

Updating Conventional Soil Maps through Digital Soil Mapping

Lin Yang

State Key Lab. of Resources and Environment Information System
Institute of Geographical Sciences and Resources Research
Chinese Academy of Sciences
Beijing 100101, China

You Jiao

Agrifoods Development Branch
Dep. of Natural Resources
Corner Brook, NF A2H6J8, Canada
and
Potato Research Centre
Agriculture and Agri-Food Canada
Fredericton, NB E3B4Z7, Canada

Sherif Fahmy

Potato Research Centre
Agriculture and Agri-Food Canada
Fredericton, NB E3B4Z7, Canada

A-Xing Zhu*

State Key Lab. of Resources and Environment Information System
Institute of Geographical Sciences and Resources Research
Chinese Academy of Sciences
Beijing 100101, China
and
Dep. of Geography
Univ. of Wisconsin
550 N. Park St.
Madison, WI 53706
and
State Key Lab. of Remote Sensing Science
Institute of Remote Sensing Applications
Chinese Academy of Sciences
Beijing 100101, China

Sheldon Hann

Potato Research Centre
Agriculture and Agri-Food Canada
Fredericton, NB E3B4Z7, Canada

James E. Burt

Dep. of Geography
Univ. of Wisconsin
550 N. Park St.
Madison, WI 53706

Feng Qi

Dep. of Geology and Meteorology
Kean Univ.
1000 Morris Ave.
Union, NJ 07083

Conventional soil maps, as the major data source for information on the spatial variation of soil, are limited in terms of both the level of spatial detail and the accuracy of soil attributes. These soil maps, however, contain valuable knowledge on soil–environment relationships. Such knowledge can be extracted for updating conventional soil maps through the use of available high-quality data on environmental variables and data analysis techniques. We developed a method to update conventional soil maps using digital soil mapping techniques without additional field work, which can be used in situations where the study area contains no or few soil profile descriptions at points. The basis of the method is that soil polygons on a conventional soil map correspond to landscape units, which can be considered as combinations of environmental factors. Such environmental combinations were approximated through fuzzy clustering on the environmental factors. We extracted the knowledge on soil–environment relationships by relating the environmental combinations to the mapped soil types. The extracted knowledge was then used for soil mapping using the Soil Land Inference Model (SoLIM) framework. This method was demonstrated through a case study for updating a conventional 1:20,000 soil map of Wakefield, NB, Canada. The case study showed that the updated digital soil map contained much greater spatial detail than the conventional soil map. Field validation indicated that the accuracy of the updated soil map was much higher than the conventional soil map at the level of soil associations with drainage classes, indicating that the proposed method is an effective approach to updating conventional soil maps.

Abbreviations: DEM, digital elevation model; FCM, fuzzy *c*-means; TWI, topographic wetness index.

Information on the spatial distribution of soils is increasingly required for watershed management and ecological modeling applications (Zhu et al., 2001; Brus et al., 2008; Miao et al., 2010). At present, conventional soil maps produced in the past decades are the major data sources for these applications. In the United States, the State Soil Geographic database (STATSGO) and Soil Survey Geographic database (SSURGO) created by the NRCS are the most commonly used conventional soil databases. The SSURGO maps, compiled at scales between 1:12,000 and 1:63,360, are the most detailed products of conventional soil mapping. Similarly in Canada, soil surveys at scales from 1:10,000 to 1:250,000 have been published for most of the agricultural areas and many surrounding areas. These represent the most detailed soil inventory information in the National Soil Database in Canada. Such conventional soil maps are widely available and used extensively for many applications. Due to the limitations of conventional mapping techniques and the cartographic model used, however, these conventional soil maps are limited in terms of both the level of spatial detail and the accuracy of the soil attributes (Zhu, 1997; Zhu et al., 2001). There is a need, therefore, to update conventional soil maps to provide detailed and accurate soil information (Brus et al., 1992; Rossiter, 2008).

The traditional way to update a soil map is to obtain new field samples from new soil surveys. This can be very costly, however, considering the extensive fieldwork needed to obtain the number of field soil samples required. Recently emerged digital soil mapping technology, which uses advanced geographic infor-

Soil Sci. Soc. Am. J. 75:1044–1053

Published online 26 Apr. 2011

doi:10.2136/sssaj2010.0002

Received 1 Jan. 2010.

*Corresponding author (axing@lreis.ac.cn).

© Soil Science Society of America, 5585 Guilford Rd., Madison WI 53711 USA

All rights reserved. No part of this periodical may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Permission for printing and for reprinting the material contained herein has been obtained by the publisher.

mation acquisition and processing techniques (Zhu et al., 2001; McBratney et al., 2003), can play an important role in updating conventional soil maps (Rossiter, 2008). By using more detailed ancillary data and advanced data analysis techniques, it is possible to improve the spatial detail and accuracy of conventional soil maps without much additional field sampling.

A few studies have investigated methods to update conventional soil maps through digital soil mapping. Brus et al. (2008) used Bayesian Maximum Entropy to estimate the probabilities of the occurrences of soil categories and predicted soil spatial variation using existing field observation data in the soil information system (legacy data) as hard information and the conventional soil map as soft information. Many samples are needed, however, for calibration of the probability model. Vitharana et al. (2008) evaluated the potential of detailed observations (electrical conductivity) from a proximal soil sensor to upgrade the 1:20,000 soil map of Belgium. The relationship between electrical conductivity with depth of the Tertiary clay substratum was modeled using 60 calibration points and the detailed map of the depth of the Tertiary clay substratum generated with regression kriging was then used to upgrade the 1:20,000 soil map. This study demonstrated the potential of combining existing soil maps with proximal soil sensing technology, but it also needed abundant field samples to establish the relationships between electrical conductivity with depth of the Tertiary clay substratum. In a more recent application, Kempen et al. (2009) used a multinomial logistic regression approach to update the 1:50,000 Dutch soil map using legacy soil sample data and ancillary environmental data. The accuracy of the updated map, as assessed by the validation samples, was 58%, which was 6% higher than that of the conventional soil map. Because the methods investigated in most previous studies on updating conventional soil maps have relied on a large amount of field sample data, these approaches may have limited uses for areas where soil samples are absent or sparse. The objective of this study was to develop a method to update conventional soil maps without the requirement of field samples, which is appropriate in situations where the soil database (soil information system) does not contain point observations but only conventional soil maps.

