

Optimality modelling and similarity calculation. For each study area, the soil map was overlaid with the environmental covariate layers to conduct data mining. For categorical environmental variable (i.e. parent material), the dominant values were regarded as the optimal environmental values, and the value of each location is compared with the optimal type to determine the optimality. For continuous environmental variables (i.e. all the terrain attributes), both the parametric and non-parametric methods were experimented with to fit the frequency distributions. The fitted curves were then used to calculate the optimality.

After the optimality values at the individual environmental variable level are determined for each location, these values are integrated using the fuzzy MIN operator, as described in 'Similarity calculation', to obtain the overall environmental optimality. The integrated optimality is used as the final similarity of the soil at each location to the prescribed soil type.

Mapped soil properties. The soil properties mapped in the Raffelson Watershed are sand and silt percentage (soil texture) of soil *A* horizon. In Pleasant Valley, the soil properties mapped include sand, silt percentage of soil *A* horizon and depth to soil *C* horizon.

The representative soil property values were also obtained from the SSURGO database. For a soil complex with more than one major soil components (soil types), the representative value of each soil type was retrieved, and a weighted average approach was used to derive its final representative soil property value. The weights were the proportions of different soil types in the complex. In the Raffelson Watershed, the representative sand and silt percentages of *A* horizon were calculated based on the representative percentage of samples that passes #10 (2.0 mm) sieve and #200 (0.074 mm) sieves recorded in the SSURGO database. In Pleasant Valley, the representative sand and silt percentages of *A* horizon were calculated using the same method as the Raffelson Watershed, and the representative value for depth to *C* horizon was calculated by summing all the depth values above the *C* horizon.

Evaluation methods. The prediction results from the *dmSoil* approach were compared with that from the linking method. Visual evaluation was first conducted to see whether the soil property maps from the *dmSoil* approach could better capture the continuous soil property variation of the two study areas.

In addition to visual evaluation, field validation samples were used to quantitatively validate the results. In the Raffelson Watershed, a total of 49 field samples were collected. Large portion of the 49 samples were collected along a few transects to cover different landscape positions and different soil types. Laboratory analysis was conducted to determine the soil texture in the *A* horizon for all samples. In Pleasant Valley, a total of 48 samples had laboratory analysis done for soil texture in the *A* horizon, and 50 samples were collected for depth to the *C* horizon measured. Similar to the Raffelson Watershed, many samples were collected along a few transects to cross different landscape positions and different soil types. A few validation points fall in soil polygons labeled as 'stony and rocky land' in Pleasant Valley. Since there are no representative property values for those polygons, the linking method cannot predict soil properties at those sample locations. Therefore, those validation points were excluded from the comparison.

The field samples along transects were used to evaluate whether the *dmSoil* approach could better capture the continuous soil property variation along transects. The whole field sample sets were used to assess the prediction accuracy. Specifically, the root mean squared error

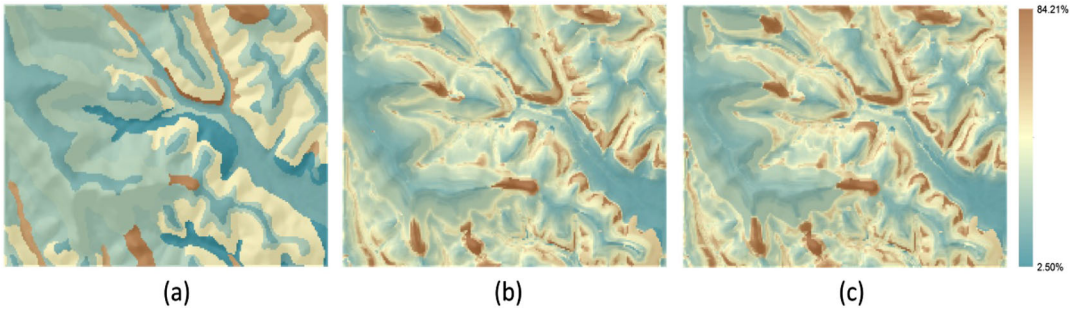

Figure 4. Maps of sand percentage of *A* horizon in the Raffelson Watershed from (a) the linking method, (b) the *dmSoil* approach with parametric curve fitting and (c) the *dmSoil* approach with non-parametric curve fitting

(RMSE) was calculated from field samples for both the *dmSoil* approach and the linking method. The smaller the RMSE, the better the prediction accuracy is. The equation of RMSE is as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (v_i - v'_i)^2}{n}} \quad (4)$$

where v_i is the ith observed value, v'_i is the ith predicted value and n is the number of validation samples. In addition to RMSE, a one-tailed paired samples t-test was performed for each of the mapped soil properties to see whether the prediction errors (absolute differences between the predicted and observed values) of the *dmSoil* approach are significantly lower than the prediction errors of the linking method in the field validation sets. The p values from the t-test were used for comparison.

