DOI: http://dx.doi.org/10.12989/cac.2012.10.5.457

Regression-based algorithms for exploring the relationships in a cement raw material quarry

Bulent Tutmez*[1] and Ahmet Dag[2]

[1]*Department of Mining Engineering, Inonu University, 44280 Malatya, Turkey*
[2]*Department of Mining Engineering, Cukurova University, 01100 Adana, Turkey*

(Received July 11, 2011, Revised December 7, 2011, Accepted May 10, 2012)

Abstract. Using appropriate raw materials for cement is crucial for providing the required products. Monitoring relationships and analyzing distributions in a cement material quarry are important stages in the process. CaO, one of the substantial chemical components, is included in some raw materials such as limestone and marl; furthermore, appraising spatial assessment of this chemical component is also very critical. In this study, spatial evaluation and monitoring of CaO concentrations in a cement site are considered. For this purpose, two effective regression-based models were applied to a cement quarry located in Turkey. For the assessment, some spatial models were developed and performance comparisons were carried out. The results show that the regression-based spatial modelling is an efficient methodology and it can be employed to evaluate spatially varying relationships in a cement quarry.

Keywords: cement; quarry; regression-based modelling; spatial relationship.

1. Introduction

Cement is a formed construction product composed by blending various raw materials and firing them at a high temperature for providing precise chemical proportions of silica, lime, alumina and iron in the final product, known as cement clinker. Availability of the main components of cement including limestone, clay, mudstone and shale, has vital importance for an effective manufacturing (Wilson and Kosmatka 2011). Such raw material deposits (sedimentary rocks) are common lithologies and may vary considerably in their chemistry. The evaluation of the distribution of chemical components in these deposits is an essential to provide uniform sources for production.

The continuous quality of cement production is possible only if the raw mix possesses ideal composition and furthermore if variations in this composition remain within the narrowest possible range. CaO content of cement raw material is 65% and limiting values is between 60% and 69% (Labahn 1983). The raw material composition is usually characterized by certain ratios, called standards. They are in fact proportioning formulas into which the percentages of the various oxides to compute the optimum lime (CaO) content of the mix, the so-called Lime Standard. This standard indicated as CaO content of mix, is the most promising characteristic of mix. The lime standard obtains a criterion for calculating the optimum lime content. The measured content of CaO present in the raw material is measured as a percentage of maximum CaO content which can be integrated by the acidic oxides (SiO_2, Al_2O_3, Fe_2, O_3) in the most lime-rich clinker phases under technical

* Corresponding author, Ph.D., E-mail: bulent.tutmez@inonu.edu.tr

limitations of burning and cooling (Barnes and Bernsted 2001).

Assessment of the relationships in a spatial system such as limestone or clay quarry needs some modelling tools (Asad 2011). In addition, selection of the effective analysis tool is the corner stone of successful (may be robust and accurate) estimations (Onur *et al.* 2008). In spatial data analysis, spatial regression models (Goovaerts 1997, Schabenberger and Gotway 2005) have been widely employed in different problems. All statistical methods for spatial data have to take the correlations of the observations into consideration to provide accurate, meaningful conclusions. Therefore, spatial correlation-based appraisal of a cement quarry can be the most reliable approach to evaluate the distributions of chemical components (Almeida *et al.* 2004).

In this work, two powerful algorithms such as regression kriging (RK) (Wackernagel 1998, Hengl *et al.* 2007) and geographically weighted regression (GWR) (Fotheringham *et al.* 2002) are applied to a quarry in Turkey. The main objective of the study is to analyse the CaO distributions based on spatial relationships. The spatial analysing methods are used to detect the relationships in the deposit and modelling the relationships via regression-based algorithms on a comparative manner composes the frame of the paper.

The rest of the paper is structured as follows. Section 2 states the problem and methods used in the study. Section 3 gives the real case study. Section 4 presents the results and discussion and finally Section 5 concludes the paper.