MATERIALS AND METHODS

In conventional soil mapping, soil experts first construct a soil-landscape model across the area through intensive field work. The soil-landscape model describes the relationships between the local soils and unique landscape units that can be defined by specific environmental conditions. Soil scientists draw soil polygons based on the perceived distribution of landscape units through aerial photo interpretation. The resulting soil polygons for a soil type thus correspond to the spatial distributions of the landscape units (Hudson, 1992; Moran and Bui, 2002; Qi and Zhu, 2003; Qi et al., 2008). Similar to the manual process of identifying landscape units based on the combinations of environmental conditions through photo interpretation, fuzzy clustering of the environmental variables could generate environmental combinations that could be used to approximate the landscape units identified by soil experts (Yang et al., 2007). We thus assumed that there is a correspond-

ing relationship between the clustered environmental combinations and the mapped soil polygons. Therefore, the basic idea of this study was to associate the soil types on a conventional soil map with the environmental combinations generated using fuzzy clustering, then extract and quantify the knowledge of soil-environment relationships based on the assumption that soils that occur at locations with a high fuzzy membership of the environmental clusters represent the typical soil type associated with the environmental clusters. This knowledge could then be used to produce updated soil maps of the area in conjunction with detailed and accurate environmental data, which is increasingly available. We expected that the updated soil map would have not only a greater spatial detail but also a higher accuracy than the conventional soil map.

The method consists of five major steps: (i) development of an environmental database; (ii) generation of environmental combinations (clusters); (iii) association of environmental clusters with soil types; (iv) extraction and quantification of knowledge on soil-environment relationships; and (v) soil inference.

Development of Environmental Database

The first step is to choose environmental covariates that are relevant to the local soils and construct a geographic information system database of these variables. Most commonly used environmental covariates include parent material, vegetation, and topographic variables such as elevation, slope gradient, slope aspect, planform curvature, profile curvature, and wetness index (McSweeney et al., 1994; Zhu et al., 1997). The environmental database may vary from area to area because of the difference in pedogenesis across different areas.

Generation of Environmental Combinations (Clusters)

We used a fuzzy *c*-means (FCM) clustering approach to generate environmental combinations in this step. Fuzzy *c*-means clustering is an unsupervised classifier that has been extensively used in soil science and terrain analysis (de Bruin and Stein, 1998; Lark, 1999; Burrough et al., 2000; English, 2001; Hanesch et al., 2001; Yang et al., 2007). It optimally partitions a data set (such as the developed environmental database) into a set of classes and computes the membership of each data element in each of the classes (Bezdek et al., 1984). The centroids of the classes are identified by minimizing the fuzzy partition error (Bezdek et al., 1984):

$$J_m(\mathbf{U}, \mathbf{v}) = \sum_{k=1}^n \sum_{i=1}^c (u_{ik})^m d_{ik}^2 \quad y \in Y \quad [1]$$

$$d_{ik}^2 = \|y_k - v_i\|_{\mathbf{A}}^2 = (y_k - v_i)^{\top} \mathbf{A} (y_k - v_i) \quad [2]$$

where Y is the data set, n is the number of objects in Set Y , c is the number of clusters, m is a weighting exponent, \mathbf{U} is a $c \times n$ matrix, $\mathbf{U} = (u_{ik})$ where u_{ik} is the membership of the k th object belonging to the i th cluster, \mathbf{v} is a vector of cluster centers, d_{ik} is the weighted distance between y_k and v_i , and \mathbf{A} is a weighting matrix for performing distance calculation, such as Euclidean, diagonal, and Mahalanobis distances. The fuzzy partition error, J_m , can be described as a weighted measure of the squared distance between pixels and class centroids and so is a measure of the total squared errors as minimized with respect to each cluster (Ahn et al.,

1999); J_m decreases as the clustering improves, which means that pixels tend to be overall closer to their representative centroids.

Optimal fuzzy clusterings of Y are defined as pairs of (\hat{U}, \hat{V}) that locally minimize J_m (Bezdek et al., 1984). The equations for \hat{u}_{ik} and \hat{v}_i are

$$\hat{v}_i = \frac{\sum_{k=1}^n (\hat{u}_{ik})^m y_k}{\sum_{k=1}^n (\hat{u}_{ik})^m} \quad 1 \leq i \leq c \quad [3]$$

$$\hat{u}_{ik} = \left[\sum_{j=1}^c \left(\frac{\hat{d}_{ik}}{\hat{d}_{jk}} \right)^{2/(m-1)} \right]^{-1} \quad 1 \leq k \leq n, 1 \leq i, j \leq c \quad [4]$$

It is possible to find J_m by iterating between these two equations until changes between successive iterations are below a prescribed tolerance (Bezdek et al., 1984).

It is often difficult to know the number of classes that best describe the structure in the data set when using FCM clustering. To determine the optimum number of clusters, two cluster validity measures, the partition coefficient (F) and the normalized entropy (H) (Bezdek et al., 1984), were utilized in our study to judge the effectiveness of the clustering results:

$$F_c(\hat{u}) = \sum_{k=1}^n \sum_{i=1}^c \frac{\hat{u}_{ik}^2}{n} \quad [5]$$

$$H_c(\hat{u}) = - \sum_{k=1}^n \sum_{i=1}^c \frac{\hat{u}_{ik} \log_a(\hat{u}_{ik})}{n} \quad [6]$$

The partition coefficient F will take the values of $1/c$ to 1, while entropy H ranges from zero to $\log_a(c)$ (Ahn et al., 1999). While F measures the amount of overlap between clusters and is inversely proportional to the overall average overlap between pairs of fuzzy sets (Ahn et al., 1999), H is a scalar measure of the amount of fuzziness in a given fuzzy partition U (Bezdek et al., 1984). Normally, H increases and F decreases as the number of clusters increases. We can examine the improvement in F or H across adjacent clustering to determine the optimal number of clusters, c (English, 2001; Yang et al., 2007). When the increment of H ($H_{c_1} - H_{c_1-1}$) with cluster number changing from $(c_1 - 1)$ to c_1 is smaller than the increment with cluster number changing from $(c_1 - 2)$ to $(c_1 - 1)$ and that from c_1 to $(c_1 + 1)$, the current clustering can be considered as a satisfying partition of the data set and c_1 is a possible optimal cluster number. Several possible optimal cluster numbers could be obtained this way. The optimal cluster number should be determined based on information about environmental niches and the soil variation of the study area.

When running FCM on an environmental database, if categorical variables (such as parent material) are included in the environmental database, these variables will be used to stratify the whole study area into uniform strata. The FCM classification will then be applied to each stratum. The results of FCM clustering include a cluster centroids file and fuzzy membership maps of the environmental combinations (clusters). The cluster centroids file lists the environmental factor values of the cluster centroids. The fuzzy membership map for a given environmental cluster contains the membership in that environmental cluster for each pixel across the area.

Association of Environmental Clusters with Soil Types

The conventional soil map is used to relate the generated environmental clusters with mapped soil types. First, each fuzzy membership map of an environmental cluster is converted into a binary map to obtain areas representative of each environmental cluster. For each environmental cluster, pixels with fuzzy membership values >0.5 are assigned to 1 and considered to be the representative area of the environmental cluster, while others are assigned to 0. Second, the derived representative areas of environmental clusters are overlaid with the conventional soil map. Each environmental cluster is related to the soil type that has the greatest area of intersection with the representative area of the environmental cluster. It is very possible that one soil type can be related to multiple environmental clusters. In that case, each of the environmental clusters is perceived to be an instance of the soil type (Zhu, 1999). In the rare case, there are two soil types with similar sizes of intersection area for one environmental cluster. To determine whether the environmental cluster is related to a transitional type between the two soil types or one of the two soil types, its cluster centroid is compared with those of environmental clusters related to the two soil types. If the environmental cluster centroid is numerically in the middle of the other two cluster centroids, this environmental cluster is considered to be a transitional type. Otherwise, if its cluster centroid is numerically closer to one of the two environmental clusters, this environmental cluster is related to the soil type of that environmental cluster.

