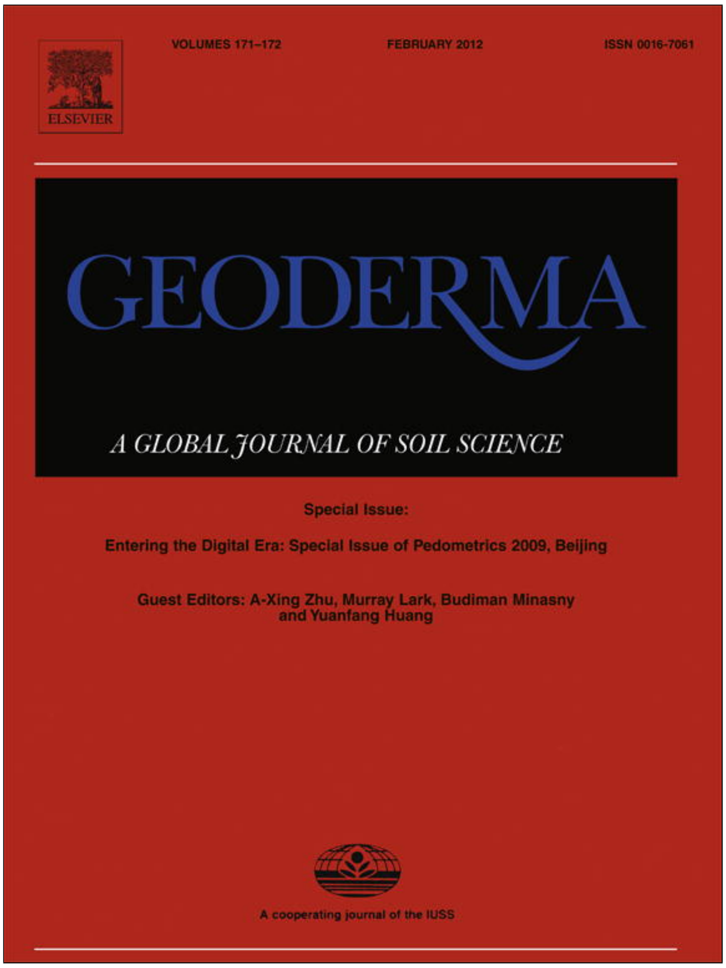


VOLUMES 171–172 FEBRUARY 2012 ISSN 0016-7061

ELSEVIER

GEODERMA

A GLOBAL JOURNAL OF SOIL SCIENCE

Special Issue:

Entering the Digital Era: Special Issue of Pedometrics 2009, Beijing

Guest Editors: A-Xing Zhu, Murray Lark, Budiman Minasny
and Yuanfang Huang

A cooperating journal of the IUSS

Geoderma 171-172 (2012) 24–34

Contents lists available at ScienceDirect

Geoderma

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

Sensitivity of digital soil maps based on FCM to the fuzzy exponent and the number of clusters

Xiao-Lin Sun [a,b,f], Yu-Guo Zhao [a], Hui-Li Wang [c], Lin Yang [d], Cheng-Zhi Qin [d], A-Xing Zhu [d,e,*], Gan-Lin Zhang [a,**], Tao Pei [d], Bao-Lin Li [d]

[a] State Key Laboratory of Soil and Sustainable Agriculture, Institute of Soil Science, Chinese Academy of Sciences, Nanjing 210008, China
[b] Joint Open Laboratory of Soil and Environment, Institute of Soil Science, Chinese Academy of Sciences and Hong Kong Baptist University, Hong Kong, China
[c] Guangxi Academy of Forestry Sciences, Nanning 530002, China
[d] State Key Laboratory of Environment and Resources Information System, Institute of Geographic Sciences and Resources Research, Chinese Academy of Sciences, Beijing 100101, China
[e] Department of Geography, University of Wisconsin-Madison, Madison, WI 53706, USA
[f] College of Agriculture, Guangxi University, Nanning 530004, China

ARTICLE INFO

Article history:
Received 30 April 2010
Received in revised form 3 March 2011
Accepted 24 March 2011
Available online 19 April 2011

Keywords:
Fuzzy c-means clustering
Fuzzy validity function
Digital soil mapping
Sensitivity

ABSTRACT

Fuzzy c-means clustering (FCM) has been used frequently in digital soil mapping. One of the key issues in applying FCM is the determination of the appropriate classification parameters of the fuzzy exponent (m) and the number of clusters (c). To determine the optimal selection of appropriate m and c values, in this study, we first used two simulated datasets to demonstrate the sensitivity of three commonly used validity functions to m and c. These two simulated datasets contained overlapping clusters and hierarchical clusters, respectively. The three studied validity functions were fuzzy performance index (*FPI*), compactness and separation (S) and a derivative of the objective function with respect to the fuzzy exponent ($-[(\delta J_E/\delta m)c^{0.5}]$). Then, a case study mapping soil organic matter (SOM) based on memberships from FCM clustering terrain attributes was conducted to investigate the sensitivity of soil maps to m and c. The results of the study on the simulated datasets showed that the three validity functions were sensitive in differing degrees to the structures of the clustered datasets under a wide range of m, but the sensitivities and the range of m were different for different validity functions and depended on the clustered datasets. The results from the case study of the soil mapping showed that soil maps based on FCM clustering were sensitive to m and c, but only the spatial variations of SOM presented on the maps were significantly sensitive to c. Furthermore, mapping accuracy was slightly sensitive to m and c. It is concluded that there was a range of optimal m over which digital soil maps did not change very much, but this was not certain for c, given that the spatial variation presented on the maps changed significantly with c.

1. Introduction

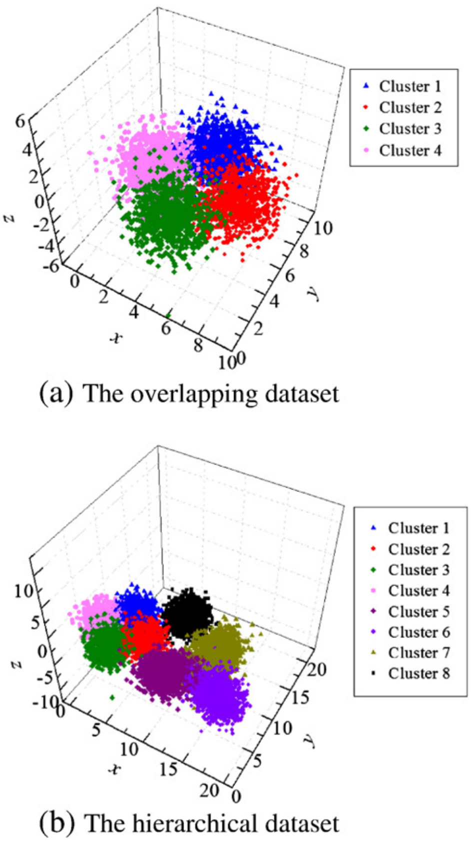
Fuzzy c-means (FCM) clustering is frequently used to classify soil observations or environmental attributes for digital soil mapping (Lark, 1999; McBratney and Odeh, 1997; Odeh et al., 1992; Zhu et al., 2010). The advantage of using FCM in soil mapping is that FCM can capture the fuzzy nature of the soil and the landscape (Lark, 1999).

FCM is a type of unsupervised continuous classification that is very easy to implement. Generally speaking, a user must only set several parameters to run an FCM. Two key parameters are the fuzzy exponent (m) and the number of clusters (c) (Odeh et al., 1992). The fuzzy exponent is set to define fuzziness between clustered clusters, and the number of clusters, as the name suggests, is set to control how many clusters will be used. An appropriate fuzzy exponent and number of clusters are very important for FCM clustering because the two parameters determine the final clustering outputs, i.e., the centers of clusters and the memberships of every object belonging to the clusters (Lagacherie et al., 1997; Pal and Bezdek, 1995). Therefore, almost every research using FCM must focus on the optimal selection of these two parameters. At present, a number of validity functions have been proposed for optimally selecting these parameters. For example, in soil mapping, commonly used validation functions are partition coefficient, partition entropy, fuzzy performance index (*FPI*), modified partition entropy (*MPE*) and the derivative of objective function with respect to the fuzziness exponent (i.e., $-[(\delta J_E/\delta m)c^{0.5}]$) (Amini et al., 2005; Lagacherie et al., 1997; Lark, 1999; Odeh et al., 1992; Tan et al., 2006; Zhu et al., 2010).

