Local spatial regression models : a comparative analysis on soil contamination

Stoch Environ Res Risk Assess
DOI 10.1007/s00477-011-0532-2

ORIGINAL PAPER

Local spatial regression models: a comparative analysis on soil contamination

Bulent Tutmez · Uzay Kaymak · A. Erhan Tercan

Abstract Spatial data analysis focuses on both attribute and locational information. Local analyses deal with differences across space whereas global analyses deal with similarities across space. This paper addresses an experimental comparative study to analyse the spatial data by some weighted local regression models. Five local regression models have been developed and their estimation capacities have been evaluated. The experimental studies showed that integration of objective function based fuzzy clustering to geostatistics provides some accurate and general models structures. In particular, the estimation performance of the model established by combining the extended fuzzy clustering algorithm and standard regional dependence function is higher than that of the other regression models. Finally, it could be suggested that the hybrid regression models developed by combining soft computing and geostatistics could be used in spatial data analysis.

Keywords Local regression modelling · GWR · Fuzzy clustering · Regional dependence function

B. Tutmez (✉)
School of Engineering, Inonu University,
Malatya 44280, Turkey
e-mail: bulent.tutmez@inonu.edu.tr

U. Kaymak
School of Industrial Engineering, Eindhoven University
of Technology, P. O. Box 513, 5600 MB Eindhoven,
The Netherlands

A. E. Tercan
Department of Mining Engineering, Hacettepe University,
Ankara 06800, Turkey

1 Introduction

In spatial data analysis, each measurement is associated with a location and there is at least an implied connection between the location and the measurement. In addition, spatial relationships are concerned with different values for any property, which is measured at a set of irregularly distributed geographic locations in an area. Spatial relationships between some variables can be modelled in different ways using statistical models. In recent years, a variety of useful regression models have been developed to explore the spatial nature of variables (Gao et al. 2006).

Although some global approaches have been employed for evaluating uncertainties in the systems, they have the shortcoming that they can mask geographical variations in relationships. The aim is to construct a regional model on the basis of locations with measurements and then to use this model for regional estimations at any desired point within the area (Şen 2009). For this purpose, local regression models have been proposed to permit the exploration of spatial relationships in datasets (Atkinson and Naser 2010; Harris et al. 2011).

Local modelling has been employed widely in some disciplines for several decades. However, in some disciplines, such as geosciences, environment, ecology and geography, a focus on methods that account for local variation and spatially heterogeneous effects has been a comparatively new development (Lloyd 2006; Waller et al. 2007). Recently, geographically weighted regression (GWR), which is a useful and effective methodology for locally modelling relationships, has been developed by (Fotheringham et al. 1998).

Many spatial datasets have high levels of uncertainty. The treatment of uncertainty in analysis is going through a paradigm shift from a probabilistic framework to a

Published online: 12 October 2011

generalized framework (Tutmez and Tercan 2007). In recent years some hybrid (Lee 2000), soft computing (Wong et al. 2001) and machine learning algorithms (Kanevski et al. 2009) have been proposed for modelling complex and vague systems. These methods are based on less restrictive assumptions, and are flexible in modelling non-linearity and non-constant variable structures (Bogardi et al. 2003; Sousa and Kaymak 2002).

In this study, in addition to traditional GWR model, four new model structures are proposed for spatial data analysis and system modelling. The main motivation of the models is obtaining regression weights and bandwidths (search radii) from spatial analysis. For this purpose, objective function based fuzzy clustering algorithms and regional dependence functions are used for analysing the spatial system. The performance of the hybrid model established by combining the extended fuzzy clustering algorithm and standard regional dependence function (SRDF) is exhibited in the paper. The results indicate that the hybrid frameworks could be found reliable methodologies for spatial data analysis.

The rest of the paper is organized as follows. The next section describes the methodological frame and the methods such as locally weighted regression, GWR, and the proposed hybrid structure. After that, an experimental comparative study to appraise the spatially varying data by the weighted areal regression models is given. Five areal models are established and their estimation capacities are assessed via some indicators. Based on the results and a discussion, the superiority of the hybrid regression models developed by integrating soft computing and geostatistics and their contribution to spatial data analysis is presented.

2 Regression modelling for spatial data

Regression analysis is employed to estimate the quantitative functional relationships between response variable and one or more predictor variables from the measured data. A common feature of this procedure is that it is applied globally, that is, to the entire site under study. However, it is often desirable to examine the relationship at a more local scale. In this section we first introduce the general frame of the regression procedure and then give some alternative local regression models.

2.1 Local regression estimate

From a general statistical point of view, the regression is employed to describe a relation between a predictor variable (or variables) X and a response variable Y. It is possible to account for correlated observations by considering a structure of the following kind in the model (Waller and Gotway 2004). If the vector of response variables is multivariate normal, we can express the model as follows:

$$Y = \mu + e, \tag{1}$$

where μ is the vector of area means, which can be modelled in different ways and e is the vector of random errors, which we assume is normally distributed with zero mean and covariance matrix V (Bivand et al. 2008).

The classical regression equation, in matrix form, can be given by:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon \tag{2}$$

where the vector of parameters to be estimated, β, is constant in space, this can be taken

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y}. \tag{3}$$

Spatial evaluation is applied at a 'global' level in such a way that one set of results is generated from the models, representing one set of relationships, which is assumed to apply equally across the study area (Harris et al. 2011). An assumption of global analysis is that the relationship under study is spatially constant, and thus, the relationships being characterized are 'stationary' over space. However, in most cases, the relationship varies in space. If the local coefficients vary in space, it can be taken as an indication of non-stationary. For this purpose local spatial procedures, which should deal with the spatial non-stationary of empirical relationships, must be considered.

2.2 Geographically weighted regression (GWR) model

Fotheringham et al. (1998) proposed GWR model for local estimation of the parameters given by (3). In the mechanics of GWR, the observations are weighted in accordance with their distance from the kernel centre. The parameters for GWR may be estimated by solving Eq. 4

$$\hat{\boldsymbol{\beta}}(u_i, v_i) = \left[\mathbf{X}^T\mathbf{W}(u_i, v_i)\mathbf{X}\right]^{-1}\mathbf{X}^T\mathbf{W}(u_i, v_i)\mathbf{y}, \tag{4}$$

where $\hat{\boldsymbol{\beta}}$ represents an estimate of β, and $\mathbf{W}(u_i, v_i)$ is an n by n matrix whose off-diagonal elements are zero and diagonal elements are geographical weights of each of the n observed data for regression point I (Fotheringham et al. 2002).