Extraction and Quantification of Knowledge on Soil–Environment Relationships

Knowledge of the relationships between the soil and the environmental conditions is embedded in associations between environmental clusters and mapped soil types. This knowledge can be extracted and quantified in order to be useful in digital soil mapping (Zhu et al., 2001; Qi and Zhu, 2003). This is achieved through the construction of fuzzy membership functions. These fuzzy membership functions describe how similarity between a local soil and the typical case of a given soil type will change as environmental conditions change (Zhu, 1999). For a given soil type, each environmental variable has a corresponding fuzzy membership function. The reason that we constructed functions of environmental variables and did not use functions of fuzzy memberships generated using FCM to determine soil–environment relationships is that the functions of environmental variables do not rely on the fuzzy membership of environmental clusters and can be used for digital soil mapping and updates in the future or in an area with similar environmental conditions. In addition, there might be more than one fuzzy cluster corresponding to one soil type. Using the membership function from fuzzy classification would present problems for this situation.

The construction of fuzzy membership functions follows the methods described in Zhu et al. (2010). Detailed discussion of the method can be found there, but the general process is described below. Membership functions are approximated using Gaussian curves. Two pieces of information are needed to define a Gaussian curve: the environmental conditions (typical values) at which the membership in the given soil type is 1.0 and the environmental conditions (crossover) at which the membership value in the given soil type is 0.5 (Zhu, 1999; Zhu et al., 2010).

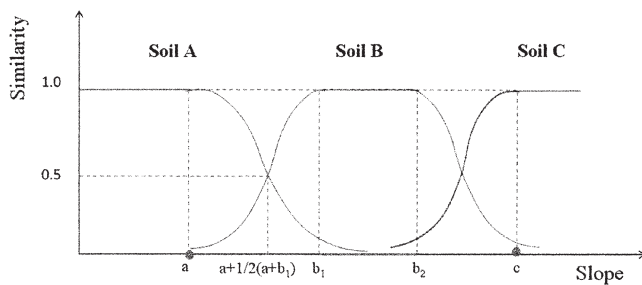


Fig. 1. An illustration of generating knowledge.

The first can be approximated with the environmental conditions at locations where the local soils are typical or representative of a soil type (Zhu et al., 2010). Based on the assumption that the higher the fuzzy membership value of an environmental cluster at a pixel, the more typical is the local soil at that pixel to the associated soil type, the pixels with high membership values in the environmental cluster are considered to be the locations of typical instances of the associated soil type. Operationally, we can use memberships >0.9 as a cut-off value for locating areas that are representative of the corresponding soil type.

The second piece of information is determined by ordering the environmental clusters along each environmental variable, and the midpoint between typical values of two adjacent environmental clusters along this environmental variable is considered to be the crossover point of the membership functions for both environmental clusters (Zhu et al., 2010). As illustrated in Fig. 1, the crossover point in the overlap region of the membership functions for Soil Types A and B along the slope gradient variable is $1/2(a + b_1)$. With this process, the fuzzy membership function with respect to an environmental variable for a soil type can be established (Zhu et al., 2010).

Soil Inference

The last step is to create an updated digital soil map using the extracted relationships (in the form of fuzzy membership functions) and environmental data layers. Our study adopted the soil land inference model (SoLIM) approach (Zhu et al., 2001) for soil inference. The SoLIM approach is a knowledge-based approach for soil mapping. It combines knowledge of the relationships between the soil and the environment with the environmental conditions to infer the spatial distribution of soils (Zhu et al., 1996; Zhu, 1997, 1999; Zhu et al., 2001). With SoLIM, the inference engine scans across all the pixels in the area to compute their similarities to each soil type. For a given pixel, the inference engine looks up its values for the environmental factors and then matches them with the fuzzy membership functions to compute the corresponding similarity values (Zhu and Band, 1994). Once the similarity values for all pixels are computed, a set of fuzzy membership maps in all soil types is generated, and each of them shows the spatial variation of membership in a certain soil type across the landscape. A raster map of soil types could also be created by hardening the fuzzy membership maps. Hardening is achieved by assigning each location the label of the soil map unit bearing the highest membership value at that point (Zhu, 1997). For convenience, we refer to the new soil map updating approach as the FCM–SoLIM approach.

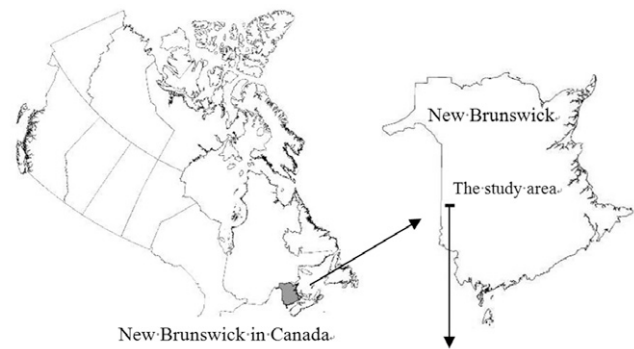


Fig. 2. Location and digital elevation model of the study area.

CASE STUDY

The Study Area

The study area is located in Wakefield, NB, Canada (Fig. 2), with a total area of 39 km^2 . The elevation ranges from about 40 to about 180 m above mean sea level. Annual precipitation averages 925 mm. Annual average growing degree days calculated on a 5°C base is 1649. Parent materials of the area are mainly morainal deposits (Fahmy and Rees, 1996). Landforms of this area consist mainly of hummocks and gentle areas between hummocks due to glacial effects. The topography is predominately rolling terrain. Slopes vary from level to 18.1° . Cultivated land accounts for approximately 55% of the total area and the remaining area is covered with forests. Most of the cultivated lands are in potato (*Solanum tuberosum* L.), grain, hay fields, or pasture land. Forest types mainly include cedar, spruce, fir, larch, maple, birch, and beech species (Fahmy and Rees, 1996).

The Conventional Soil Map

The conventional soil map at a scale of 1:20,000 for the study area was produced by the Land Resource Division Centre for Land and Biological Resources Research (Fahmy and Rees, 1996). During the soil survey and mapping process, preliminary soil boundaries were drawn on color aerial photographs at a scale of 1:12,500 by soil experts and then verified by field checking.