However, different validity functions often suggest different optimal combinations of these two parameters (Fecher and Schmidt, 2003; Pal and Bezdek, 1995; Srinivas et al., 2008; Xie and Beni, 1991). Sometimes, there are no clear answers for their optimal combination (Fecher and Schmidt, 2003; Lagacherie et al., 1997; Triantafilis et al., 2009).

* Correspondence to: Tel.: +1 608 262 0272; fax: +1 608 265 3991.
** Correspondence to: Tel.: +86 25 86881279; fax: +86 25 86881000.
E-mail addresses: azhu@wisc.edu (A.-X. Zhu), glzhang@issas.ac.cn (G.-L. Zhang).

0016-7061/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.geoderma.2011.03.016



(a) The overlapping dataset

(b) The hierarchical dataset

Fig. 1. Simulated datasets.

Interestingly, de Bruin and Stein (1998) suggested that external criteria should be used instead of internal validity functions to select the two parameters. However, the optimal selection of the values of these two parameters remains difficult. In fact, more studies on this issue have been called for by Fecher and Schmidt (2003), McBratney and de Gruijter (1992), Pal and Bezdek (1995) and Xie and Beni (1991). In addition, FCM is being widely used in digital soil mapping. The sensitivity of digital soil maps based on FCM to these two parameters must be addressed.

In this research, we are specifically interested in knowing whether there is a range of m and c values over which the produced digital soil maps do not vary greatly. First, this study demonstrated the sensitivity of three commonly used validity functions, i.e., *FPI*, compactness and separation (S) and the function $-[(\delta J_E/\delta m)c^{0.5}]$, to the parameters using two simulated datasets. Then, the sensitivity of the digital soil maps based on FCM to the two parameters were investigated using a case study of soil mapping.

2. Materials and methods

2.1. Simulated datasets

Many dataset configurations have been presented in FCM researches, for example, the four notional configurations of soil individual datasets in McBratney and de Gruijter (1992) and the Normal-4 and Iris datasets in Pal and Bezdek (1995). In these dataset configurations, a dataset can have overlapping clusters, hierarchical clusters, or equally distributed data. However, an equally distributed dataset can be treated as a special case of the overlapping dataset (McBratney and de Gruijter, 1992). Thus, this study simulated two datasets of overlapping clusters and hierarchical clusters, respectively.

The dataset of overlapping clusters contains four clusters, i.e., Clusters 1, 2, 3 and 4 in Fig. 1. Each cluster consisted of 1000 objects that were characterized by three attributes, x, y and z. Attribute values of the objects in each cluster were randomly generated using normal distributions defined by the parameters of mean (μ) and variance (σ^2), which are listed in Table 1. These parameters were chosen to shape the clusters into a sphere and position them as overlapping one-by-one (Fig. 1(a)).

The dataset of hierarchical clusters contains eight clusters, including the above four overlapping clusters, i.e., Clusters 1–4. The objects of the other four clusters, i.e., Clusters 5–8 in Fig. 1(b), were also characterized by x, y and z. The values of these attributes were also randomly generated using normal distributions with mean and variance parameters that are defined in Table 1. Similarly, these parameters were chosen to shape the clusters and position them in a hierarchical structure. Therefore, Clusters 5, 6 and 7 were ellipsoidal and Cluster 8 was spherical (Fig. 1(b)). Clusters 5, 6 and 7 contained 1500 objects each, and Cluster 8 contained 2000 objects. Different numbers of objects and different shapes of these clusters were arbitrarily selected because clusters in a hierarchical dataset can vary greatly in the number of objects and shapes; an example of this is the notional hierarchical distributed clusters in McBratney and de Gruijter (1992). After the creation of the clusters, Clusters 5 and 6 were revolved 60° counterclockwise and clockwise, respectively, around their designated centers to make Clusters 5, 6 and 7 systematic and to form an isolated combined cluster relative to the other clusters (Fig. 1(b)). Likewise, Clusters 1–4 were separated from the others and formed another combined cluster. Cluster 8 was isolated from the others. Therefore, we can argue that both $c = 8$ and $c = 3$ are favored for this hierarchical dataset.

2.2. Soil mapping case study

The Heshan area has been described in Zhu et al. (2010). It is located close to the middle-west boundary of Heilongjiang Province in Northeast China (Fig. 2). The climate is cool temperate and monsoon. The landforms are colluvial plateaus, alluvial plains and lacustrine plains. The elevation is between 100 and 366 m, with an average slope of 1.9° (Fig. 2). The parent material for soil formation is mainly silt loam loess except for the fluvial deposits in valley bottoms. The land cover of this area was originally composed of sparsely forested meadows, shrub meadow grasslands and forb meadows. In the last half century, the area has been intensively farmed, and soybean and wheat are the main agricultural products. The soils are black soil and meadow soil, which are referred to as Mollisols and Inceptisols, respectively, in U.S. Soil Taxonomy.

The soil sampling included regular sampling, purposive sampling, subjective sampling and transect sampling (Zhu et al., 2010). Forty-four samples were collected by regular sampling at an interval of

Table 1
Parameters for simulating the clusters.

	Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5		Cluster 6		Cluster 7		Cluster 8	
	μ	σ^2	μ	σ^2	μ	σ^2	μ	σ^2	μ	σ^2	μ	σ^2	μ	σ^2	μ	σ^2
x	5	1	7.5	1	5	1	2.5	1	12.5	1.5	17.5	1.5	15	1	10	1
y	7.5	1	5	1	2.5	1	5	1	2.5	1	2.5	1	6.8	1.5	10	1
z	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1

μ: mean. σ^2: variance.

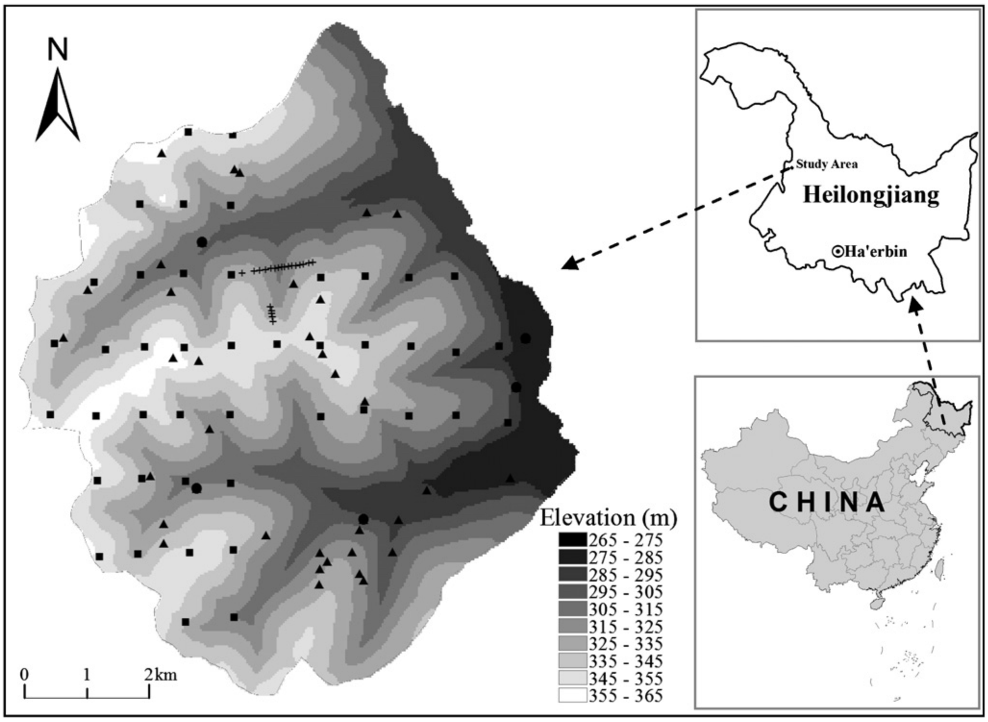


Fig. 2. Heshan study area and sampling sites based on the local DEM. Squares are regular sampling sites, triangles are purposive sampling sites, circles are subjective sampling sites and pluses are transect sampling sites.

1100 m in the north and 740 m in the east (Fig. 2). Thirty-five samples were collected from typical landforms that were identified using FCM clustering terrain attributes, i.e., purposive sampling. Five samples were collected subjectively by soil surveyors to cover unique landforms that were not included in previous samplings (Fig. 2). Transect sampling was used in two transects, with one comprising of 16 sampling sites along a distance of 1200 m across a valley and the other comprising five sampling sites along a distance of 500 m along a slope (Fig. 2). The soil organic matter (SOM) content of Soil Layer A in all of these sampling sites was measured using the potassium dichromate colorimetric method. Considering that the 84 samples collected using regular sampling, purposive sampling and subjective sampling systematically covered the soil formation environment, they were used for mapping. The 21 samples collected using transect sampling were used to validate the mapping.