In a standard GWR analysis, instead of $\mathbf{W}(u_i, v_i)$, $\mathbf{W}(\mathrm{i})$ can be used as weighting scheme based on the proximity of the regression point i to the data points around i without an explicit relationship being stated. There are many weighting schemes which express w_{ij} as a continuous function of distance d_{ij}. In practice, the following Gaussian function is used extensively.

$$w_{ij} = \exp\left[-\frac{1}{2}\left(\frac{d_{ij}}{b}\right)^2\right] \quad (5)$$

where d_{ij} is the Euclidean distance between the location of measurement i and the centre of the kernel j, and b is the bandwidth of the kernel. If i and j coincide, the weighting of data at that point will be unity and the weighting of the other data will decrease according to a Gaussian curve as the distance between i and j increases (Fotheringham et al. 2002).

As can be seen in (5), the weighting has a critical importance and it completely depends on the bandwidth of the function. If b is too small, insufficient data fall within the smoothing window, and a noisy fit, or large variance, will result (Paez et al. 2002). On the other hand, if b is too large, the local model may not fit the data well within the smoothing window, and important features of the mean function may be distorted or lost completely. That is, the fit will have large bias (Loader 1999). From an ideal methodological view, one might like to define a separate bandwidth for each estimation point.

2.3 Standard regional dependence function based regression (SRDFR) model

In this proposed method, the diagonal weight matrix $\mathbf{W}(u_i, v_i)$ in (4) is derived from the spatial relationships between the variables. In Geostatistics, the spatial variability in any phenomenon within a site can be measured by comparing the relative change between two locations. Two numerical values z(x) and $z(x + h)$ at two points x and $x + h$ separated by the vector h are spatially correlated. As the distance between these values increases, one would expect that the spatial correlation decreases and vice versa. This correlation can be modeled by tools such as the semivariogram (Goovaerts 1997).

In this study, the point cumulative semimadogram (PCSM) function (Tutmez et al. 2007) is employed for evaluating the spatial relationships between the observations that are mostly irregularly spaced. To overcome a notational complexity, we consider a measured target variable Z which is the dependent variable (Y) in regression framework. The mathematical expression can be given as follows

$$\gamma(h_i) = \frac{1}{2}\sum_{i=1}^{N-1} |Z_c - Z_i| \quad (6)$$

where $\gamma(h_i)$ is the PCSM value; Z_c and Z_i are the measured values at pivot location and other adjacent locations, respectively.

Madograms are particularly useful for establishing the range parameter. The PCSM takes the absolute difference between the measurements Z_m and Z_{m+h}. The PCSM is obtained by the successive summation of the semimadograms for irregularly spaced distances. The traditional experimental variograms may be noisy to infer the anisotropy features and range values due to squared experimental deviations. In such cases, madograms are useful because absolute experimental deviations have less influence on measure of spatial variability than squared deviations.

The PCSM can be obtained from data by the following steps (Tutmez and Hatipoglu 2007):

(a) Calculate distance between the concerned location and the remaining locations.If there are N locations, the number of different distances is $N - 1$, $h_i(i = 1, \ldots, N - 1)$.
(b) For each pair (pivot and any other location), compute the half of absolute differences between data values. By this way, each distance will have its half of absolute value.
(c) Plot distances versus corresponding successive cumulative sums of half of absolute differences. By using this procedure, a non-decreasing function which is the sample PCSM at the pivot location is obtained.
(d) Apply previous steps by considering different pivot locations, to give N sample PCSMs.

The PCSM leads to a non-decreasing function with distance. To weight the measured values, the SRDF (Şen and Şahin 2001; Tutmez and Hatipoglu 2007) can be applied as a suitable tool. The SRDF provides weights for different regional locations depending on the distance from the pivot location. This non-increasing function value is computed by the following steps:

– Find the maximum PCSM value (γ_m).
– Divide all the PCSM values by (γ_m). The result is a scaled form of the sample PCSM values within interval [0, 1].
– Subtract the dimensionless PCSM values from one at each distance.

2.4 Fuzzy c-means clustering based regression (FCMR) model

The model follows the least squares form given by (4). In the model, the observations are included in the clusters by their maximum memberships which are obtained from the clustering application, that is, this model does not need bandwidth values. The weights are provided by the SRDF values which are derived from the members of the clusters.

Methodologically, fuzzy clustering algorithms employ fuzzy partitioning such that a given data point can belong to several groups with the degree of belongingness

specified by membership grades between 0 and 1. Let $\{\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_N\}$ be a set of N data objects represented by a n-dimensional feature vectors $\mathbf{x}_k = [x_{1k}, \ldots, x_{nk}]^T \in R^n$. A fuzzy clustering algorithm partitions the data $\mathbf{X}$ into M fuzzy clusters, forming a fuzzy partition in $\mathbf{X}$. A fuzzy partition can be conveniently represented as a matrix $\mathbf{U}$, whose elements $u_{ik} \in [0, 1]$ represent the membership degree of $\mathbf{x}_k$ in cluster i. Hence, the ith row of $\mathbf{U}$ contains values of the ith membership function in the fuzzy partition.

Objective function based fuzzy clustering algorithms such as fuzzy c-means (FCM) algorithm (Bezdek et al. 1984) minimizes an objective function of the type:

$$J(X; U, V) = \sum_{i=1}^{M} \sum_{k=1}^{N} (u_{ik})^m d^2(\mathrm{x}_k, \mathbf{v}_i) \tag{7}$$

where $\mathrm{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_M]$, $\mathbf{v}_i \in R^n$ is M-tuple cluster prototypes (centers) which have to be computed, and $m \in (1, \infty)$ is a weighting exponent which defines the fuzziness of the clusters. There are some constraints for the algorithm as follows:

$$\sum_{i=1}^{M} u_{ik} = 1, \quad \forall k, \; 0 < \sum_{k=1}^{N} u_{ik} < N, \quad \forall i. \tag{8}$$

The general structure of the distance measure employed is given by

$$d^2(\mathbf{x}_k, \mathbf{v}_i) = (\mathbf{x}_k - \mathbf{v}_i)^T \mathbf{A}_i (\mathbf{x}_k - \mathbf{v}_i) \tag{9}$$

where the norm matrix $\mathbf{A}_i$ is a positive-define symmetric matrix. The FCM algorithm uses the Euclidian distance measure.

