



# Soil polygon disaggregation through similarity-based prediction with legacy pedons

LIU Feng<sup>1\*</sup>, GENG Xiaoyuan<sup>2</sup>, ZHU A-xing<sup>3,4</sup>, Walter FRASER<sup>2</sup>, SONG Xiaodong<sup>1</sup>, ZHANG Ganlin<sup>1</sup>

<sup>1</sup> State Key Laboratory of Soil and Sustainable Agriculture, Institute of Soil Science, Chinese Academy of Sciences, Nanjing 210008, China;

<sup>2</sup> Science and Technology Branch, Agriculture and Agri-Food Canada, Ottawa K1A 0C5, Canada;

<sup>3</sup> School of Geography, Nanjing Normal University, Nanjing 210046, China;

<sup>4</sup> Department of Geography, University of Wisconsin-Madison, Madison WI 53706, USA

**Abstract:** Conventional soil maps generally contain one or more soil types within a single soil polygon. But their geographic locations within the polygon are not specified. This restricts current applications of the maps in site-specific agricultural management and environmental modelling. We examined the utility of legacy pedon data for disaggregating soil polygons and the effectiveness of similarity-based prediction for making use of the under- or over-sampled legacy pedon data for the disaggregation. The method consisted of three steps. First, environmental similarities between the pedon sites and each location were computed based on soil formative environmental factors. Second, according to soil types of the pedon sites, the similarities were aggregated to derive similarity distribution for each soil type. Third, a hardening process was performed on the maps to allocate candidate soil types within the polygons. The study was conducted at the soil subgroup level in a semi-arid area situated in Manitoba, Canada. Based on 186 independent pedon sites, the evaluation of the disaggregated map of soil subgroups showed an overall accuracy of 67% and a Kappa statistic of 0.62. The map represented a better spatial pattern of soil subgroups in both detail and accuracy compared to a dominant soil subgroup map, which was commonly used in practice. Incorrect predictions mainly occurred in the agricultural plain area and the soil subgroups that are very similar in taxonomy, indicating that new environmental covariates need to be developed. We concluded that the combination of legacy pedon data with similarity-based prediction is an effective solution for soil polygon disaggregation.

**Keywords:** legacy pedon data; similarity-based prediction; spatial disaggregation; conventional soil maps

**Citation:** LIU Feng, GENG Xiaoyuan, ZHU A-xing, Walter FRASER, SONG Xiaodong, ZHANG Ganlin. 2016. Soil polygon disaggregation through similarity-based prediction with legacy pedons. *Journal of Arid Land*, 8(5): 760–772. doi: 10.1007/s40333-016-0087-7

Conventional soil maps are an important soil information source for land management, soil carbon sequestration and climate change prediction (Sauchyn, 2001; Ju and Chen, 2005; Schut et al., 2011). A soil polygon generally represents a mixture of soil types. It associates with soil property and interpretive information typically derived from laboratory analysis and soil surveyors. However, geographic locations of the soil types within the polygon are not specified (Geng et al., 2010; Ashtekar and Owens, 2013). The lack of individual soil type spatial specificity restricted the spatial accuracy of site-specified agricultural management and environmental

\*Corresponding author: LIU Feng (E-mail: fliu@issas.ac.cn)

Received 2016-01-22; revised 2016-04-20; accepted 2016-04-22

© Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences, Science Press and Springer-Verlag Berlin Heidelberg 2016

modelling. Therefore, spatially disaggregating or downscaling (McBratney, 1998) are required to improve the quality of soil maps.

To avoid the cost involved in conducting a new soil survey, legacy soil data from a traditional soil survey were favorite for soil polygon disaggregation. Among the different forms of legacy data (soil maps, reports and pedons), legacy soil maps were frequently utilized for the disaggregation (Smith et al., 2010; Yang et al., 2011; Subburayalu and Slater, 2012; Collard et al., 2014; Du et al., 2014; Nauman and Thompson, 2014; Odgers et al., 2014). Smith et al. (2010) extracted rules on soil-environmental relationships from a legacy soil map and then used the rules to disaggregate soil polygons. Huang et al. (2016) obtained soil-environmental knowledge from a legacy soil map using a spatial data mining method and then produced a detailed soil type map.

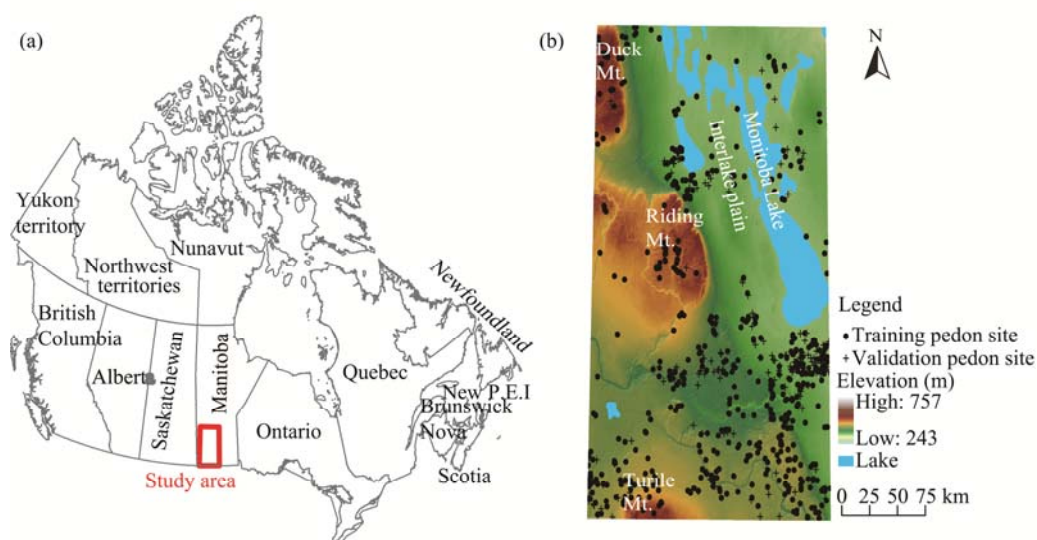
However, legacy pedon data have rarely been used for soil polygon disaggregation because of the challenge of using such data. Most pedon data were not originally designed for soil mapping even though it has been acknowledged that they are useful for this purpose (Geng et al., 2010). The sampling of the pedons usually does not follow a statistical criteria (Carré et al., 2007; Minasny et al., 2008; Arrouays et al., 2014), leading to under- or over-sampling in the geographic and attribution spaces. Some areas may have dense pedon sites while others may have sparse pedon sites. Some soil types may have many pedon sites while others may have a few. These limits the applications of statistical and geostatistical methods which are commonly used in digital soil mapping (Myers, 1994; Boruvka et al., 2002; Diem and Comrie, 2002; Penížek and Borůvka, 2004; Lagacherie, 2008). An alternative is a similarity-based approach (Holt, 1999; Zhu et al., 2010; Pla et al., 2013), which makes soil predictions at unsampled locations based on environmental similarities of the sampled locations. This approach has no requirement regarding the number and distribution of pedon sites. Moreover, the prediction process is transparent and traceable, allowing an easy interpretation of the prediction results.

