Local models for the analysis of spatially varying relationships in a lignite deposit

Bulent Tutmez[1] A. Erhan Tercan[2] Uzay Kaymak[3] Christopher D. Lloyd[4]

1 School of Engineering, Inonu University, Malatya 44280, Turkey
2 School of Engineering, Hacettepe University, Ankara 06532, Turkey
3 Econometric Institute, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR, Rotterdam, the Netherlands
4 School of Geography, Archaeology and Palaeoecology, Queen's University, Belfast, Northern Ireland, UK
Email: btutmez@inonu.edu.tr

Abstract *— Relationships between geographically referenced variables are usually spatially heterogeneous and, to account for such variations, local models are necessary. This paper compares the Geographically Weighted Regression (GWR) model, usually used to integrate and examine the spatial heterogeneity of a relationship, and the Fuzzy Clustering-Based Least Squares (FCBLS) model for the analysis of spatially varying relationships. Both models use the same model parameters and bandwidth values derived from the Akaike Information Criterion. The results show that FCBLS outperforms the GWR model.*

Keywords*— GWR, local model, weighted regression, fuzzy clustering, lignite.*

1 Introduction

Spatial measures contain both attribute and locational information. By their nature, local analyses focus on differences across space whereas global analyses focus on similarities across space. Recently, a variety of useful regression models have been developed to explore the spatial nature of variables [1].

Local modelling has been employed widely in some disciplines for several decades. For example, in image processing local filters have long been used to smooth or sharpen images. However, in some disciplines, like the geosciences in general and geography in particular, a focus on methods that account for local variation has been a comparatively recent development [2]. A variety of localized modelling techniques have been proposed to capture spatial heterogeneity. Geographically Weighted Regression (GWR) is one technique which is being increasingly widely used to explore local spatial variations in relationships, [3]. GWR is a useful and effective technique for locally modelling relationships by calibrating a spatially varying coefficient regression model.

Spatial statistics was developed based on probability and classical statistics. However, many spatial datasets have high levels of uncertainty, and in some cases analyses are dependent on 'soft' data, which may be more qualitative than quantitative in nature [4]. On the other hand, soft approaches like fuzzy computing have desirable features for spatial data analysis [5]: they are based on less restrictive assumptions, and are flexible in modelling non-linearity and non-constant variable structures.

Lignite quality parameters considered in this study have crucial importance in the production of energy in power plants and thus modeling of these parameters is useful in making decisions and in planning future production levels of operating plants [6]. In this study, the spatial relationships between the lignite quality parameters are investigated by local models. To meet this aim, the two local regression methods, GWR and fuzzy clustering based Least Squares, have been applied for local modelling of lignite reserve parameters derived from a deposit. First, both the methods have been used for developing weighted local models and then the performance comparisons have been conducted using a variety of performance indices.

2 Problem and methods

2.1 Description of problem

In general, spatial models and methods of spatial evaluation have been applied at a 'global' level, meaning 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 [7]. Although global approaches have proved useful they have the shortcoming that they can mask geographical variations in relationships. Owing to this realization, local regression model approaches have been developed that permit the exploration of spatially varying relationships in datasets [8].

The problem of local estimation considered in this study can be formulated as follows: we have a region and lignite quality values $\{y_\alpha\}, \alpha = 1,..., n,$ at n sampled locations. We aim to estimate the values $\{y_\beta\}, \beta = 1,..., N$ at unsampled locations. To model a system by a spatially weighted method, the identification data and the weights could be arranged as in the following matrices:

ISBN: 978-989-95079-6-8

$$\mathbf{X} = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{bmatrix}, \ \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \ \mathbf{W}_i = \begin{bmatrix} w_{i1} & 0 & 0 & 0 \\ 0 & w_{i2} & 0 & 0 \\ 0 & 0 & w_{i3} & 0 \\ 0 & 0 & 0 & w_{iN} \end{bmatrix} \quad (1)$$

2.2 Geographically Weighted Regression (GWR)

An assumption of global regression is that the relationship under study is spatially constant, and thus, the relationships being measured are assumed to be ‘stationary’ over space. However, in most cases, the relationship varies in space. GWR, as a refinement to traditional regression methods, explicitly deals with the spatial non-stationary of empirical relationships. The method provides a weighting of information that is locally specific, and allows regression model parameters to vary in space [3].