Results and evaluation

Visual evaluation of the derived maps. The soil property maps generated from the *dmSoil* approach were visually compared with the maps from the linking method. Figure 4 shows the maps of predicted sand percentage of soil *A* horizon in the Raffelson Watershed. Though the general spatial patterns of the maps from the linking method and the *dmSoil* approach are similar, the results from the *dmSoil* approach have greater spatial continuity and more spatial details. The map from the linking method indicates that there is no variation of sand percentage within each of the polygons. This uniform distribution within a soil polygon cannot be realistic in this area. In contrast, the maps from the *dmSoil* approach clearly show spatial gradation of sand percentage within a polygon. The map from the linking method shows sharp changes at polygon boundaries, which is also not realistic in this area. The maps from the *dmSoil* approach exhibit continuous changes at the

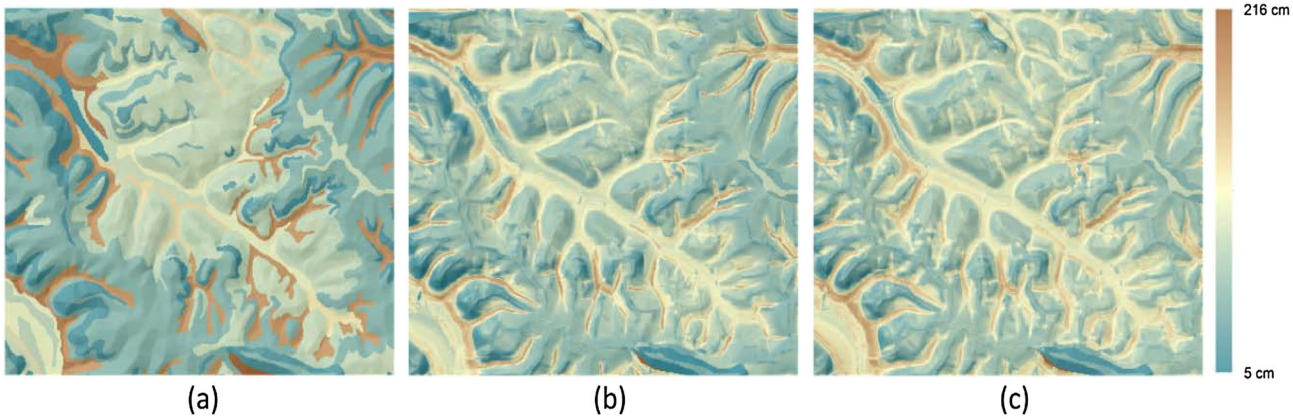

Figure 5. Maps of depth to *C* horizon in Pleasant Valley from (a) the linking method, (b) the *dmSoil* approach with parametric curve fitting and (c) the *dmSoil* approach with non-parametric curve fitting

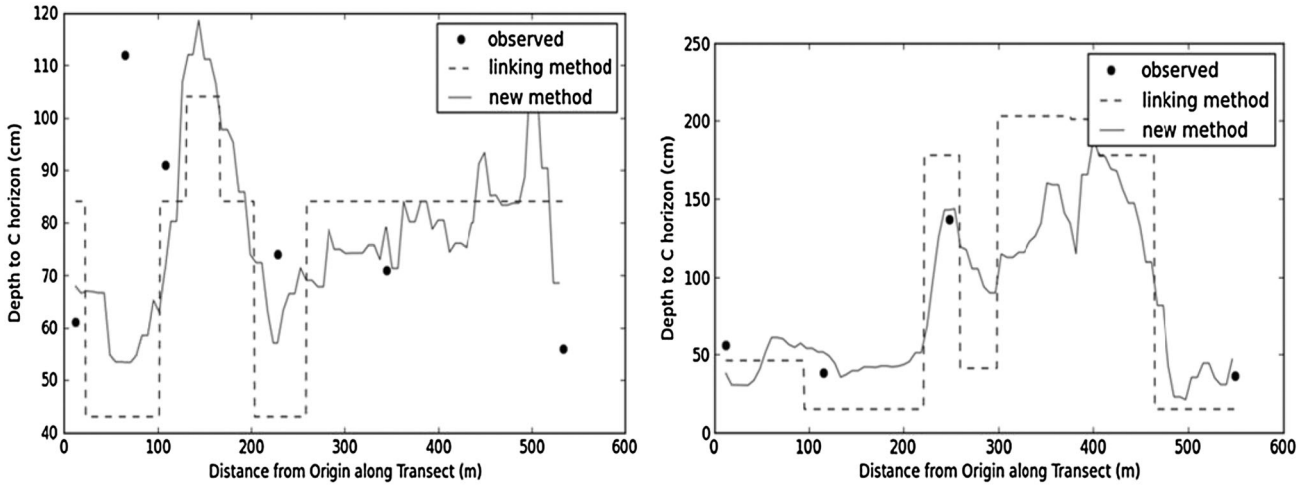

Figure 6. Observed *versus* predicted depth to *C* horizon at two transects in Pleasant Valley

polygon boundaries and the transitions between different neighbourhood soil polygons are captured. It can also be noted that the maps from the parametric curve fitting and non-parametric curve fitting exhibit no significant difference.