In this study, the drainage class with soil association was the target soil unit to update. The soil association is the equivalent of a soil catena that consists of soils developed on the same parent material but different topographic positions and therefore possess different drainage characteristics (Fahmy and Rees, 1996). Soil associations modified by drainage classes identify different soil series (Fahmy et al., 1986). There were nine soil associations in the study area: Caribou (Ca), Carleton (Cr), Fen (Fe), Interval (In), Riverbank (Ri), Grand Falls (Gf), Thibault (Th), Wakefield (Wk), and Green Road (Gr). Three drainage classes were included: well drained

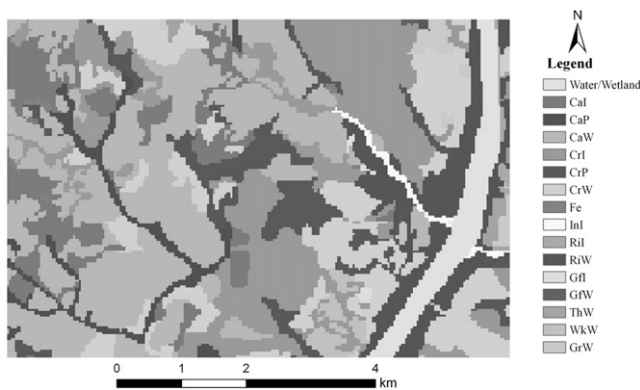


Fig. 3. The 30- by 30-m raster soil map with soil association and drainage class as the soil unit created from the 1:20,000 soil map.

(W), imperfectly drained (I), and poorly drained (P). Some soil associations have all three of the drainage classes and some do not. There are 15 soil units (soil association with drainage class) in total in the study area.

A 30- by 30-m-resolution raster soil map (Fig. 3) with the 15 soil units was created from the original 1:20,000 soil map for comparison with the updated digital soil map. The first two letters in a soil unit name (Fig. 3) are the abbreviation of the soil association and the last letter represents the soil drainage condition. For example, CaI represents an imperfectly drained Caribou soil. One soil association, Fe, is not modified by drainage class because it is an organic soil and always occurs under very poor drainage conditions.

Environmental Data

Soils in this area were influenced deeply by glaciation. Parent material and terrain characteristics that show obvious glacial impacts were chosen to characterize the environmental conditions in this study area. No detailed parent material map is available for the study area. We created an alternative parent material layer from the 1:20,000 soil map (Fig. 3) using information on the surficial deposit modes and the local lithology of soil polygons. Surficial deposits are the result of past and present weathering within a geologic environment (Rampton et al., 1984), and the modes of deposition refer to these surficial materials or regolith.

Table 1. The surficial deposits with lithology information for the Wakefield site.

Surficial deposit mode	Lithology	Parent material unit	Soil map unit	Coverage of total area	
				%	
Compact tills	calcareous siltstones, calcareous sandstones, and calcareous slates	202	CrI	21	
			CrP	7.3	
			CrW	15.1	
			GrW	1.1	
Noncompact tills	acidic, dark reddish brown material, conglomerate, and sandstone argillaceous limestones (minor limestones)	203	WkW	1.4	
			301	CaI	8.5
				CaP	3.5
Glaciofluvial deposit	calcareous siltstones, calcareous sandstones, and calcareous slates grey lithic-feldspathic sandstones (minor quartzose sandstones, polymictic conglomerates, quartz pebble conglomerates, and red mudstones) metaquartzites, slates, metasiltsstones, metasandstones, metaconglomerates, and metawackes	302	ThW	0.5	
			406	RiI	0.4
				RiW	7.6
			410	GfI	0.1
				GfW	0.1
Organic deposit	undifferentiated	514	Fe	0.3	
Alluvial deposit	undifferentiated	614	InI	0.7	

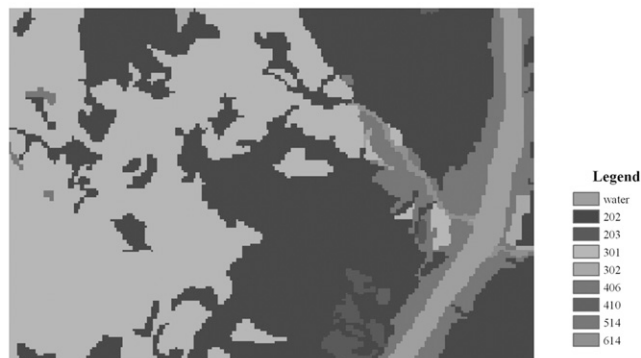


Fig. 4. Parent material map of the study area; the numbers in the legend are the parent material units.

There were a total of five deposit modes in the study area, including compact till, noncompact till, glaciofluvial deposits, organic deposits, and alluvial deposits. Each deposit mode could be divided into homogeneous units, which were called *parent material units* (each was assigned to a unique identification number), based on the lithology types in it. An initial parent material map was then created by aggregating soil polygons with the same deposit mode and lithology type. We verified and further refined the initial parent material map using field observations of parent material information provided by a geologist from the New Brunswick Department of Natural Resources (Seaman, 2000). The final eight parent material units are listed in Table 1 and the map is shown in Fig. 4. The soil units and their percentage of the total area in each parent material unit are also included in Table 1.

Four topographic variables (slope gradient, planform curvature, profile curvature, and topographic wetness index) were used in this study to characterize the terrain. Information on these topographic variables was derived from a 30- by 30-m-resolution digital elevation model (DEM), which was created from elevation points published by Service New Brunswick (www.snb.ca/gdam-igec/e/2900e_1c.asp; verified 18 Jan. 2011) using the TINLATTICE tool in ArcInfo (Esri, Redlands, CA) (Fig. 2). Layers of slope, planform curvature, and profile curvature were then derived from the DEM with the 3DMapper software (Burt and Zhu, 2004) (Fig. 5).

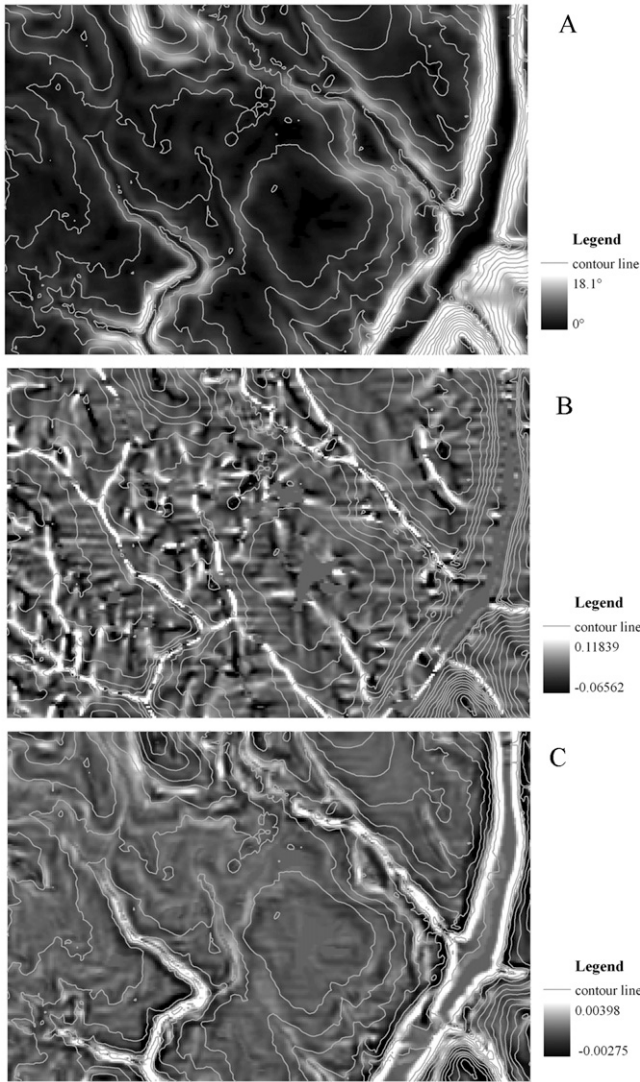


Fig. 5. The three terrain parameters of the Wakefield site: (a) slope; (b) planform curvature; and (c) profile curvature.