Classical regression equation in matrix form can be given by:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon \quad (2)$$

where the vector of parameters to be estimated, β, is constant over space and is estimated by

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-\mathbf{1}}\mathbf{X}^{\mathbf{T}}\mathbf{Y}\,. \quad (3)$$

The local estimation of the parameters with GWR is as given by (3), but the difference is that the observations used in GWR are weighted in accordance with their distance from the kernel centre. Fig. 1 indicates a spatial kernel in GWR. The parameters for GWR may be estimated by solving

$$\hat{\beta}(u_i, v_i) = \left[\mathbf{X}^T\mathbf{W}(u_i, v_{i,})\mathbf{X}\right]^{-1}\mathbf{X}^T\mathbf{W}(u_i, v_i)\mathbf{Y}\,, \quad (4)$$

where $\hat{\beta}$ 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 whose diagonal elements denote the geographical weighting of each of the *n* observed data for regression point *i* [9]. There are many weighting schemes which express w_{ij} as a continuous function of distance between *i* and *j*, d_{ij}. In this study, the following Gaussian function has been used.

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

where d_{ij} is the Euclidean distance between the location of measurement *i* and the centre of the kernel and *b* is the bandwidth of the kernel.

2.3 Fuzzy Clustering-Based Least Squares

A class of fuzzy clustering algorithms can be used to approximate a set of data by local linear models. Each of these models is represented by a fuzzy subset in the data set available for identification [10]. Most analytical fuzzy clustering algorithms are based on optimization of the basic c-means objective function, or some modification of it.

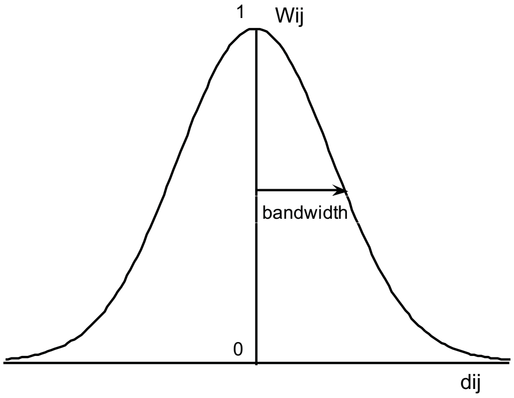


Figure 1: A Gaussian kernel in GWR.

In the present study, we used a general structure which is the fuzzy c-means functional [11] for constructing the weighted least squares model from the fuzzy partitions.

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

$$\boldsymbol{\theta}_{\mathbf{i}} = \left[\mathbf{a}_i^T, b_i\right]^T. \quad (6)$$

Appending a unitary column to **X** gives the extended regressor matrix $\mathbf{X_e}$:

$$\mathbf{X_e} = [\mathbf{X,1}]. \quad (7)$$

Assuming that each cluster represents a local linear model of the system, the consequent parameter vectors $\boldsymbol{\theta}_{\mathbf{i}}$, $i = 1,2,...,c$, can be estimated independently by the least-squares method. The membership degrees of the fuzzy partition serve as the weights expressing the relevance of the data pair x_k, y_k to that local model [10]. If the columns of $\mathbf{X}_e$ are linearly independent, then

$$\boldsymbol{\theta}_i = \left[\mathbf{X}_e^T\mathbf{W}_i\mathbf{X}_e\right]^{-1}\mathbf{X}_e^T\mathbf{W}_i\mathbf{y} \quad (8)$$

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

$$\mathbf{a}_i = \lfloor\theta_1, \theta_2, ..., \theta_p\rfloor, \ b_i = \theta_{p+1}\,. \quad (9)$$

3 Implementation

3.1 Data set

The study area, the Sivas-Kalburcayiri field, is one of the most important lignite reserves in Turkey [6]. Lignite seams in this field are used to feed coal to a power plant. In this study, one of the lignite quality parameters, Calorific Value (CV), has been estimated from the independent parameters which are spatial coordinates (*x*,*y*), ash and sulphur contents.