The maps of predicted depth to *C* horizon in Pleasant Valley are shown in Figure 5. Similar to Figure 4, the results from the *dmSoil* approach show a smoother variation of depth to *C* horizon from valleys to ridges. The small variations of depth on slopes, which are missing in the result from the linking method, are also captured. Again, no significant difference can be seen on the maps from the parametric curve fitting and non-parametric curve fitting.

Evaluation on spatial continuity along transects. As mentioned before, we have field samples collected along a few transects that cross different landscape positions. At those transects, we compared the observed soil property values at the sampled locations, the predictions from the linking method and the predictions from the *dmSoil* approach. As an example, the comparisons of depth to *C* horizon at two transects in Pleasant Valley are shown in Figure 6.

The predictions from the *dmSoil* approach appear to be more continuous while the predictions from linking method suffer from abrupt changes in boundary and no variations within a soil type. The predictions from the *dmSoil* approach are generally closer to the field observed values. This indicates the *dmSoil* approach does a better job at approximating the real-world soil property variation than the linking method.

Assessment of prediction accuracy with all validation samples. The RMSE from the linking method and the *dmSoil* approach for the field validation samples are shown in Table I. The *dmSoil* approach always yields lower RMSE compared to the linking method. The accuracies of *dmSoil* approach on field validation samples are consistently higher than that of the linking method for different mapped soil properties. Between the two curve fitting method (parametric and non-parametric), no consistent trend, i.e. which curve fitting method consistently yields better result, can be observed from the table.

Table II shows the results from the one-tailed paired samples *t*-test. For sand and silt percentage of soil *A* horizon in Raffelson Watershed, the prediction errors of the *dmSoil* approach are significantly lower than those of the linking method at the 95% confidence level. In Pleasant Valley, the differences between the prediction errors of the *dmSoil* approach and those of the linking method for the sand and silt percentage in soil *A* horizon are not statistically significant (p values >0.05). For depth to *C* horizon, however, the prediction error is significantly lower than those of the linking method at the 95% confidence level.

In order to make a more detailed comparison, scatterplots of observed *versus* predicted values at validation sample locations were generated for the two study areas. Figure 7 shows the scatterplots of observed *versus* predicted depth to *C* horizon from the linking method and the *dmSoil* approach (non-parametric) in Pleasant Valley. The scatterplot from the linking method shows apparent discontinuities and the lack of within-class variation (dots distribute along horizontal lines). In contrast, the scatterplot from the *dmSoil* approach shows

Table I. RMSE on the validation samples: the linking method *versus* the *dmSoil* approach

	Raffelson Watershed		Pleasant Valley		
	Sand in *A* Horizon (%)	Silt in *A* Horizon (%)	Sand in *A* Horizon (%)	Silt in *A* Horizon (%)	Depth to *C* Horizon (cm)
Linking method	19.68	17.86	24.80	25.92	39.48
dmSoil (parametric)	17.43	15.78	20.47	20.95	28.83
dmSoil (non-parametric)	17.37	15.64	21.67	22.02	26.73

Table II. One-tailed paired samples *t*-test results (p values) for comparing the prediction errors of the linking method and the *dmSoil* approach on validation samples

	Raffelson Watershed		Pleasant Valley		
	Sand in *A* horizon	Silt in *A* horizon	Sand in *A* horizon	Silt in *A* horizon	Depth to *C* horizon
dmSoil (parametric) *versus* linking method	0.0081	0.0016	0.2526	0.0972	0.0170
dmSoil (non-parametric) *versus* linking method	0.0152	0.0010	0.4685	0.2231	0.0123