The objective of this study was to examine: (1) the utility of legacy pedon data for disaggregating soil polygons; (2) the effectiveness of the similarity-based approach for making use of legacy pedon data for the disaggregation at soil subgroup level. Soil subgroup was selected because it is a basic taxonomic unit in regional and national soil maps.

## 1 Materials and methods

### 1.1 Study area

The study area is located in the southwestern Manitoba, Canada (49°–52°N, 98°–101°W; Fig. 1).



**Fig. 1** Location of study area (a) and distribution of training and valuation pedon sites on the background of Digital Elevation Model (b)

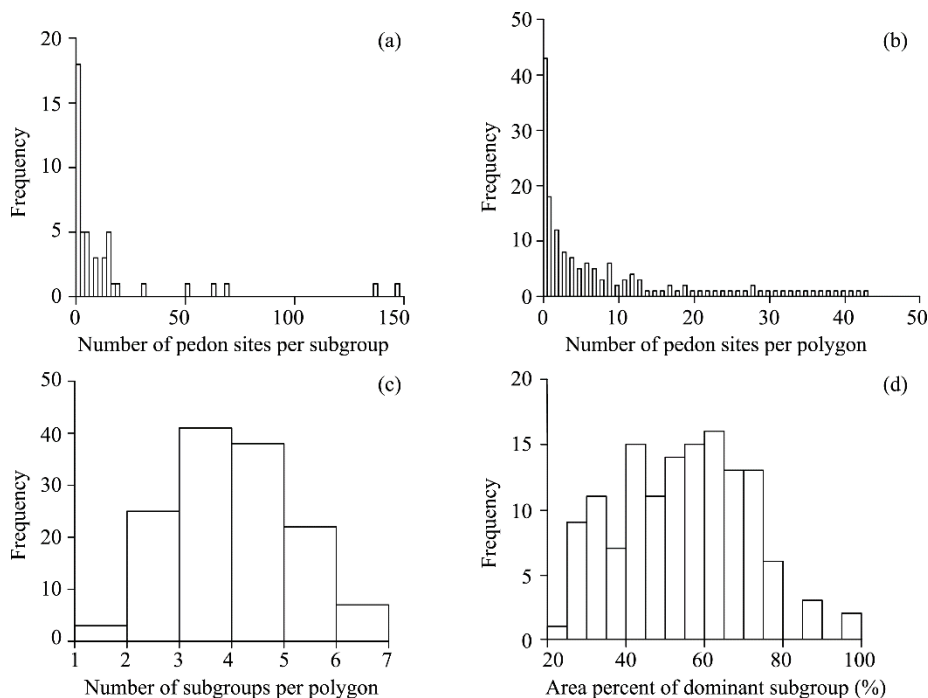
It covers an area of  $7 \times 10^4 \text{ km}^2$  and is approximately 330 km north to the border with the United States (the U.S.) and 220 km away from Saskatchewan in the east. It is a semi-arid area with an annual mean temperature of  $2.4^\circ\text{C}$  and a mean annual precipitation of 474 mm (meteorological data of 1971–2000 at Brandon station; [http://climate.weatheroffice.gc.ca/climate\\_normals](http://climate.weatheroffice.gc.ca/climate_normals)). The temperature decreases from south to north and the precipitation decreases from southeast to northwest (Welsted et al., 1996). The surficial geology of the area consists of unconsolidated surface deposits resulting from glaciation and deposition of glacial sediments, including tills, glaciofluvial and glaciolacustrine materials. Manitoba escarpment divides the area into the interlake plains in the northeast area and the highlands in the southwestern part. The relief ranges from nearly flat in the plains to rolling and hilly in the highlands. The elevation ranges from 240 to 760 m. Slope gradient in most of the area is less than 5% with a mean of 1.3%. Natural vegetation comprise grasses and deciduous tree species. Farmland is the major landuse type especially in the low relief areas. Dominant crops include wheat, barley, oats, canola, peas and forage. According to the Canadian soil classification (Soil Classification Working Group, 1998), seven major soil orders including the Chernozemic, Luvisolic, Regosolic, Brunisolic, Organic, Gleysolic and Vertisolic soils were identified in this area. They are equivalent to the soil orders Mollisols, Alfisols, Entisols, Inceptisols, Histosols, Alfisols and Vertisols, respectively, in the U.S. soil taxonomy (Soil Survey Staff, 2014).

## 1.2 Data sets

### 1.2.1 Legacy pedon data

There are 746 legacy pedons distributed in this study area according to the Manitoba Soil Description Database developed by the Manitoba Land Resource Unit, Agriculture and Agri-Food, Canada. Most of the pedons were surveyed in the 1980s. The description of the pedons includes geographical coordinates, soil order, great group, subgroup, soil properties and landscape characteristics. The coordinates were recorded using three different coordinate reference systems including the Canadian Military Grid Reference (Universal Transverse Mercator), the Dominion Land Survey and the Canadian National Topographic System.

Figure 2a shows a histogram of the frequency distribution of the number of pedon sites in each



**Fig. 2** Histograms of the legacy pedon sites and soil subgroups

soil subgroup, indicating an imbalance between different soil subgroups. A stratified random strategy was thus adopted to separate the pedons for training and validation (Wang et al., 2012). The stratification was performed for each soil subgroup to ensure that all soil subgroups were included in both training and validation datasets. Then, within each stratum (i.e. soil subgroup), a simple random selection was performed to separate its pedons into two parts: 75% for training and 25% for validation. The resultant training set consists of 560 pedon sites which were used to develop the prediction model and the validation set with 186 pedon sites were used to estimate the prediction accuracy. The training set had a bigger proportion of pedon sites than the validation set in order to ensure a well-trained model. Figure 1 shows the spatial distribution of the training and the validation pedon sites.