The locations of the 67 records of the field have been employed in the model. The data set was divided into two subsets randomly: the training set (60%: 40 records) and the validation set (40%:27 records), respectively. For the analyses, data conditioning is necessary. 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 [7],

$$x_{ij}^{L} = \frac{x_{ij} - \min_j(x_{ij})}{\max_j(x_{ij}) - \min_j(x_{ij})}\ . \qquad (10)$$

3.2. Local Models

GWR model

For the local analysis of the lignite parameters, GWR analyses were undertaken using a fixed spatial kernel with Gaussian function in which the bandwidth was determined by minimization of the Akaike Information Criterion (AIC) [12].

In the literature, various approaches have been used for ascertaining an optimal bandwidth. A method of deriving the bandwidth which provides a trade-off between goodness-of-fit and degrees of freedom is to minimize the AIC [12], used in this study. In [9], the AIC has been modified for GWR as follows:

$$AIC_c = 2n\log_e\left(\hat{\sigma}\right) + n\log_e\left(2\prod\right) + n\left\{\frac{n + tr(\mathbf{S})}{n - 2 - tr(\mathbf{S})}\right\} \qquad (11)$$

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

In this application, the GWR model was fitted using a computer software program, *SAM 3.0.* [13]. In addition, the fixed bandwidth value has been determined from AIC using '*spgwr*' package in *R* [14]. Table 1 summarizes the model fitting statistics for training data. Similarly, the test statistics which were obtained from the fixed bandwidth are given in Table 2. In these tables, r^2 relates to observations against their estimates and OLS denotes the Ordinary Least Squares optimization.

Table 1. GWR fitting statistics for training data.

Spatial function:	Gaussian
Fixed Bandwidth (h):	0.191 distance units.
Number of Locations to Fit Model (n):	40
Coefficient of Determination (r^2):	0.811 (OLS: 0.785)

Table 2. GWR fitting statistics for testing data.

Spatial function:	Gaussian
Fixed Bandwidth (h):	0.191 distance units.
Number of Locations to Fit Model (n):	27
Coefficient of Determination (r^2):	0.718 (OLS: 0.655)

FCBLS model

To analyse the spatial relationships by clustering based least-squared optimization, firstly a clustering operation was conducted. The optimal number of clusters was determined experimentally using an index method which has been presented in [6]. Because the data variability is crucial in spatial data modelling, the method aims to reproduce variability of the sample data in the CV of cluster centres with a minimum number of clusters as follows:

$$\text{Minimize } n_c \quad under \quad Std\left[CV(x)\right] \approx Std\left[CV(c)\right] \qquad (12)$$

where n_c is the optimal number of clusters, *Std* is the standard deviation of CV values. In this application, the appropriate numbers of clusters was determined as two. The cluster centres are depicted in Fig. 2.