2.5 Extended fuzzy clustering based regression (EFCR) model

This model combines fuzzy clustering and local least squares regression for analyzing spatially varying data. The review of this hybrid structure discussed in Tutmez (2009, 2011) is presented as follows:

The main difference of the present model is the use of a new fuzzy clustering algorithm. In this study, we employ the extended fuzzy clustering (e-FCM) for the structure identification. The algorithm proposed by Kaymak and Setnes (2002). Extended fuzzy clustering approach has some potential advantages. First, the (point) prototypes in traditional fuzzy clustering are extended to hypervolumes whose size is determined automatically from the data. It covers a potential for using e-FCM in regional analysis. In addition, the merging approach offers a more automated and computationally less expensive way of determining the right partition (region).

Volume prototypes extend the cluster prototypes from points to regions in the clustering space (Krishnapuram and Kim 2000). The data points $\mathbf{x}_k$ that fall within the hypersphere, i.e. $d(\mathbf{x}_k, \mathbf{v}_i) \leq r_i$, are components of the volume prototype $\tilde{\mathbf{v}}_i$ and have by definition a membership of 1.0 in that particular cluster. The size of the volume prototypes is thus computed by the radius r_i.

The radii r_i, $i = 1, \ldots, M$ can be computed by considering the fuzzy cluster covariance matrix (Kaymak and Setnes 2000)

$$\mathbf{P}_i = \frac{\sum_{k=1}^{N} u_{ik}^m (\mathbf{x}_k - \mathbf{v}_i)(\mathbf{x}_k - \mathbf{v}_i)^T}{\sum_{k=1}^{N} u_{ik}^m}. \tag{10}$$

The volume of the cluster can be obtained from the determinant $|\mathbf{P}_i|$ of the cluster covariance matrix. The covariance matrix $\mathbf{P}_i$ is a positive definite and symmetric matrix and it can be decomposed such that $\mathbf{P}_i = \mathbf{Q}_i \Lambda_i \mathbf{Q}_i^T$, where $\mathbf{Q}_\mathrm{i}$ is orthonormal and $\mathbf{\Lambda}_i$ is diagonal with nonzero elements $\lambda_{i1}, \ldots, \lambda_{in}$. The volume prototypes extend a distance of $\sqrt{\lambda_{ij}}$, $j = 1, 2, \ldots, n$ along each eigenvector q_{ij}. For the multidimensional case, the size of radius in each direction is computed by measuring the distances along the transformed coordinates as follows:

$$\sqrt{\Lambda_i} Q_i^T \mathbf{A}_i \mathbf{Q}_i \sqrt{\mathbf{\Lambda}_i} \tag{11}$$

where $\sqrt{\mathbf{\Lambda}_i}$ represents a matrix whose elements are equal to the square root of the elements of $\mathbf{\Lambda}_i$. Applying (11) for the size of the cluster prototypes one obtains

$$\mathbf{R}_i = \sqrt{\mathbf{\Lambda}_i} \mathbf{Q}_i^T \mathbf{I} \mathbf{Q}_i \sqrt{\mathbf{\Lambda}_i} = \mathbf{\Lambda}_i. \tag{12}$$

As seen in Fig. 1, different values for the radius are provided depending on the direction one selects. Commonly a value between the maximal and minimal diagonal elements of $\mathbf{\Lambda}_i$is employed as the radius. In the E-FCM algorithm, the selection of the mean radius thus corresponds to following averaging operation (Kaymak and Setnes 2002):

$$r_i = \sqrt{\prod_{j=1}^{n} \lambda_{ij}^{1/n}} = \sqrt{|\mathbf{P}_i|^{1/n}}. \tag{13}$$

Consequently, this selection for the radius leads to a spherical prototype that preserves the volume of the cluster.

In addition to cluster radii, determining the number of clusters is other main motivation of the E-FCM algorithm. For this application, a cluster merging approach, which is based on similarity (Frigui and Krishnapuram 1996), has been proposed by Kaymak and Setnes (2000). In fuzzy clustering, the similarity of two fuzzy sets could be quantified by a fuzzy inclusion measure. Given two fuzzy

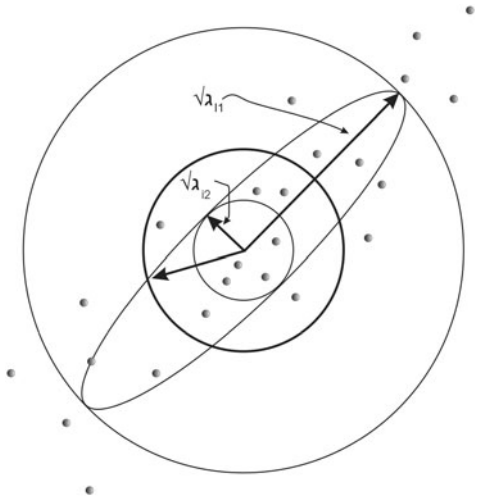

Fig. 1 The E-FCM radius and the cluster volume for a two dimensional case (After Kaymak and Setnes 2002)

clusters, $u_i(\mathbf{x}_k)$ and $u_j(\mathbf{x}_k)$, defined pointwise on $\mathbf{X}$, the fuzzy inclusion measure is defined as

$$I_{ij} = \frac{\sum_{k=1}^{N} \min(u_{ik}, u_{jk})}{\min(\sum_{k=1}^{N} u_{ik}, \sum_{k=1}^{N} u_{jk})}. \tag{14}$$

The inclusion measure denotes the ratio of the cardinality of the intersection of the two fuzzy sets to the cardinality of one of them. This parameter considers the contribution to similarity from all data points, both from those within the volume prototypes and those outside (Setnes 2001).

As an additional user-defined parameter, the merging threshold $\alpha \in [0, 1]$ is a critical parameter for the extended clustering algorithm. Kaymak and Setnes (2002) proposed an adaptive threshold that depends on the number of clusters in partition at any time:

$$\alpha^{(l)} = \frac{1}{M^{(l)} - 1}. \tag{15}$$

where M is the number of clusters. Clusters are merged when the change in maximum cluster similarity from iteration $(l-1)$ to iteration (l) is below a predefined threshold $\in_1$ and the similarity is above the threshold α.

2.6 Extended fuzzy clustering and regional dependence based regression (EFCRDR) model

From a methodological view, any approach for appraising spatial data needs to recognize that such data have the fundamental property of spatial dependence or spatial autocorrelation (Haining et al. 2010). The methodology presented in this section addresses a bridging between fuzzy clustering and geostatistics that is a distinctive methodology within the field of spatial statistics. The hybrid methodology has several features that distinguish it from the methodologies for analyzing spatial variation associated with regional data. It determines the regions (clusters) automatically from the data by computationally less expensive way. In addition, the estimated values of regression model are obtained from the spatial dependence measures which are the central part of geostatistical analysis.

