Geoderma 158 (2010) 199–206

Contents lists available at ScienceDirect

Geoderma

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

Prediction of soil properties using fuzzy membership values

A-Xing Zhu [a,b], Feng Qi [c,*], Amanda Moore [d], James E. Burt [b]

[a] State Key Lab of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, 100101, China
[b] Department of Geography, University of Wisconsin-Madison, Madison, Wisconsin 53706, USA
[c] Department of Geology and Meteorology, Kean University, 1000 Morris Ave., Union, New Jersey 07083, USA
[d] USDA-NRCS, 339 Busch's Frontage Road, Suite 301, Annapolis, Maryland, 21409, USA

ARTICLE INFO

Article history:
Received 5 October 2009
Received in revised form 17 February 2010
Accepted 5 May 2010
Available online 13 June 2010

Keywords:
Soil property
Fuzzy membership
Weighted average
Regression

ABSTRACT

Detailed information on the spatial variation of soils is desirable for many agricultural and environmental applications. This research explores three approaches that use soil fuzzy membership values to predict detailed spatial variation of soil properties. The first two are weighted average models with which the soil property value at a location is the average of the typical soil property values of the soil types weighted by fuzzy membership values. We compared two options to determine the typical property values: one that uses the representative values from existing soil survey and the other that uses the property value of a field observation typical of a soil type. The third approach is a multiple linear regression in which the soil property value at a location is predicted using a regression between the soil property and fuzzy membership values. We compared this to multiple linear regression with environmental variables. In a case study in the Driftless Area of Wisconsin, the models were also compared with a predictive model based on existing soil survey. The results showed that regression with environmental variables works well for areas where the soil–terrain relationship is relatively simple but regression with fuzzy membership values is an improvement for areas where soil–terrain relationships are more complicated. From the perspectives of data requirement and model simplicity as well as accuracy of prediction the weighted average with maximum fuzzy membership option has obvious advantages.

1. Introduction

Spatial distribution of soil properties provides essential information for agricultural and environmental management applications. The physical and chemical properties of soils are commonly documented in soil surveys at the soil component level and can be mapped to display spatial distributions of such properties. Soil property maps generated from conventional soil survey maps, however, are no longer sufficient in many cases because they often do not represent the spatial variability of soil properties at the level of detail required by many environmental applications (Band and Moore, 1995) and may result in a "mismatch of aggregation level (s)" between the soils and other environmental data layers (De Gruijter et al., 1997; Zhu, 2008; Zhu et al., 2008a).

Statistical methods have been used to predict detailed spatial variations of soil properties (Moore et al., 1993; Gessler et al., 1995; McKenzie and Ryan, 1999; Gessler et al., 2000; Park and Burt, 2002). Quantitative relationships between certain soil properties and environmental variables are usually developed for a local area through multiple linear regressions. These techniques often rely heavily on the assumption of linearity and do not consider spatial correlation of soil observations. However, the relationships between soil property variation and the underlying environmental variables can be very complex (Lark, 1999) and the assumption of linearity is often difficult to meet. Geostatistical methods have also been investigated to take into account the spatial autocorrelation of observed values in field samples (McBratney and Webster, 1986; Odeh et al., 1992, 1994; Odeh and Chittleborough, 1992; McBratney et al., 2000; Heuvelink and Webster, 2001; Hengl et al., 2004). These methods were found to outperform multiple regressions, especially when auxiliary information is available and is incorporated through regression-kriging or kriging with external drift (Odeh et al., 1994; McBratney et al., 2000; Bishop and McBratney, 2001). Geostatistical methods, however, are limited in that they often require large amount of field observations to account for complex landscape types or require the assumption of stationarity and thus are best suited for modeling soil spatial variation over small areas which have extensive field observations. This presents significant challenges to their application over large and diverse landscapes.

This research explores the possibilities of using soil fuzzy membership values (Zhu, 1997; Zhu et al., 2001) to predict soil property values in areas where the relationship between soil property

* Corresponding author. Tel.: +1 908 737 3668; fax: +1 908 737 3699.
E-mail addresses: azhu@wisc.edu (A.-X. Zhu), fqi@kean.edu (F. Qi), Amanda.Moore@md.usda.gov (A. Moore), jeburt@wisc.edu (J.E. Burt).

0016-7061/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.geoderma.2010.05.001

values and environmental attributes is perceived to be complex and non-linear. For example, a certain soil property might increase from a summit to backslope position and then decrease from backslope to foot slope or depression positions. Soil properties tend to change gradually within well-defined landscape units but change more quickly in transition zones between landscape positions. The fuzzy membership values to a set of soil types for a local soil (such as the soil similarity vector in the Soil Land Inference Model (SoLIM) approach (Zhu, 1997)) can be viewed as a non-linear transformation of environmental variables based on expert knowledge of soil–landscape relationships. The premise of this research is that the inherent non-linearity as captured by a set of soil fuzzy membership values can be used to describe and model non-linear variation in soil property values. In this study, we use the SoLIM approach to derive the set of membership values for mapped locations.

2. Methods and materials

2.1. Soil similarity vector and SoLIM

SoLIM is a predictive approach to soil mapping (Zhu, 1997; Zhu et al., 2001). It is based on the concept that the autocorrelation of soil formative factors results in the development of natural entities of soil on soil–landscape units (Hudson, 1992). Soils are thus predictable from environmental conditions that define the soil–landscape units. The core of SoLIM is a similarity model (Zhu, 1997) for representing soil spatial variation under fuzzy logic. With the similarity model, soil at location (i,j) is represented by an n-element similarity vector (referred to as soil similarity vector), $\boldsymbol{S}_{ij} = (S_{ij}^1, S_{ij}^2, \ldots S_{ij}^k \ldots S_{ij}^n)$, where n is the number of prescribed soil classes (such as taxonomic units) over the area and S_{ij}^k is an index that measures the similarity between the local soil at i,j to a typical soil class k. Such measure is predicted based on the similarity between the environmental conditions of a typical soil class k and that at the local site From a fuzzy logic perspective, this similarity value is the same as the fuzzy membership of the local soil to the soil class.