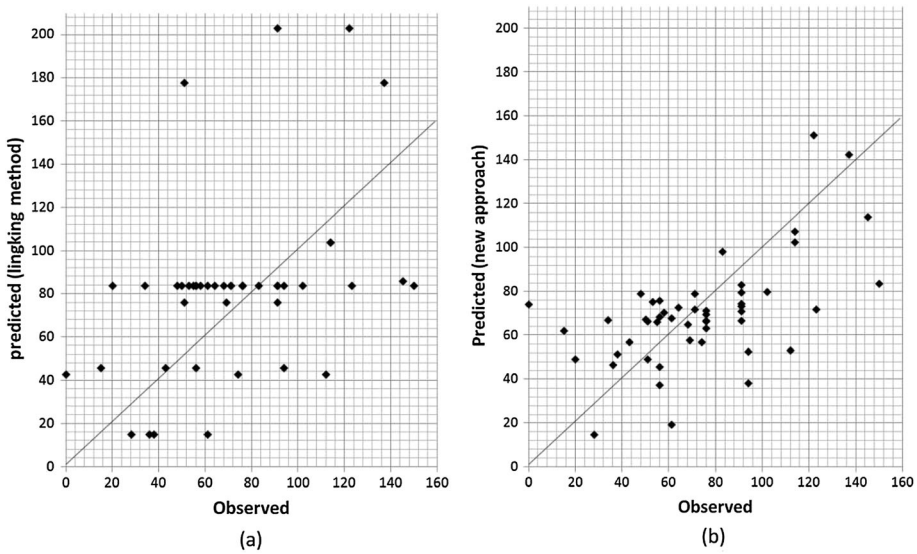

Figure 7. Scatterplots of observed *versus* predicted depth to *C* horizon at validation sample locations in Pleasant Valley: (a) the linking method; (b) the *dmSoil* approach with non-parametric curve fitting

stronger correlation between the observed and predicted soil property values.

DISCUSSION

Analysis of the results

The case study results demonstrate the soil property information from the *dmSoil* approach is able to better capture the continuous variation nature of soil properties and to improve prediction accuracy in terms of RMSE. The *t*-test results in Raffelson watershed show that the prediction errors from the *dmSoil* approach are significantly lower than those from the linking method for both the sand and silt percentage of soil *A* horizon. In Pleasant Valley, the accuracy improvement is statistically significant for depth to *C* horizon but not for sand and silt percentage of soil *A* horizon. One potential reason is that the soil *A* horizon texture is relatively more prone to be affected by human activities (e.g. cultivation) than soil *C* horizon.

In the experiments, both parametric and non-parametric methods were tested to fit the frequency distribution for continuous variables when conducting the data mining. The results did not show an apparent difference between the two. It was originally expected that the non-parametric method, as it does not assume the shape and can better approximate the frequency decay, would outperform the parametric method. One possible reason that they show comparable performance is the non-parametric method is more sensitive to noise in the data while the parametric method has some generalization effects and filters out most of small and unnecessary details. The advantage of the parametric method is it is less computationally intensive, especially for large datasets.

It can be seen from the scatterplots in Figure 7 that there is an overestimation (more points lie above 45 degree line) when the observed values are small and an underestimation (more points lie below 45 degree line) when the observed values are large. The major reason is the smoothing effect from the similarity weighted approach. For each location, we use its similarities to different soil types as weights to derive the final soil property. Due to the nature of this weighting procedure, the predicted values tend to be smoothed out, which leads to over-estimation for small values and under-estimation for large values.

The impacts of input data on the results

The major inputs to the *dmSoil* approach are environmental covariate data and soil category maps. Both of them have big impact on the prediction result.

As we use environmental optimality to indicate soil similarities, the selected environmental variables should be good indicators to soil variations. For the two study areas, a clear correlation can be found between the soil variation and topography. Therefore, it is important to include terrain attributes into the environmental variable set. Either missing important environmental variables or including unnecessary ones will deteriorate the results.

The soil category map should be able to capture the environmental conditions that each soil type corresponds to. This requires that the soil category map was generated based on a soil-landscape model (Hudson, 1992) and the delineated soil polygons could characterize the soil type variation to a large extent. The soil category maps used in this research come from the SSURGO database. Though there are some inaccuracies/errors due to map generali-

zation (e.g. smooth polygon boundaries) and manual boundary delineation in soil survey (e.g. some mismatches between soil polygons and terrain positions), the soil maps are in general of good quality, which ensures the improvement of the *dmSoil* approach compared to the linking method.

In soil category maps, some map units contain a mixture of more than one major soil types. They are represented as soil complexes. The existence of soil complexes affected our results in two aspects. First, when calculating the similarities, we had to treat a soil complex as a 'soil type' and extract the environmental conditions for the soil complex instead of single soil types. Each soil type in the soil complex, however, has its own niche and their environmental condition might be quite different from each other. The second problem comes from the derivation of representative soil property values. For a soil complex, the weighted average approach was used to obtain the final representative soil property values where the weights are proportions of different soil types in the soil complex. The final representative values are not really 'representative' but only reflect average status. In this research, we mainly focused on soil category maps with map units dominated by a single component (soil type) and did not address these issues. One possible solution is to apply spatial disaggregation (Bui and Moran, 2001; Häring *et al.*, 2012; Nauman and Thompson, 2014) to spatially differentiate different soil type in a soil complex before data mining. Using the disaggregated soil category maps is likely to yield better prediction results.

Uncertainty analysis

To evaluate the results, the derived soil property values at field sample locations from the *dmSoil* approach were compared with those from the linking method. It should be realized that the derived soil property values contain prediction uncertainties, so the value-to-value comparisons (e.g. Figures 6 and 7) also have uncertainties.

The *dmSoil* approach relies on the data mining result (the optimality of environmental conditions for each soil type) and the representative soil property value for each soil type to map the continuous soil property variation. Both of them would introduce uncertainties to the prediction results.