The topographic wetness index (TWI), which combines the local upslope contributing area with slope characteristics, is commonly used to quantify topographic control on hydrologic processes (Sørensen et al., 2006). It helps to indicate the spatial distribution of soil moisture and surface saturation (Rodhe and Seibert, 1999; Schmidt and Persson, 2003; Zinko et al., 2005). It is therefore correlated with soil drainage conditions, which are essential in separating the soil units in our study. We calculated TWI using (Beven and Kirkby, 1979)

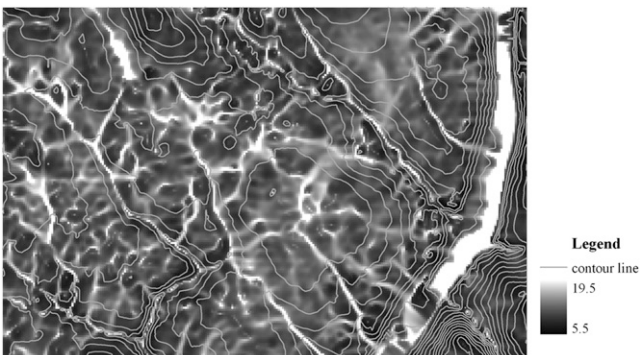


Fig. 6. Topographic wetness index of the Wakefield site.

Table 2. Cluster centroids of four environmental classes in Parent Material Unit 202.

Class	Slope	Planform curvature	Profile curvature	Topographic wetness index
1	1.94	-0.00591	-0.00007	8.91
2	11.74	-0.00244	-0.00044	7.28
3	1.60	0.01298	0.00007	10.88
4	5.27	-0.00171	-0.00020	8.34

$$TWI = \ln \left(\frac{a}{\tan \beta} \right) \quad [7]$$

where a is the cumulative upslope area draining through a point (per unit contour length) and β is the slope gradient at that point. Because of the gentle terrain in the study area, a multiple-flow-direction strategy MFD-md (Qin et al., 2007) was used to calculate the upslope drainage area. The MFD-md method performed better than the single-flow-direction strategy (D8) or the multiple-flow-direction strategy constructed by Quinn et al. (1991) in modeling the effect of local terrain on flow partitioning (Qin et al., 2006, 2007). The TWI map is shown in Fig. 6.

RESULTS AND EVALUATION

Results

The study area was first stratified using the parent material layer (Fig. 4). The FCM clustering of the four topographic variables was performed in each of these stratified areas. The FCM clustering was not conducted for parent material units with only one soil type mapped (Units 203, 302, 514, and 614 in our case study). We used Parent Material Unit 202 as an example to illustrate the process of obtaining environmental clusters and extract knowledge about soil–environment relationships.

By examining the performance of the normalized entropy and partition coefficient across adjacent clustering as described above, we determined three possible optimal numbers of clusters in this unit: four, six, or eight. Four was eventually selected as the optimal number of clusters because four soil types occur in this unit on the conventional soil map. The centroids for four environmental clusters are listed in Table 2.

Fuzzy membership maps of the environmental clusters were produced next. Figure 7 illustrates one such fuzzy membership map. It shows the spatial variation of membership for Environmental

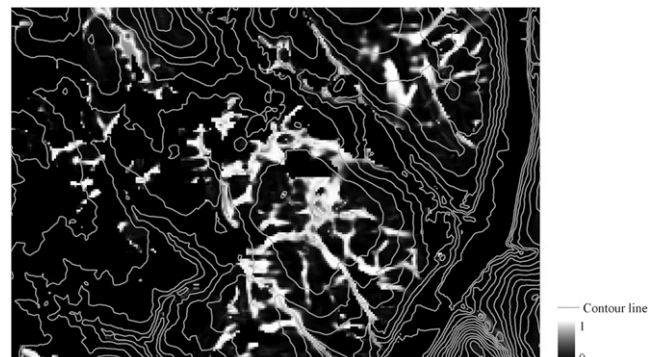


Fig. 7. Fuzzy membership of Cluster 3 in Parent Material Unit 202.

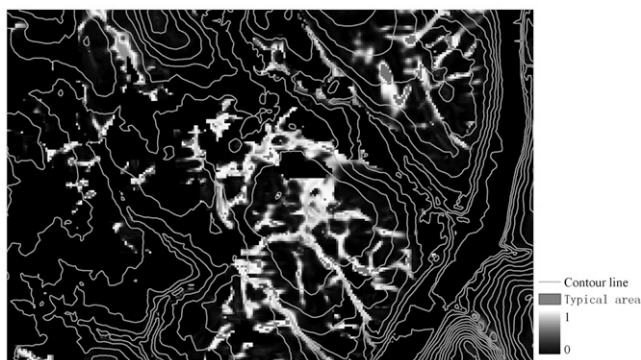


Fig. 8. Typical area (red) of Cluster 3 in Parent Material Unit 202.

Cluster 3 in Parent Material Unit 202. The whiter the color in this figure, the higher the membership in this environmental cluster.

Associations between the environmental clusters and soil types were next established by measuring the overlapping areas of both units. Clusters 1, 2, and 3 were uniquely associated with imperfectly drained Carleton, well-drained Carleton, and poorly drained Carleton soils, respectively, on the conventional soil map. Cluster 4, however, was found to be associated with two closely related soils: the well-drained Carleton and the imperfectly drained Carleton. To determine to which soil type Environmental Cluster 4 should be related, the TWI was considered as the decisive variable in this study because it could be used to distinguish different soil drainage conditions, which was essential in determining the soil mapping units in this area. We compared the TWI value (Table 2) of its centroid with two other clusters that were identified as well-drained Carleton and imperfectly drained Carleton to determine the soil type for Cluster 4 or to decide whether it is just a transition type. The TWI value for the centroid of Cluster 1 (associated with imperfectly drained Carleton) was 8.91. The TWI for the centroid of Cluster 2 (associated with well-drained Carleton) was 7.28. The TWI value for the centroid of Cluster 4 was 8.34, which is in between those of Clusters 2 (well-drained Carleton) and 1 (imperfectly drained Carleton). We thus considered Cluster 4 as a transition between the imperfectly drained and well-drained Carleton soils. The transitional classes would not be used in the knowledge extraction stage.

