Computers & Geosciences 43 (2012) 34–41

Contents lists available at SciVerse ScienceDirect

Computers & Geosciences

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

Evaluating geo-environmental variables using a clustering based areal model

Bulent Tutmez [a,*], Uzay Kaymak [b], A. Erhan Tercan [c], Christopher D. Lloyd [d]

[a] Department of Mining Engineering, Inonu University, Malatya 44280, Turkey
[b] School of Industrial Engineering, Eindhoven University of Technology, P.O. Box 513 5600 MB, Eindhoven, The Netherlands
[c] Department of Mining Engineering, Hacettepe University, Ankara 06800, Turkey
[d] School of Geography, Archaeology and Palaeoecology, Queen's University, Belfast, UK

ARTICLE INFO

Article history:
Received 9 November 2011
Received in revised form
18 February 2012
Accepted 21 February 2012
Available online 1 March 2012

Keywords:
Spatial relationship
GWR
Fuzzy clustering
Local analysis
Geo-environmental

ABSTRACT

Global regression models do not accurately reflect the spatial heterogeneity which characterises most geo-environmental variables. In analysing the relationships between such variables, an approach is required which allows the model parameters to vary spatially. This paper proposes a new framework for exploring local relationships between geo-environmental variables. The method is based on extended objective function based fuzzy clustering with the environmental parameters estimated through on a locally weighted regression analysis. The case studies and prediction evaluations show that the fuzzy algorithm yields well-fitted models and accurate predictions. In addition to an increased accuracy of prediction relative to the widely-used geographically weighted regression (GWR), the proposed algorithm provides the search radius (bandwidth) and weights for local estimation directly from the data. The results suggest that the method could be employed effectively in tackling real world kernel-based modelling problems.

1. Introduction

The environmental and geological sciences deal with the spatial behaviours of natural phenomena. Geo-environmental data feature complex spatial pattern at different level scales owing to a combination of several spatial phenomena or various influencing factors of different origins (Kanevski et al., 2004). Uncertainty and irregularity are common properties of measurements of these phenomena. In many cases, it is common to assume that measurements are independent and identically distributed, but this may not be the case when working with spatial data (Cressie, 1993; Bivand et al., 2008). Spatially varying relationships between environmental variables are common and are a result of complex processes. These spatial relationships can be evaluated in different approaches by statistical models. Recently, a variety of models have been proposed to explore variations in relationships between variables (Schabenberger and Gotway, 2005; Gao et al., 2006).

The exploration of relationships between multiple variables is often approached in a regression framework. Multiple linear regression analysis allows the assessment of the strength and nature of the relationship between one variable and a set of independent variables, and such approaches are very widely used in the analysis of environmental variables, and more generally throughout the physical and social sciences. However, limitations of the standard approaches have led to the development of more robust methods which are appropriate in particular contexts. In cases where the variables are spatially referenced, a standard ordinary least square approach may not be suitable because positive spatial autocorrelation means that the assumption of independence of the samples is violated. Alternative approaches such as generalised least squares exist which account for the spatial structure in variables (Lloyd, 2006).

Spatial measures cover both attribute and location information. Areal analyses concentrate on differences across space whereas global analyses concentrate on similarities across space owing to their nature. Areal identification has been used widely in many disciplines such as image processing (local filters) for several decades. However, in some disciplines, such as the geosciences, the environmental sciences, ecology and geography, a motivation on approaches that account for areal variation has been a comparatively recent development. For the purposes of prediction of a dependent variable given a set of independent variables, local regression approaches can offer considerable benefits in terms of an increase in prediction accuracy over standard global models (Lloyd, 2006). Geographically Weighted Regression (GWR) is one approach which is being used increasingly widely to explore areal spatial variations in relationships (Fotheringham et al., 1998). It is an adaptive and effective method for modelling relationships locally by calibrating a spatially varying coefficient regression model.

[*] Corresponding author. Tel.: +90 422 3774773; fax: +90 422 3410046.
E-mail address: bulent.tutmez@inonu.edu.tr (B. Tutmez).

0098-3004/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cageo.2012.02.019

Spatial statistical methods were advanced based on probability and classical statistics. On the other hand, many spatial datasets have high levels of uncertainty, and in some cases, analyses depend on 'soft' data, which may be more qualitative than quantitative in nature (Tutmez et al., 2009). On the other hand, soft approaches like fuzzy computing have desirable characteristic features for spatial data analysis (Wong et al., 2001). The soft computing based algorithms are developed on the ground of less restrictive assumptions, and are flexible in appraising non-linearity and non-constant variable structures.