In the data mining process, it is assumed that the environmental frequency curves constructed from all the pixels within a soil type can approximate the optimality of environmental conditions for the soil type. However, the pixels of a soil type in the soil category map may not be able to well represent the population of the soil type. Therefore there is uncertainty when using the environmental frequency curve to approximate the optimality.

The representative soil property value for each soil type is a single value. In fact, often times the representative soil property value of a soil type is given as a value range. Though we can use the min-point value of the range as the representative soil property value, the actual representative soil property value can be any value within the range. This brings uncertainty to the derived representative soil property values.

Both uncertainties can be quantified using a simulation approach. In the data mining process, based on the bootstrapping idea in statistics (Mooney and Duval, 1993), multiple rounds of random sampling (with replacement) on the pixels within a soil type can be conducted and an ensemble of frequency curves can be derived. This ensemble of frequency curves could capture the uncertainty in data mining procedure. Similarly, a random sampling can be conducted to generate an ensemble of representative soil property values within the range of possible representative values. This ensemble of values could capture the uncertainty of representative soil property values. The ensemble of frequency curves and representative soil property values can then be fed into the similarity weighted procedure to generate multiple realizations of soil property maps, which could provide a measure of final prediction uncertainty.

Implication for hydrological modelling

The *dmSoil* approach provides an easy and efficient way to derive information on continuous soil property variation for hydrological modelling. The only input data are soil category maps and environmental covariates that indicate soil spatial variation. Soil category maps are already widely used by hydrologists. Environmental covariates can usually be easily obtained from a DEM and/or remote sensing data of the study area.

The derived soil property information from the *dmSoil* approach is compatible with remote sensing data and other input data for hydrological models, such as topographic data, in terms of data model (raster) and spatial resolution. More importantly, it shows better co-variation with other parameters (such as slope gradient) of hydrological models. As an example, the co-variation of depth to *C* horizon with slope gradient along two transects in Pleasant Valley are shown in Figure 8. The transect in Figure 8(a) is composed of a valley (left), a big slope (middle) and a smaller slope (right). It can be seen that the slope gradient change on the two slopes does not have an impact on depth to *C* horizon derived from the linking method. In contrast, the depth variation from the *dmSoil* approach shows a negative correlation with slope gradient. Slope gradient indicates the strength of erosion. The greater the slope gradient is, the stronger the erosion and the thinner the soil would be. This negative correlation agrees with current understanding of the physical processes. In Figure 8(b), the depth from the linking method shows no variation (middle)

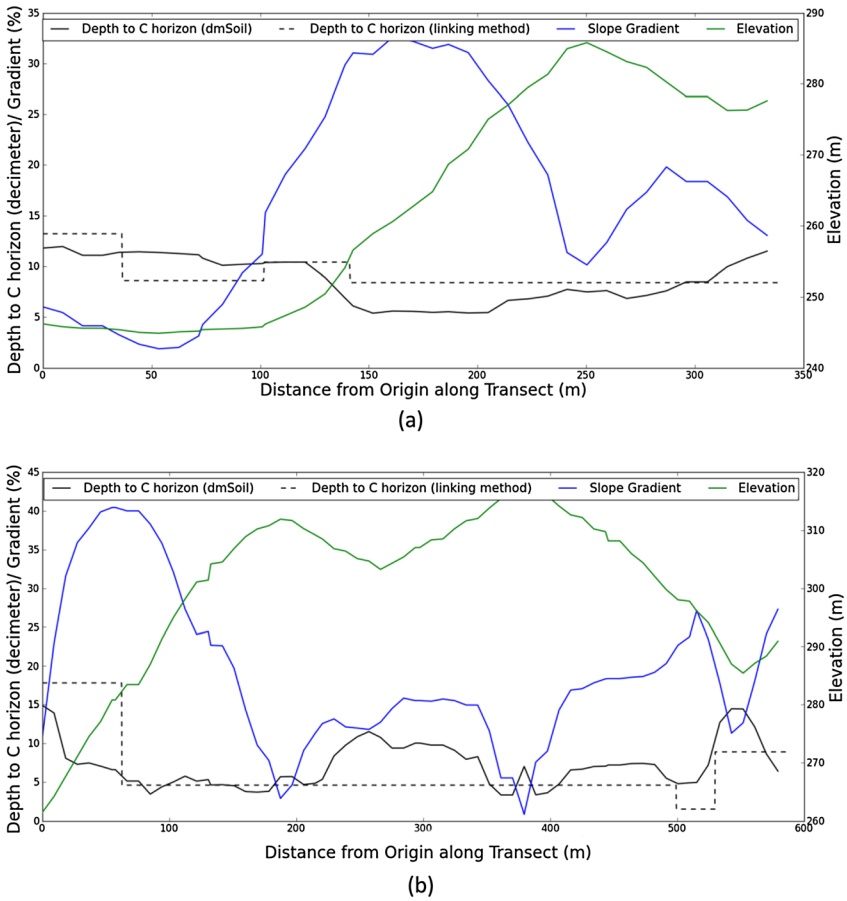

Figure 8. Co-variations of predicted depth to soil *C* horizon with slope gradient along two transects in Pleasant Valley. Slope gradient change (blue line) does not have a clear impact on depth to *C* horizon derived from the linking method (dashed black line). The predicted depth to soil *C* horizon from the *dmSoil* approach (solid black line) has a clear co-variation with slope gradient

even though dramatic slope gradient change can be observed. Similar to Figure 8(a), the depth from the *dmSoil* approach shows a clear co-variation with slope gradient.