The model first defines the structure of the system by extended fuzzy clustering and then builds a weighted regression model that uses the SRDF described. Determination of the regression weights from spatial analysis is the corner stone of the model.

In this method, each cluster represents a local linear model. The consequent parameter vectors $\boldsymbol{\theta}_\mathbf{i}$, $i = 1, 2, \ldots, c$, can be estimated independently by the least-squares method (Babuska 1998). The inputs of the regression model are given as follows:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_N^T \end{bmatrix}, \ \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \qquad \mathbf{W}_i = \begin{bmatrix} w_{i1} & 0 & 0 & 0 \\ 0 & w_{i2} & 0 & 0 \\ 0 & 0 & w_{i3} & 0 \\ 0 & 0 & 0 & w_{iN} \end{bmatrix}. \tag{16}$$

In the proposed model, the output parameters for the ith cluster, a_i and b_i are connected by a single parameter vector $\boldsymbol{\theta}_\mathbf{i}$ as follows:

$$\boldsymbol{\theta}_\mathbf{i} = \left[\mathbf{a}_i^T, b_i\right]^T. \tag{17}$$

Adding a unitary column to $\mathbf{X}$ gives the extended predictor matrix $\mathbf{X}_\mathbf{e}$:

$$\mathbf{X}_\mathbf{e} = [\mathbf{X}, \mathbf{1}]. \tag{18}$$

The SRDF values of the observations serve as the weights expressing the relevance of the data pair x_k, y_k to that local model. If the columns of $\mathbf{X}_e$ are linearly independent, then

$$\theta(x_k, y_k) = \left[\mathbf{X}_e^T \mathbf{W}(x_k, y_k) \mathbf{X}_e\right]^{-1} \mathbf{X}_e^T \mathbf{W}(x_k, y_k) \mathbf{y} \tag{19}$$

is the least-squares solution of $\mathbf{y} = \mathbf{X}_\mathbf{e}\boldsymbol{\theta} + \varepsilon$ where the kth data pair (x_k, y_k) is spatially weighted by w_{ik}. The parameters $\mathbf{a}_i$ and b_i are given by:

$$\mathbf{a}_i = [\theta_1, \theta_2, \ldots, \theta_p], \ b_i = \theta_{p+1}. \tag{20}$$

3 Case studies

3.1 Data set

The case studies have been conducted using Meuse data set (Rikken and van Rijn 1993; Burrough and McDonell 1998)

which is comprised of four heavy metals measured in the top soil in a flood plain along the river Meuse, near Stein in the Netherlands. The data set has a sample of 155 locations and top soil heavy metal concentrations (ppm), along with a number of soil and landscape variables. For the case studies, spatial coordinates and two heavy metals (cadmium, copper) content have been considered. For the applications, we have determined cadmium as the dependent (response) and copper as the independent (predictor) variable. These metals are a serious human health hazard MacBride (1994). Table 1 presents summary statistics of the data used in the case studies.

To provide some standard indices and avoid negative values, data conditioning is necessary for clustering applications (Jain and Dubes 1998). In the present study, scaling was carried out by the local metric (L-metric) rescaling, in which the minimum and maximum values of x_{ij} for each j are respectively mapped onto zero and one respectively,

$$x_{ij}^{L} = \frac{x_{ij} - \min_j(x_{ij})}{\max_j(x_{ij}) - \min_j(x_{ij})} \tag{21}$$

3.2 Case study 1: GWR model

The GWR analyses have been carried out by a fixed spatial kernel with the Gaussian function. Various approaches have been used for ascertaining an optimal bandwidth. The following modified Akaike Information Criterion (AIC) (Fotheringham et al. (2002), which obtains a trade-off between goodness-of-fit and degrees of freedom, was used to provide bandwidth:

$$AIC_c = 2n\log_e(\hat{\sigma}) + n\log_e(2\Pi) + n\left\{\frac{n + tr(\mathbf{S})}{n - 2 - tr(\mathbf{S})}\right\} \tag{22}$$

where n is the sample size, $\hat{\sigma}$ is the estimated standard deviation of the error term, and tr($\mathbf{S}$) denotes the trace of the hat matrix $\mathbf{S}$ which maps $\hat{y}$ on to y (i.e., $(\hat{y} = \mathbf{S}\mathbf{y})$)

The GWR model was fitted using *R* routines. In addition, the fixed bandwidth value was determined as 1.41 from AIC using the '*spgwr*' package in *R* (Bivand et al. 2008). By using the Gaussian spatial function, coefficient of determination (CoD) has been computed as 0.858.

Table 1 Summary statistics for variables

	Copper	Cadmium
Min.	14.0	0.2
Median	31.0	2.1
Mean	40.3	3.2
Max.	128.0	18.1

Table 2 Spatial measure for location no. 1

Location	Distance	Cadmium	PCSM	Distance ratio	SRDF Weighting
1	0.000	0.642	0.000	0.000	1.000
2	0.022	0.469	0.086	0.017	0.998
3	0.038	0.352	0.231	0.030	0.994
8	0.066	0.145	0.480	0.052	0.987
:					:
:					:
148	1.272	0.162	37.442	1.000	0.000

3.3 Case study 2: SRDFR model

In the second case study, the weight matrix is obtained from the spatial relationships between the variables. For this purpose, the spatial variability is modelled by point semimadogram function. Functional analyses were carried out based on the distance measures between pivot locations and other locations.

The weights for different locations depending on the distance from the pivot location have been calculated by the SRDF. Using this algorithm, for example, the SRDF analysis can be conducted for the location no: 1, the other location weightings with respect to the pivot location (no: 1) can be obtained easily. The last column in Table 2 includes the SRDF weighting which can also be taken from the graph in Fig. 2. The structure in Fig. 2 explains the spatial dependence in terms of distance. The closest location to the pivot contributes the highest weight, and the furthest ones relatively contribute the least weights.