2.2. Methods

We examined three fuzzy membership-based approaches to the prediction and mapping of soil property values using the similarity vectors produced with SoLIM. The first two approaches use a fuzzy membership-weighted average model in which the soil property value at a location is the weighted average of the typical soil property values of the prescribed soil types with the weights being the fuzzy membership values (similarity values) (Eq. (1)) (Zhu et al., 1997).

$$V_{ij} = \frac{\sum_{k=1}^{n} s_{ij}^k v^k}{\sum_{k=1}^{n} s_{ij}^k} \quad (1)$$

where V_{ij} is the predicted soil property value at location i,j, $S_{i,j}^k$ is the fuzzy membership value in soil type k for the soil at the given location, and v^k is the typical soil property value for soil type k. This model is based on the assumption that the higher the membership of a local soil in a given soil series the closer the property values at that location will be to the typical property values of the series. We tested two options with this model: one uses representative values (RVs) from existing soil survey as the typical soil property values (v^k) of the prescribed soil types (referred to as the *weighted average-RV model*) and the other uses the property values observed at a field location where the fuzzy membership of the local soil to the given soil type, as determined by SoLIM, is the highest among all field observations (*weighted average-maximum membership or weighted average-MM model*). The third approach we propose in this study uses the similarity values in a statistical model. The incorporation of similarity measures in terms of fuzzy membership or taxonomic distance has been reported previously (Odeh and Chittleborough, 1992; Carre and Gigard, 2002). We tested a simple regression model in which the soil property value at a location is predicted using a multiple linear regression between observed soil properties at sampling locations and the fuzzy membership values of these soils to all prescribed soil types as determined by SoLIM (*regression-fuzzy membership model*).

In order to explore the hypothesis that models involving soil similarity vectors will better predict soil property variation than models using only environmental data for landscapes where non-linearity is high, the proposed models were compared with a predictive model based on multiple linear regression with environmental variables (*regression-environmental variable model*). We used only a multiple linear regression model here instead of other statistical methods (GLS regression, regression-kriging, etc.) to be consistent with and thus comparable to our regression-fuzzy membership model, which employs only a multiple linear regression. The explanatory variables we used include environmental variables that are commonly used in soil property predictions (Moore et al., 1993; Gessler et al., 1995) and field observations of soil property values. In our case study, the environmental variables used are topographic variables including elevation, slope, aspect, planform curvature, and profile curvature. While other variables likely have some influence on soil formation in the study area (for example, measures of slope positions, wetness indices, etc.), only these five were used in the inference of soil similarity vectors with SoLIM in a previous study (Smith et al., 2006). Therefore, for a fair comparison, our regression model was also developed using only these explanatory variables.

For the sake of comparison, we also compared our model prediction results to soil properties mapped on a conventional soil map (referred to as the *soil map model*). The soil map model uses the documented typical soil property values from an existing soil survey to approximate the soil properties at specific locations. If a location is enclosed within a soil polygon based on the existing soil survey, the local soil is considered to exhibit the same property values as the map unit corresponding to that polygon. The typical values of the soil properties for these map units were determined based on the US Map Unit Interpretation Record (MUIR) database (Soil Survey Staff, 1997).

2.3. Study area and data

The study site is Thompson farm, WI. It is a watershed in the 'Driftless Area' of southwestern Wisconsin (Fig. 1). The Driftless Area

Fig. 1. Location of the study area in Dane County, Wisconsin (the area to the west of the dashed line is the Driftless Area).

is the name applied to that portion of Wisconsin, Minnesota, Iowa, and Illinois not glaciated during the recent Wisconsin glaciation (12,000–90,000 years before present). Rather than being shaped by glacial processes, the primary mechanisms for recent landscape evolution in the Driftless Area have been fluvial, resulting in a well-drained, deeply dissected terrain (Dott and Attig, 2004). The watershed in our study consists of two distinct but related sections: the Galena uplands and the St. Peter backslopes (Fig. 2). The Galena uplands have gently rolling terrain consisting of a thin layer of loess over clayey residuum underlain by fractured dolomite. The soil classes in this section differ primarily in terms of depths to the bedrock layer. The St. Peter backslopes, on the other hand, are marked by steeper terrain, more variable soil types, and occur in places where stream channels have cut through the dolomite to expose the sandstone below.

Two representative transects were established over these two sections in the watershed: one on the gently rolling summit (the "Galena transect"), and one on the steep backslope (the "St. Peter transect") for our study (Fig. 2). Based on preliminary field investigations and an earlier soil mapping application, these transects were designed to capture the maximum amount of soil variation possible in the study area. The Galena transect starts from a convex position on the summit and extends across the summit, down the shoulder and into a concave drainage way. The St. Peter transect starts on a shoulder, extends down the steep south-facing backslope to a concave footslope and into the drainage way. It resumes at the base of the north-facing slope and extends through the footslope, up the backslope, and terminates 10 meters past the transition from backslope to shoulder.