The co-variation has a big impact on hydrological models. Soil depth, together with soil saturated hydraulic conductivity, decides soil transmissivity. Slope gradient is usually used to approximate the effective hydraulic gradient of the saturated zone. Therefore, the co-variation between soil depth and slope gradient is central to a landscape level solution to Darcy's Equation in hydrological modelling. The characterization of the spatial co-variations direct influences the approximation of shallow through flow and other hydrologic fluxes.

Take TOPMODEL (Beven *et al*., 1995; Beven, 1997) as an example, the soil depth, soil saturated hydraulic conductivity, slope gradient and upslope contributing area (which can be obtained from topographic data) are used to calculate the soil-topographic index (Beven, 1986; Quinn *et al*., 1995; Ambroise *et al*., 1996). This index is applied to model the relationship between lateral moisture flux and lateral flow capacity. It measures the soil moisture deficit which is directly related to the run-off generation. A better characterization of the co-variation between soil depth and slope gradient can help to better depict the spatial pattern of soil moisture deficit and therefore has the potential to improve the simulation results.

The *dmSoil* approach has already been partially implemented as a module in EcohydroLib which is series of Python scripts for performing eco-hydrology data preparation workflows. The library is available to hydrological modelling community through GitHub https://github.com/selimnairb/EcohydroLib (Miles and Band, 2013).

Mapping continuous soil property variations

This paper provides a way to derive information on continuous soil property variation from existing soil category maps without extensive sampling and knowledge acquisition efforts. Though the initial goal is to provide detailed soil property information that is compatible with other inputs in hydrological modelling, the *dmSoil* approach can also be used in other fields that need detailed soil information, such as ecological modelling.

Soil property mapping is the major focus of many soil mapping efforts that are currently carried out in the world, such as GlobalSoilMap.net (Sanchez *et al.*, 2009; Hartemink *et al.*, 2010). As mentioned earlier, research work has been done on field sample based spatial prediction, expert knowledge based spatial prediction, direct observation using remote sensing and proximal sensing technologies as well as deriving soil property information from legacy soil maps. This research contributes a new method to make use of legacy soil category maps for mapping continuous soil property variation.

Existing methods of using legacy soil category maps for mapping soil property mainly include linking a representative soil property value to each soil type (referred to as the linking method in this research) and linking a soil property distribution to each soil type. The problem of linking a single representative value is the discontinuity issue. Linking a probability distribution of soil property to each soil type can generate ensemble realizations of soil property maps, which can capture the prediction uncertainties. However, it is a statistical simulation that purely relies on the probability distribution of soil properties. The derived maps may look very 'unnatural'.

Soils often show strong correlations with other environmental variables. With the development of spatial data capture techniques (e.g. remote sensing), high quality environmental data are increasingly available. Different from existing methods of using existing soil maps, the *dmSoil* approach makes use of these environmental covariate data together with soil category map for mapping soil properties. In the new method, soil–environmental relationships are extracted from soil category maps and an environmental similarity weighted procedure is used to predict the soil property at each location. The generated soil property maps can capture the continuous spatial gradation of soil properties. They have better correlation with soil environmental covariates and show natural spatial autocorrelation effect (smooth transitions) due to the spatial autocorrelation of environmental covariate data (e.g. continuous elevation change). In addition, as discussed in 'Uncertainty analysis' the uncertainties of the *dmSoil* approach can also be quantified though a simulation based approach (generating multiple realizations of soil property maps using an ensemble of frequency curves and representative soil property values). The *dmSoil* approach provides an effective method to map continues soil property information if soils in a study area have clear correlations with some other environmental variables.

Different from the approach presented in this paper, which is mainly based on soil–environment relationship, a recent work on area-to-point Kriging (Kerry *et al.*, 2012) uses the Geostatitical framework to predict soil property variation from soil category maps. A semivariogram is estimated from areal data (soil category maps) to model the soil continuum. Further study should be conducted to compare the *dmSoil* approach proposed in this research and the Kriging method.