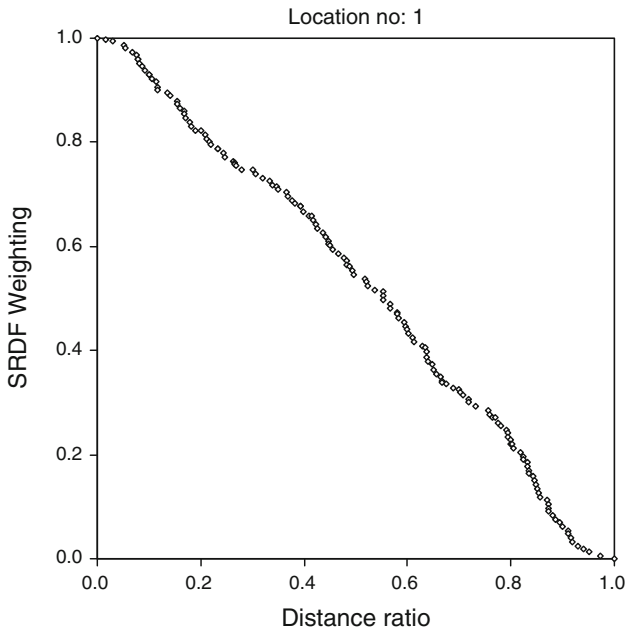

Fig. 2 SRDF graph for station no. 1

The SRDF values have been used in the least squares system given in (4) for the geographical weighting $\mathbf{W}(u_i, v_i)$ of the observations. In this manner, the SRDFR method provided an alternative solution by determining the weights from a spatial function.

3.4 Case study 3: FCMR model

In the FCMR model, the weights of the least squares system have been obtained from the SRDF directly. In this model, first a FCM clustering application has been conducted and then by using the belongingness with maximum memberships, the members of the clusters have been determined. Because the local estimations have been carried out by all members of the clusters, the system has not required any bandwidth value.

The optimal number of clusters was defined experimentally using Xie–Beni index (Xie and Beni 1991) which is a well-known clustering validity method. For this application, the appropriate numbers of clusters was determined as five. The cluster centres are depicted in Fig. 3.

In the following step, the SRDF values have been calculated for each location using the observations within the related cluster. The SRDF weights have been employed in the least squares system and local estimations have been carried out.

3.5 Case study 4: EFCR model

For this model, the extended fuzzy clustering, application was implemented. The details of the application procedure

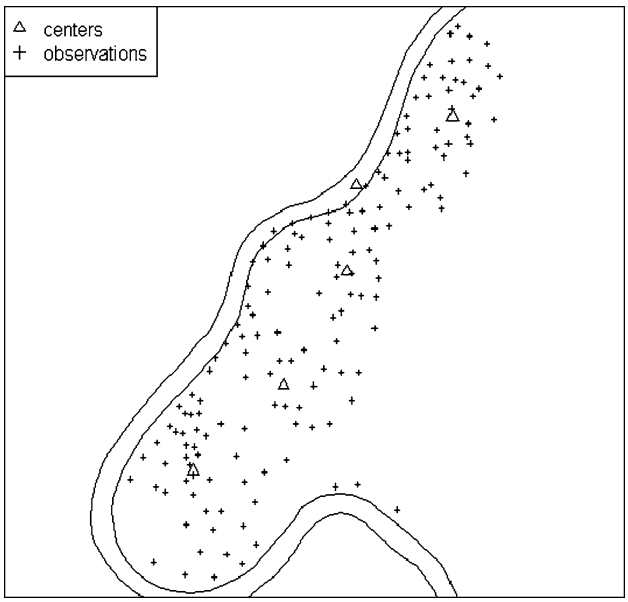

Fig. 3 Meuse data and cluster centers for FCMR model

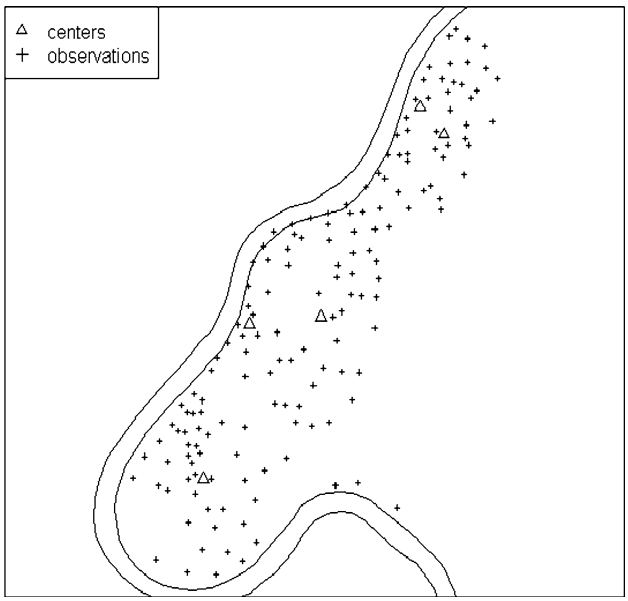

Fig. 4 Meuse data and cluster centers for EFCR model

were introduced in Tutmez (2009, 2011). The optimal number of clusters was determined experimentally using a similarity-based cluster merging approach and an adaptive threshold. The threshold $\alpha \in [0, 1]$ deals with some characteristics such as cluster size and the clustering parameters, like the fuzziness m. The fuzziness parameter m is selected as 1.6. As a result of the application, the appropriate numbers of clusters was found as five. These centres are indicated in Fig. 4. Note that there is a difference between the positions of the cluster centers of FCMR and EFCR model that has been reasoned from the preferred cluster validity index.

By the clustering application, the fuzzy cluster covariance matrix and volume prototypes have been computed to determine the bandwidth (radius) values. In addition, by using the information obtained from the clustering, Gaussian type membership functions have been established. The function values (membership values) have been employed in the least square systems as the weights. Figure 5 shows the input memberships considered in the model.

In the model, the bandwidths of the Gaussian functions have had different values for each cluster as [0.066; 0.086; 0.122; 0.097; 0.085]. In addition, the CoD has been determined as 0.858.

3.6 Case study 5: EFCRDR model

The EFCRDR model enabled to make a connection between fuzzy clustering and spatial data. The EFCRDR model first described the structure of the system by extended fuzzy clustering and then established a weighted