Soil profile descriptions were made and soil samples were taken at five meter intervals along each transect to ensure a sufficient number of samples and to conform to the resolution of the GIS data layers for the study site, 32 sites for the Galena transect and 43 for the St. Peter transect were included. The location of each sample point was recorded with a GPS receiver. Soil pits were excavated to a depth of 50–90 cm and augured to 150 cm or bedrock, whichever was shallower. The soil at each location was described according to procedures developed by the USDA-Natural Resources Conservation Service (NRCS) (Schoeneberger et al., 1998). Horizon designation, horizon thickness, Munsell color, soil texture, soil structure, percent coarse fragments, depth to bedrock or weathered bedrock, and clay films were observed and recorded. Other soil features of interest that could aid in the classification of the soil profiles were also recorded. After the soil profile descriptions were completed, samples were taken from the A and Bt_1 horizons. Soil profiles were classified to the series level by NRCS soil scientist Chanc Vogel (Richland Center, Wisconsin USDA Service Center). Particle size distribution in A-horizon and Bt_1 horizon samples was subsequently determined using a laboratory procedure combining hydrometer analysis and sonic sifting analysis (Knox, 1994). The following soil properties were determined for use in each of the models: A-horizon soil texture (sand and silt content), Bt_1-horizon soil texture (sand and silt content), depth to Bt_1-horizon, loess thickness, and depth to weathered bedrock.

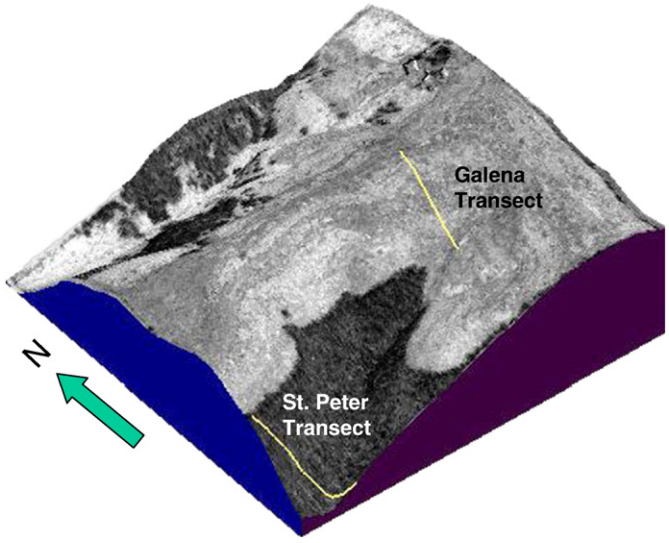

Fig. 2. Topography of the study area and the location of field transects.

Fuzzy membership maps of soil series for the area created with SoLIM in a previous study (Smith et al., 2006) were used to derive the soil similarity vectors for locations along the two transects. Soil inference in that previous study used only terrain variables based on the local soil–landscape model developed by soil experts. The terrain variables used to capture the landscape units were elevation, slope gradient, planform curvature, and profile curvature. Prototype-based inference (Qi et al., 2006) was utilized to generate fuzzy membership maps of all prescribed soil types determined by local soil experts. For each sample point along the two transects, the fuzzy membership values to all soil types are combined to form the similarity vector which was then used in the fuzzy membership-based approaches to predict soil properties.

2.4. Model development

In order to compare the performances of soil property prediction models on different types of landscapes, models were developed independently for each of the two transects because they occur in areas with different topographic conditions and have different sets of soil series developed. For the *weighted average-RV model*, typical values of the soil properties for the soil series mapped in the study watershed (v^k) were determined based on the MUIR database and the representative profile descriptions recorded in the existing soil survey (Glocker and Patzer, 1978). Specific soil texture information (percent sand and silt) was calculated from sieve-size data from the MUIR. Depth to Bt_1-horizon was determined by examining the representative profile descriptions reported in the existing soil survey and recording the depth to the upper boundary of the argillic horizon. Depth to weathered bedrock was determined by examining the representative profile descriptions reported in the existing soil survey which record either the depth to weathered bedrock or 152 cm, whichever was greater. Loess thickness was determined by examining the representative profile descriptions and noting the depth at which a change in parent material occurred. Once the typical soil property values (v^k) were obtained for each soil type, fuzzy membership values of the sample site ($S_{i,j}^k$) were retrieved from the fuzzy membership maps based on the recorded GPS location. Predicted soil property values were then calculated for all sites on each transect using Eq. (1).

For the *weighted average-MM model*, the weights ($S_{i,j}^k$) remain the same as those used in the *weighted average-RV model* but typical soil property values (v^k) were obtained from the field samples with highest fuzzy membership values in their respective soil types. This is done in the following way: first field observation locations were intersected with fuzzy membership maps for each soil series and fuzzy membership values at each location were recorded. Then, the field observation with the highest fuzzy membership value for each soil type was considered to be "typical" and the measured soil properties associated with this location were assumed to be representative of the soil type. Predicted soil property values of all locations were then calculated using Eq. (1).

In the case of the *regression-fuzzy membership model*, explanatory variables in the regression models are fuzzy membership values of all soil series in the study area whose fuzzy membership value for at least one transect point is greater than zero. Correlation matrices were generated for each transect using the statistical software package R.

Linear regression models were developed for all possible combinations of the explanatory variables: 31 separate regression models were developed for the Galena transect and 256 separate regression models were developed for the St. Peter transect. In order to find the "best" model for each soil property and each transect, the adjusted R^2 statistic was used to reduce overfitting and the significance of each explanatory variable was also considered. The model with the highest adjusted R^2 value and only explanatory variables with significant ($\alpha = 0.05$) regression coefficients was selected.

The implementation of the *regression-environmental variable model* is largely similar to the *regression-fuzzy membership model* except that the explanatory variables now are the four environmental variables chosen for this study site: elevation, slope, aspect, planform curvature, and profile curvature. Data layers of these variables were calculated using a 10-foot resolution DEM and a 120-foot neighborhood size (Smith et al., 2006). Transect points were overlaid onto the various topographic layers and the values of each layer were recorded for each transect point. Correlation matrices and regression models were developed using the same approach as with the *regression-fuzzy membership model*.

Finally, the *soil map model* simply uses the typical soil property values based on the existing Soil Survey of Dane County, WI to approximate the soil property values at sites visited along the transects. First sample locations were intersected with the existing polygon map to determine the soil map unit each observation is associated with (Fig. 3). Then typical values of the soil properties for the soil series in each of these map units were determined based on MUIR database and the representative profile descriptions recorded in the existing survey as what we did for the *weighted average-RV model*.