2. Method

2.1 Statement of problem

Because the measured concentrations of a chemical variable are described by the coordinates, the sample data obtained from the quarry should be considered as spatially varying data. In this structure, each measurement is associated with a location and there is at least one implied connection between the location and the measurement of chemical variable.

In a general spatial modelling approach, if we know the actual values of input variables and models are known, a model can be used to appraise the chemical variable of interest, z. For the estimation of a concentration at a sample point, some neighbouring locations and weighting structures are used.

2.2 Regression kriging

Regression Kriging (RK) is a geostatistical analysis technique to examining the distribution of regionalized variables in a spatial system. In general, a geostatistical technique comprises of three main stages: (1) exploration of the data to characterize its spatial continuity; (2) structural analysis to build a semivariogram model and (3) application of kriging for estimation.

The semivariogram is a statistical function which denotes how the data vary spatially across the area of interest. The variation between points is measured using the semivariance. Pooling together pairs of data at geographic distance h, the experimental semivariogram $\gamma(h)$ of the sample can be written as

$$\gamma(h)=\frac{1}{2N(h)}\sum_{i=1}^{N(h)}[Z(x_i)-Z(x_i+h)]^2 \tag{1}$$

where $N(h)$ is the number of sample pairs within the distance interval $h;$ $Z(x_i)$, $Z(x_i+h)$ is the sample value at two points separated by the distance interval h. Once the experimental semivariogram function has been computed from the sampled values at different locations, the next step is to fit a parametric semivariogram by a method such as the weighted least squares method (Cressie 1993).

In geostatistical theory, it is assumed that a random function $Z(\mathbf{x})$ includes a trend parameter $f(\mathbf{x})$ that can be modelled as a linear function of smoothly varying secondary variable. The random function can be modelled as a combination of trend a random variable

$$Z(x)=f(\mathbf{x})+\varepsilon(\mathbf{x}) \tag{2}$$

where $\varepsilon(\mathbf{x})$ is the random variable of mean zero and with a variogram that presents its spatial relationships. Estimation of the variogram can obtain a prediction of z at unsampled site, $\mathbf{x}_0$. The model is an integration of the trend prediction, $f(\mathbf{x}_0)$, and a kriged estimate of $\varepsilon(\mathbf{x}_0)$.

The information obtained from semivariograms used to appraise the observations in the site was considered by kriging techniques. Ordinary kriging is a process of estimating variable values at unmeasured location as follows

$$Z^*(x_0)=\sum_{i=1}^{n}\lambda_i Z(x_i) \tag{3}$$

where $Z^*(x_0)$ is the kriged or predicted value at location x_0, $Z(x_i)$ is the known value that used for estimating value at location x_0 and λ_i is kriging weight which is the solution of the kriging system (Goovaerts 1997).

As given in Eq. (4), a spatial interpolation of an unknown location (k_0) using the measured values $z(k_1),z(k_2),\ldots,z(k_n)$ can be made via summing the predicted drift and residuals

$$\hat{z}(k_0)=\hat{m}(k_0)+\hat{e}(k_0) \tag{4}$$

where the first component, deterministic part (drift) $\hat{m}$ can be computed by linear regression, and the second component, residual (error) $\hat{e}$ can be predicted by ordinary kriging as follows (Stacey *et al.* 2006)

$$\hat{z}(k_0)=\sum_{j=0}^{p}\hat{\beta}_j+q_j(k_0)+\sum_{i=1}^{n}\lambda_i\cdot e(k_i);\qquad q_0(k_0)=1 \tag{5}$$

In Eq. (5), $\hat{\beta}_j$ denotes the estimated drift model coefficients, $q_j(k_0)$ is the predictor at location k_0, p is the number of predictors, λ_i is kriging weight calculated by the spatial dependence function and $e(k_i)$ is residual of the regression. The regression coefficient vector β is estimated from a least squares technique such as ordinary least squares (OLS) or, preferably, generalized least squares using the spatial relationships between the variables (Cressie 1993, Hengl e*t al.* 2007)

$$\hat{\beta}_{GLS}=(\mathbf{q}^T\cdot\mathbf{C}^{-1}\cdot\mathbf{q})^{-1}\cdot\mathbf{q}^T\cdot\mathbf{C}^{-1}\cdot\mathbf{z} \tag{6}$$

where $\hat{\boldsymbol{\beta}}_{GLS}$ gives a vector of $p+1$ predicted drift model coefficients, $\mathbf{C}$ represents the covariance matrix of the errors, $\mathbf{q}$ indicates the matrix of independent variables, and z denotes the vector of observed values of the target variable.