In addition to the four terrain attributes used in Zhu et al. (2010), i.e., slope gradient, contour curvature, profile curvature and topographic wetness index, slope position was also used in this study because Qin et al. (2009) suggested that the inclusion of this attribute is beneficial to digital soil mapping. All the terrain attributes were derived from the local 10 m digital elevation model (DEM) which was constructed based on the 1:10,000 topography map. Details about calculating these terrain attributes can be found in Zhu et al. (2010) for the first four and in Qin et al. (2009) for slope position.

The mapping techniques used in this study were similar to those in Lark (1999). First, the clustered memberships were transformed by a symmetric log-ratio, considering that the membership values were compositional (Lark, 1999). Then, empirical best linear unbiased predictor with residual maximum likelihood (REML-EBLUP) (Lark et al., 2006) was used instead of a regression model by maximum likelihood (Lark, 1999) to construct soil maps based on transformed memberships. Considering that not all of the terrain attributes influence the spatial distribution of SOM (Chai et al., 2008; Sun et al., 2011), REML-EBLUP was applied based on those memberships that were statistically significantly correlated with SOM. An analysis of REML-EBLUP was performed using geoR (Ribeiro and Diggle, 2001) in R (R Development Core Team, 2006).

2.3. Implementation of FCM clustering

The FCM clustering algorithm of Bezdek et al. (1984) is the most commonly used fuzzy classification procedure. The procedure involves an iterative process of partitioning a dataset into a given number of clusters, calculating the membership of each object in the dataset belonging to each cluster and calculating the center of each cluster. The process stops when the objective function in Eq. (1) converges to a minimum:

$$J(M,c)=\sum_{i=1}^{n}\sum_{j=1}^{c}u_{ij}^{m}d_{ij}^{2}\left(x_i,C_j\right) \tag{1}$$

$$u_{ij}=\frac{d_{ij}^{-2/(m-1)}}{\sum\limits_{j=1}^{c}d_{ij}^{-2/(m-1)}} \tag{2}$$

$$d_{ij}^{2}=\left(x_{iv}-C_{jv}\right)'A\left(x_{iv}-C_{jv}\right),\quad A=A_E=\left|\begin{array}{cccc}1&0&\cdots&0\\0&1&\cdots&0\\\vdots&\vdots&\ddots&\vdots\\0&0&\cdots&1\end{array}\right|,$$

$$A=A_D=\left|\begin{array}{cccccc}\sigma_1^2&0&\cdots&\cdots&\cdots&0\\0&\sigma_1^2&\cdots&\cdots&\cdots&\vdots\\\vdots&\cdots&\ddots&\cdots&\cdots&\vdots\\\vdots&\cdots&\cdots&\sigma_k^2&\cdots&\vdots\\\vdots&\cdots&\cdots&\cdots&\ddots&\vdots\\0&\cdots&\cdots&\cdots&\cdots&\sigma_p^2\end{array}\right|^{-1},\ \text{or}$$

$$A=A_M=\left|\begin{array}{cccccc}\sigma_{11}^2&\sigma_{12}^2&\cdots&\cdots&\cdots&\sigma_{1p}^2\\\sigma_{21}^2&\sigma_{22}^2&\cdots&\cdots&\cdots&\vdots\\\vdots&\cdots&\ddots&\cdots&\cdots&\vdots\\\vdots&\cdots&\cdots&\sigma_{kv}^2&\cdots&\vdots\\\vdots&\cdots&\cdots&\cdots&\ddots&\vdots\\\sigma_{p1}^2&\cdots&\cdots&\cdots&\cdots&\sigma_{pp}^2\end{array}\right|^{-1} \tag{3}$$

where J denotes the objective function, M denotes the classification result, c is the number of clusters, n is the number of objects in the dataset, u_{ij} is the membership value of object x_i belonging to cluster C_j, m is the fuzzy exponent, d_{ij} is the distance between object x_i and cluster C_j, x_{iv} is the value of the attribute v of object x_i, C_{jv} is the value of the attribute v of cluster C_j, A is the distance metric, which could be the Euclidean norm A_E, diagonal norm A_D or Mahalonobis' norm A_M, σ_v is the standard deviation of attribute v, and σ_{kv}^2 is the covariance between attribute k and attribute v.

In theory, the value of m can be chosen to be between 1 and infinity, but in practice, it was frequently selected to be approximately 2.0 and usually equal to 2.0 (Lark, 1999; Odeh et al., 1992; Pal and Bezdek, 1995). c ranges from 2 to n, but it was usually less than 15 in practice. The Euclidean norm weights equally the variables to be analyzed and is most suitable for the general shape of a sphere. The diagonal norm can compensate for distortions in the assumed spherical shape caused by disparities in variances among attributes. Mahalonobis' norm can not only compensate for the distortions like the diagonal norm but can also account for the statistical correlations between attributes (Odeh et al., 1992).

In this study, FCM clustering was performed using the FuzMe program (Minasny and McBratney, 2002). m was set at [1.1, 3.5] for simulated data and at [1.1, 2.5] for data in the case study of soil mapping, both with an increment of 0.1. c was set at [2, 15] for all data with an increment of 1. The Euclidean norm was used in the simulated datasets because it is the best method for classifying them. Mahalonobis' norm was used for the case study of soil mapping considering that there could be correlations between terrain attributes, e.g., slope position and slope gradient.

2.4. Validity functions

FPI and *MPE* are the most frequently used validity functions in FCM clustering, and have been used in studies by Fecher and Schmidt (2003), Tan et al. (2006), Srinivas et al. (2008) and Triantafilis et al. (2009). These functions actually performed very similarly in many studies, including Amini et al. (2005), Odeh et al. (1992) and Tan et al. (2006). Therefore, only *FPI* was used in this study. *FPI* is computed as:

$$FPI = 1-\frac{cF-1}{c-1}\quad , \; F = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c}\left(u_{ij}\right)^2 \tag{4}$$

where F denotes the partition coefficient and measures the degree of overlapping between clusters. Considering that F has a monotonically decreasing tendency with c, it was normalized as shown in Eq. (4) and replaced by *FPI* to validate the FCM clustering results. *FPI* has a value between zero and one. A value of zero indicates that there is no overlapping between clusters, i.e., no shared objects by clusters, whereas a value of one indicates complete and even overlapping of all clusters, namely, that all objects are evenly shared by all clusters. In general, local minimized *FPI* along a series of c for a given m indicates the optimal clustering for this given m (Odeh et al., 1992).

Although *FPI* and *MPE* have been widely used, they have been criticized for ignoring the actual geometric properties of clustering results (Fecher and Schmidt, 2003; Xie and Beni, 1991). To cope with this defect of *FPI* and *MPE*, Xie and Beni (1991) devised another validity function called the compactness and separation index:

$$S = \frac{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c}u_{ij}^{m}d_{ij}^{2}\left(x_i, C_j\right)}{\min\left(d_{kj}^{2}\left(C_k, C_j\right)\right)} \tag{5}$$

In Eq. (5), the numerator measures the compactness of a clustering result, and the denominator measures the separation of a clustering result. Smaller compactness and larger separation give a smaller S, which indicates a better clustering result. Therefore, a clustering that has the smallest S against a series of c for a given m will be identified as the best clustering result for the given m. However, S has one defect; it will decrease monotonically with c when c is large (Xie and Beni, 1991). Fortunately, the monotonic decrease of S only occurs when c is very large, and, in practice, the largest c set for running FCM is usually small.

McBratney and Moore (1985) devised another validity function based on the assumption that the most valid c is associated with an extremum of the objective function's derivative. This validity function is a derivative of the objective function with respect to m, i.e., the function $-[(\delta J_E/\delta m)c^{0.5}]$:

$$\frac{\delta J(M,C)}{\delta m} = \sum_{i=1}^{n}\sum_{j=1}^{c}u_{ij}^{m}\ \log\left(u_{ij}\right)d_{ij}^{2}\left(x_i, C_j\right) \tag{6}$$

Plotting this function for different c against a series of m was used to identify the best combination of m and c, i.e., the one that has the lowest maximal value for this function.