2.5. Assessment measures

Once the soil property values predicted by each model were calculated, they were compared to the observed soil property values in order to assess model performances. Several measures were used for quantitative assessment of the models, including mean absolute error (MAE), R^2, and agreement coefficient (AC). MAE measures model precision. It was calculated based on Eq. ((2):

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} \left| (v_i - v_i') \right|}{n} \tag{2}$$

where v_i is the observed soil property value, v_i' is the predicted soil property value, and n is the number of observations. R^2 describes the amount of variability in the predicted value Y explained by the explanatory variables $X_1, X_2, \ldots, X_n$. As R^2 approaches 1, the amount of variation in Y explained by the model increases. The AC index is defined by Willmott (1984) as:

$$\mathrm{AC} = 1 - \frac{n \cdot \mathrm{RMSE}^2}{\mathrm{PE}} \tag{3}$$

where n is the number of observations and PE the potential error variance defined as:

$$\mathrm{PE} = \sum_{j=1}^{n} \left(|P_i - \overline{O}| + |O_i - \overline{O}| \right)^2 \tag{4}$$

given that $\overline{O}$ is the observed mean, and P_i and O_i are the estimated and observed value, respectively. AC values vary between 0 and 1, where 1 indicates perfect agreement and 0 means complete disagreement between the estimated and observed values (Willmott, 1984).

3. Results and discussion

The above assessment measures were used to evaluate two types of model performances in our current study. First, prediction accuracies were compared among the two weighted average models

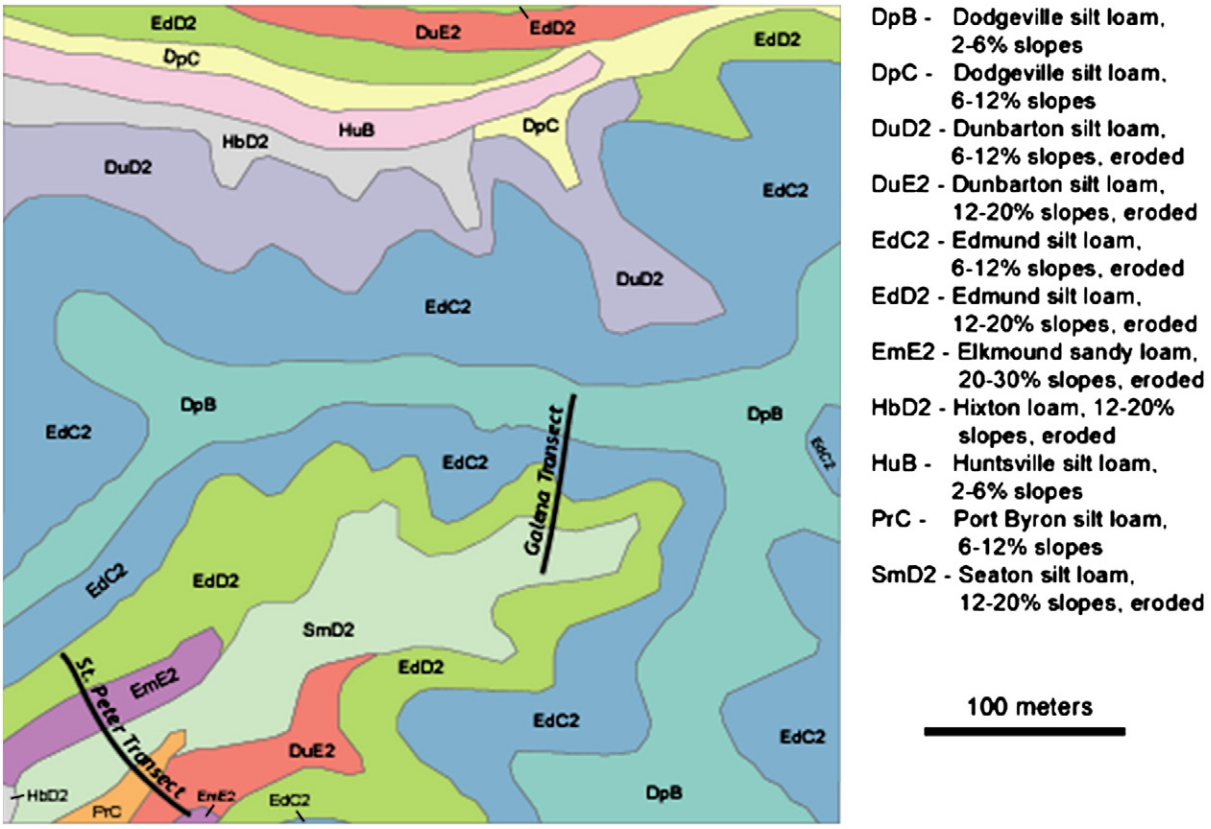

Fig. 3. Transect locations with respect to existing soil survey delineations.

Table 1
MAE for all selected models — Galena transect.

Property	Soil map model	Weighted average-RV	Weighted average-MM	Regression-environmental variable	Regression-fuzzy membership
A-horizon sand	5.53	5.56	6.68	0.74	0.73
A-horizon silt	6.72	6.51	2.18	1.44	2.02
Bt_1-horizon sand	6.12	2.62	1.15	1.02	1.14
Bt_1-horizon silt	9	6.14	4.71	3.67	5.56
Depth to Bt_1	10.16	9.37	8.68	7.67	8.02
Loess thickness	30.32	24.35	14.81	9.46	14.29
Depth to Cr	29.37	24.26	16.23	11.56	15.36

and the soil map model. Both the weighted average-RV model and soil map model did not use any field observation in model development. The weighted average-fuzzy membership model used one observation per soil class in model development and the used field sites were excluded from the sample set when used for testing the model's prediction accuracies. Second, the two regression models (environmental variable-based and membership-based) were compared using the accuracy measures computed based on the field observations. Since the two models used all field points in model development, the performance measures were considered those of model development due to lack of an independent validation set. Tables 1–6 below, list the computed MAE, R^2, and AC values from all models across the 7 soil properties used along the two transects together for simplicity.