### 1.2.2 Conventional soil map

The conventional soil map to be disaggregated in this study was polygon-based at the scale of 1:1,000,000. It was compiled in the 1990s. The mapping units were soil subgroups. Polygons of the map usually contain one or more soil types. Figure 2b shows a histogram of the frequency distribution of the number of pedon sites per soil polygon. There were 55 polygons without pedon site and 26 polygons had more than 10 pedon sites. On average, each polygon had approximately five pedon sites. Figure 2c shows a histogram of the frequency distribution of the number of soil subgroups per polygon. There are a total of 152 soil polygons in the area. Almost all polygons contain more than three soil subgroups. Figure 2d shows a histogram of the frequency distribution of the area percent of dominant soil subgroups within polygons. The average area percent of all dominant soil subgroups is 55%, with a standard deviation of 16%.

### 1.2.3 Environmental variables

According to the SCORPAN (soil, climate, organisms, topography, parent material, age and space) concept (McBratney et al., 2003), environmental variables were assembled to characterize soil formative environments in this area. The annual mean temperature and the mean annual precipitation over 1950–2000 were used to represent climatic conditions, which were obtained from the WorldClim database at 1,000-m resolution (Hijmans et al., 2005). Shortwave infrared (SWIR) surface reflectance (500-m resolution) and diurnal land surface temperature (LST) difference (1,000-m resolution) from the Terra/Moderate Resolution Imaging Spectroradiometer (MODIS) on 9 May 2006 were selected to represent surficial deposits. This selection was based on their utility in distinguishing different surficial deposits (Dobos et al., 2000; Boettinger et al., 2008). A 16-day composite MODIS normalized difference vegetation index (NDVI) at 250-m resolution was acquired in July 2006 to represent vegetation conditions (Vandandorj et al., 2015). These MODIS data were provided by the Reverb Echo System (<http://reverb.echo.nasa.gov>). Elevation, slope gradient and surface curvature were used to represent terrain conditions. They were derived from the 90-m Shuttle Radar Topography Mission digital elevation model (DEM) which was provided by the Consortium for Spatial Information (version 4.1; <http://www.cgiar-csi.org/>). The surface curvature was calculated based on the algorithm proposed by Park and van de Giesen (2004) using the terrain analysis tool developed by Qin et al. (2009). It can reflect comprehensive curvature characteristics while other curvature indices (profile or plan curvature) generally reflect partial curvature information from specific directions (Shary et al., 2002). These environmental variables were gridded at a 90-m pixel size in order to make use of the detailed terrain information. Local terrain controls the way in which water and soil materials move through and over the land surface, thus determining soil erosion and deposition and exerting significant control on soil development (Moore et al., 1993).

## 1.3 Method

We assumed that similar environmental conditions at two sites would develop similar soil types and the more similar the environmental conditions, the more similar the soil types. With this assumption, a similarity-based prediction method was developed for soil polygon disaggregation at soil subgroup level through three steps of: (1) computing environmental similarities between pedon sites and each location in this area; (2) aggregating the similarities to derive similarity

distribution map for each soil subgroup; and (3) allocating candidate soil subgroups within polygons.

### 1.3.1 Computing environmental similarity

The study area was spatially discretized into a regular grid with 90 m×90 m pixel size. Each pixel in the grid was a prediction unit. Environmental values of the pixels and pedon sites were extracted from the data layers of annual precipitation, annual mean temperature, elevation, slope gradient, curvature, SWIR, diurnal LST difference and NDVI. The environmental similarity computation included three steps.

First, we computed the univariate-based similarity between each pedon site and each pixel using each individual environmental variable and a Gaussian function as Eq. 1,

$$S_{ij,m}^p = \exp\left[-\frac{(x_{p,ij} - x_{p,m})^2}{2\sigma_p^2}\right]. \quad (1)$$

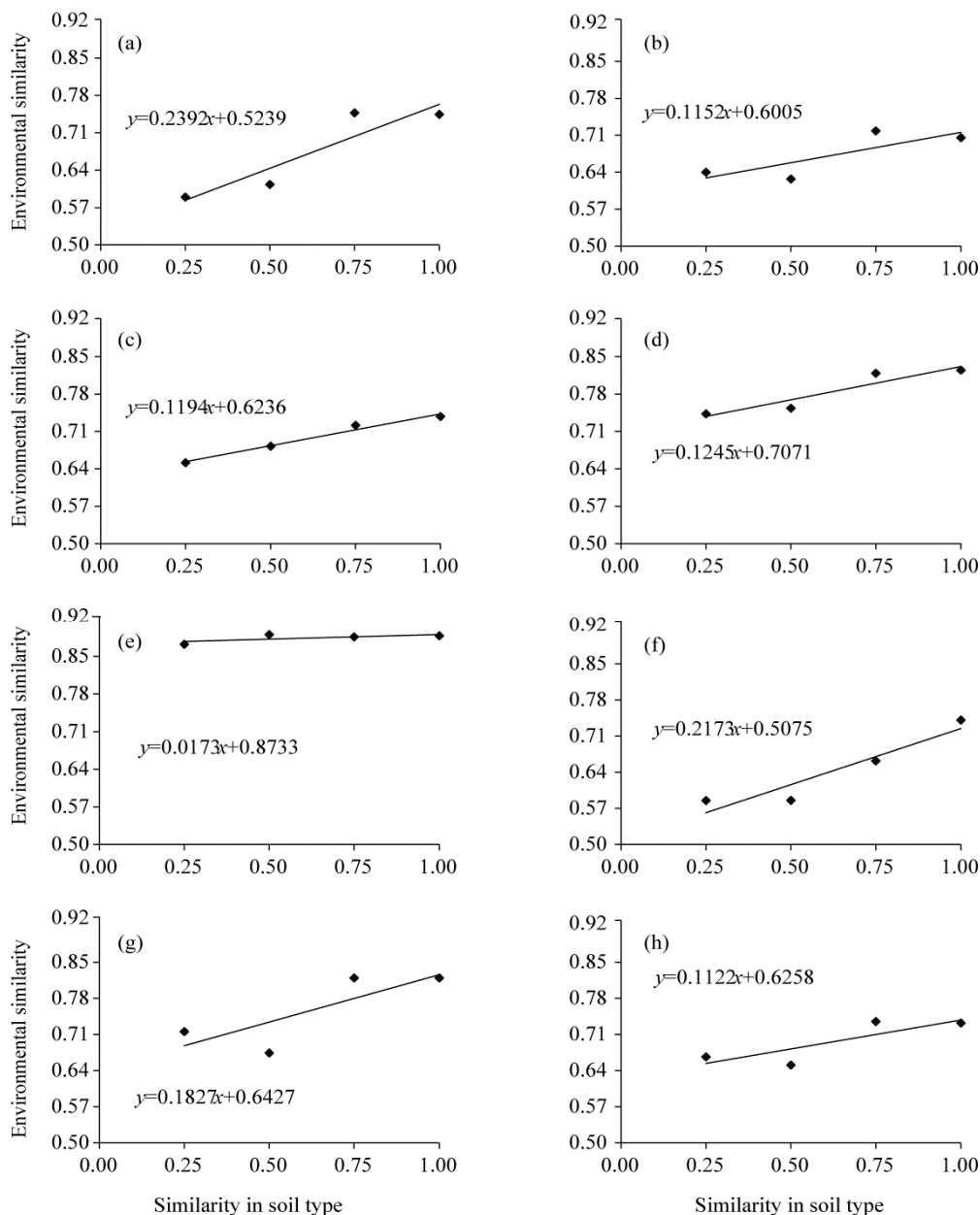
Where  $x_{p,m}$  and  $x_{p,ij}$  are the values of the  $p^{\text{th}}$  environmental variable at a pedon site ( $m$ ) and a pixel ( $ij$ ),  $\sigma_p$  is standard deviation of the data layer of the variable and  $S_{ij,m}^p$  is the univariate-based similarity between the pedon site and the pixel-based on this variable. The similarity would be equal to 1 for identical environmental values between the pedon site and the pixel, while it would be close to 0 for extremely different values.