In the system presented in Eq. (5), first the drift model coefficients are estimated using least squares, then the covariance function of the errors is determined to provide the GLS coefficients (Hengl *et al.* 2007). In addition, the interpolated residuals by kriging are added to the fitted drift and thus, the estimated values are obtained. The model can be written as follows (Christensen 2010)

$$\hat{z}(k_0) = \mathbf{q}_0^T \cdot \hat{\boldsymbol{\beta}}_{GLS} + \boldsymbol{\lambda}_0^T \cdot (\mathbf{z} - \mathbf{q}, \hat{\boldsymbol{\beta}}_{GLS}) \tag{7}$$

where $\hat{z}(k_0)$ is the estimated value at location k_0, $\mathbf{q}_0$ is the vector of $p+1$ estimators, q is the matrix of predictors at all measured locations, and λ_0 is the vector of n kriging weights used to take the errors.

2.3 Geographically weighted regression (GWR)

Since the spatial natural systems have heterogeneous properties, the relationships can be varied in space. If the coefficients vary in space, it can be taken as an indication of non-stationary. Therefore, spatial procedures, which should cope with the spatial non-stationary of empirical relationships, should be considered.

In the matrix form of regression equation, the vector of parameters to be estimated, β, is constant over space and which is estimated by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \tag{8}$$

On the other hand, GWR has been proposed to analyse spatially varying relationships based on areal modelling perspective. GWR has a kernel-based modelling structure. In the mechanics of GWR, the observations are weighted in accordance with their distance from the kernel centre (Fig. 1). The parameters for GWR can be estimated by solving Eq. (9) as follows

$$\hat{\beta}(u_i, v_i) = [\mathbf{X}^T\mathbf{W}(u_i, v_i)\mathbf{X}]^{-1}\mathbf{X}^T\mathbf{W}(u_i, v_i)\mathbf{y} \tag{9}$$

where $\hat{\beta}$ represents an estimation 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

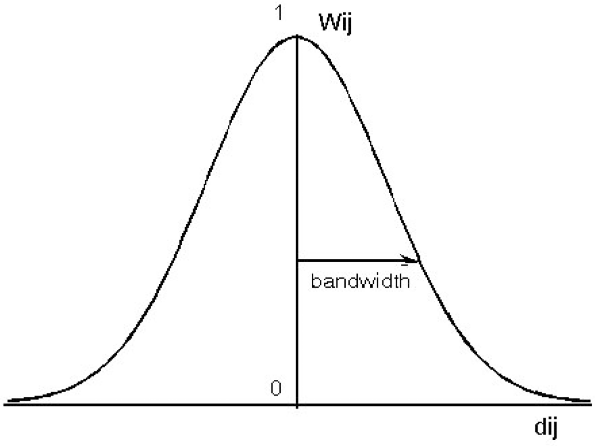

Fig. 1 A spatial kernel

for regression point *I* (Fotheringham *et al.* 2002).

Sometimes, instead of $\mathbf{W}(u_i, v_i)$, $\mathbf{W}(i)$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.

$$w_{ij} = \exp\left[-\frac{1}{2}\left(\frac{d_{ij}}{b}\right)^2\right] \tag{10}$$

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.* 1998).

The weighting is a crucial step in the modelling process and it completely relies 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). However, 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. Thus, the fit will have large bias. From an ideal methodological view, one might like to define a separate bandwidth for each estimation point (Tutmez *et al.* 2012).

3. Experimental studies

In this section, the regression models introduced in the previous section are applied to a real quarry. The distribution of CaO concentrations in the field is considered based on a spatial regression perspective.