The results show that on both landscape types, the *weighted average-MM models* perform better than simple *weighted average-RV models*. R^2 values are mostly higher and MAE values are generally lower for the MM models. The differences are more obvious for the St. Peter transect and the type of landscape represented by this transect. This indicates that the soil property values associated with the transect observation with the highest fuzzy membership value for each series may be a better representation of local soil property values than the typical values presented in the soil survey. The tables show that, especially along the St. Peter transect, the models that used information from the existing soil map (the *soil map model* and *weighted average-RV model*) both have significantly higher MAEs and lower R^2 values than the *weighted average-MM model*. Looking at the scatter plots for *A*-horizon sand values (Figs. 4 and 5), we see the tendency for predicted soil property values to be stratified by soil type with the *soil map model* and *weighted average-RV model*, which contrasts the linear trend seen in the plots with the two regression models (Figs. 6 and 7). This stratification can be explained as an artifact of the polygon data model used for existing soil maps, which only represents the typical property within an entire map unit instead of more detailed local variability. Because of the areal coverage of a soil map unit is often large, the impact of a misclassification can be profound in detailed soil property prediction. It is noted, however, that on the Galena landscape, R^2 values are sometimes the highest with the *soil map model* (A-horizon Silt, for example). With the polygon-based soil map model, soil property value within a polygon is often an average of the local variations. In cases when local variation is great and an attempt to predict the variation within the unit might result in poorer agreement between the predicted and observed (as indicated by the R^2), the best estimate of a property value for a spatial unit could well be the mean value for the unit, which might explain the case with A-horizon silt in this study.

With regard to the two regression models, results from the Galena transect (Tables 1–3) shows that the *regression-environmental variable model* generally outperforms (evidenced by lower MAE, higher R^2 and AC values for most of the soil properties) the regression-fuzzy membership variable model. This, however, is not echoed by the results from the St. Peter transect (which has a stronger relief and more complexity). Tables 4–6 show apparent lower MAE and higher AC values with the *regression-fuzzy membership model* than the *regression-environmental variables model* for most soil properties.

Table 2
R^2 for all selected models — Galena transect.

Property	Soil map model	Weighted average-RV	Weighted average-MM	Regression-environmental variable	Regression-fuzzy membership
A-horizon sand	0.1727**	−0.0094*	0.0241*	0.0785*	0.0620*
A-horizon silt	0.5331****	0.4426****	0.5054****	0.6791****	0.4593****
Bt_1-horizon sand	0.0035*	0.0441**	0.0900*	0.1525*	0.1735**
Bt_1-horizon silt	0.0918*	0.1304**	0.1384**	0.3991***	0.1221**
Depth to Bt_1	−0.02*	0.1587**	−0.0056*	0.3509***	0.3178***
Loess thickness	0.4484****	0.4457****	0.4451****	0.6750****	0.3645****
Depth to Cr	0.4303****	0.2812***	0.2401***	0.5316****	0.3005***

* Not significant at the 0.05 level.
** Significant at the 0.05 level.
*** Significant at the 0.01 level.
**** Significant at the 0.001 level.

Table 3
AC for all selected models — Galena transect.

Property	Soil map model	Weighted average-RV	Weighted average-MM	Regression-environmental variable	Regression-fuzzy membership
A-horizon sand	0.21	0.40	0.045	0.47	0.44
A-horizon silt	0.54	0.66	0.777	0.91	0.81
Bt_1-horizon sand	0.33	0.41	0.522	0.64	0.60
Bt_1-horizon silt	0.56	0.70	0.194	0.80	0.49
Depth to Bt_1	0.42	0.58	0.350	0.76	0.72
Loess thickness	0.69	0.81	0.569	0.91	0.73
Depth to Cr	0.72	0.79	0.472	0.85	0.71

Table 4
MAE for all selected models — St. Peter transect.

Property	Soil map model	Weighted average-RV	Weighted average-MM	Regression-environmental variable	Regression-fuzzy membership
A-horizon sand	18.35	15.38	7.62	6.72	4.81
A-horizon silt	17.97	15.10	6.2	4.61	3.65
Bt_1-horizon sand	15.89	16.21	10.98	10.97	4.3
Bt_1-horizon silt	14.68	16.94	9.33	8.01	3.59
Depth to Bt_1	13.35	12.24	10.86	9.11	9.28
Loess thickness	56.29	59.78	5.98	4.88	3.45
Depth to Cr	17.72	20.22	17.94	15.48	9.37

Table 5
R^2 for all selected models — St. Peter transect.

Property	Soil map model	Weighted average-RV	Weighted average-MM	Regression-environmental variable	Regression-fuzzy membership
A-horizon sand	0.19***	0.1671***	0.5271****	0.6039****	0.8015****
A-horizon silt	0.2468****	0.3031****	0.5634****	0.7459****	0.8369****
Bt_1-horizon sand	0.1033**	0.1003**	0.3612****	0.4627****	0.8938****
Bt_1-horizon silt	0.1466***	0.0659*	0.2766****	0.4408****	0.8592****
Depth to Bt_1	0.2287****	0.3255****	0.2308****	0.3880****	0.2970****
Loess thickness	0.113**	0.2921****	0.7801****	0.8202****	0.8956****
Depth to Cr	0.664****	0.7970****	0.8965****	0.8309****	0.9173****

* Not significant at the 0.05 level.
** Significant at the 0.05 level.
*** Significant at the 0.01 level.
**** Significant at the 0.001 level.