Second, weight of an individual environmental variable in distinguishing soil subgroups was determined through analyzing the response relationships between the differences in soil types and that in the values of the variable. According to our assumption, similar soil types would have high environmental similarity while different soil types would have low environmental similarity. Their difference in the similarity can be an indicator of the ability of an environmental variable to distinguish soil types. The bigger the difference is, the stronger the ability of the variable would be. The variable with strong ability should be assigned a high importance or weight. Based on this idea, we defined four soil type similarity levels between the pedons in an increasing order: different soil order, same soil order but different great group, same great group but different subgroup and same soil subgroup. To facilitate data analysis, we arbitrarily set the four levels of the soil type similarity values as 0.25, 0.50, 0.75 and 1.00, respectively. The basis for the arbitrary setting includes two aspects. One is that soil classification is a hierarchical system which generally consists of order, great group, subgroup, family and series (Soil Classification Working Group, 1998). The order level reflects the difference in dominant soil-forming processes. The great group level forms by subdividing each order, which reflects the strengths of dominant processes. The subgroup forms by subdividing each great group, which reflects the type and horizons that indicate conformity to the central concept of the great group. The other is that the uniform setting for all variables can ensure that this arbitrary has little influence on the accuracy of the resulting relative importance. Then, for an individual variable, a matrix of univariate-based similarity between the pedon sites was computed by Eq. 1. The matrix was summarized to generate an average environmental similarity for each soil type similarity level. We finally performed a linear regression analysis for the average environmental similarity values against the soil type similarity values at the four levels (Fig. 3). The slope of the regression line indicated the ability of the variable to distinguish soil subgroups. The slope values for all variables were normalized to make the sum of the slope values equal to 1. The normalized slope values were used to approximate weights of the variables.

Third, with the weights of the variables, the univariate-based similarities were aggregated to generate an overall environmental similarity between each pedon site and each pixel using the following linear weighting function:

$$S_{ij,m} = \sum_{p=1}^N (w_p \times S_{ij,m}^p). \quad (2)$$

Where  $S_{ij,m}$  is the value of environmental similarity between a pixel ( $ij$ ) and a pedon site ( $m$ ),  $w_p$  is the weight of the  $p^{\text{th}}$  environmental variable, and  $N$  is the number of the environmental variables. Through this procedure, the values of environmental similarity of all pixels in this area to all pedon sites were obtained.



**Fig. 3** Fittings of regression lines between the similarities in soil types and the similarities in the environmental variables of annual mean temperature (a), mean annual precipitation (b), shortwave infrared surface reflectance (c), diurnal land surface temperature difference (d), normalized difference vegetation index (e), elevation (f), slope gradient (g) and surface curvature (h)

### 1.3.2 Deriving similarity distributions of soil subgroups

According to our assumption, the environmental similarity values of the pixels to the pedon sites were used to approximate the similarity in soil subgroups between the pixels and the pedon sites.

For a given pixel, the values of its similarity to the pedon sites with the same soil subgroup were aggregated to determine its similarity to the corresponding soil subgroup. Fuzzy maximum operation (Zhu et al., 1996) was adopted for the determination. Namely, among the similarity values of the pixel to the pedon sites with the same soil subgroup, the biggest value was considered as the similarity of the pixel to the corresponding soil subgroup. When all pixels were exhausted, we obtained spatial distribution map of soil similarity to each soil subgroup over the study area.

### 1.3.3 Allocating candidate soil subgroups within polygons

Within a given soil polygon, there are one or more soil subgroups recorded in the soil database. We considered them as candidate soil subgroups for the pixels within this polygon. A hardening process (Zhu et al., 2001) was performed on the similarity maps for soil subgroups to allocate candidate soil subgroups within this polygon. It was accomplished by assigning a pixel the candidate soil subgroup that had the biggest soil similarity to the pixel among all candidates within this polygon. Through the hardening process, each pixel within the polygon was labeled with only one soil subgroup. When all soil polygons were exhausted, we obtained a disaggregated map of soil subgroups over this area.

### 1.3.4 Evaluation criteria

Based on the 186 independent pedon sites, we evaluated the disaggregated map of soil subgroups. Overall accuracy and Kappa statistic (Sim and Wright, 2005) were used for the evaluation. The former is a simple ratio between the correctly predicted number of pedon sites and the overall number of the validation pedon sites. The latter is a coefficient of agreement, which takes into account agreement that could have occurred by chance. The value of the Kappa statistic ranges between -1 (perfect disagreement) and 1 (perfect agreement). Landis and Koch (1977) proposed the strength of agreement for the Kappa statistic as, slight for 0–0.20, fair for 0.21–0.40, moderate for 0.41–0.60, substantial for 0.61–0.80 and almost perfect for 0.81–1.00.

In addition, due to the lack of spatial specificity relating to individual soil types in the conventional soil maps, it is a common practice to portray only the most dominant soil type for each polygon even though most polygons contain many different soil types. This dominant value method spatially misrepresents soil data. The soil subgroup map produced by this method is called ‘dominant soil subgroup map’. Based on the same validation sites, we used the dominant map as a benchmark to evaluate the disaggregated soil subgroup map in order to examine the improvement in soil spatial representation.