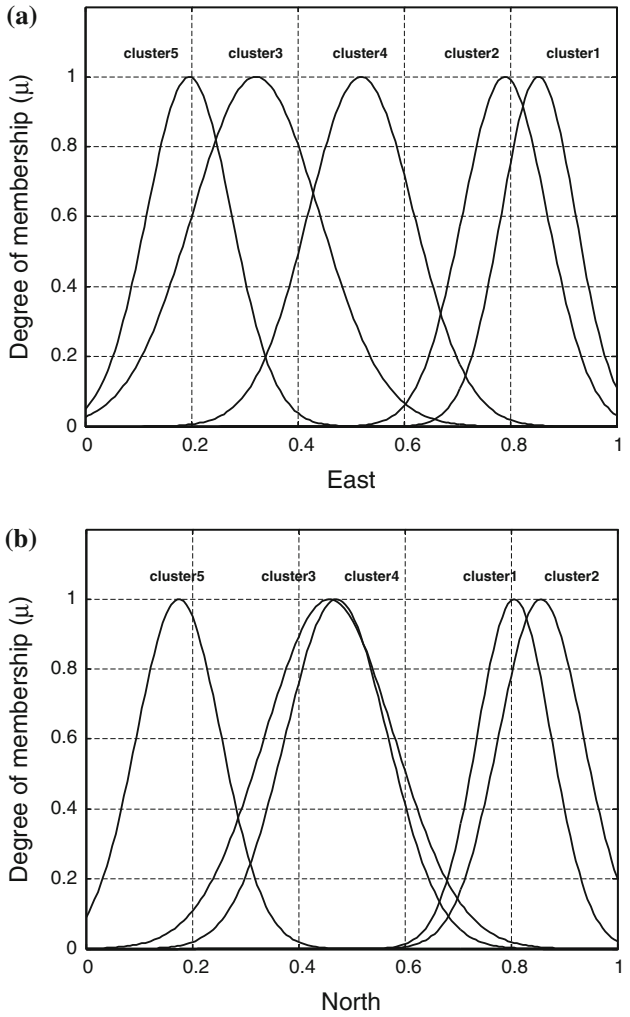

Fig. 5 Input membership functions for EFCR model

regression model via the SRDF values. By using the same parameters given in case study 4, the appropriate numbers of clusters was determined as five. As stated before, the main difference of the EFCRDR model from the EFCR model is determination of the local regression weights.

In the EFCRDR model, the SRDF values of the observations have served as the weights expressing the relevance of the location that local model. To obtain these weights some spatial measures have been done. Figure 6a gives a measure that has been performed for a sample location (no: 70). By using the information taken from the spatial measure, the following steps were taken and the SRDF graph produced is given in Fig. 6b.

- Maximum PCSM value (γ_m) has been determined.
- All the PCSM values is divided by (γ_m). In this manner, the result was a scaled form of the sample PCSM values within limits of zero and one.
- Finally, the dimensionless PCSM values are subtracted from one at each distance.

4 Results and discussions

To evaluate the performance of the local models provided by the case studies, we have plotted the estimated cadmium concentrations against the measured (actual) concentrations. Figure 7 illustrates the results of the models together with the cross-correlations between estimated and measured values. The large determination coefficient (r^2) shows that the model has good estimation capability. In addition, the model performances have been indicated by three additional effective indices which are variance account for (VAF), root mean square error (RMSE) and standard deviation (Std) as in Table 3.

As seen in Fig. 7, the EFCRDR model performed best. In addition to the accuracy, the EFCRDR model has not

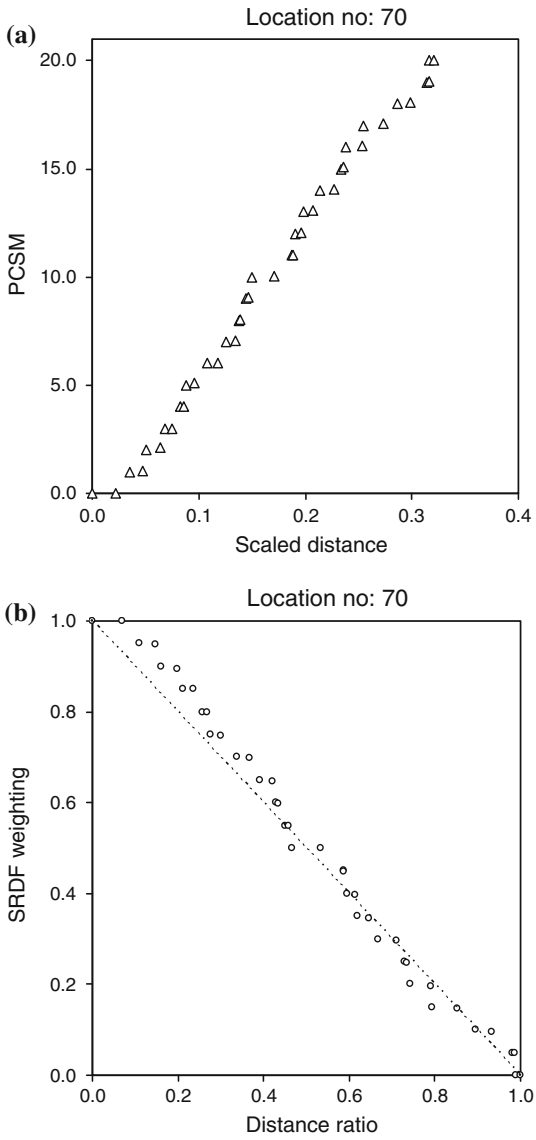

Fig. 6 Spatial measures for location no 70. **a** Experimental PCSM and **b** SRDF graph