3.1 Data and structure identification

The raw material quarry of Adana Cement Factory was considered for the case study (Fig. 2). The quarry occurred with marl, marly-limestone and limestones. Data used in this study belong to marl units which are more favourable than marly-limestone and limestone (Alkan 2007). The data comprising of 67 measurements were randomly divided into two subsets: the training set (55 samples) and the validation set (12 samples), respectively.

Structure identification (variable and feature selection) has crucial importance for the recent modelling problems (Guyon and Elisseeff 2003). In the model, in addition to spatial positions of measurements (coordinates), thickness was selected as auxiliary input variable. From these inputs, CaO concentrations were estimated. To examine the relationships between coordinates and CaO values, scatter diagrams were designed. As can be seen in Figs. 3 and 4, there are clearly big effects of coordinates on CaO concentrations. Thus, a spatial analysis of the data is necessary.

3.2 RK model

A regression kriging model should comprise of two main parts which are deterministic part (drift) and residual term. First the deterministic part of variation is estimated, and then the variogram function of the errors is employed to provide the GLS coefficients. Next, the residuals are re-

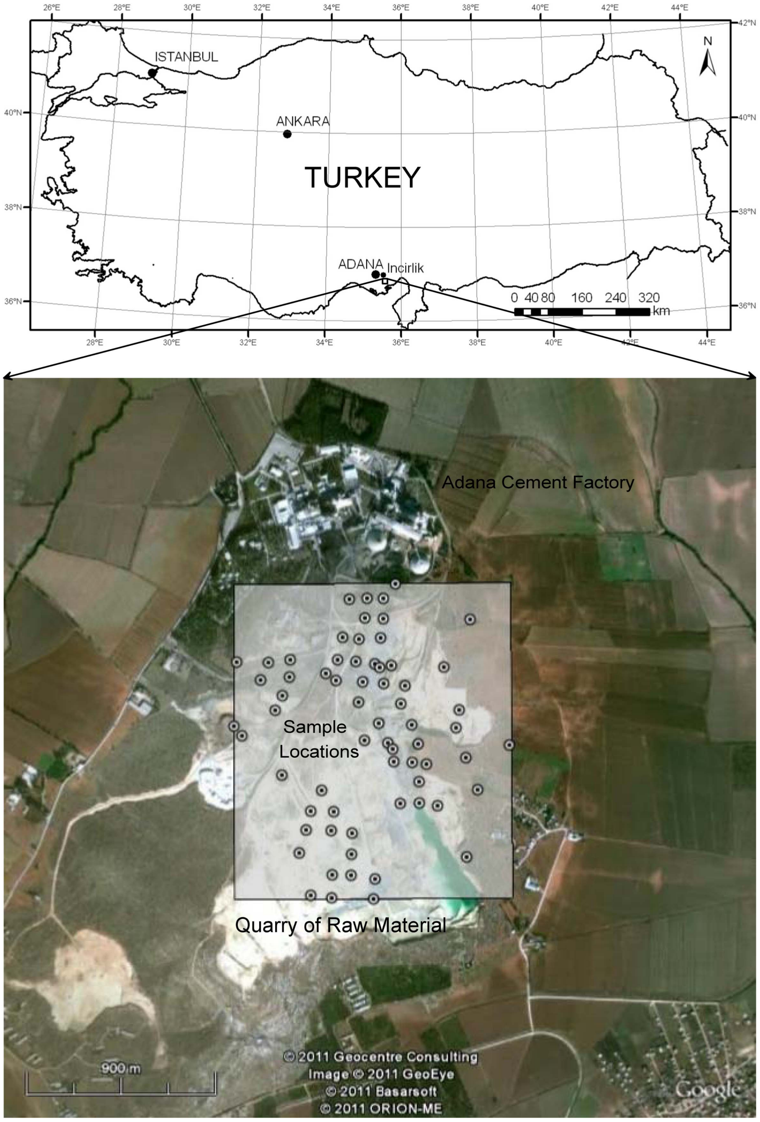

Fig. 2 Map of Quarry

computed, from which an updated variogram function is determined, and so on. Fig. 5 indicates the relationships between residuals and fitted values. In addition, the final experimental variogram of residuals is given in Fig. 6.