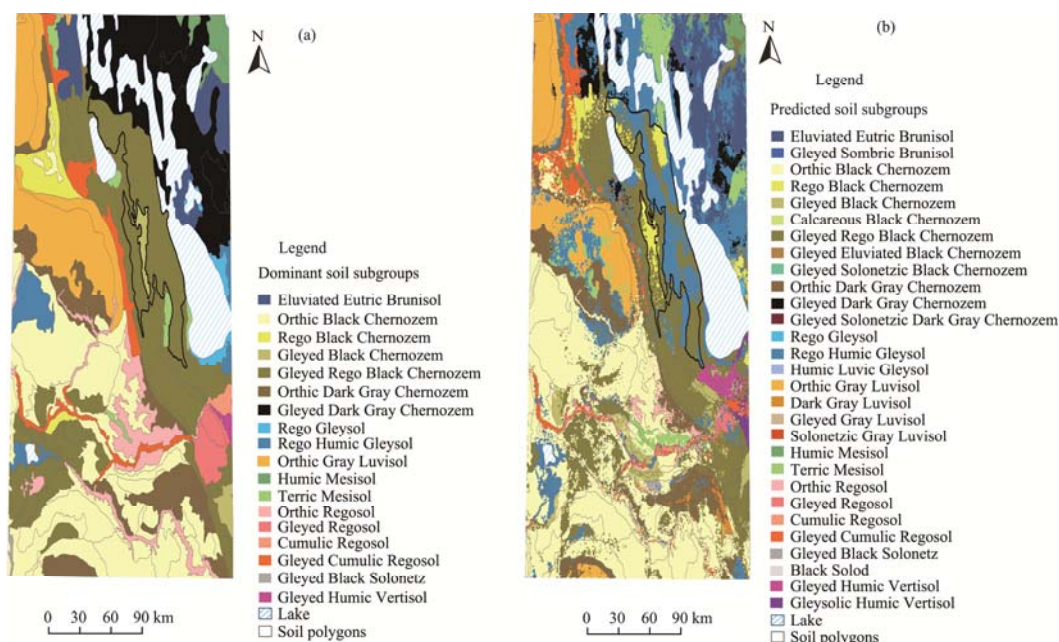
## 2 Results

Figure 4b shows the disaggregated map of soil subgroups over this area. The scale of the disaggregated map is about 1:300,000. Out of the 186 independent validation pedon sites, the proposed disaggregation method correctly predicted the soil subgroups for 125 pedon sites (an overall accuracy, 67%; a Kappa statistic value, 0.62), while the dominant soil subgroup map only correctly identified 84 pedon sites (overall accuracy, 45%; Kappa statistic value, 0.41). The difference of 22% in overall accuracy and 0.21 in Kappa statistic value indicated an improvement of approximate 49% relative to the dominant soil subgroup map commonly used in practice. In previous studies on soil polygon disaggregation, Grinand et al. (2008), Subburayalu and Slater (2012), and Odgers et al. (2014) produced much lower overall accuracy than our study. But, MacMillan (2004) and Li et al. (2012) achieved the accuracy greater than 65%. Huang et al. (2016) reported a total accuracy of 75% and Kappa statistic value of 70% for an updated soil map over a small hilly area.

The pattern of soil subgroups in the disaggregated map is generally consistent with that in the dominant map (Fig. 4). But, the disaggregated map provides considerable details than the dominant map, largely reducing soil spatial misrepresentation. There are 29 soil subgroups mapped on the disaggregated map while there are only 18 soil subgroups appeared on the dominant map. A total of 11 soil subgroups are absent in the dominant map due to their relatively



small area percentages within polygons and the use of the dominant value mapping method. The disaggregated map represents geographic locations of both the most dominant soil subgroups and other soil subgroups while the dominant map exaggerates geographic extent of the most dominant soil subgroup to the whole polygon. The exaggeration could be very large as most soil polygons in this area contain three to seven soil subgroups (Fig. 2c). Take an example of the soil polygon located immediately west of the Lake Manitoba, whose border is highlighted by bold black line in Fig. 4. According to the soil database, there are 5 soil subgroups recorded in this polygon: Gleyed Rego Black Chernozem, Rego Humic Gleysol, Rego Black Chernozem, Orthic Black Chernozem and Terric Mesisol. Their area percentages within the polygon are 35%, 34%, 25%, 3% and 3%, respectively. On the dominant map, Gleyed Rego Black Chernozem was selected to represent soil conditions of the entire polygon and its areal percentage was exaggerated from 35% to 100%. On the disaggregated map, the first three (major) soil subgroups were mapped with areal percentages in the polygon as 42%, 45% and 13%, respectively. These predicted areal percentages, to a large extent, reflect the reality of the composition of soil subgroups within this polygon.



**Fig. 4** Dominant map of soil subgroups commonly used in practice (a) and the disaggregated map of soil subgroups generated in this study (b)

Table 1 shows a part of the cross-tabulation between the observed and predicted soil subgroups at the validation pedon sites. It was found that 77% of the wrong prediction sites (i.e. 47 out of 61) were highly similar in soil taxonomy, especially those soil subgroups with gley features. For example, Gleyed Black Chernozem was not differentiated from Gleyed Rego Black Chernozem at 13 sites. Gleyed Rego Black Chernozem was predicted as Rego Humic Gleysol at 10 sites. Gleyed (Rego) Black Chernozem was not differentiated from Gleyed (solon) Humic Vertisol at 6 sites and from Gleyed (Cumelic) Regosol at 9 sites. Orthic Black Chernozem was not differentiated from Gleyed (Rego) Black Chernozem at 9 sites. This is not surprising and can be explained by the interpretation ability of the environmental variables used in this study. They failed to capture subtle differences in soil formative environments between these soil subgroups. This is in line with Dobos et al. (2000) who reported low ability of soil prediction techniques in differentiating the soil types that were very similar in soil taxonomy.