2.5. Sensitivities of validity functions and soil maps to m and c

Validity functions from FCM clustering the simulated datasets were plotted against m and c to show their relationship to m and to c. Then, the sensitivity of the validity functions to m and c were assessed based on their plotted behaviors. Meanwhile, the squared Euclidean distance between a clustered center of a clustering result with c equal to the known c of a simulated dataset and its corresponding designed center was computed to examine the performances of these validity functions for finding the best m.

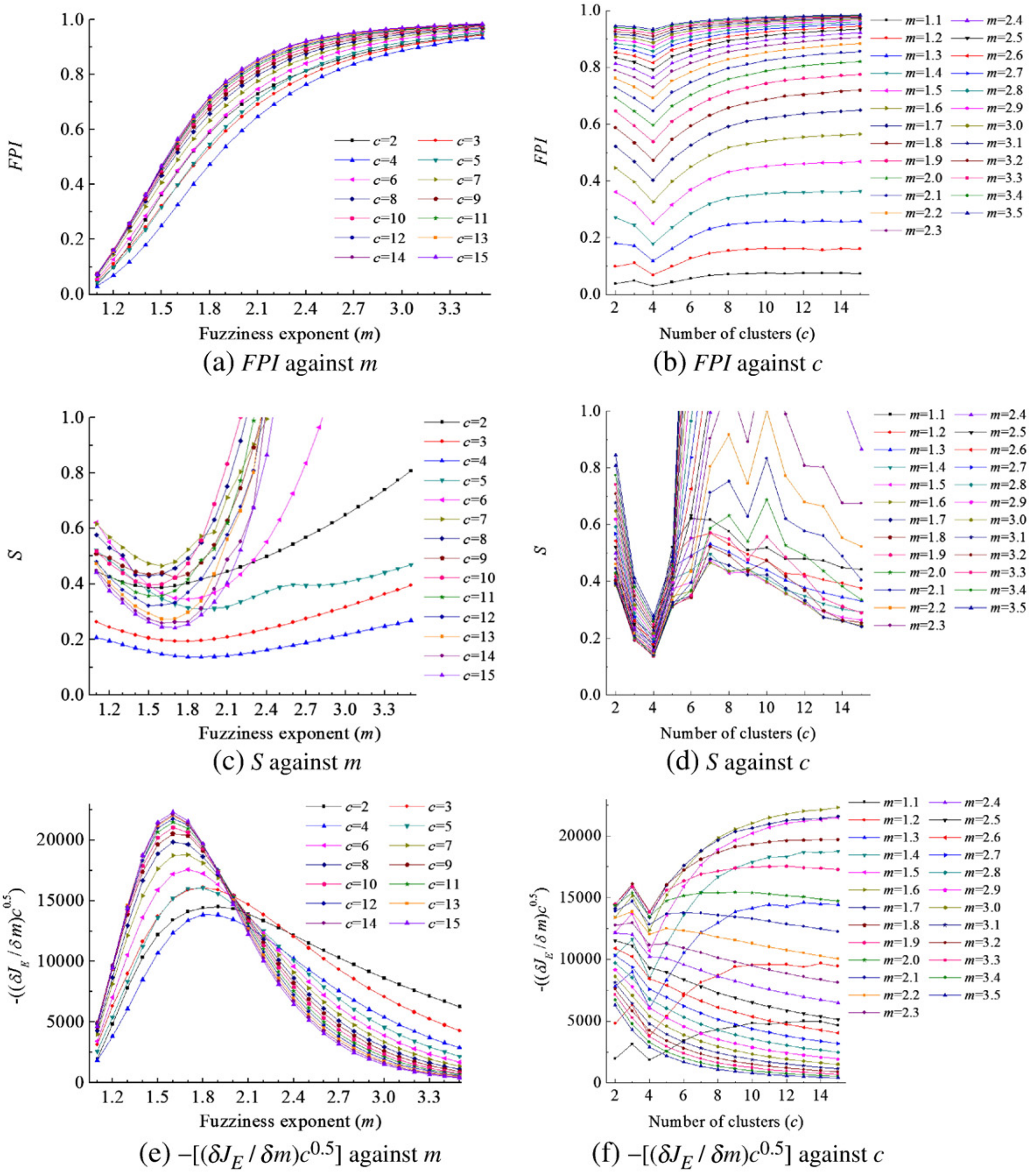
Sensitivities of soil mapping accuracy to m and c were also investigated by plotting the behaviors of mapping accuracy indexes against the two parameters. The maps produced in this study were validated according to Brus et al. (2011). Considering the small number of validation samples in this study ($n=21$), cross-validation on the samples for mapping was also conducted to obtain mapping errors, in addition to mapping errors from the validation samples. All of the mapping errors and all of the absolute mapping errors were then used to compute global mean error and global mean absolute error, respectively, through block kriging. This process took into account the spatial correlations of the mapping errors and the absolute mapping errors (Brus et al., 2011).

3. Results

3.1. Sensitivities of validity functions to m and c

3.1.1. Simulated overlapping dataset

Fig. 3 shows the behavior of the validity functions from FCM clustering the simulated overlapping dataset to m and to c. In Fig. 3(a), *FPI* increased monotonically with m for any given c. The gradient increased first as m increased from 1.1 to 1.3–1.6 (depending on c), and then decreased as m continued to increase. This suggests that the sensitivity of *FPI* to m was the highest with an m of approximately 1.4 (1.4 was the mode of m at which the increasing gradient of *FPI* for a given c was largest) and decreased as m decreased or increased from approximately 1.4. The sensitivity of *FPI* to m was closely related to c, as shown in Fig. 3(a), such that the curves of *FPI* were in a regular sequence of c. Among these c, the sensitivity of *FPI* to m was always the weakest for $c=4$, i.e., the designated number of clusters for this dataset (Fig. 3(a)). In Fig. 3(b), for each m, *FPI* was smallest at $c=4$, indicating the high sensitivity of *FPI* to the structure of the clustered dataset. The curves of *FPI* over the series of c varied gradually among m. The range of *FPI* over the series of c for a given m increased first as m increased from 1.1 to 1.7, and then decreased as m continued increasing until it reached $m=3.5$. Therefore, the sensitivity of *FPI* to c



(a) *FPI* against *m*

(b) *FPI* against *c*

(c) *S* against *m*

(d) *S* against *c*

(e) $-[(\delta J_E/\delta m)c^{0.5}]$ against *m*

(f) $-[(\delta J_E/\delta m)c^{0.5}]$ against *c*

Fig. 3. Plots of validity functions from clustering the overlapping dataset against *m* and *c* (*S* larger than 1 were not displayed in (c) and (d)).

was strongest for $m = 1.7$, and decreased gradually as *m* increased or decreased from $m = 1.7$.

Similarly, the sensitivities of *S* to *m* and to *c* were inferred from Fig. 3(c) and (d). As shown in Fig. 3(c), the sensitivity of *S* to *m* for a given *c* was the lowest when *m* was approximately 1.7 (1.7 was the mode of *m* at which *S* was lowest for a given *c*) and increased as *m* decreased or increased from approximately 1.7. The sensitivity of *S* to *m* was slightly more complexly related to *c* when compared with the sensitivity of *FPI* to *m* because the curves of *S* in Fig. 3(c) were in a more complex sequence of *c*. It was also the weakest for $c = 4$, i.e., the designated number of clusters in this dataset. Fig. 3(d) shows that *S* was prominently minimized at $c = 4$ for any given *m*, suggesting that *S* was highly sensitive to the structure of the clustered dataset. According to the range of *S* over the series of *c* for a given *m*, the sensitivity of *S* to *c* was weakest for $m = 1.5$ and increased gradually as *m* increased or decreased from $m = 1.5$.