For kriging estimation, the experimental variogram was stated by a spherical structure and range value was defined as 4. By using the spatial structure, weighted estimations of residuals were carried out. Finally, the results derived from kriging interpolation were added to drift and estimated values were provided.

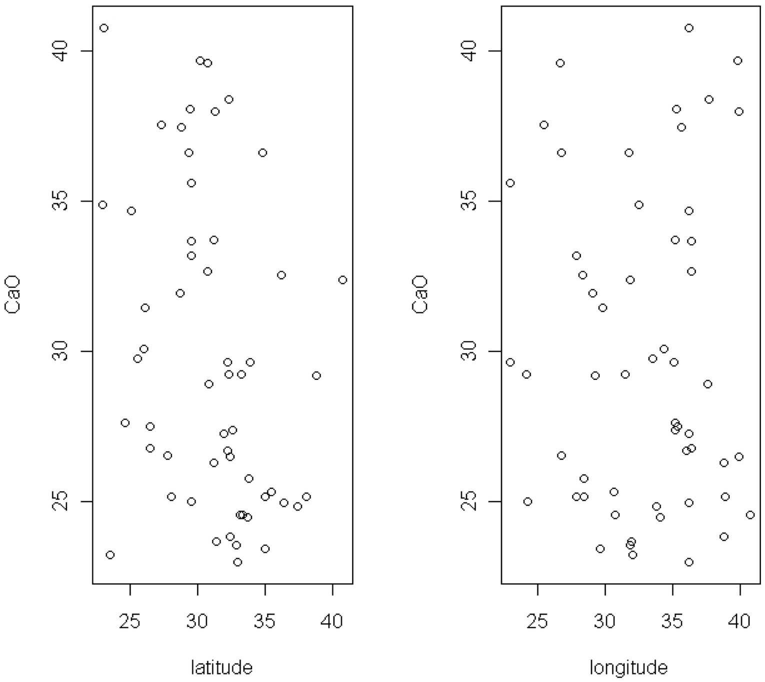

Fig. 3 Training data

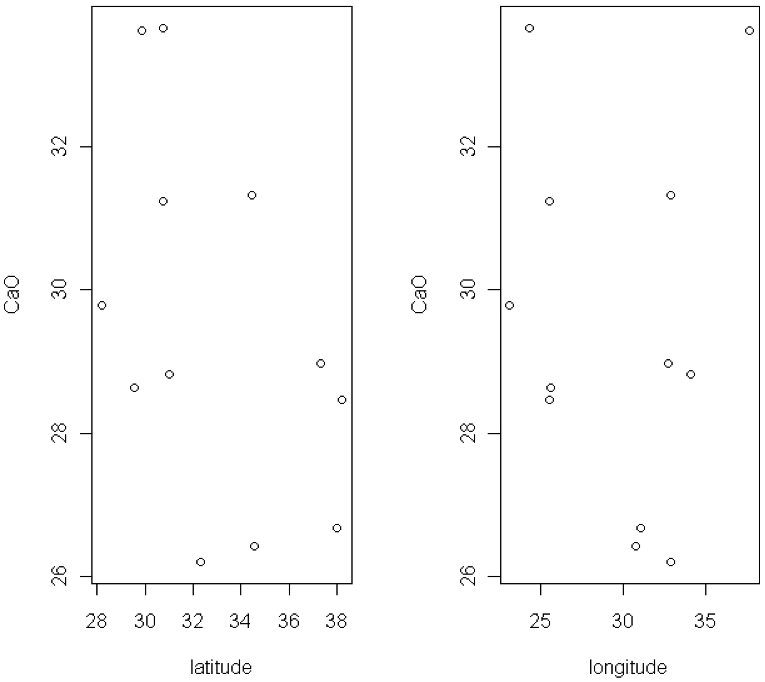

Fig. 4 Testing data

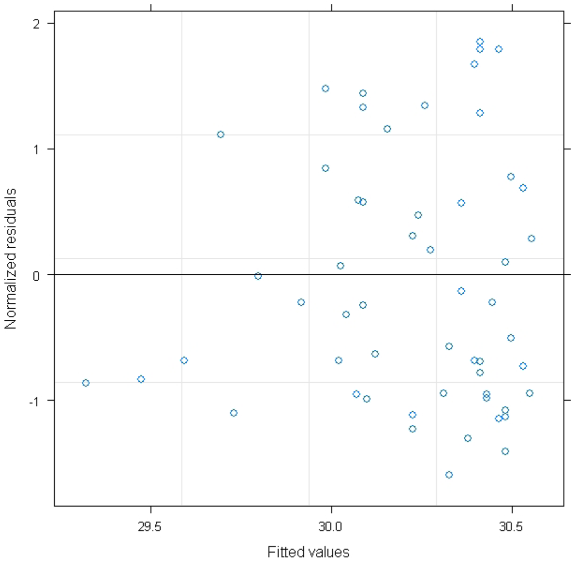

Fig. 5 Relationships between fitted values and residuals

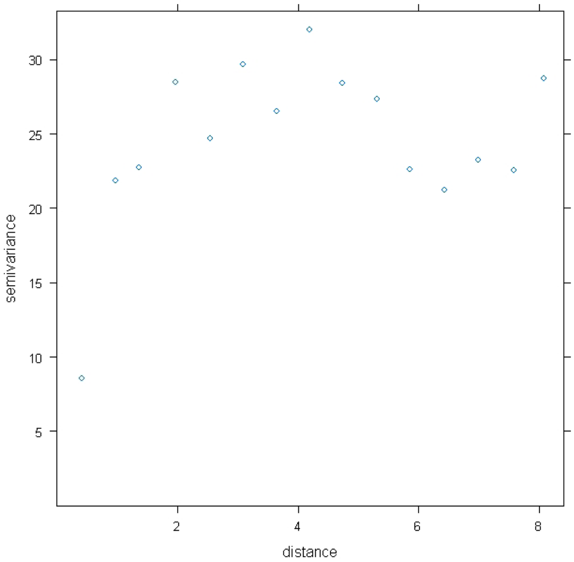

Fig. 6 Variogram function of residuals

3.3 GWR model

The GWR analysis has been carried out using a Gaussian distance decay function with a fixed spatial kernel. The bandwidth may be selected manually, or an optimal bandwidth can be identified using an approach such as cross-validation. A method of deriving the bandwidth which provides a trade-off between goodness-of-fit and degrees of freedom is effective to minimize the Akaike Information Criterion (AIC). The AIC has been modified by Fotheringham *et al.* (2002) for GWR as follows

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

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** which maps $\hat{\mathbf{y}}$ onto $\mathbf{y}$ (i.e., $(\hat{\mathbf{y}} = \mathbf{sy})$).

In this application, the GWR model was fitted using *R* routines. In addition, the fixed bandwidth value was determined using AIC with the '*spgwr*' package in *R* (Bivand *et al.* 2008). Table 2

Table 1 Some inputs for GWR model

	Training	Testing
Spatial function	Gaussian	Gaussian
Fixed bandwidth (h)	4.793 distance units.	4.793 distance units
Number of locations to fit model (n)	55	12

summarizes some inputs for the training and test data. In these tables, r^2 relates to observations against their estimates and OLS denotes the Ordinary Least Squares optimization.

4. Results and discussion

For appraising the performances of the developed regression-based spatial models, the relationship between the estimated CaO concentrations and the actual (measured) CaO concentrations was considered. Because the kriging models produce exact estimations of training data, the results are assessed by testing data. Fig. 7 illustrates the results of two models together with the actual (measured) CaO concentrations. As can be followed from Fig. 7, RK model has relatively better estimations compared to with GWR model.