It was also found that approximate 70% of the wrong prediction sites were located in the plain area mostly under cultivation. The plain has a very low terrain relief with a slope gradient less than 2%. The cropping, to some extent, weakens the correlation between soils and vegetation conditions. The effectiveness of terrain and vegetation variables in predicting soil subgroups was



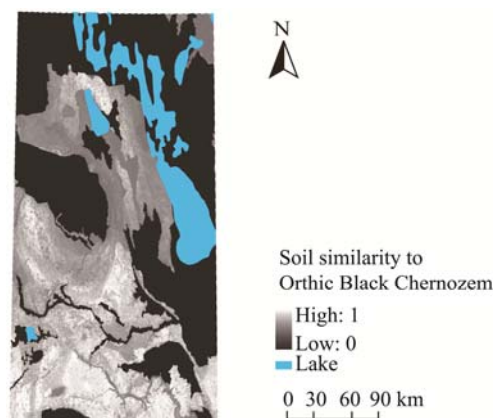
thus reduced in this area. This is consistent with the findings of previous studies. Odeh and McBratney (2000) stated that terrain attributes can predict soils in regions of medium to high relief but the effectiveness becomes low in areas with low relief. Scull et al. (2005) reported that the classification tree model failed to produce satisfactory soil type prediction accuracy in a low relief area mainly because terrain variables did not effectively indicate the distribution of soil types. Odgers et al. (2014) argued that covariates that are effective in strongly dissected landscapes may not be as effective as in flat areas. Yahiaoui et al. (2015) found that topographical variables still showed low correlations with soil salinity in an alluvial plain although soil samples were collected from a topographical perspective. These suggest that new covariates are needed for the plain areas.

**Table 1** Cross-tabulation between the observed and predicted soil subgroups at the validation pedon sites\*

Predicted soil subgroups	Observed soil subgroups							
	O.BLC	GL.BLC	GLR.BLC	R.HG	GL.R	GLCU.R	GL.HV	GLC.HV
O.BLC	28	0	3	0	0	0	0	0
GL.BLC	2	9	9	0	1	1	1	0
GLR.BLC	4	4	29	7	2	1	2	0
R.HG	0	0	3	9	0	0	0	0
GL.R	0	2	0	0	1	0	0	0
GLCU.R	0	1	1	0	0	2	0	0
GL.HV	0	0	1	0	0	0	3	0
GLC.HV	0	0	2	0	0	0	0	1

Note: \*, Due to the big size of the complete cross-tabulation table, only the soil subgroups that are difficult to differentiate from each other are listed here; O.BLC, Orthic Black Chernozem; GL.BLC, Gleyed Black Chernozem; GLR.BLC, Gleyed Rego Black Chernozem; R.HG, Rego Humic Gleysol; GL.R, Gleyed Regosol; GLCU.R, Gleyed Cumulic Regosol; GL.HV, Gleyed Humic Vertisol; GLC.HV, Gleysolic Humic Vertisol.

In addition, the proposed disaggregation method also produced similarity map for each soil subgroup over the study area. Figure 5 is one such soil similarity map, which illustrates spatial distribution of soil similarity to the soil subgroup, Orthic Black Chernozem. The white color represents high similarity to this soil subgroup and the black color represents low similarity to this soil subgroup. The whiter the color, the higher the similarity to this soil subgroup is. The similarity maps actually construct a multi-dimensional soil subgroup similarity space in which the similarity to each soil subgroup was considered as one dimension. Each pixel in the area thus can



**Fig. 5** Distribution of soil similarity to soil subgroup, Orthic Black Chernozem

find its location in this soil subgroup similarity space. Namely, the soil type of a pixel can be described as a set of values of similarity to all soil subgroups. MacMillan (2004) noted that this is a realistic way to describe soil type of a pixel considering that few pixels can be expected to be totally representative of any given soil type.

### 3 Discussion

#### 3.1 Legacy pedon data

Ashtekar and Owens (2013) pointed out that it is important for the digital soil mapping community not to lose sight of the wealth of legacy soils data acquired by historical soil surveys. Our study demonstrated the utility of the legacy pedon data for disaggregating conventional soil maps. They are a valuable soil data source for soil data updating using soil mapping techniques. The legacy pedon data contain implicit soil-environmental relationships at specific locations. They have the advantage of capturing local details of soil patterns but the disadvantage of neglecting overall characteristics of soil patterns especially when the number of soil samples are limited and not representative. In contrast, the legacy soil maps contain implicit soil-environmental relationships over specific spatial extent. They have the advantage of capturing overall characteristics of soil pattern but the disadvantage of neglecting local details especially when the map scale is small. With currently available soil formative data such as DEM and remotely sensed images, the implicit soil-environmental relationships can be used to allocate soil types to specific landscape locations within polygons. Therefore, it could be expected that integrating the two types of legacy soil data would improve current disaggregation performance.

Inconsistency in the legacy pedon data, if exists, should be considered when using such data. Different soil classification systems, such as the Genetic Soil Classification of China, the Chinese Soil Taxonomy (Cooperative Research Group on Chinese Soil Taxonomy, 2001) and the World Reference Base for Soil Resources (IUSS (International Union of Soil Science) Working Group WRB, 2014), may be involved in the definition of soil types of the pedons. In this situation, soil type names should be cross-referenced to one classification system for the disaggregation of conventional soil map. In addition, legacy pedon data can also be used for mapping soil physical and chemical properties. For this purpose, two types of inconsistency should be considered. First, the sampling of pedons generally involves dividing a pedon into horizons at the vertical dimension. The vertical divisions among the pedons can be very different, making it difficult to know soil property values at a specified depth. Second, the pedon data are often collected from different soil survey projects at different historical stages. It is possible that different laboratory analysis methods were used in measuring the soil properties. In both situations, data harmonization is necessary for preparing the legacy pedon data for soil property mapping. For the pedon data with inconsistent depth divisions, depth functions such as weighted average and equal-area quadratic spline functions can be used to derive soil properties at a set of standardized depth intervals (Malone et al., 2009). For the pedon data derived from different laboratory analysis methods, the specifications of the GlobalSoilMap.net has provided some regression equations for harmonizing multi-source pedon data to a reference standard (GlobalSoilMap Science Committee, 2015).