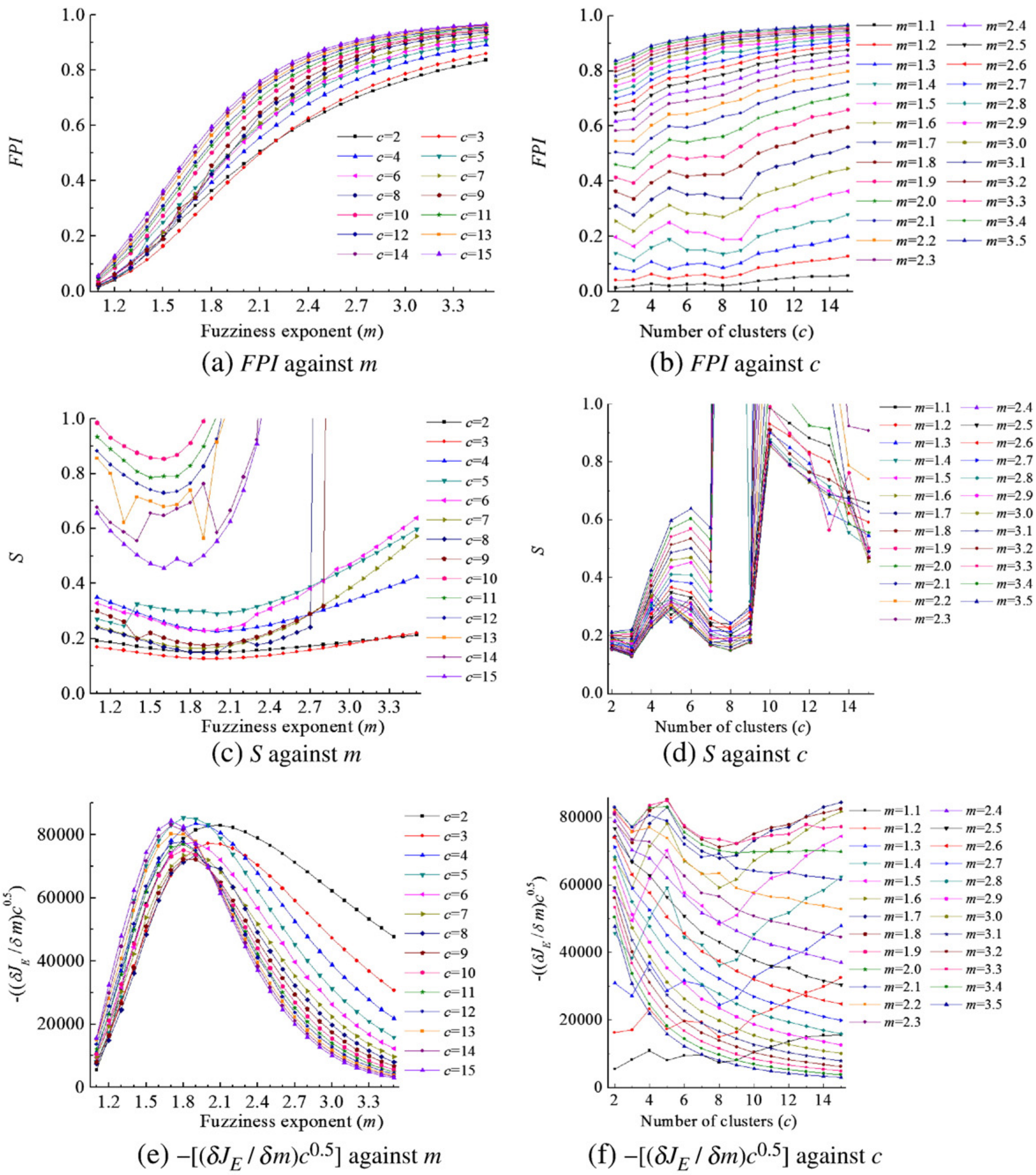
Fig. 3(e) shows that, for all *c*, the sensitivity of the function $-[(\delta J_E/\delta m)c^{0.5}]$ to *m* was the lowest at an *m* of approximately 1.7 (1.7 was the mode of *m* at which the function was highest for a given *c*) and the sensitivity increased as *m* increased or decreased from approximately 1.7. The function also correctly identified $c = 4$ as the most valid *c* to cluster the overlapping dataset. Fig. 3(e) shows that the lowest maximization of this function occurred at $c = 4$ and $m = 1.8$. Therefore, the function was also highly sensitive to the structure of the clustered dataset. The sensitivity of the function to *m* was related not only to *c* but also to *m*. For an *m* smaller than 2.0, the sensitivity of this function to *m* was the weakest at $c = 4$, whereas for an *m* greater than 2.2, its sensitivity at $c = 4$ was the strongest. Fig. 3(f) shows that there was a large difference between the curves of the function for a small *m* and for a large *m*; however, the curves for different *m* values varied gradually amongst *m*. For *m* values between 1.1 and 2.3, the function was minimized at $c = 4$. For other *m* values, the function decreased as c increased. According to the range of this function over the series of c for a given m, the sensitivity of the function to c became stronger as m increased from 1.1 to 1.5, then weaker as m increased from 1.5 to 2.0, again stronger as m increased from 2.0 to 2.8, and it finally became weaker as m continued increasing.

3.1.2. Simulated hierarchical dataset

Behavior of the three validity functions to m and to c from clustering the simulated hierarchical dataset (Fig. 4) were mostly similar to those from clustering the simulated overlapping dataset (Fig. 3) with the exception of two major distinctions. The results for this case are summarized in this section.

The first distinction is that the validity functions sometimes behaved discontinuously as shown in Fig. 4. For example, Fig. 4(a) shows that $c = 5$ when $m = 1.1$, 1.2 or 1.3, and Fig. 4(c) shows that $c = 8$ when $m = 1.9$, 2.1, 2.2 and other values. The three validity functions behaved discontinuously at the same time, but the discontinuities could be avoided by repeating the FCM clustering a number of times. Therefore, the discontinuity could be attributed to the complex structure of this dataset and the initial memberships for FCM clustering. Because repeating FCM clustering to avoid the discontinuities is very time-consuming, this was not done in the study.

The second but more important distinction is that the validity functions behaved more complexly with c (Fig. 4(b), (d) and (f)). Undoubtedly, this is due to the complex structure of this dataset. The sensitivity of the validity functions to c in Fig. 4 was much smaller than that in Fig. 3, suggesting that sensitivity of these validity functions to the structure of the dataset was much lower in this case. Fig. 4(b) shows that *FPI* identified several c values as the most valid c for clustering this dataset, including not only the two arguable designed c for this dataset (i.e., 3 and 8) but also 5 and 9. In Fig. 4(d), S indicated that 3 was the best c value for most m values, and 8 was the best for several for other m values. Fig. 4(e) shows that the



(a) *FPI* against m

(b) *FPI* against c

(c) S against m

(d) S against c

(e) $-[(\delta J_E / \delta m)c^{0.5}]$ against m

(f) $-[(\delta J_E / \delta m)c^{0.5}]$ against c

Fig. 4. Plots of validity functions from clustering the hierarchical dataset (S larger than 1 were not displayed in (c) and (d)).

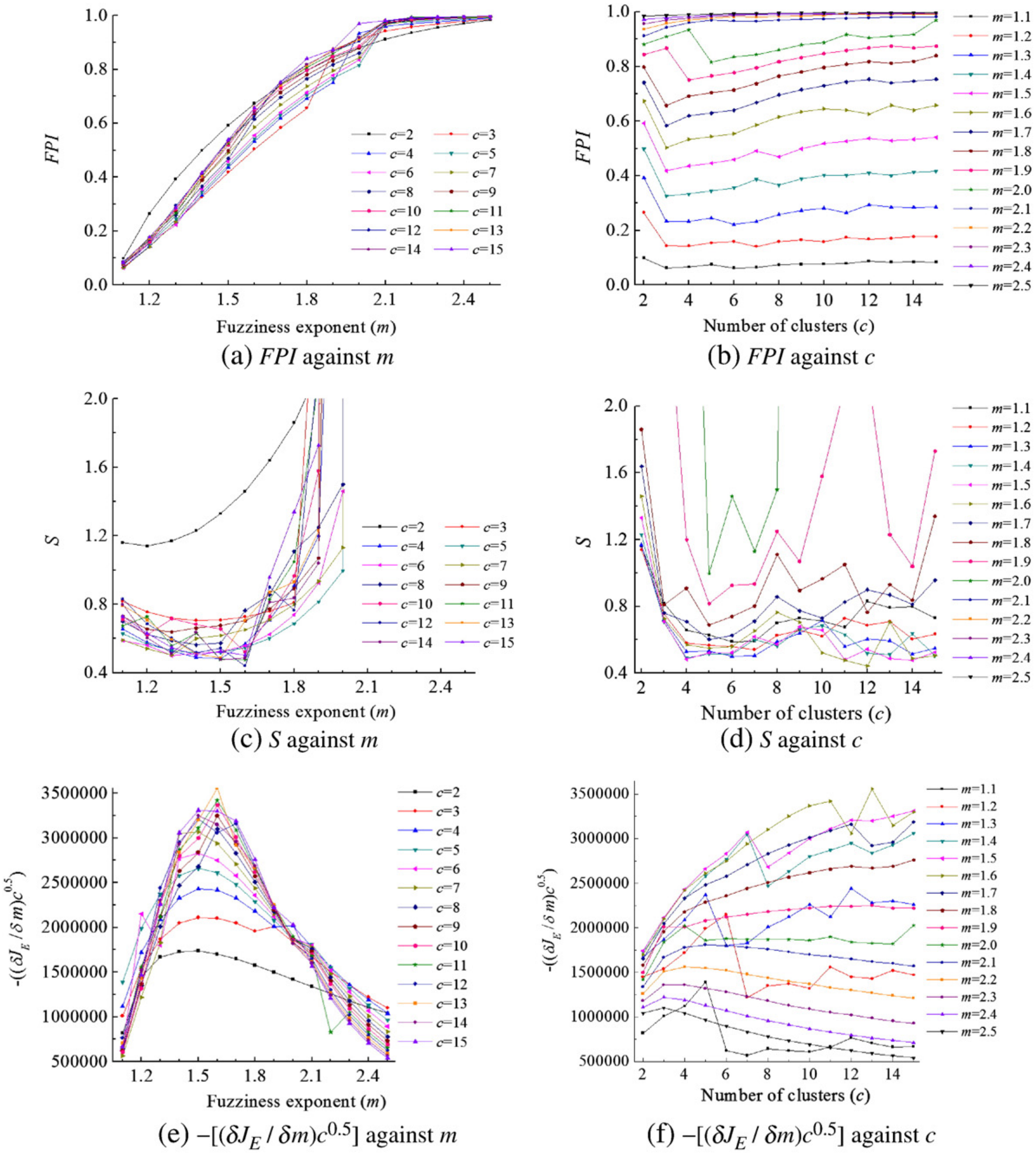
function $-[(\delta J_E/\delta m)c^{0.5}]$ indicated that $c=9$, rather than $c=3$ or $c=8$, at $m=1.9$ was the best parameter to cluster the dataset according to the general rule guiding the use of this function to find the best parameters. However, if Fig. 4(f) was the basis for selecting the most valid value of c, then $c=3$, $c=8$ and other c values would be chosen. Our results show that performance of these validity functions for identifying the optimal clustering result was seriously degraded in this case.