To show the performances more clearly, the relative errors of the estimations are presented in Fig. 8. Relative Error (RE) is well-known performance indicator that can be stated as follows

$$RE = 100 \times \frac{\left|y - y^*\right|}{y} \tag{12}$$

where y and y^* denote measured and estimated CaO values, respectively. Because the average RE errors are smaller than 10%, it can be expressed that both methods can be accepted as successful (Bardossy and Fodor 2004). In particular, RK outperforms than GWR model. In addition to estimation capacities, smoothing degree of the estimations was handled. Because data variability is important in spatial data analysis, reproducing the variability in the estimation values are checked. The variability (standard deviation) and average RE values are summarized in table. The results indicate that RK algorithm considers data variability and smoothing more.

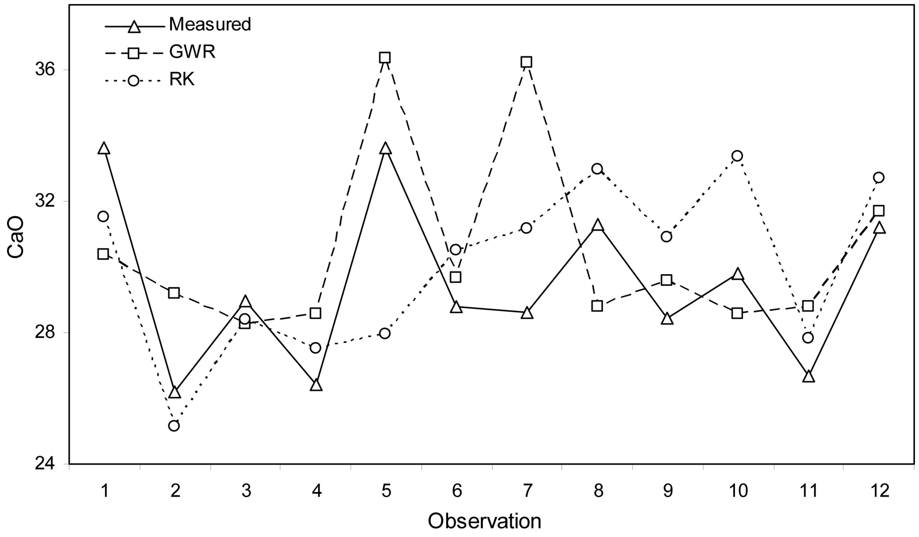

Fig. 7 Measured and estimated test values

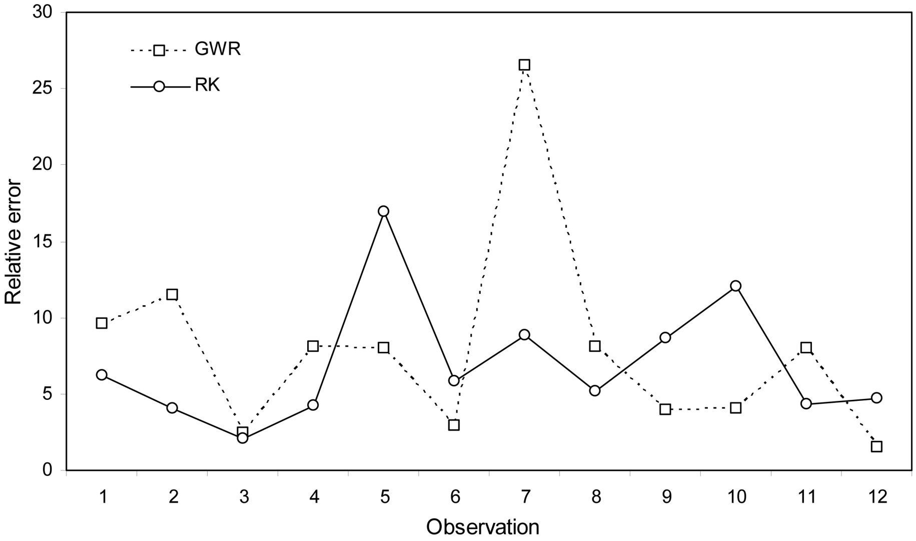

Fig. 8 Relative errors

Table 2 Test performances of models

	Measured	GWR	RK
Average standard deviation	2.554	2.858	2.584
Average relative error	-	7.891	6.911

5. Conclusions

Chemical composition of cement is crucial for manufacturing and the evaluation of the composition in a quarry is also necessary for obtaining suitable raw materials. In this study, the relationships in a cement raw material quarry were explored by spatial modelling. Two spatial data analysis methods were employed and some estimation was conducted.