CONCLUSION

This paper presented *dmSoil*, a new approach to map the continuous spatial variation of soil properties from soil category maps to overcome the problem of the linking method that is commonly used in hydrological modelling. The *dmSoil* approach extracts soil environment relationships from soil category maps using data mining techniques. Based on the data mining result, a similarity model is applied to map the variation of soil properties. Compared to the soil property map derived from the linking method, the prediction from the *dmSoil* approach can better capture the variations within a polygon as well as the transitions between polygons. The spatial continuity and level of spatial details of the derived soil property information are improved. In addition, the inputs of the *dmSoil* approach are readily available for hydrologists. The approach itself is easy to be deployed and integrated in current model workflows. The proposed approach provides an efficient way to derive quality soil property information for hydrological modelling.

ACKNOWLEDGEMENTS

Funding for the research reported in this paper was provided by the National Natural Science Foundation of China (Project Nos. 41023010 and 41431177), Program of International S&T Cooperation, Ministry of Science and Technology of China (Project No. 2010DFB24140), Director's Basic Research Fund of Institute of Geographic Sciences and Natural Resources, Chinese Academy of Sciences (Project No. Y3W30020YZ), Natural Science Research Program of Jiangsu (Project No. 14KJA170001) and PAPD. The support received by A-Xing Zhu through the Vilas Associate Award, the Hammel Faculty Fellow, and the Manasse Chair Professorship from the University of Wisconsin-Madison and through the 'One-Thousand Talents' Program of China is greatly appreciated.

REFERENCES

Ambroise B, Beven K, Freer J. 1996. Toward a generalization of the TOPMODEL concepts: Topographic indices of hydrological similarity. *Water Resources Research* **32**(7): 2135–2145.

Anderson RM, Koren VI, Reed SM. 2006. Using SSURGO data to improve Sacramento Model a priori parameter estimates. *Journal of Hydrology* **320**(1): 103–116.

Band LE, Moore ID. 1995. Scale: landscape attributes and geographical information systems. *Hydrological Processes* **9**(3-4): 401–422.

Barnes EM, Sudduth KA, Hummel JW, Lesch SM, Corwin DL, Yang CH, Daughtry C, Bausch WC. 2003. Remote- and ground-based sensor