The sensitivity of these validity functions to m was also lower in this case. Fig. 4(a) generally shows that the sensitivity of *FPI* to m for a given c was generally highest at an m value of approximately 1.6 (1.6 was the mode of m at which the increasing gradient of *FPI* for a given c was largest), and that the value (1.4) shown in Fig. 3(a). In addition, the increasing gradients of *FPI* were relatively smaller in Fig. 4(a) when compared with Fig. 3(a). Fig. 4(b) shows that the range of m for *FPI* to indicate the most valid c for optimally clustering the dataset, i.e., roughly from 1.3 to 2.2, was much shorter than the range shown in Fig. 3(b), i.e., the entire range of m, which was 1.1 to 3.5. Similar phenomena are also shown in Fig. 4(c)–(f) for S and the function $-[(\delta J_E/\delta m)c^{0.5}]$.

3.2. Sensitivity of digital soil maps to m and c

Fig. 5 shows the three fuzzy validity functions from clustering the dataset of terrain attributes of the Heshan study area. Fig. 5(b) shows that *FPI* indicated several combinations of m and c for clustering the dataset, e.g., $c=3$ and $m=1.2$–1.8. Fig. 5(d) shows that S indicated that $m=1.6$ and $c=12$ were the best parameters for clustering the dataset. Fig. 5(e) shows that the function $-[(\delta J_E/\delta m)c^{0.5}]$ indicated that $c=2$ and $m=1.5$ were the best parameters. It appears that the three validity functions gave completely different solutions for how to cluster the dataset.

The correlations between cluster memberships and SOM were analyzed based on the sampling dataset that was used for mapping. We found that no memberships of clustering results with a value of m



(a) *FPI* against *m*

(b) *FPI* against *c*

(c) *S* against *m*

(d) *S* against *c*

(e) $-[(\delta J_E\,/\,\delta m)c^{0.5}]$ against *m*

(f) $-[(\delta J_E\,/\,\delta m)c^{0.5}]$ against *c*

Fig. 5. Variations of the three validity functions from clustering the dataset of the terrain attributes of the Heshan study area.

Table 2
Summaries of mapping accuracies.

	Mean	Min	Max	Standard deviation
Mean error (g kg^{-1})	−0.01	−0.06	0.04	0.03
Mean absolute error (g kg^{-1})	0.78	0.80	0.89	0.04
Global mean error (g kg^{-1})	−0.01	−0.05	0.08	0.02
Global mean absolute error (g kg^{-1})	0.81	0.74	0.93	0.04

greater than 2.0 were statistically significantly correlated with SOM. This can be explained by Fig. 5(b), which shows that *FPI* for values of *m* larger than 2.0 were all close to 1, indicating that the clustering results from values of *m* greater than 2.0 were completely fuzzy, and all of the cluster memberships were equal. Therefore, only clustering results that were generated using a value of *m* less than or equal to 2.0 were applied for soil mapping.

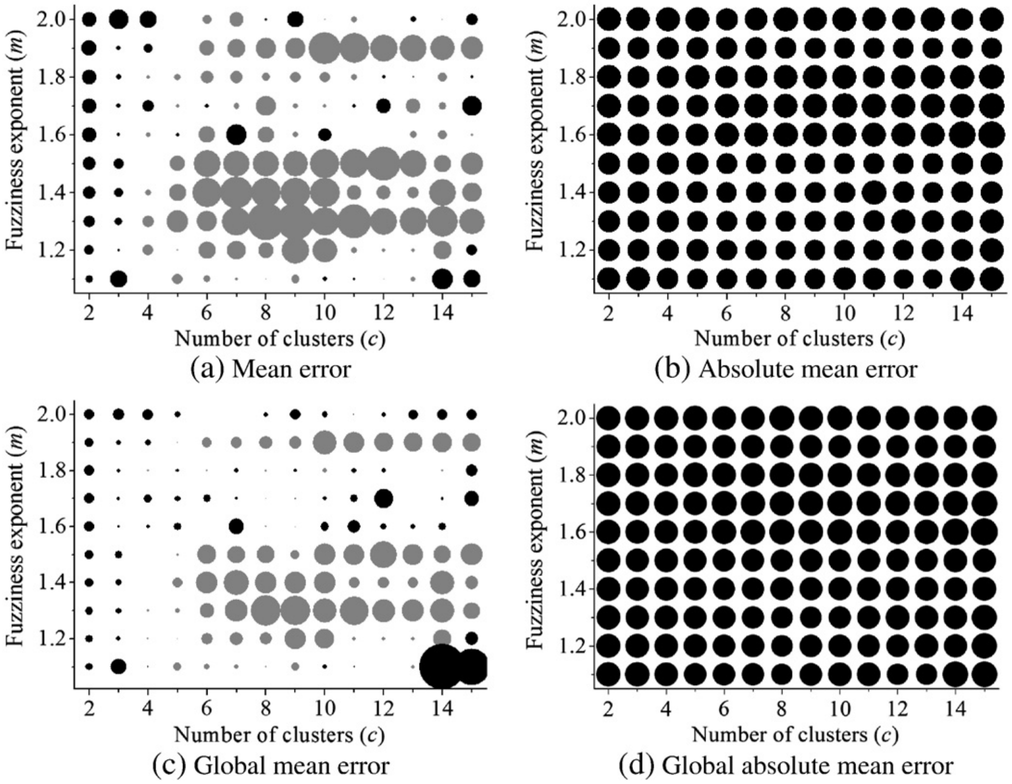
Mapping accuracy indexes, i.e., mean error, mean absolute error, global mean error and global mean absolute error, are summarized in Table 2 and plotted against *m* and *c* in Fig. 6 to show their variations with *m* and *c*. Thirteen maps are presented in Fig. 7, and their means and standard deviations are shown in Table 3. Although the variations of mean error and global mean error with *m* and *c* (Fig. 6(a) and (c)) were noticeable, they were actually small when compared with the mean predicted values in Table 3. For example, the mean and the standard deviation of the global mean error in Table 2, −0.01 g kg^{-1} and 0.03 g kg^{-1}, accounted for very small proportions of the average of the means in Table 3, 4.64 g kg^{-1}. More evidently, Table 2 and Fig. 7(b) show that the variations of mean absolute error and global mean absolute error with *m* and *c* were very small. Therefore, the sensitivity of the mapping accuracy to *m* and to *c* was very low.