With the exception of the R^2 value for the depth to Bt_1-horizon, R^2 values for all other *regression-fuzzy membership models* are not only very high but are much larger than those for the *regression-environmental variable models*. This difference shows that the explanatory variables in regression models based on fuzzy membership values explain more of the variation in soil properties than regression models based only on the environmental variables involved. Implicit in this outcome is the idea that the transformation of simple environmental characteristics to more complicated fuzzy membership values via SoLIM's fuzzy inference engine increases the predictive power of those variables. Comparing the results on the two transects we can observe that, in predicting soil property values, linear regression models based on the terrain attributes may be limited to areas with gentle landscapes. For the St. Peter landscapes in this area, the relationships between soil property and terrain attributes can be highly non-linear and non-linear transformation of the terrain variables would benefit the soil property prediction (SoLIM inference is an example of such non-linear transformations).

As aforementioned, the two regression models both used the field samples for model development and thus the performance measures cannot be directly used to compare with the other three models. One would expect that the performance measures of the regression models (based on the dataset for model development) should be generally, if not always, better than those based on an independent validation dataset. Even so, we note that the *weighted average-MM model* produced comparable and only slightly lower accuracy values than the two regression models. Considering the fact that this model was tested against independent samples while the regression models were tested against samples that were used for model development, we believe that the performance of the *weighted average-MM model* is at least comparable to, if not higher than, those of the two regression-based models. In addition, from the field data requirement perspective, the *weighted average-MM model* has clear advantages over the statistical models in that it requires only one field sample per soil series or soil category. Furthermore, in actual applications of the weighted average-MM model for soil property prediction, the search for highest fuzzy membership to the mapped soil series do not have to be limited to the sample points on the transect as used in this study (designed and collected mostly for the development of the regression models). Purposive sampling could be conducted as in Zhu et al. (2008b). The most typical site of a soil series (the site that has the highest membership) could be identified from the entire area and targeted samples could be collected to obtain the most typical properties for use in the weighted average computation.

One other observation we made from the testing results is that the R^2 values for all the models tested, are relatively low for the Galena transect. The landscape in question is gently rolling, with maximum calculated slopes (at a 120-foot neighborhood size) of 10.9% and a difference in elevation from transect top to bottom of only 15.3 meters. The soil series mapped on this upland (*Brownbeth*, *Dodgeville*, *Dubuque*, *Edmund*, and *Frankville*) all have silty surfaces and clayey subsurfaces, varying only in thickness of the surface soil and depth to bedrock. *Brownbeth*, *Dubuque*, and *Edmund* all occur on shoulder positions and are separated only by small differences in

Table 6
AC for all selected models — St. Peter transect.

Property	Soil map model	Weighted average-RV	Weighted average-MM	Regression-environmental variable	Regression-fuzzy membership
A-horizon sand	0.62	0.56	0.76	0.88	0.95
A-horizon silt	0.58	0.53	0.78	0.93	0.96
Bt_1-horizon sand	0.61	0.56	0.68	0.81	0.97
Bt_1-horizon silt	0.58	0.50	0.60	0.79	0.97
Depth to Bt_1	0.64	0.73	0.56	0.74	0.22
Loess thickness	0.09	0.09	0.93	0.95	0.93
Depth to Cr	0.90	0.91	0.93	0.95	0.95

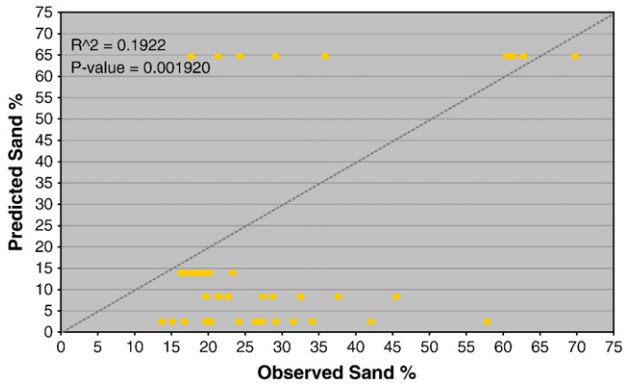

Fig. 4. A-horizon sand values predicted with the *soil map model* vs. observed A-horizon sand values — St. Peter transect.

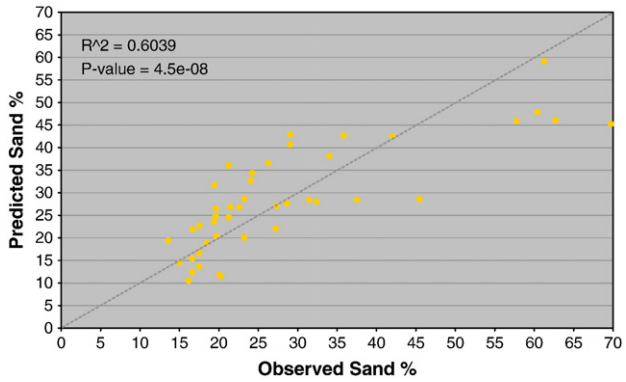

Fig. 6. A-horizon sand values predicted with the *regression-environmental variable model* vs. observed A-horizon sand values — St. Peter transect.

curvature (*Edmund* vs. *Dubuque* and *Brownbeth*) and slope gradient (*Brownbeth* vs. *Dubuque*). *Frankville* is separated from the potentially adjacent *Brownbeth*, *Dubuque*, and *Edmund* by curvature while *Dodgeville* is separated from Edmund by ridge width and from *Brownbeth* and *Dubuque* by slope gradient. In this landscape, however, these differences in landscape position can be quite subtle and difficult to determine which in turn increases the likelihood of misclassifying the soil and results in low R^2 values for all the models. We realize that the inclusion of other environmental variables may bring benefits to models that use environmental information in the soil mapping or soil property inference process. For example, SoLIM could include variables like measures of relative landform positions or even the simple wetness index measure; the regression-environmental variable model could also add these additional measures to its explanatory variable set. Since our SoLIM map is from a previous study which did not use these variables we did not exercise this option in our study. To be compatible for comparison, we used the same set of variables with the regression-environmental variable model as well. Future studies should consider these additional explanatory options. It is also worth noting that the limited performance of these models, especially the models of subsurface soil properties, could also be related to inconsistent identification and sampling of the Bt_1-horizon, mixing of the surface and subsurface soils from agricultural uses, and differential erosion. Lower soil horizons in this area could reflect development under conditions that differed quite markedly from those implied by the current surface configuration. Therefore the present surface might be a poor predictor of the subsurface attributes (Pennock and de Jong, 1987).