Figure 8 shows the area of the typical Cluster 3 (pixels with membership values >0.99) in red. The typical environmental conditions of the representative pixels of Cluster 3 were extracted to construct the fuzzy membership functions. For each soil type, four fuzzy membership curves were developed, each corresponding to one of the four environmental factors. Figure 9 shows the fuzzy membership function of soil CrP for TWI as an example. Because we used the representative pixels of the environmental clusters to generate the typical environmental conditions for the soil types, the environmental values when similarities are equal to 1.0 are not identical to the centroids of the corresponding clusters.

A fourth soil type, well-drained Greenroad, occurs in this parent material unit on the original map. It covers, however, only 1.1% of the total study area. Because of its limited coverage, FCM could not generate an environmental cluster for it. In our study, we combined this soil type with the well-drained Carleton,

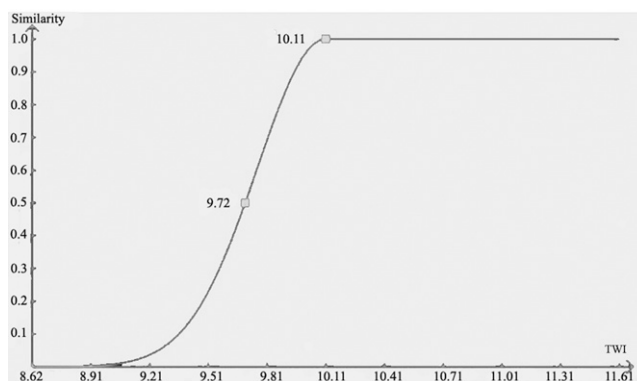


Fig. 9. The fuzzy membership curve of the CrP soil map unit for topographic wetness index (TWI).

considering their similar pedological properties. Environmental clusters generated using FCM and fuzzy membership functions describing the relationship between the soil and the environmental variables for soil types in other parent material units were obtained in the same way as illustrated above.

Using the FCM–SoLIM approach, the above fuzzy membership functions were combined with the data on the four environmental variables to derive the fuzzy membership maps of the CrI, CrP, CrW, CaI, CaP, CaW, RiI, RiW, GfI, and GfW soils. These fuzzy membership maps of soil types were used to produce a map showing the spatial distribution of soil types through hardening (Zhu, 1997). For parent material units that only contained one soil type, the polygons of soil types in those units on the conventional soil map were considered as the spatial extent of these soil types. Therefore, soil polygon layers for WkW, ThW, Fe, and InI were obtained directly from the conventional soil map. The final updated digital soil map for the study area is shown in Fig. 10.

Evaluation of the Soil Maps

The updated map (Fig. 10) shows much more spatial detail than the original soil map (Fig. 3). Field validation points were collected to evaluate the accuracy of the updated map and the conventional soil map. The validation points were collected using a stratified sampling strategy and were intended to cover all the soil types in the area. Because information on detailed soil distributions is not available otherwise, the updated soil map from the FCM–

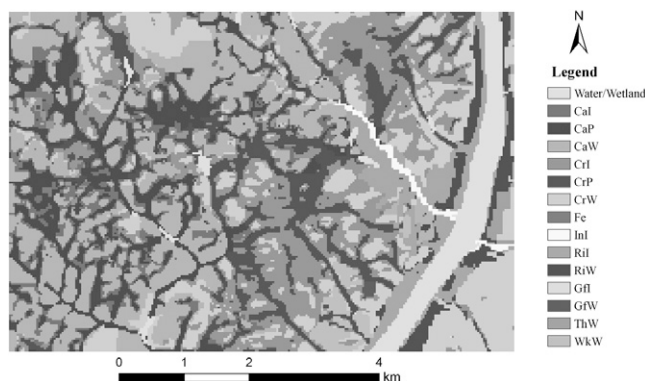


Fig. 10. The updated digital soil map using the fuzzy c-means–Soil Land Inference Model (FCM–SoLIM) approach.

SoLIM approach was used to stratify the study area. To guarantee that all strata (soil types) on the updated map were covered and to ensure that the number of validation points was proportional to the extent of each soil type, we used the following sampling design strategy: at least one point was located in each stratum and the number of samples per stratum was proportional to the area covered of the stratum. In other words, soil types with large areas received more points than soil types with small areas. The locations of samples were initially determined randomly in each stratum. Due to the limitation of accessibility and resources available for this project, points that were located in inaccessible areas were replaced by points that were accessible based on the judgment of local soil experts. This means that the validation sample cannot be analyzed as a probability sample, i.e., design-based estimation of the overall accuracy is impossible (Brus and de Gruijter, 1997). We estimated the overall accuracy by model-based inference, assuming that the classification errors were independent, so that the average of the classification-error indicator was an unbiased estimate of the overall accuracy. A total of 37 points were selected for the purpose of evaluation (Fig. 11). Soil association and soil drainage conditions were identified by local soil experts through the examination of the soil profile at each site in the field. Thirteen soil types were identified at these points; 11 of them were shown on the conventional soil map and the other two were not.

The field-observed soil types were compared with the soil types obtained from both the conventional soil map (Fig. 3) and the updated soil map (Fig. 10). The accuracy of the conventional soil map was 45.9%, while the accuracy of the updated soil map was 64.9%. The 19% difference in absolute accuracy represents a relative improvement of about 45%. This is obviously based on a relatively small sample (37 field observations), and it is important to know if the difference could have occurred by chance. We conducted a hypothesis test to assess the statistical significance of the 19% improvement.

Under the assumption of independent classification errors, the number of correctly classified soil points satisfies a binomial distribution. We then tested the hypothesis that the overall accuracy of the updated map is not larger than that of the conventional soil map. Under this hypothesis, the number of correctly classified soil points, X , is binomially distributed with parameters $n = 37$ (the number of validation points) and $p = 17/37$ (the proportion of successes for the conventional map), which can be

expressed as $X \sim B(37, 0.459)$. The probability, p , with k correctly classified soil points is calculated using the following formula:

$$p(X=k) = B(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k} \quad [8]$$

The probability of having 24 or more correctly classified soil points was calculated as $p(X \geq 24) = B(24; 37, 0.459) + B(25; 37, 0.459) + B(26; 37, 0.459) + \dots + B(37; 37, 0.459) = 0.01595$, which is < 0.05 . Thus, the hypothesis can be rejected, which indicates that the updated map is more accurate than the conventional soil map at the significant level of 0.05. Therefore, we can conclude that the accuracy of the conventional soil map was improved through the updating process from the results of the above statistical method.

It should also be noted that there are three validation points for which the inferred soil types are Fe, ThW, and WkW, respectively. The spatial distributions of the three soil types in the updated soil maps were directly obtained from the conventional soil maps. When these points are excluded from the validation data set, the accuracy of the updated soil map using the FCM-SoLIM-based approach is 67.6%.