The thirteen maps shown in Fig. 7 were based on five *m* values, i.e., 1.2, 1.4, 1.6, 1.8 and 2.0, and three *c* values, i.e. 3, 5, 7, 9, 11 and 13. The five maps at *c* = 3 changed gradually along with *m*. At other *c* values, the gradual trend with *m* was not good, and generally only occurred for the first two values, i.e., at *m* = 1.2 and 1.4. The maps at *m* = 1.6, 1.8 and 2.0 and at *c* = 7, 9 11 and 13 were not similar to each other. The means in Table 3 also show a similar trend. However, an analysis of variance on the predicted means and standard deviations in Table 3 showed that neither the differences in the predicted means nor the differences in the standard deviations were statistically significant among the *m* values (p = 0.76 and p = 0.06, respectively). Although the maps were sensitive to *m*, especially when *c* was large, the sensitivity was not significant. Similarly, the maps changed gradually with *c* at small values of *m*, i.e., *m* = 1.2 and 1.4, whereas the maps at other large *m* values changed abruptly with *c*. An analysis of variance of the predicted means and standard deviations in Table 3 showed that only the differences in the standard deviations of the maps were statistically significant among *c* values (p < 0.01). These results suggest that the soil maps were more sensitive to *c* than to *m*, but overall, only the spatial variations of the SOM presented on the maps were sensitive to *c*.

4. Discussion

4.1. Sensitivity of validity functions to m and to c

The sensitivity of the validity function to *m* and to *c* was generally the only criterion for selecting *m* and *c* values for FCM clustering. The results of this study first show that the validity functions, *FPI*, *S* and the function $-[(\delta J_E/\delta m)c^{0.5}]$, were sensitive in differing degrees to the structure of the dataset that was being clustered, and the sensitivity was determined by the structure of the dataset. For a dataset of overlapping clusters, the validity functions all correctly identified the optimal clustering result. At *m* = 1.7, *FPI* was most sensitive to the optimal clustering result, whereas at *m* = 1.5, *S* and the function $-[(\delta J_E/\delta m)c^{0.5}]$ were the most sensitive. For the dataset containing hierarchical clusters, the sensitivities of these validity functions to the optimal clustering result were substantially lower,



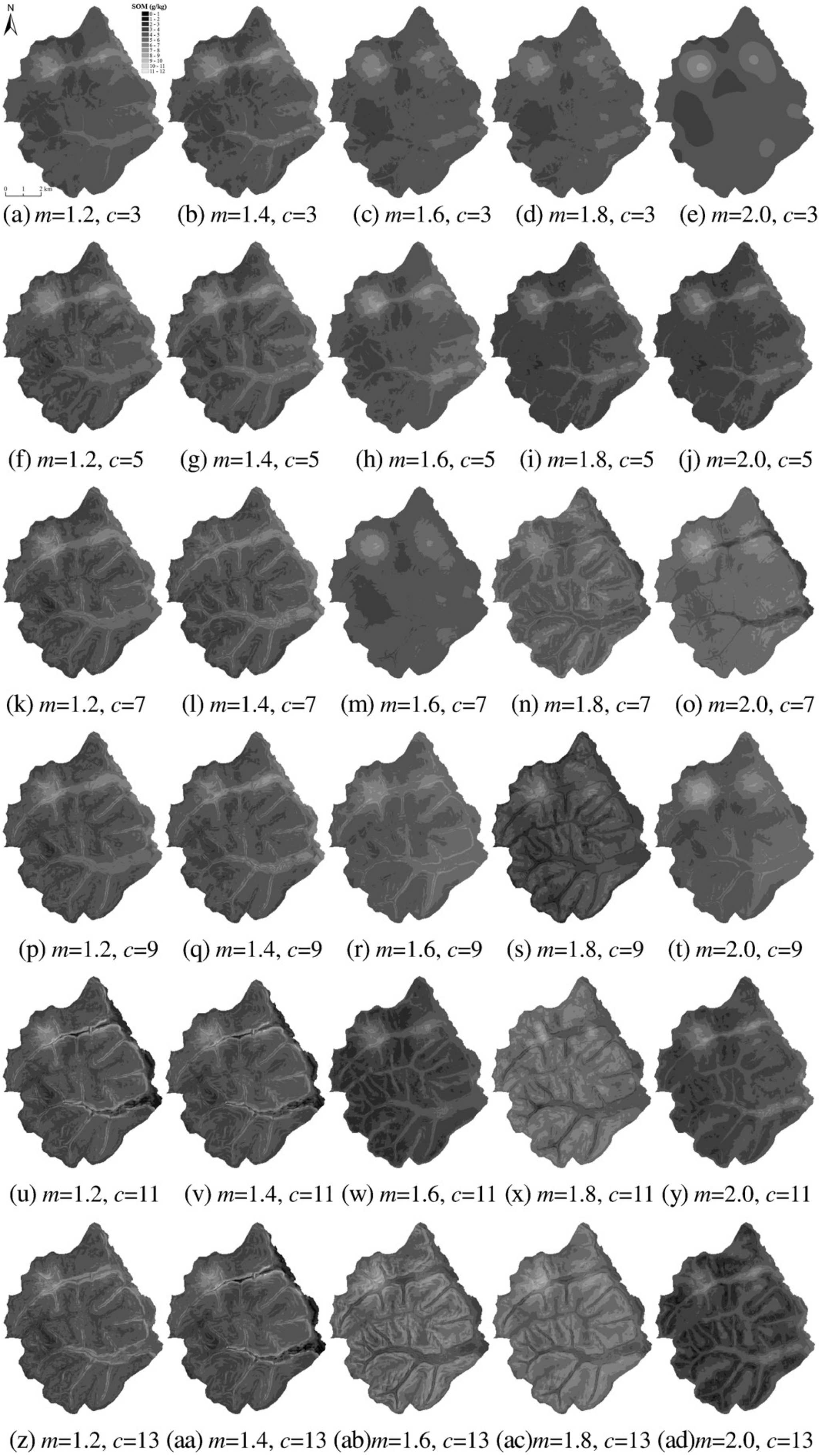
(a) Mean error

(b) Absolute mean error

(c) Global mean error

(d) Global absolute mean error

Fig. 6. Bubble plots of mapping accuracy indexes. The sizes of the bubbles represent the values of the accuracy index, which are summarized in Table 2. A gray color represents a negative value, whereas a black color represents a positive value.



(a) m=1.2, c=3 (b) m=1.4, c=3 (c) m=1.6, c=3 (d) m=1.8, c=3 (e) m=2.0, c=3

(f) m=1.2, c=5 (g) m=1.4, c=5 (h) m=1.6, c=5 (i) m=1.8, c=5 (j) m=2.0, c=5

(k) m=1.2, c=7 (l) m=1.4, c=7 (m) m=1.6, c=7 (n) m=1.8, c=7 (o) m=2.0, c=7

(p) m=1.2, c=9 (q) m=1.4, c=9 (r) m=1.6, c=9 (s) m=1.8, c=9 (t) m=2.0, c=9

(u) m=1.2, c=11 (v) m=1.4, c=11 (w) m=1.6, c=11 (x) m=1.8, c=11 (y) m=2.0, c=11

(z) m=1.2, c=13 (aa) m=1.4, c=13 (ab)m=1.6, c=13 (ac)m=1.8, c=13 (ad)m=2.0, c=13

Fig. 7. Predicted SOM maps based on different FCM clustering results.

Table 3
The mean and standard deviation (between brackets) of SOM (g kg^{-1}) on produced maps.

	$m = 1.2$	$m = 1.4$	$m = 1.6$	$m = 1.8$	$m = 2.0$
$c = 3$	4.58 (0.67)	4.60 (0.74)	4.54 (0.56)	4.57 (0.52)	4.54 (0.55)
$c = 5$	4.54 (0.77)	4.61 (0.79)	4.71 (0.73)	4.04 (0.68)	4.03 (0.67)
$c = 7$	4.56 (0.78)	4.64 (0.77)	4.60 (0.52)	5.15 (0.60)	5.38 (0.65)
$c = 9$	4.57 (0.78)	4.61 (0.78)	5.06 (0.64)	3.97 (0.73)	4.94 (0.69)
$c = 11$	4.38 (0.93)	4.41 (0.87)	3.72 (0.75)	5.48 (0.68)	4.13 (0.73)
$c = 13$	4.52 (0.82)	4.38 (0.94)	5.34 (0.92)	5.62 (0.78)	3.72 (0.85)

unfortunately giving a consequence that the non-optimal clustering results were also indicated as optimal.