Lastly, we would like to point out that the purpose for us to stratify the study area to the two different landscapes (Galena and St. Peter) was to compare the performances of soil property prediction models on different types of landscapes. Models were developed independently for each of the two transects because different sets of soil series developed on distinct landscape units. It should also be noted how such stratification should be beneficial to soil mapping and soil property predictions on potentially larger areas. Through the manual stratification, we could effectively impose a partitioning similar to what might be achieved using a regression tree data mining approach in which separate regression equations are produced for different strata, based on analysis of variance measures within strata defined by various different covariates.

4. Conclusions

The intent of this study was to assess the ability of models that incorporate soil fuzzy membership to predict soil properties. The conclusion from the comparison of the predictive models is: (1) the models based on regression with environmental variables (*regression-environmental variable model*) would be appropriate for use on gently rolling landscapes where soil-environment relationships are simple and stable over space; (2) the models based on regression with fuzzy membership values (*regression-fuzzy membership model*) work well over areas where soil environmental relationships are more complicated and the non-linear transformation of the environmental variables (such as that by SoLIM as in our study) helps to improve the performance of linear regression; and (3) the model based on the weighted average of fuzzy membership values and fuzzy maximum soil property values (*weighted average-maximum membership model*) can serve as a viable option for soil property prediction over large areas due to its good performance and limited amount of model development points needed. The work reported in this paper also confirms the findings of Zhu et al. (1997) with regard to predictive models based on existing soil survey maps. Soil property values predicted based on the existing soil survey tends to be stratified by the

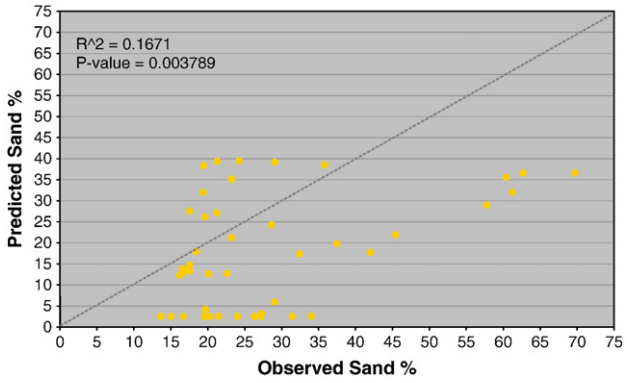

Fig. 5. A-horizon sand values predicted with the *weighted average-RV* vs. observed A-horizon sand values — St. Peter transect.

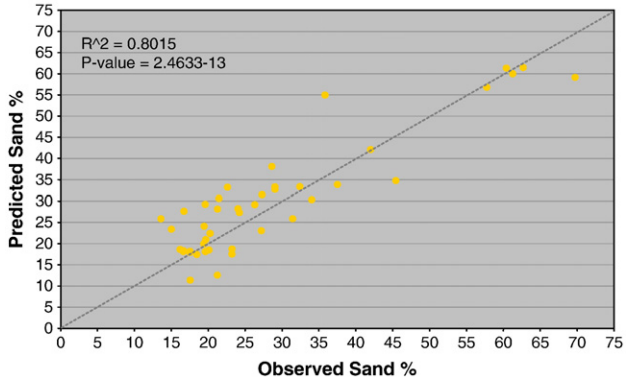

Fig. 7. A-horizon sand values predicted with the *regression-fuzzy membership model* vs. observed A-horizon sand values — St. Peter transect.

typical property values associated with soil map units. The results often do not reflect the detailed soil spatial variability locally. The soil map, however, performs adequately when the information needed is averaged properties over a region (which is how it was designed to be used) and not local variability.

Acknowledgements

The research reported here was supported by a project from the National Natural Science Foundation of China (project no. 40971236), the National Basic Research Program of China (project no. 2007CB407207), and the National Key Technology R&D Program of China (2007BAC15B01). This research is also supported by the Natural Resources Conservation Services, United States Department of Agriculture through its Graduate Studies Program to Amanda Moore, the Chinese Academy of Sciences through its "One-Hundred Talents" program, the University of Wisconsin-Madison Vilas Associate Award, the Hamel Faculty Fellow Award to Prof. A-Xing Zhu, and the Kean University through the Untenured Faculty Research Initiative Award to Prof. Feng Qi.