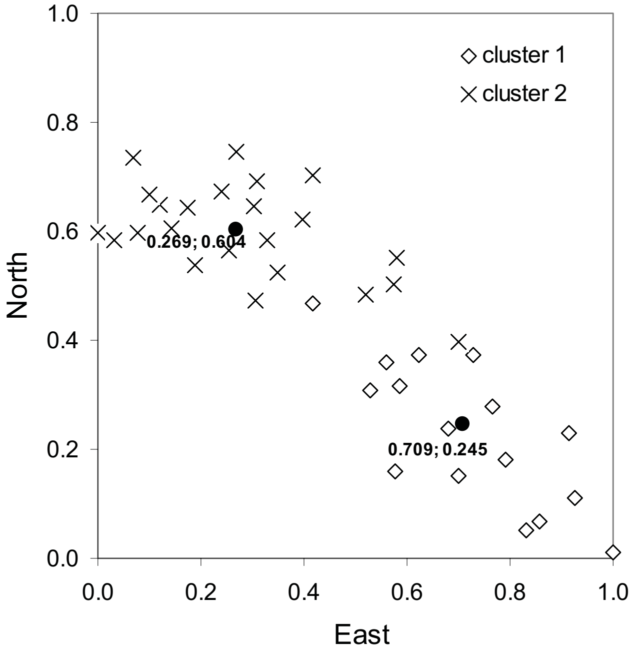


Figure 2: Data and cluster centres.

By using the information taken from the clustering, Gaussian type membership (weighting) functions were adopted. Fig.3 shows the input memberships considered in the model. To optimize the least square systems, weights taken from the Gaussian memberships have been used. Note that the Gaussian functions developed in the model used the same bandwidths derived from AIC. The consequent parameters are summarized in Table 3.

Table 3: Consequent parameters for FCBLS model.

Cluster	*Ash*	*Sulphur*	*Constant*
Cluster 1	-0.045	0.957	0.031
Cluster 2	-0.403	0.532	0.0247

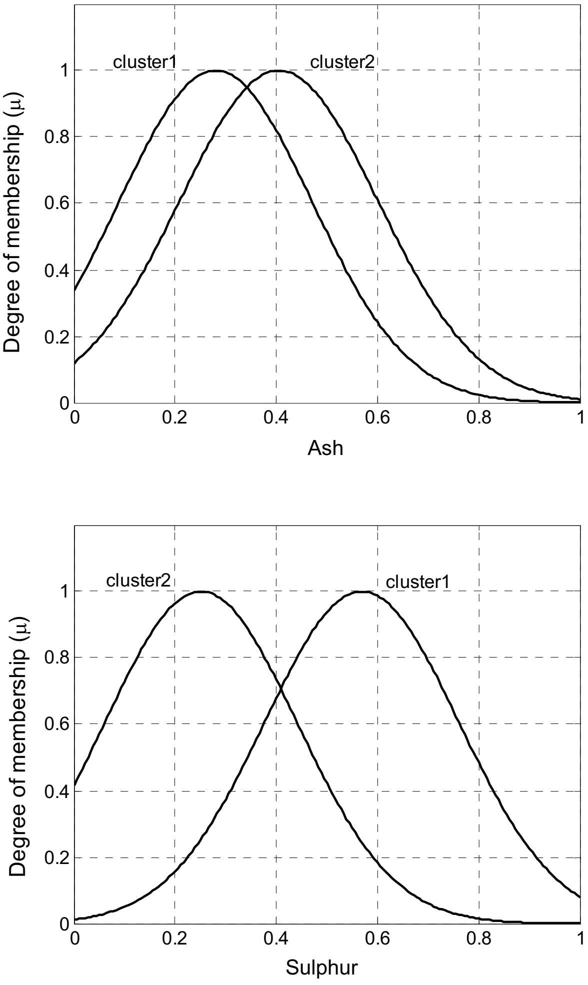
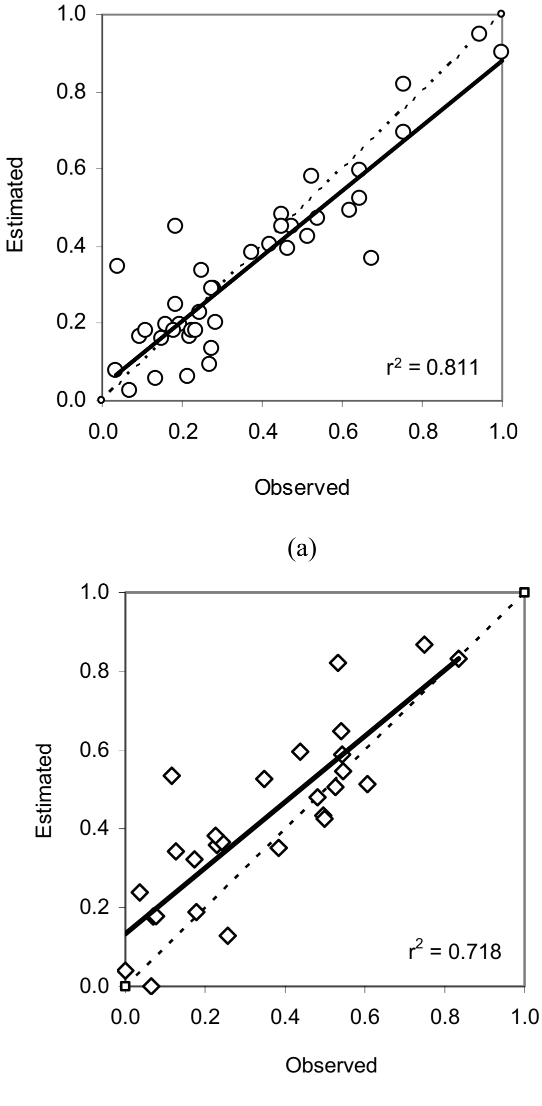