Second, a wide range of m values existed for optimally clustering a dataset, but the range was different for different validity functions and also depended on the dataset that was being clustered. As shown in Figs. 3 and 4, the validity functions gave the optimal clustering results for a series of m values. Among the three validity functions, the function $-[(\delta J_E/\delta m)c^{0.5}]$ identified the smallest range of m values followed by *FPI* and *S*. For example, in the study on the dataset of overlapping clusters, the range that was given by the function $-[(\delta J_E/\delta m)c^{0.5}]$ was roughly 1.1 to 2.3 (Fig. 3(f)). It was roughly 1.1 to 3.5 for *FPI* (Fig. 3(b)), and the range was greater than 1.1 to 3.5 for *S* (Fig. 3(d)). The sensitivities of the three validity functions to m were lower in the case of the hierarchical dataset than in the case of the overlapping dataset, suggesting that the range of m for the validity functions to identify the optimal clustering result also depended on the dataset being clustered.

4.2. Sensitivity of soil maps to m and c

In the case study of soil mapping, different combinations of m and c were identified by different validity functions as the best for clustering the dataset of terrain attributes. This makes it very difficult to use them when applying FCM in digital soil mapping. However, the mapping accuracy based on all the clustering results was not very different regardless of whether the clustering results were optimal in terms of the validity functions. This means that the sensitivity of mapping accuracy to m and c was very low. Therefore, concerning mapping accuracy, there is a wide range of m and c values that can be safely used in FCM for soil mapping.

Conversely, the soil maps that were based on different FCM clustering results were more sensitive to m and c than mapping accuracy. Sensitivity was much higher if m and c were large. This could be due to the higher sensitivity of FCM clustering to m and c at large m values (less than 2.0). For example Fig. 5(a) shows that the variance of *FPI* for c values were larger at large m values, and, generally, the sensitivity of *FPI* to m was larger at larger c values than at smaller ones (except $c = 2$, which was not used to produce maps). However, only the sensitivity of the spatial variations to c that were presented on the maps was statistically significant. Soil maps were more sensitive to c than to m, especially at medium and large m values (1.6 to 2.0 in this study). This is determined by the nature of the two parameters. In FCM, m is responsible for distributing the membership of an object to a cluster. A small change of m induces small changes to memberships, and small changes in memberships can only impose small changes on the produced soil maps. In contrast, a small change of c caused memberships to change more. For example, two memberships of 0.5 at $c = 2$ become three memberships of 1/3 at $c = 3$.

5. Conclusions

This study demonstrated that three commonly used validity functions, i.e., *FPI*, *S* and the function $-[(\delta J_E/\delta m)c^{0.5}]$, were sensitive in differing degrees to the structures of the datasets under a wide range of m values, but the sensitivities and the range of m were different for different validity functions and dependent on the clustered datasets. Soil maps based on FCM clustering were sensitive to m and c, especially when m and c were not small, for example, a value of 1.6 to 2.0 for m and a value of 5 to 13 for c. However, only the spatial variations of SOM presented on the maps were statistically significantly sensitive to c. Moreover, mapping accuracy was slightly sensitive to m and c. We concluded that there is a range of optimal m values over which digital soil maps did not change very much, but this was not certain for c because the spatial variation on the maps changed significantly with c. In addition, the range of m values varies in different situations for soil mapping. This research only presented one case study due to a limited budget. Future studies should focus on identifying the relationship between the range of m values and the situation for soil mapping.

Acknowledgements

Financial support from the Basic Research Program of Jiangsu province (Grant no. BK2008058), the Natural Science Foundation of China (40625001 and 40971236), the National Key Technology R&D Program of China (2007BAC15B01), the Vilas Associate Program and the Hamel Faculty Fellow at the University of Wisconsin-Madison are gratefully acknowledged. We are also grateful to Dr. Murray Lark (Rothamsted Research), Dr. Scott M. Lesch (University of California, Riverside) and Dr. Dick Brus (Wageningen University and Research Centre) for providing the methods used in the present study.

References

Amini, M., Afyuni, M., Fathianpour, N., Khademi, H., Flhler, H., 2005. Continuous soil pollution mapping using fuzzy logic and spatial interpolation. Geoderma 124, 223–233.

Bezdek, J.C., Ehrlich, R., Full, W., 1984. FCM: The fuzzy *c*-means clustering algorithm. Computers & Geosciences 10, 191–203.

Brus, D., Kempen, B., Heuvelink, G.B.M., 2011. Sampling for validation of digital soil maps. European Journal of Soil Science 62, 394–407.

Chai, X., Shen, C., Yuan, X., Huang, Y., 2008. Spatial prediction of soil organic matter in the presence of different external trends with RMEL-EBLUP. Geoderma 148, 159–166.

de Bruin, S., Stein, A., 1998. Soil-landscape modelling using FCM of attribute data derived from a Digital Elevation Model (DEM). Geoderma 83, 17–33.

Fecher, M., Schmidt, J.M., 2003. Fuzzy clustering as a means of selecting representative conformers and molecular alignments. Journal Chemistry Information Computer Science 43, 810–818.

Lagacherie, P., Cazemier, D.R., van Gaans, P.F.M., Burrough, P.A., 1997. Fuzzy k-means clustering of fields in an elementary catchment and extrapolation to a larger area. Geoderma 77, 197–216.

Lark, R.M., 1999. Soil-landform relationships at within-field scales: an investigation using continuous classification. Geoderma 92, 141–165.

Lark, R.M., Cullis, B.R., Welham, S.J., 2006. On spatial prediction of soil properties in the presence of a spatial trend: the empirical best linear unbiased predictor (E-BLUP) with REML. European Journal of Soil Science 57, 787–799.

McBratney, A.B., de Gruijter, J.J., 1992. A continuum approach to soil classification by modified fuzzy-k-means with extragrades. Journal of Soil Science 43, 159–175.

McBratney, A.B., Moore, A.W., 1985. Application of fuzzy sets to climatic classification. Agricultural and Forest Meteorology 35, 165–185.

McBratney, A.B., Odeh, I.O.A., 1997. Application of fuzzy sets in soil science: fuzzy logic, fuzzy measurements and fuzzy decisions. Geoderma 77, 85–113.

Minasny, B., McBratney, A.B., 2002. FuzME version 3.0, Australian Centre for Precision Agriculture, The University of Sydney, Australia. http://www.usyd.edu.au/su/agric/acpa, 2002.

Odeh, I.O.A., McBratney, A.B., Chittleborough, D.J., 1992. Soil pattern recognition with fuzzy-c-means: application to classification and soil-landform interrelationships. Soil Science Society of American Journal 56, 505–516.

Pal, R.N., Bezdek, J.C., 1995. On cluster validity for the fuzzy c-means model. IEEE Transactions on Fuzzy Systems 3, 370–379.

Qin, C.Z., Zhu, A.X., Shi, X., Li, B.L., Pei, T., Zhou, C.H., 2009. Quantification of spatial gradation of slope positions. Geomorphology 110, 152–161.

R Development Core Team, 2006. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria3-900051-07-0. URL http://www.R-project.org.

Ribeiro Jr., P.J., Diggle, P.J., 2001. GeoR: a package for geostatistical analysis. R-NEWS (ISSN: 1609-3631) 1 (2), 15–18.

Srinivas, V.V., Tripathi, S., Ramachandra Rao, A., Govindaraju, R.S., 2008. Regional flood frequency analysis by combining self-organizing feature map and fuzzy clustering. Journal of Hydrology 348, 148–166.

Sun, X.L., Zhao, Y.G, Zhang, G.L., Wu, S.C., Man, Y.B., Wong, M.H., 2011. Application of a digital soil mapping method in predicting soil orders on mountain areas of Hong Kong based on legacy soil data. Pedosphere 21, 339–350.

Tan, M.Z., Xu, F.M., Chen, J., Zhang, X.L., Chen, J.Z., 2006. Spatial prediction of heavy metal pollution for soils in peri-urban Beijing, China based on fuzzy set theory. Pedosphere 16, 545–554.

Triantafilis, J., Kerridge, B., Buchanan, S.M., 2009. Digital soil-class mapping from proximal and remotely sensed data at the field level. Agronomy Journal 101, 841–853.

Xie, X.L., Beni, G., 1991. A validity measure for fuzzy clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence 13, 841–847.

Zhu, A.X., Yang, L., Li, B.L., Qin, C.Z., Pei, T., Liu, B., 2010. Construction of membership functions for predictive soil mapping under fuzzy logic. Geoderma 155, 164–174.