DISCUSSIONS

The proposed method in this study used only the conventional soil map as a source of knowledge of soil-environment relationships but generated a more detailed and accurate updated soil map (Fig. 10 vs. Fig. 3). There are at least three reasons for this. First, we used more accurate and detailed environmental data, including a refined parent material map and terrain data layers derived from a 30-m-resolution DEM. The TWI variable played an important role in the inference. Spatial variation of the TWI indicated soil drainage conditions based on landform positions and thus improved accuracy of the soil mapping units involving drainage classes. Second, we used a FCM clustering method to generate environmental clusters, which were related to soil types. This process is fundamentally similar to the process of how local soil experts identify soil-landscape units through photo interpretations but more consistent throughout the area and thus less error prone. Third, we used the SoLIM framework to construct quantified knowledge of relationships between a soil and its environmental conditions and infer soil distributions automatically using that knowledge. It has been shown that the SoLIM framework can help obtain consistent knowledge of soil-landscape models and produce more accurate soil maps than the conventional soil survey and manual mapping process by soil experts (Zhu et al., 2001).

In this study, the way we used the conventional soil map was to interpret environmental clusters generated by fuzzy clustering on environmental factors. The specific method was to overlay the conventional soil map and the spatial distribution maps of environmental clusters and compare the intersection areas to determine to which soil types the environmental clusters are related. After the associations between soil types and environmental clusters are determined, the process of knowledge extraction is independent of the conventional soil map. It is true that there are errors (inclusion and misplacement of boundaries) in the conventional soil map.

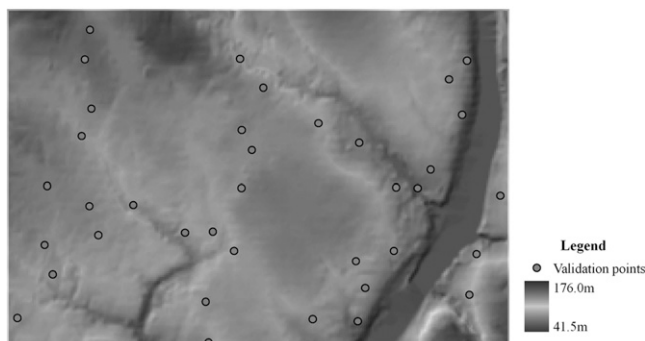


Fig. 11. Validation points at the Wakefield site.

With the assumption that errors in the conventional soil map are not so large as to overwhelm implicit relationships between soil types and environmental clusters, however, the errors will have little influence on the updated soil maps. The 19% increase of mapping accuracy is considered a good improvement.

In our case study, the soil type (well-drained Greenroad) in Parent Material Unit 202 was combined with other soil types owing to its limited coverage. This was the only soil from the original map that didn't occur in the resultant soil map. If the update is required to have the complete set of soil types, an alternative solution is to use its soil polygon in the conventional soil map as its spatial distribution in the updated soil map. This amounts to assuming that the conventional map perfectly represents that soil type.

The proposed method is appropriate in situations where the study area contains no or few point observations. When soil profile descriptions at georeferenced points are available in soil information systems for some areas, these point data could also be useful for updating conventional soil maps. Through overlaying soil profile points with fuzzy membership maps of the generated environmental clusters, environmental clusters could be related to the soil types of points falling into a high-fuzzy-membership area. These points could be considered as the typical instances of the associated soil types and used to extract knowledge of the soil-environment relationships. For environmental clusters for which no points fall into a high-fuzzy-membership area, conventional soil maps can serve as a source to establish their relationship with soil types.

CONCLUSIONS

The study proposed a method to update conventional soil maps using digital soil mapping techniques. Environmental clusters generated by fuzzy clustering were associated with soil types on conventional soil maps. The associations contained knowledge about the relationships between soils and their environmental conditions, which were used to infer soil spatial distributions under the SoLIM framework. The result in our case study in Wakefield, Canada, showed that the updated soil map was more detailed and accurate than the 1:20,000 conventional soil map. The spatial detail seen in the distribution of the predicted soil types in the updated soil maps is commensurate with the resolution of the environmental data layers that are available. This is important for modern environmental modeling and site-specific management activities. Therefore, we conclude that this method can serve as an effective model to update soil maps in areas where legacy soil maps are available.

ACKNOWLEDGMENTS

This research was supported by the National Natural Science Foundation of China (Projects no. 40971236 and 41001298), the 2008 MOE-AAFC Ph.D. Research Program of the China Scholarship Council and Agriculture & Agri-Food Canada, the Vilas Associate Program, the Hamel Faculty Fellowship at the University of Wisconsin-Madison, and the Open Fund of State Key Laboratory of Remote Sensing Science, Institute of Remote Sensing Applications, Chinese Academy of Sciences. We also thank the China Scholarship Council, the Ministry of Education of China, and the Potato Research Centre of Agriculture and Agri-Food Canada. We also like to thank Dr. Lien Chow, Dr. Xianzhou Nie, Dr. Bernie Zebarth, Dr.

Zhisheng Xing, Dr. Yang Yu, and Dr. Xiuqing Li of the Potato Research Centre, Agriculture Agri-Food Canada, and Dr. Fanrui Meng of the University of New Brunswick for their support.