Figure 3: Gaussian memberships.

4 Results and conclusions

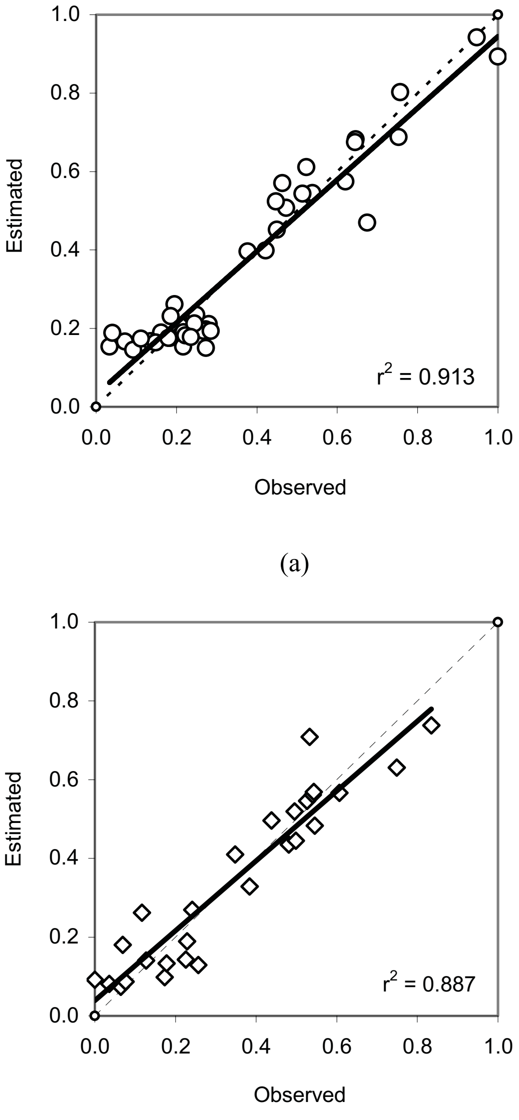
To assess the performance of the local models detailed in section 3.2, we have plotted the observed calorific values against the estimated calorific values. Fig. 4 illustrates the GWR model results together with the cross-correlations between estimated and observed data both for the training set and the validation set, respectively. Similarly, the results of the FCBLS model are depicted in Fig. 5.



(b)

Figure 4: Scatter plot for GWR model: (a) training data and (b) testing data.

The large determination coefficient (r^2) shows that the model has good estimation capability. As observed from the determination coefficient, the FCBLS method outperforms the method based on GWR. In addition to r^2, the performance of the models has been assessed using the performance index, namely, the Variance Accounted For (VAF). In multivariate analysis, the measure of VAF plays a central role. Table 4 gives the VAF values.