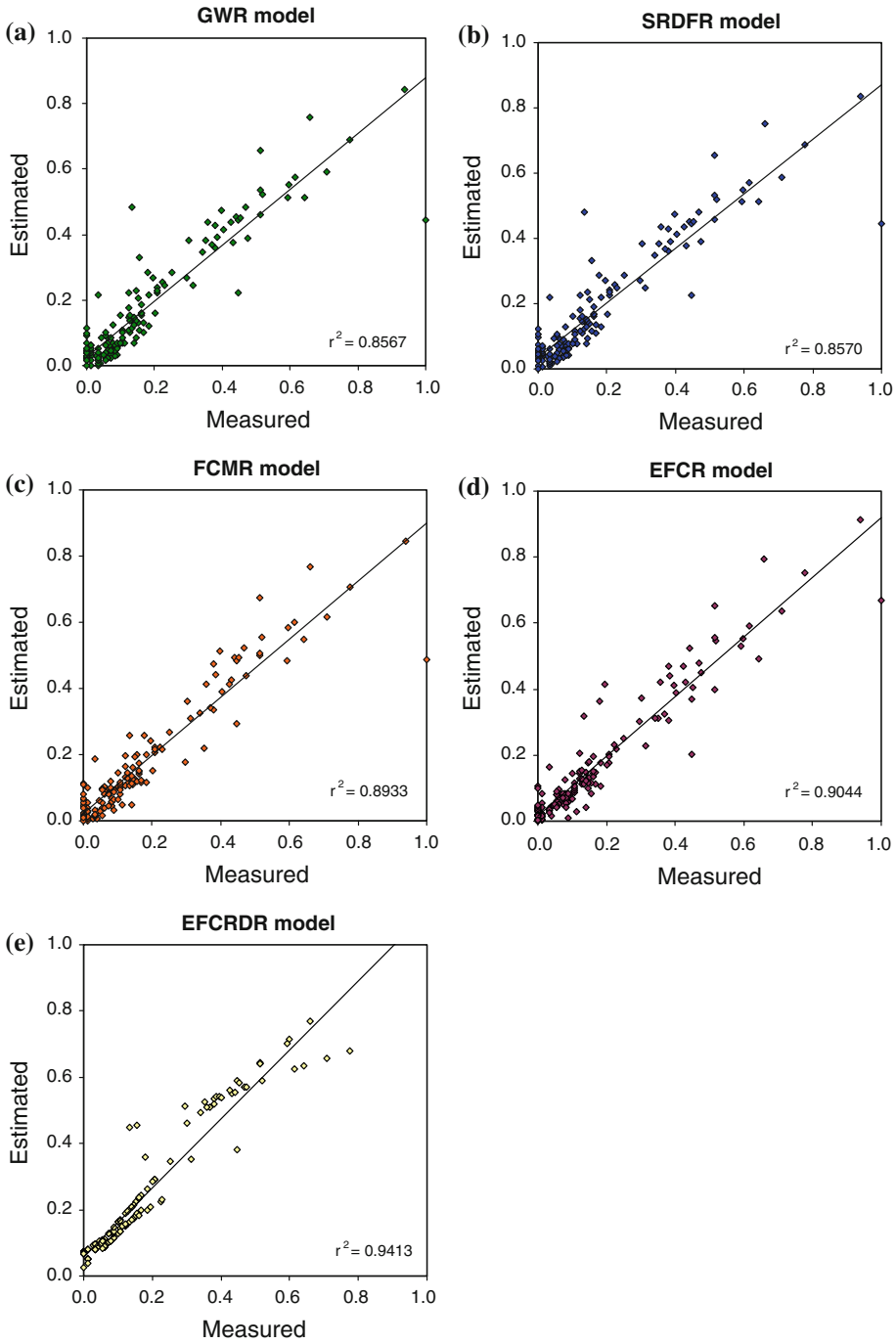

Fig. 7 Scatter plots for the models **a** GWR model, **b** SRDFR model, **c** FCMR model, **d** EFCR model and **e** EFCRDR model

produced any outlier values. On the other hand, the EFCRDR model yields some over estimations. This finding clearly indicates an ensemble of the local models, one of which is consistently biased and others provide mis-specified slope parameters but otherwise surprisingly good linear fit. This could be resourced from the identification of the SRDF. Figure 8 presents the estimation errors for both the EFCR and the EFCRDR model. Based on the error variability indicated in Fig. 8, the over estimation can be said to be a disadvantage of the EFCRDR model. This issue

Table 3 Performance measures of models

Models	VAF	RMSE	Std
Measured	–	–	0.197
GWR	85.67	0.074	0.181
SRDFR	85.66	0.074	0.178
FCMR	89.29	0.065	0.183
EFCR	90.45	0.062	0.187
EFCRDR	93.22	0.077	0.209

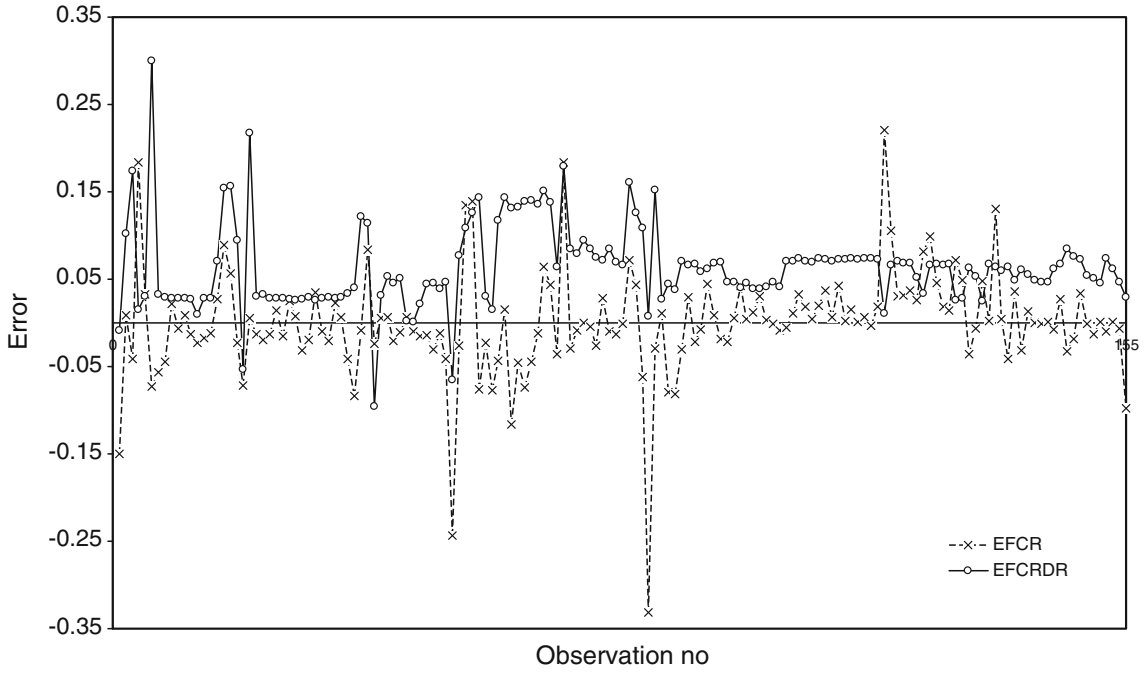

Fig. 8 Error analysis for EFCR and EFCRDR models

also can be seen with relatively high RMSE and Std values as presented in Table 3.

Note that the data set has a sample of 155 locations and is limited in number. The difference between the performances of the models could be easily observed by some extended data sets. Another important point is that the fuzzy clustering based models have used only five clusters for the interpolations. The number of clusters can be increased depending on an upper estimate in the fuzzy clustering based algorithms. However, in practice, there can be a reverse relationship between the accuracy and generality of the models. This point should be considered in structure identification.

In addition to analysis of relationships between the variables, spatially varying coefficients are also important for geographically referenced data analysis. In the present study, instead of the coefficient based GWR analysis, the structure has been handled as a prediction model. However, Griffith (2008) showed that maps of the GWR coefficients tend to exhibit multicollinearity as well as strong positive autocorrelation.

5 Conclusions

In the presented study, a comparative experimental study has been presented to discuss the spatially varying relationships. For this purpose, five local regression models were developed and the performances of the models have been compared by using a geographically referenced data set.