References

Band, L.E., Moore, I.D., 1995. Scale: landscape attributes and geographical information systems. Hydrol. Process. 9, 401–422.
Bishop, T.F.A., McBratney, A.B., 2001. A comparison of prediction methods for the creation of field-extent soil property maps. Geoderma 103, 149–160.
Carre, F. And, Gigard, M.C., 2002. Quantitative mapping of soil types based on regression kriging of taxonomic distances with landform and land cover attributes. Geoderma 110, 241–263.
De Gruijter, J.J., Walvoort, D.J.J., van Gaans, P.F.M., 1997. Continuous soil maps—a fuzzy set approach to bridge the gap between aggregation levels of process and distribution models. Geoderma 77, 169–195.
Dott, R.H., Attig, J.W., 2004. Roadside Geology of Wisconsin. Mountain Press Publishing Company, Missoula, Montana.
Gessler, P.E., Moore, I.D., McKenzie, N.J., Ryan, P.J., 1995. Soil–landscape modeling and spatial prediction of soil attributes. Int. J. Geogr. Inf. Syst. 9, 42–432.
Gessler, P.E., Chadwick, O.A., Chamran, F., Althouse, L., Holmes, K., 2000. Modeling soil–landscape and ecosystems properties using terrain attributes. Soil Sci. Soc. Am. J. 64, 2046–2056.
Glocker, C.L., Patzer, R.A., 1978. Soil Survey of Dane County. United States Department of Agriculture—Soil Conservation Service, Wisconsin.
Hengl, T., Heuvelink, G.B.M., Stein, A., 2004. A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma 120, 75–93.
Heuvelink, G.B.M., Webster, R., 2001. Modeling soil variation: past, present, and future. Geoderma 100, 269–301.
Hudson, B.D., 1992. The soil survey as paradigm-based science. Soil Sci. Soc. Am. J. 56, 836–841.
Knox, J.C., 1994. Laboratory procedure for particle size analyses: hydrometer and sonic sifting. Unpublished lab material, Analysis of the Physical Environment. University of Wisconsin-Madison.
Lark, R.M., 1999. Soil–landform relationships at within-field scales: an investigation using continuous classification. Geoderma 92, 141–165.
McBratney, A.B., Webster, R., 1986. Choosing functions for semi-variograms of soil properties and fitting them to sampling estimates. J. Soil Sci. 37, 617–639.
McBratney, A.B., Odeh, I.O.A., Bishop, T.F.A., Dunbar, M.S., Shatar, T.M., 2000. An overview of pedometric techniques for use in soil survey. Geoderma 97, 293–327.
McKenzie, N.J., Ryan, P.J., 1999. Spatial prediction of soil properties using environmental correlation. Geoderma 89, 67–94.
Moore, I.D., Gessler, P.E., Nielsen, G.A., Peterson, G.A., 1993. Soil attribute prediction using terrain analysis. Soil Sci. Soc. Am. J. 57, 443–452.
Odeh, I.O.A., McBratney, A.B., Chittleborough, D.J., 1992. Soil pattern recognition with fuzzy c-means: applications to classification and soil landform interrelationships. Soil Sci. Soc. Am. J. 56, 505–516.
Odeh, I.O.A., Chittleborough, D.J., 1992. Fuzzy-c-means and Kriging for mapping soil as a continuous system. Soil Sci. Soc. Am. J. 56, 1848–1854.
Odeh, I.O.A., McBratney, A.B., Chittleborough, D.J., 1994. Spatial prediction of soil properties from landform attributes derived from a digital elevation model. Geoderma 63, 197–214.
Park, S.J., Burt, T.P., 2002. Identification and characterization of pedomorphological processes on a hillslope. Soil Sci. Soc. Am. J. 66, 1897–1910.
Pennock, D.J., de Jong, E., 1987. The influence of slope curvature on soil erosion and deposition in Hummock terrain. Soil Sci. 144, 209–217.
Qi, F., Zhu, A.X., Harrower, M., Burt, J.E., 2006. Fuzzy soil mapping based on prototype category theory. Geoderma 136, 774–787.
Schoeneberger, P.J., Wysocki, D.A., Benham, E.C., Broderson, W.D., 1998. Field Book for Describing and Sampling Soils, Version 1.1. Natural Resources Conservation Service, USDA, National Soil Survey Center, Lincoln, NE.
Smith, M., Zhu, A.X., Burt, J.E., Stiles, C., 2006. Effects of DEM resolution and neighborhood size on soil survey. Geoderma 137, 58–69.
Staff, Soil Survey, 1997. National Map Unit Interpretation Record (1997). Natural Resources Conservation Service, USDA. National Soil Survey Center, Lincoln, NE. URL: http://soils.usda.gov/soils/survey/nmuir/index.html.
Willmott, C.J., 1984. On the evaluation of model performances in physical geography. In: Gaile, G.L., Willmott, C.J. (Eds.), Spatial Statistics and Models. D. Reidel Publi., Dordrecht, The Netherlands, pp. 443–460.
Zhu, A.X., 1997. A similarity model for representing soil spatial information. Geoderma 77, 217–242.
Zhu, A.X., 2008. Spatial scale and neighborhood size in spatial data processing for modeling the natural environment. In: Mount, N.J., Harvey, G.L., Aplin, P., Priestnall, G. (Eds.), Representing, Modeling and Visualizing the Natural Environment: Innovations in GIS 13. CRC Press, Florida, pp. 147–165.
Zhu, A.X., Band, L.E., Vertessy, R., Dutton, B., 1997. Deriving soil property using a soil land inference model (SoLIM). Soil Sci. Soc. Am. J. 61, 523–533.
Zhu, A.X., Hudson, B., Burt, J., Lubich, K., Simonson, D., 2001. Soil mapping using GIS, expert knowledge, and fuzzy logic. Soil Sci. Soc. Am. J. 65, 1463–1472.
Zhu, A.X., Burt, J.E., Smith, M., Wang, R.X., Gao, J., 2008a. The impact of neighbourhood size on terrain derivatives and digital soil mapping. In: Zhou, Q., Lees, B., Tang, G. (Eds.), Advances in Digital Terrain Analysis. Springer-Verlag, New York, pp. 333–348.
Zhu, A.X., Yang, L., Li, B., Qin, C., English, E., Burt, J.E., Zhou, C.H., 2008b. Purposefully sampling for digital soil mapping. In: Hartemink, A.E., McBratney, A.B., Mendonca Santos, M.L. (Eds.), Digital Soil Mapping with Limited Data. Springer-Verlag, New York, pp. 233–245.