(a)

(b)

Figure 5: Scatter plot for FCBLS model: (a) training data and (b) testing data.

Table 4: VAF measures for the models.

Model	Training	Testing
GWR	82.06	72.57
FCBLS	91.30	88.73

In addition to performance indices, the estimates can be presented in map form. Figs. 6-8 show the observed values, GWR estimations and FCBLS estimations, respectively. Based on the performance evaluations and the maps, it can be argued that the fuzzy clustering based regression model (FCBLS) outperforms the geographically weighted regression model in this case.

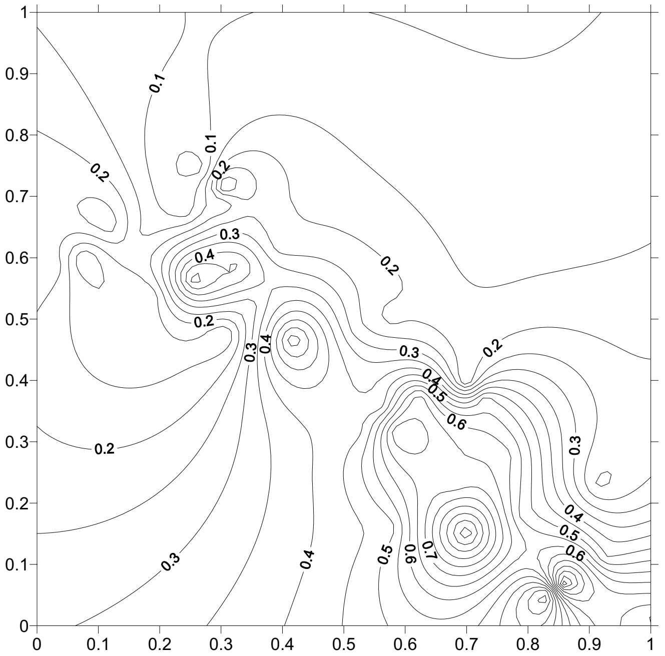


Figure 6: Contour map for observed data.

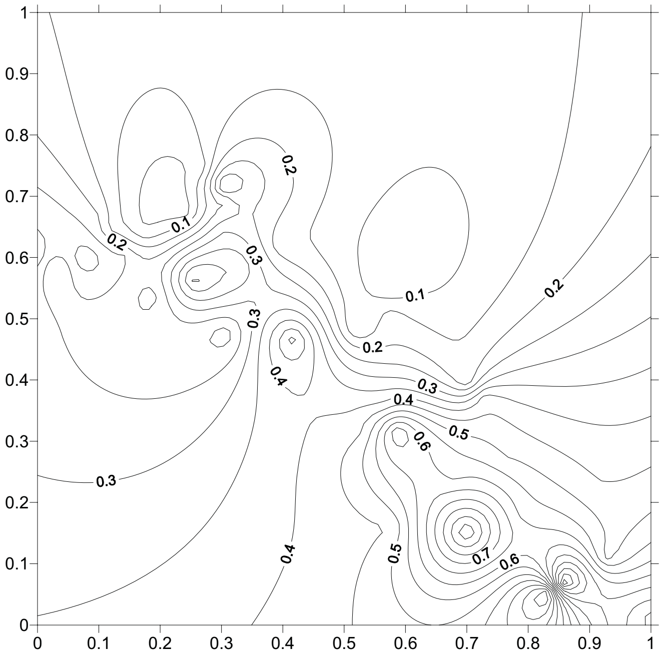


Figure 7: Contour map for the GWR model.