REFERENCES

- Ahn, C.W., M.F. Baumgardner, and L.L. Biehl. 1999. Delineation of soil variability using geostatistics and fuzzy clustering analyses of hyperspectral data. *Soil Sci. Soc. Am. J.* 63:142-150.
- Beven, K.J., and N.J. Kirkby. 1979. A physically based variable contributing area model of basin hydrology. *Hydrol. Sci. Bull.* 24:43-69.
- Bezdek, J.C., R. Ehrlich, and W. Full. 1984. FCM: The fuzzy *c*-means clustering algorithm. *Comput. Geosci.* 10:191-203.
- Brus, D.J., and J.J. de Gruijter. 1997. Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). *Geoderma* 80:1-59.
- Brus, D.J., J.J. de Gruijter, and A. Breeuwmsma. 1992. Strategies for updating soil survey information: A case study to estimate phosphate sorption characteristics. *J. Soil Sci.* 43:567-581.
- Brus, D.J., P. Bogaert, and G.B.M. Heuvelink. 2008. Bayesian maximum entropy prediction of soil categories using a traditional soil map as soft information. *Eur. J. Soil Sci.* 59:166-177.
- Burrough, P.A., P.F.M. van Gaansa, and R.A. MacMillan. 2000. High-resolution landform classification using fuzzy *k*-means. *Fuzzy Sets Syst.* 113:37-52.
- Burt, J.E., and A.X. Zhu. 2004. 3DMapper 4.02. Terrain Analytics, Madison, WI.
- de Bruin, S., and A. Stein. 1998. Soil-landscape modelling using fuzzy *c*-means clustering of attribute data derived from a digital elevation model (DEM). *Geoderma* 83:17-33.
- English, E.M. 2001. Assisting knowledge-based inferential soil mapping: The application of fuzzy *c*-means clustering to expose environmental niches. M.S. thesis. Univ. of Wisconsin, Madison.
- Fahmy, S.H., and H.W. Rees. 1996. Soils of the Woodstock-Florenceville area, Carleton County, New Brunswick. Vol. 3. New Brunswick Soil Surv. Rep. 14. Research Branch, Agriculture and Agri-Food Canada, Ottawa, ON.
- Fahmy, S.H., H.W. Rees, and J.K. MacMillan. 1986. Soils of New Brunswick, a first approximation. New Brunswick Dep. of Agriculture and Aquaculture, Fredericton, NB, Canada.
- Hanesch, M., R. Scholger, and M.J. Dekkers. 2001. The application of fuzzy *c*-means cluster analysis and non-linear mapping to a soil data set for the detection of polluted sites. *Phys. Chem. Earth A* 26:885-891.
- Hudson, B.D. 1992. The soil survey as a paradigm-based science. *Soil Sci. Soc. Am. J.* 56:836-841.
- Kempen, B., D.J. Brus, G.B.M. Heuvelink, and J.J. Stoorvogel. 2009. Updating the 1:50,000 Dutch soil map using legacy soil data: A multinomial logistic regression approach. *Geoderma* 151:311-326.
- Lark, R.M. 1999. Soil-landform relationships at within-field scales: An investigation using continuous classification. *Geoderma* 92:141-165.
- McBratney, A.B., M.L. Mendonca Santos, and B. Minasny. 2003. On digital soil mapping. *Geoderma* 117:3-52.
- McSweeney, K., B.K. Slater, R.D. Hammer, J.C. Bell, P.E. Gessler, and G.W. Petersen. 1994. Towards a new framework for modeling the soil-landscape continuum. p. 127-145. *In* R. Amundson et al. (ed.) Factors of soil formation: A fiftieth anniversary retrospective. *SSSA Spec. Publ.* 33. SSSA, Madison, WI.
- Miao C.Y., J.R. Ni, A.G. L. Borthwick. 2010. Recent changes of water discharge and sediment load in the Yellow River basin, China. *Progress Phys. Geography* 34:541-561.
- Moran, C.J., and E.N. Bui. 2002. Spatial data mining for enhanced soil map modeling. *Int. J. Geogr. Inf. Sci.* 16:533-549.
- Qi, F., and A.X. Zhu. 2003. Knowledge discovery from soil maps using inductive learning. *Int. J. Geogr. Inf. Sci.* 17:771-795.
- Qi, F., A.X. Zhu, T. Pei, C.Z. Qin, and J.E. Burt. 2008. Knowledge discovery from area-class resource maps: Capturing prototype effects. *Cartogr. Geogr. Inf. Syst.* 35:223-237.
- Qin, C.Z., A.X. Zhu, B.L. Li, T. Pei, and C.H. Zhou. 2006. Review of multiple flow direction algorithms based on gridded digital elevation models. (In Chinese with English abstract.) *Earth Sci. Front.* 13:90-98.
- Qin, C.Z., A.X. Zhu, T. Pei, B.L. Li, C.H. Zhou, and L. Yang. 2007. An adaptive approach to selecting a flow-partition exponent for a multiple-flow-direction algorithm. *Int. J. Geogr. Inf. Sci.* 21:443-458.
- Quinn, P., K. Beven, P. Chevalier, and O. Planchon. 1991. The prediction of hillslope flow paths for distributed hydrological modeling using digital

- terrain models. *Hydrol. Processes* 5:59–79.
- Rampton, V.N., R.C. Gauthier, J. Thibault, and A.A. Seaman. 1984. Quaternary geology of New Brunswick. Memoir 416. Geol. Surv. of Canada, Ottawa, ON, Canada.
- Rodhe, A., and J. Seibert. 1999. Wetland occurrence in relation to topography: A test of topographic indices as moisture indicators. *Agric. For. Meteorol.* 98–99:325–340.
- Rossiter, D.G. 2008. Digital soil mapping as a component of data renewal for areas with sparse soil data infrastructures. p. 3–14. *In* A. Hartemink et al. (ed.) *Digital soil mapping with limited data*. Springer, Dordrecht, the Netherlands.
- Schmidt, F., and A. Persson. 2003. Comparison of DEM data capture and topographic wetness indices. *Precis. Agric.* 4:179–192.
- Seaman, A.A. 2000. Till geochemistry of the New Brunswick part of the Woodstock (NTS 21 J/04) map area, Carleton County: Site and geochemical data. Open File 2000–3. New Brunswick Dep. of Nat. Resour. and Energy, Minerals and Energy Div., Fredericton, NB, Canada.
- Sørensen, R., U. Zinko, and J. Seibert. 2006. On the calculation of the topographic wetness index: Evaluation of different methods based on field observations. *Hydrol. Earth Syst. Sci.* 10:101–112.
- Vitharana, U.W.A., T. Saey, L. Cockx, D. Simpson, H. Vermeersch, and M. Van Meirvenne. 2008. Upgrading a 1/20,000 soil map with an apparent electrical conductivity survey. *Geoderma* 148:107–112.
- Yang, L., A.X. Zhu, B.L. Li, C.Z. Qin, T. Pei, B.Y. Liu, R.K. Li, and Q.G. Cai. 2007. Extraction of knowledge about soil–environment relationship for soil mapping using fuzzy *c*-means (FCM) clustering. *Acta Pedol. Sin.* 44:16–23.
- Zhu, A-X. 1997. A similarity model for representing soil spatial information. *Geoderma* 77:217–242.
- Zhu, A-X. 1999. A personal construct-based knowledge acquisition process for natural resource mapping. *Int. J. Geogr. Inf. Sci.* 13:119–141.
- Zhu, A-X., and L.E. Band. 1994. A knowledge-based approach to data integration for soil mapping. *Can. J. Rem. Sens.* 20:408–418.
- Zhu, A-X., L.E. Band, B. Dutton, and T. Nimlos. 1996. Automated soil inference under fuzzy logic. *Ecol. Modell.* 90:123–145.
- Zhu, A-X., L.E. Band, R. Vertessy, and B. Dutton. 1997. Derivation of soil properties using a Soil Land Inference Model (SoLIM). *Soil Sci. Soc. Am. J.* 61:523–533.
- Zhu, A-X., B. Hudson, J.E. Burt, K. Lubich, and D. Simonson. 2001. Soil mapping using GIS, expert knowledge, and fuzzy logic. *Soil Sci. Soc. Am. J.* 65:1463–1472.
- Zhu, A-X., L. Yang, B.L. Li, C.Z. Qin, T. Pei, and B.Y. Liu. 2010. Construction of membership functions for predictive soil mapping under fuzzy logic. *Geoderma* 155:166–174.
- Zinko, U., J. Seibert, M. Dynesius, and C. Nilsson. 2005. Plant species numbers predicted by a topography based groundwater-flow index. *Ecosystems* 8:430–441.