#### 3.2 Similarity-based method

Conventional soil surveyors observed a soil landscape in the field and then interpreted the soil landscape as soil spatial distribution according to personal experiences. To confirm the interpretation and identify the exact boundary of soil type distribution, they subjectively selected some locations for soil sampling and legacy pedon data were obtained. Because the distribution of the legacy pedons does not follow a statistical design, the use of legacy pedon data poses challenges on statistical and geostatistical methods that commonly used in digital soil mapping. This study demonstrated the potential of the similarity-based prediction method to make use of the legacy pedon data for soil polygon disaggregation. This potential can be attributed to its three advantages. First, it has no requirement on the distribution of pedon sites in both geographical and attribution spaces because the prediction is performed in the attribution space and relies on similar pedon sites. Individual pedon site contributes in a locally limited way to the prediction results. That is, a pedon site in a soil subgroup does not have a global influence on the prediction results for other soil subgroups. Thus, the method is applicable to the legacy pedon data that suffer from under- or over-sampling problems. Second, it has the capability of modeling complex

and nonlinear relationships between soil and environment. The similarity-based prediction is a local approximation. A collection of local approximations have large flexibility to deal with any relationships. This is necessary for soil polygon disaggregation over such a large area where soil-environmental relationships exhibit high heterogeneity. Third, it is easy to interpret the prediction results due to the relatively transparent prediction process. From the prediction process, the most similar pedon site to a pixel can be known, on which the prediction at the pixel is primarily based. It is thus possible for soil scientists to check how a certain soil subgroup prediction is made at a location. This makes it more attractive than other black-box machine learning techniques such as artificial neural networks.

However, some limitations should be mentioned for the use of this method. First, the performance of the method is affected by the effectiveness of environmental covariates. Terrain and vegetation variables may not be effective for differentiating soil subgroups in agricultural plain areas. For such areas, new soil covariates may need to be developed. Second, it is important to define appropriate weights for environmental variables in soil prediction. Although we developed an idea for determining the weights in this study, further improvements are required.

## 4 Conclusions

This study proposed an approach to spatially disaggregate soil polygons based on the combination of legacy pedon data with the similarity-based prediction method. We come to the following conclusions. First, legacy pedon data, which imply soil-environmental relationships at specific locations, can be effectively used for spatial disaggregation of soil polygons. It could be a necessary complement to the legacy soil maps which have been demonstrated many times in previous studies to be effective for the disaggregation. Second, the similarity-based prediction is effective for making use of the under- or over-sampled legacy pedon data for soil polygon disaggregation. Moreover, its prediction results are easy to interpret due to a relatively simple and transparent prediction process. Third, environmental covariates are critical for accurate soil subgroup predictions. It is necessary to develop new covariates which can reflect soil differences for the agricultural plain area and the soil types that are similar in taxonomy.

In addition to the map of Soil Landscapes of Canada, many national and regional soil maps such as the Soil Database of China, Soil Survey Geographic Database, the European Soil Database and the Harmonized World Soil Database also have the problem of lacking spatial specificity of soil types within soil polygons. The approach presented here provides an alternative solution for dealing with this problem and producing updated soil maps to meet site-specified agricultural management and environmental modelling.

## Acknowledgements

This study was supported by the National Natural Science Foundation of China (41130530, 91325301, 41431177, 41571212, 41401237), the Project of “One-Three-Five” Strategic Planning & Frontier Sciences of the Institute of Soil Science, Chinese Academy of Sciences (ISSASIP1622), the Government Interest Related Program between Canadian Space Agency and Agriculture and Agri-Food, Canada (13MOA01002), and the Natural Science Research Program of Jiangsu Province (14KJA170001).

## References

- Arrouays D, Grundy M G, Hartemink A E, et al. 2014. GlobalSoilMap: toward a fine-resolution global grid of soil properties. *Advances in Agronomy*, 125: 93–134.
- Ashtekar J M, Owens P R. 2013. Remembering knowledge: an expert knowledge based approach to digital soil mapping. *Soil Horizons*, 54(5), doi: 10.2136/sh13-01-0007.
- Boettinger J L, Ramsey R D, Bodily J M, et al. 2008. Landsat spectral data for digital soil mapping. In: Hartemink A E, McBratney A B, Mendonça-Santos M L. *Digital Soil Mapping With Limited Data*. Dordrecht: Springer, 193–202.
- Boruvka L, Kozak J, Nemecek J, et al. 2002. New approaches to the exploitation of former soil survey data. In: 17<sup>th</sup> World

- Congress of Soil Science Bangkok, Thailand: IUSS.
- Carré F, McBratney A B, Minasny B. 2007. Estimation and potential improvement of the quality of legacy soil samples for digital soil mapping. *Geoderma*, 141(1–2): 1–14.
- Collard F, Kempen B, Heuvelink G B M, et al. 2014. Refining a reconnaissance soil map by calibrating regression models with data from the same map (Normandy, France). *Geoderma Regional*, 1: 21–30.
- Cooperative Research Group on Chinese Soil Taxonomy. 2001. Keys to Chinese Soil Taxonomy (3<sup>rd</sup> ed.). Hefei: University of Science and Technology of China Press. (in Chinese)
- Diem J E, Comrie A C. 2002. Predictive mapping of air pollution involving sparse spatial observations. *Environmental Pollution*, 119(1): 99–117.
- Dobos E, Micheli E, Baumgardner M F, et al. 2000. Use of combined digital elevation model and satellite radiometric data for regional soil mapping. *Geoderma*, 97(3–4): 367–391.
- Du F, Zhu A X, Band L E, et al. 2014. Soil property variation mapping through data mining of soil category maps. *Hydrological Processes*, 29(11): 2491–2503.
- Geng X Y, Fraser W, VandenBygaart B, et al. 2010. Toward digital soil mapping in Canada: existing soil survey data and related expert knowledge. In: Boettinger J L, Howell D W, Moore A C, et al. *Digital Soil Mapping: Bridging Research, Environmental Application, And Operation*. Dordrecht: Springer, 325–335.
- GlobalSoilMap Science Committee. 2015. GlobalSoilMap specifications-Tiered<sup>1</sup> *GlobalSoilMap* products. Release 2.4. [2015-07-12]. <http://globalsoilmap.net/specifications>.
- Grinand C, Arrouays D, Laroche B, et al. 2008. Extrapolating regional soil landscapes from an existing soil map: sampling intensity, validation procedures, and integration of spatial context. *Geoderma*, 143(1–2): 180–190.
- Hijmans R J, Cameron S E, Parra J L, et al. 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 25(15): 1965–1978.
- Holt A. 1999. Applying case-based reasoning techniques in GIS. *International Journal of Geographical Information Science*, 13(1): 9–25.
- Huang W, Luo Y, Wang S Q, et al. 2016. Knowledge of soil-landscape model obtain from a soil map and mapping. *Acta Pedologica Sinica*, 53(1): 72–80. (in Chinese)
- IUSS Working Group WRB. 2014. World Reference Base for Soil Resources 2014: International soil classification system for naming soils and creating legends for soil maps. *World Soil Resources Reports* FAO, Rome.
- Ju W M, Chen J M. 2005. Distribution of soil carbon stocks in Canada's forests and wetlands simulated based on drainage class, topography and remotely sensed vegetation parameters. *Hydrological Processes*, 19(1): 77–94.
- Lagacherie P. 2008. Digital soil mapping: a state of the art. In: Hartemink A E, McBratney A B, Mendonça-Santos M L. *Digital Soil Mapping With Limited Data*. Dordrecht: Springer, 3–14.
- Landis J R, Koch G G. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1): 159–174.
- Li Z, Huffman T, Zhang A N, et al. 2012. Spatially locating soil classes within complex soil polygons-mapping soil capability for agriculture in Saskatchewan Canada. *Agriculture, Ecosystems and Environment*, 152: 59–67.
- MacMillan R A. 2004. Automated knowledge-based classification of landforms, soils and ecological spatial entities. Edmonton: LandMapper Environmental Solutions. [2009-03-03]. <http://www.georeference.org/Forum/e32412F39303135342F31322D426F624D61634D696C6C69616E2E646F6312-BobMacMillan.doc>.
- Malone B P, McBratney A B, Minasny B, et al. 2009. Mapping continuous depth functions of soil carbon storage and available water capacity. *Geoderma*, 154(1–2): 138–152.
- McBratney A B. 1998. Some considerations on methods for spatially aggregating and disaggregating soil information. *Nutrient Cycling in Agroecosystems*, 50(1–3): 51–62.
- McBratney A B, Mendonça Santos M L, Minasny B. 2003. On digital soil mapping. *Geoderma*, 117(1–2): 3–52.
- Minasny B, McBratney A B, Lark R M. 2008. Digital soil mapping technologies for countries with sparse data infrastructures. In: Hartemink A E, McBratney A B, Mendonça-Santos M L. *Digital Soil Mapping With Limited Data*. Dordrecht: Springer, 15–30.
- Moore I D, Gessler P E, Nielsen G A, et al. 1993. Soil attribute prediction using terrain analysis. *Soil Science Society of America Journal*, 57(2): 443–452.
- Myers D E. 1994. Spatial interpolation: an overview. *Geoderma*, 62(1–3): 17–28.
- Nauman T W, Thompson J A. 2014. Semi-automated disaggregation of conventional soil maps using knowledge driven data mining and classification trees. *Geoderma*, 213: 385–399.
- Odeh I O A, McBratney A B. 2000. Using AVHRR images for spatial prediction of clay content in the lower Namoi Valley of eastern Australia. *Geoderma*, 97(3–4): 237–254.