techniques to map soil properties. *Photogrammetric Engineering & Remote Sensing* **69**(6): 619–630.
Beven K. 1986. Runoff production and flood frequency in catchments of order n: an alternative approach. In *Scale Problems in Hydrology*. Springer: Netherlands; 107–131.
Beven K. 1997. TOPMODEL: a critique. *Hydrological Processes* **11**(9): 1069–1085.
Beven K, Lamb R, Quinn P, Romanowicz R, Freer J, Singh VP. 1995. Topmodel. In *Computer models of watershed hydrology*, Singh VP (ed). Water Resources Publication: Colorado; 627–668.
Bosch DD, Sheridan JM, Batten HL, Arnold JG. 2004. Evaluation of the SWAT model on a coastal plain agricultural watershed. *Transactions of the ASAE* **47**(5): 1493–1506.
Bui EN, Henderson BL, Viergever K. 2006. Knowledge discovery from models of soil properties developed through data mining. *Ecological Modelling* **191**(3): 431–446.
Bui EN, Moran CJ. 2001. Disaggregation of polygons of surficial geology and soil maps using spatial modelling and legacy data. *Geoderma* **103**(1): 79–94.
Chaplot V. 2005. Impact of DEM mesh size and soil map scale on SWAT runoff, sediment, and NO_3-N loads predictions. *Journal of Hydrology* **312**(1): 207–222.
Di Luzio M, Arnold JG, Srinivasan R. 2004. Integration of SSURGO maps and soil parameters within a geographic information system and nonpoint source pollution model system. *Journal of Soil and Water Conservation* **59**(4): 123–133.
Goovaerts P. 1999. Geostatistics in soil science: state-of-the-art and perspectives. *Geoderma* **89**(1): 1–45.
Grossman RB, Benham EC, Fortner JR, Waltman SW, Kimble JM, Branham CE. 1992. A demonstration of the use of soil survey information to obtain areal estimates of organic carbon. *Proceedings Resource Technology* **92**: 457–467.
Häring T, Dietz E, Osenstetter S, Koschitzki T, Schröder B. 2012. Spatial disaggregation of complex soil map units: A decision-tree based approach in Bavarian forest soils. *Geoderma* **185**: 37–47.
Hartemink AE, Hempel J, Lagacherie P, McBratney A, McKenzie N, MacMillan RA, Zhang GL. 2010. GlobalSoilMap.net–a new digital soil map of the world. In *Digital Soil Mapping*. Springer: Netherlands; 423–428.
Heuvelink GBM, Webster R. 2001. Modelling soil variation: past, present, and future. *Geoderma* **100**(3): 269–301.
Hudson BD. 1992. The soil survey as paradigm-based science. *Soil Science Society of America Journal* **56**(3): 836–841.
Isaaks EH, Srivastava RM. 1989. *Applied Geostatistics*. Oxford University Press: New York.
Jenny H. 1941. *Factors of Soil Formation: a System of Quantitative Pedology*. McGraw Hill Book Company: New York.
Kempen B, Brus DJ, Heuvelink G, Stoorvogel JJ. 2009. Updating the 1: 50,000 Dutch soil map using legacy soil data: A multinomial logistic regression approach. *Geoderma* **151**(3): 311–326.
Kerry R, Goovaerts P, Rawlins BG, Marchant, BP. 2012. Disaggregation of legacy soil data using area to point kriging for mapping soil organic carbon at the regional scale. *Geoderma* **170**: 347–358.
Li W, Zhang C, Burt JE, Zhu A, Feyen J. 2004. Two-dimensional Markov chain simulation of soil type spatial distribution. *Soil Science Society of America Journal* **68**(5): 1479–1490.
Li R, Zhu A, Song X, Li B, Pei T, Qin C. 2012. Effects of spatial aggregation of soil spatial information on watershed hydrological modelling. *Hydrological Processes* **26**(9): 1390–1404.
Loosvelt L, Pauwels V, Cornelis WM, De Lannoy GJ, Verhoest NE. 2011. Impact of soil hydraulic parameter uncertainty on soil moisture modeling. *Water Resources Research* **47**(3): W03505, doi: 10.1029/2010WR009204.
McBratney AB, Mendonça Santos ML, Minasny B. 2003. On digital soil mapping. *Geoderma* **117**(1): 3–52.
Miles B, Band L. 2013. Toward a common framework for ecohydrology model data preparation workflows: EcohydroWorkflowLib. Oral presentation. 2013 CUAHSI Conference on Hydroinformatics and Modelling, Logan, UT.
Mooney CZ, Duval RD. 1993. *Bootstrapping: A nonparametric approach to statistical inference*. Sage: Newbury Park, California.
Moore ID, Gessler PE, Nielsen GAE, Peterson GA. 1993. Soil attribute prediction using terrain analysis. *Soil Science Society of America Journal* **57**(2): 443–452.
Moran CJ, Bui EN. 2002. Spatial data mining for enhanced soil map modelling. *International Journal of Geographical Information Science* **16**(6): 533–549.
Mulder VL, de Bruin S, Schaepman ME, Mayr TR. 2011. The use of remote sensing in soil and terrain mapping - A review. *Geoderma* **162**(1-2): 1–19.
Nauman TW, Thompson JA. 2014. Semi-automated disaggregation of conventional soil maps using knowledge driven data mining and classification trees. *Geoderma* **213**: 385–399.
Qi F, Zhu A. 2003. Knowledge discovery from soil maps using inductive learning. *International Journal of Geographical Information Science* **17**(8): 771–795.
Quinn P, Beven K, Lamb R. 1995. The In (a/tan/β) index: How to calculate it and how to use it within the Topmodel framework. *Hydrological Processes* **9**(2): 161–182.
Rossel RV, Adamchuk VI, Sudduth KA, McKenzie NJ, Lobsey C. 2011. Proximal soil sensing: an effective approach for soil measurements in space and time. *Advances in Agronomy* **113**(113): 237–282.
Rossel RV, McBratney AB, Minasny B (Eds). 2010. *Proximal soil sensing*, vol. **1**. Springer: Netherlands.
Sanchez PA, Ahamed S, Carré F, Hartemink AE, Hempel J, Huising J, Zhang GL. 2009. Digital soil map of the world. *Science* **325**(5941): 680–681.
Scott DW. 1992. *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons: New York, Chicester.
Webb TH, Lilburne LR. 2005. Consequences of soil map unit uncertainty on environmental risk assessment. *Soil Research* **43**(2): 119–126.
Yang L, Jiao Y, Fahmy S, Zhu A, Hann S, Burt JE, Qi F. 2011. Updating conventional soil maps through digital soil mapping. *Soil Science Society of America Journal* **75**(3): 1044–1053.
Zhu A. 1997. A similarity model for representing soil spatial information. *Geoderma* **77**(2): 217–242.
Zhu A. 1999. A personal construct-based knowledge acquisition process for natural resource mapping. *International Journal of Geographical Information Science* **13**(2): 119–141.
Zhu A, Band LE, Dutton B, Nimlos TJ. 1996. Automated soil inference under fuzzy logic. *Ecological Modelling* **90**(2): 123–145.
Zhu A, Hudson B, Burt JE, Lubich K, Simonson D. 2001. Soil mapping using GIS, expert knowledge, and fuzzy logic. *Soil Science Society of America Journal* **65**(5): 1463–1472.
Zhu A, Mackay DS. 2001. Effects of spatial detail of soil information on watershed modelling. *Journal of Hydrology* **248**(1): 54–77.
Zhu A, Qi F, Moore A, Burt JE. 2010. Prediction of Soil Properties Using Fuzzy Membership. *Geoderma* **158**: 199–206.