ISBN: 978-989-95079-6-8

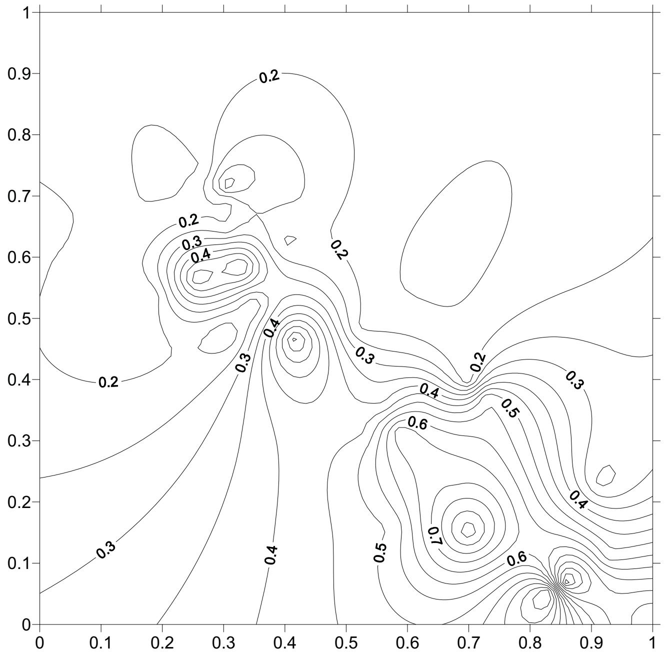


Figure 8: Contour map for FCBLS model.

Acknowledgements

This paper is part of the project (108M393) which is granted by Scientific and Technological Research Council of Turkey (TUBITAK).

References

[1] X. Gao, Y. Asami, CJ. F. Chung. An empirical evaluation of spatial regression models, *Computers &Geosciences*, 32:1040-1051,2006.

[2] C.D. Lloyd, *Local Models for Spatial Analysis*, Boca Raton: CRC Press, 2006.

[3] A.S. Fotheringham, M.E. Charlton, C. Brunsdon, Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis, *Environment and Planning A*, 30: 1905-1927, 1998.

[4] E. S. Lee, Neuro-Fuzzy estimation in spatial statistics, *J Mathematical Analysis and Applications*, 249:221-231, 2000.

[5] P. Wong, F. Aminzadeh, M. Nikravesh, *Soft Computing for Reservoir Characterization and Modeling*, Heidelberg:Physica-Verlag, 2001.

[6] B. Tutmez, A.E. Tercan, U. Kaymak. Fuzzy modeling for reserve estimation based on spatial variability, *Mathematical Geology*, 39(1):87-111, 2007.

[7] A.S. Fotheringham, C. Brunsdon, M.E. Charlton, *Qualitative Geography: Perspectives on Spatial Data Analysis*, London: SAGE Publications, 2000.

[8] C.D. Lloyd and I. Shuttleworth, Analysing commuting using local regression techniques: scale, sensitivity, and geographical patterning, *Environment and Planning A*, 37:81-113, 2005.

[9] A.S. Fotheringham, C. Brunsdon, M.E. Charlton, *Geographically Weighted Regression: the Analysis of Spatially Varying Relationships*, Chichester:Wiley, 2002.

[10] R. Babuska, *Fuzzy Modeling for Control*, USA:Kluwer Academic, 1998.

[11] J.C. Bezdek, R. Ehrlich, W. Full, FCM: the fuzzy c-means clustering algorithm, Computers&Geosciences, 10(2-3):191-203, 1984.

[12] H. Akaike, Information theory and an extension of the maximum likelihood principle, In B. Petrov, F. Csaki (eds) *2nd Symposium on Information Theory*, Budapest, 267-281, 1973.

[13] Thiago F. L.V.B. Rangel, Jose A.F. Diniz-Filho, L.M. Bini, Towards and integrated computational tool for spatial analysis in macroecology and biogeography, *Global Ecology and Biogeography*, 15:321-327, 2006.

[14] R.S. Bivand, E.J. Pebesma, V. Gomez-Rubio, Applied Spatial Data Analysis with R, New York: Springer, 2008.