In recent geo-environmental analyses, the fuzzy approach has shown to be highly suitable for exploring complex and vague systems (Bardossy and Fodor, 2004). Therefore many fuzzy computing based methods have been used for spatial estimation in geo-environmental problems (Amini et al., 2005). The fuzzy approach obtains the means to combine numerical data and linguistic statements and satisfies a more transparent representation of the system under study (Sousa and Kaymak, 2002). It is utilised to handle uncertainties and imprecision involved in the analysis of real world data.

Fuzzy clustering, as an effective spatial data analysis method, is the corner stone of the model proposed in this study. Various fuzzy clustering algorithms such as the Fuzzy c-means Algorithm (FCM) have been used extensively for different tasks like spatial data analysis and system modelling. In the present study, the extended fuzzy clustering algorithm, which has been proposed by Kaymak and Setnes (2002), is employed and combined with least squares estimation for local analysis of environmental data. The algorithm defines the number of clusters and the cluster radii (bandwidth) directly from the data given, and it is very convenient for local kernel based modelling problems (Tutmez et al., 2009). From this perspective, a novel analysis method, descriptive case studies and a performance comparison with the GWR method (based on estimation accuracy) is presented.

The outline of the paper is as follows. Section 2 outlines the problem formulation and structure of GWR which is a well-known local modelling technique. In Section 3, the general formulation of the proposed model is presented. Section 4 gives two case studies using real data sets. Section 5 discusses the performance of the different approaches. Finally, Section 6 concludes the paper.

2. Areal identification

2.1. Problem formulation

In spatial analysis, each observation is linked with a location and there is at least an implied connection between the location and the observation. In addition, spatially varying relationships are concerned with different values for any set of properties, which are often observed at a set of irregularly distributed geographic locations in an area. In this case, the objective is to establish an areal model on the basis of locations with observations and then to use this model to make estimates at any desired point within the area (Şen, 2009).

In most cases, spatial identification was applied at a 'global' level, meaning that one set of results is generated from the models, denoting one set of relationships, which is assumed to apply equally across the study area (Fotheringham et al., 2002). Although global methods have proved useful they have the drawback that they can mask geographical variations in relationships. One of the assumptions in global analysis is that the relationship under study is spatially constant, and thus, the relationships being characterised are assumed to be 'stationary' over space. Nonetheless, in most cases the relationship varies in space. Because of this realisation, areal regression models have been established that permit the exploration of spatially varying relationships in datasets (Fotheringham et al., 2002).

Statistically, it is possible to account for correlated observations by considering a structure of the following kind in the model (Schabenberger and Gotway, 2005). If the vector of response variables is multivariate normal, the following model can presented:

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

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 generic variance V (Bivand et al., 2008).

2.2. Geographically weighted regression (GWR)

When the local coefficients vary in space, it could be taken as an indication of non-stationary. As a refinement to conventional regression approaches, GWR associates with the spatial non-stationary of empirical relationships. The approach obtains a weighting of data that is locally specific, and allows regression model parameters to vary in space (Fotheringham et al., 1998). 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 over space and is estimated by

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

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. The parameters for GWR may be estimated by solving

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

where $\hat{\beta}$ represents an estimate of β, and $\boldsymbol{W}(u_i, v_i)$ is an n by n matrix whose off-diagonal elements are zero and whose diagonal elements represent the geographical weighting of each of the n observations with respect to regression point i (Fotheringham et al., 2002). There are some weighting structures which presents w_{ij} as a continuous function of distance between i and j, d_{ij}. In the present work, the following Gaussian function has been employed.

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

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

3. Extended fuzzy clustering based least squares

3.1. Clustering by fuzzy sets

Clustering is well-known multivariate data analysis technique. The main objective of a cluster analysis is to partition a given set of data or objects into clusters (subsets, groups, classes). Hard (or crisp) clustering algorithms require that each data point belongs to only one cluster. On the other hand, fuzzy clustering extends this notion to associate each data point with every cluster using a membership function. The objective of a fuzzy clustering algorithm is to partition data into clusters so that the similarity of data objects within each cluster is maximised and similarity of data objects among clusters is minimised (Liu et al., 2008).