The regression-based algorithms used in this study have revealed some successful outcomes. In addition, it was observed that in addition to estimation capacity, RK algorithm also takes into account data variability and smoothing degree. As a consequence, the regression-based spatial data algorithms could be applied to cement raw materials deposits to analyse some distributions and relationships.

Acknowledgements

The authors would like to extend their appreciation to Adana (Oyak) Cement Factory.

References

Alkan, B. (2007), *Evaluation of Adana cement raw material field by geostatistics and fuzzy approaches (MSc Thesis)*, Cukurova University, Adana, Turkey.

Almeida, J., Rocha, M. and Teixeria, A. (2004), "Spatial characterization of limestone and marl quality in a quarry for cement manufacturing", *Geostatistics Banff: 7th Inernational Geostatistics Congress*, Canada.

Asad, M.W.A. (2011), "A heuristic approach to long-range production planning of cement quarry operations", *Prod. Plan. Control.*, **22**(4), 353-364.

Bardossy, G. and Fodor, J. (2004), *Evaluation of uncertainties and risks in geology*, Springer-Verlag, Heidelberg.

Barnes, B. and Bernsted, J. (2001), *Structure and performance of cements,* Spon Press.

Bivand, R.S., Pebesma, E.J. and Gomez-Rubio, V. (2008), *Applied spatial data analysis with R,* Springer, New York.

Christensen, R. (2010), *Advanced linear modeling: multivariate, time series, and spatial data; nonparametric regression and response surface maximization*, Springer, New York.

Cressie, N. (1993), *Statistics for spatial data*, Wiley, New York.

Fotheringham, A.S., Brunsdon, C. and Charlton, M.E. (2002), *Geographically weighted regression: the analysis of spatially varying relationships*, Wiley, Chichester.

Fotheringham, A.S., Charlton, M.E. and Brunsdon, C. (1998), "Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis"*, Environ. Plan. A,* **30**(11), 1905-1927.

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

Guyon, I. and Elisseeff, A. (2003), "An introduction to variable and feature selection", *J. Mach. Learn. Res.,* **3**, 1157-1182.

Hengl, T., Heuvelink, G.B.M. and Rossiter, D.G. (2007), "About regression-kriging: from equations to case studies", *Comput. Geosci.,* **33**(10), 1301-1315.

Labahn, O. (1983), *Cement engineers handbook*, Wiesbadener Graphische Betriebe GmbH, Berlin.

Onur, A.H., Konak, G. and Karakus, D. (2008), "Limestone quarry quality optimization for a cement factory in Turkey", *J. S. Afr. I. Min. Metall.,* **108**(12), 753-757.

Paez, A., Uchida, T. and 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**(4), 733-754.

Schabenberger, O. and Gotway, C.A. (2005), *Statistical methods for spatial data analysis*, CRC Press, Boca Raton.

Stacey, K.F., Lark, R.M., Whitmor, A.P. and Milne, A.E. (2006), "Using a process model and regression kriging to improve predictions of nitrous oxide emissions from soil", *Geoderma.*, **135**, 107-117.

Tutmez, B., Kaymak, U. and Tercan, A.E. (2012), "Local spatial regression models: a comparative analysis on soil contamination", *Stoch. Env. Res. Risk A.*, **26**(7), 1013-1023.

Wackernagel, H. (1998), *Multivariate geostatistics: An introduction with applications*, Springer, Berlin.

Wilson, M.L. and Kosmatka, S.H. (2011), *Design and control of concrete mixtures*, Portland Cement Assn, Canada.

CM