The experimental studies showed that models based on the extended fuzzy clustering provide the best estimations. In particular, the hybrid model developed by combining extended fuzzy clustering and regional dependence based regression model (EFCRDR) outperforms the other regression models. The EFCRDR model firstly divided the area in different subareas (clusters) and then analysed the relationships by some regional dependence functions. In addition to the EFCRDR model, the EFCR model which is a combination of extended fuzzy clustering and conventional LSE, has produced some successful results such as lower error and better variation on predicted values.

In contrast to existing estimation strategies, the hybrid models presented in this paper integrate soft computing and geostatistics and this view covers a potential to enhance the robustness and generalisability of the models for appraising real-world spatial environmental problems efficiently and intuitively. As a consequence, it can be expressed that combining soft computing and spatial statistics provides a novel methodological perspective for local spatial data analysis.

Acknowledgments This research was supported by the Scientific and Technological Research Council of Turkey (TUBITAK Project: 108M393) and COST (European Cooperation in Science and Technology) Action IC0702.

References

Atkinson PM, Naser DK (2010) A geostatistically weighted k-NN classifier for remotely sensed imagery. Geogr Anal 42:204–225

Babuska R (1998) Fuzzy modeling for control. Kluwer, USA

Bezdek JC, Ehrlich R, Full W (1984) FCM: the fuzzy c-means clustering algorithm. Comput Geosci 10:191–203

Bivand RS, Pebesma EJ, Gomez-Rubio V (2008) Applied spatial data analysis with R. Springer, New York

Bogardi I, Bárdossy A, Duckstein L, Pongracz P (2003) Fuzzy logic in hydrology and water resources. In: Demicco RV, Klir GJ (eds) Fuzzy logic in geology. Elsevier, San Diego, pp 153–190

Burrough PA, McDonell RA (1998) Principles of geographical information systems. Oxford University Press, Oxford

Fotheringham AS, Charlton ME, Brunsdon C (1998) Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis. Environ Plan A 30:1905–1927

Fotheringham AS, Brunsdon C, Charlton ME (2002) Geographically weighted regression: the analysis of spatially varying relationships. Wiley, Chichester

Frigui H, Krishnapuram R (1996) A robust algorithm for automatic extraction of an unknown number of clusters from noisy data. Pattern Recognit Lett 17:1223–1232

Gao X, Asami Y, Chung YCF (2006) An empirical evaluation of spatial regression models. Comput Geosci 32:1040–1051

Goovaerts P (1997) Geostatistics for natural resources evaluation. Oxford University Press, New York

Griffith DA (2008) Spatial-filtering-based contributions to a critique of geographically weighted regression (GWR). Environ Plan A 40:2751–2769

Haining RP, Kerry R, Oliver MA (2010) Geography, spatial data analysis, and geostatistics: an overview. Geogr Anal 42:7–31

Harris P, Brunsdon C, Fotheringham AS (2011) Links, comparisons and extensions of the geographically weighted regression model when used as a spatial predictor. Stoch Environ Res Risk Assess 25(2):123–138

Jain A, Dubes R (1988) Algorithms for clustering data. Prentice Hall, Englewood Cliffs

Kanevski M, Pozdnoukhov A, Timonin V (2009) Machine learning for spatio-environmental data: theory, applications and software. EPFL Press, Boca Raton

Kaymak U, Setnes M (2000) Extended fuzzy clustering algorithms. ERIM reports, ERS-2000-51-LS, Rotterdam

Kaymak U, Setnes M (2002) Fuzzy clustering with volume prototypes and adaptive cluster merging. IEEE Trans Fuzzy Syst 10(6): 705–711

Krishnapuram R, Kim J (2000) Clustering algorithms based on volume criteria. IEEE Trans Fuzzy Syst 8:228–236

Lee ES (2000) Neuro-fuzzy estimation in spatial statistics. J Math Anal Appl 249:221–231

Lloyd CD (2006) Local models for spatial analysis. CRC Press, Boca Raton

Loader C (1999) Local regression and likelihood. Springer, New York

MacBride MB (1994) Environmental chemistry of soils. Oxford University Press, New York

Paez A, Uchida T, Miyamoto K (2002) A general framework for estimation and inference of geographically weighted regression models: 1. Location-specific kernel bandwidths and a test for locational heterogeneity. Environ Plan A 34:733–754

Rikken MGJ, van Rijn RPG (1993) Soil Pollution with heavy metals-an inquiry into spatial variation, cost mapping and the risk evaluation of copper, cadmium, lead and zinc in the floodplains of the Meuse West of Stein, the Netherlands. Doctoral Thesis, Utrecht University, the Netherlands

Şen Z (2009) Spatial modeling principles in earth sciences. Springer, Heidelberg

Şen Z, Şahin AD (2001) Spatial interpolation and estimation of solar irradiation by cumulative semivariograms. Sol Energy 71:11–21

Setnes, M. (2001) Complexity reduction in fuzzy systems, PhD Thesis, Delft University of Technology, the Netherlands

Sousa JMC, Kaymak U (2002) Fuzzy decision making in modelling and control. World Scientific, Singapore

Tutmez B, Hatipoglu Z (2007) Spatial estimation model of porosity. Comput Geosci 33:465–475

Tutmez B, Tercan AE (2007) Assessment of uncertainty in geological sites based on data clustering and conditional probabilities. J Uncertain Syst 1(3):206–221

Tutmez B, Tercan AE, Kaymak U (2007) Fuzzy modeling for reserve estimation based on spatial variability. Math Geol 39(1):87–111

Tutmez, B, Tercan AE, Kaymak U, Lloyd CD (2009) Local models for the analysis of spatially varying relationships in a lignite deposit. IFSA-EUSFLAT 2009, Lisbon, pp 351–356

Tutmez, B, Tercan AE, Kaymak U, Lloyd CD (2011) Evaluating spatial relationships between ecological variables using a clustering based areal model. Ecol Inform (submitted)

Waller LA, Gotway CA (2004) Applied spatial statistics for public health data. Wiley, Hoboken

Waller LA, Zhu L, Gotway CA, Gorman DM, Gruenevald PJ (2007) Quantifying geographic variations in associations between alcohol distribution and violence: a comparison of geographically weighted regression and spatially varying coefficient models. Stoch Environ Res Risk Assess 21(5):573–588

Wong P, Aminzadeh F, Nikravesh M (2001) Soft computing for reservoir characterization and modeling. Physica-Verlag, Heidelberg

Xie XL, Beni GA (1991) A validity measure for fuzzy clustering. IEEE Trans Pattern Anal Mach Intell 13(8):841–847