- Odgers N P, Sun W, McBratney A B, et al. 2014. Disaggregating and harmonising soil map units through resampled classification trees. *Geoderma*, 214–215: 91–100.
- Park S J, van de Giesen N. 2004. Soil-landscape delineation to define spatial sampling domains for hillslope hydrology. *Journal of Hydrology*, 295(1–4): 28–46.
- Penížek V, Borůvka L. 2004. Processing of conventional soil survey data using geostatistical methods. *Plant Soil and Environment*, 50(8): 352–357.
- Pla A, López B, Gay P, et al. 2013. eXiT\*CBR.v2: Distributed case-based reasoning tool for medical prognosis. *Decision Support Systems*, 54(3): 1499–1510.
- Qin C Z, Lu Y J, Bao L L, et al. 2009. Simple digital terrain analysis software (SimDTA\_1.0) and its application in fuzzy classification of slope positions. *Journal of Geo-information Science*, 11(6): 737–743. (in Chinese)
- Sauchyn D J. 2001. Modeling the hydroclimatic disturbance of soil landscapes in the southern Canadian plains: the problems of scale and place. *Environmental Monitoring and Assessment*, 67(1–2): 277–291.
- Schut P, Smith S, Fraser W, et al. 2011. Soil landscapes of Canada: building a national framework for environmental information. *Geomatica*, 65(3): 293–309.
- Scully P, Franklin J, Chadwick O A. 2005. The application of classification tree analysis to soil type prediction in a desert landscape. *Ecological Modelling*, 181(1): 1–15.
- Shary P A, Sharaya L S, Mitusov A V. 2002. Fundamental quantitative methods of land surface analysis. *Geoderma*, 107(1–2): 1–32.
- Sim J, Wright C C. 2005. The Kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical Therapy*, 85(3): 257–268.
- Smith S, Bulmer C, Flager E, et al. 2010. Digital soil mapping at multiple scales in British Columbia, Canada. In: Program and Abstracts, 4<sup>th</sup> Global Workshop on Digital Soil Mapping Rome, Italy.
- Soil Classification Working Group. 1998. The Canadian System of Soil Classification (3<sup>rd</sup> ed.). Ottawa: Agriculture and Agri-Food Canada Publication.
- Soil Survey Staff. 2014. Keys to Soil Taxonomy (12<sup>th</sup> ed.). Washington: United States Department of Agriculture.
- Subburayalu S K, Slater B K. 2012. Soil series mapping by knowledge discovery from an Ohio County Soil Map. *Soil Science Society of America Journal*, 77(4): 1254–1268.
- Vandendorj S, Gantsetseg B, Boldgiv B. 2015. Spatial and temporal variability in vegetation cover of Mongolia and its implications. *Journal of Arid Land*, 7(4): 450–461.
- Wang J F, Stein A, Gao B B, et al. 2012. A review of spatial sampling. *Spatial Statistics*, 2: 1–14.
- Welsted J, Everitt J, Stadel C. 1996. The Geography of Manitoba: Its Land and Its People. Manitoba: University of Manitoba Press.
- Yahiaoui I, Douaoui A, Zhang Q, et al. 2015. Soil salinity prediction in the Lower Cheliff plain (Algeria) based on remote sensing and topographic feature analysis. *Journal of Arid Land*, 7(6): 794–805.
- Yang L, Jiao Y, Fahmy S, et al. 2011. Updating conventional soil maps through digital soil mapping. *Soil Science Society of America Journal*, 75(3): 1044–1053.
- Zhu A X, Band L E, Dutton B, et al. 1996. Automated soil inference under fuzzy logic. *Ecological Modelling*, 90(2): 123–145.
- Zhu A X, Hudson B, Burt J, et al. 2001. Soil mapping using GIS, expert knowledge, and fuzzy logic. *Soil Science Society of America Journal*, 65(5): 1463–1472.
- Zhu A X, Liu J, Qin C Z, et al. 2010. Soil property mapping over large areas using sparse ad-hoc samples. In: 19th World Congress of Soil Science: Soil Solutions for a Changing World Brisbane, Australia.