www.czasopisma.pan.pl PAN www.journals.pan.pl
VERSITA

Arch. Min. Sci., Vol. 58 (2013), No 4, p. 1289–1300

Electronic version (in color) of this paper is available: http://mining.archives.pl

DOI 10.2478/amsc-2013-0089

DOGAN KARAKUS*

# FINDING THE BEST-FIT POLYNOMIAL APPROXIMATION IN EVALUATING DRILL DATA: THE APPLICATION OF A GENERALIZED INVERSE MATRIX

# POSZUKIWANIE NAJLEPSZEJ ZGODNOŚCI W PRZYBLIŻENIU WIELOMIANOWYM WYKORZYSTANEJ DO OCENY DANYCH Z ODWIERTÓW – ZASTOSOWANIE UOGÓLNIONEJ MACIERZY ODWROTNEJ

In mining, various estimation models are used to accurately assess the size and the grade distribution of an ore body. The estimation of the positional properties of unknown regions using random samples with known positional properties was first performed using polynomial approximations. Although the emergence of computer technologies and statistical evaluation of random variables after the 1950s rendered the polynomial approximations less important, theoretically the best surface passing through the random variables can be expressed as a polynomial approximation. In geoscience studies, in which the number of random variables is high, reliable solutions can be obtained only with high-order polynomials. Finding the coefficients of these types of high-order polynomials can be computationally intensive. In this study, the solution coefficients of high-order polynomials were calculated using a generalized inverse matrix method. A computer algorithm was developed to calculate the polynomial degree giving the best regression between the values obtained for solutions of different polynomial degrees and random observational data with known values, and this solution was tested with data derived from a practical application. In this application, the calorie values for data from 83 drilling points in a coal site located in southwestern Turkey were used, and the results are discussed in the context of this study.

**Keywords :** polynomial approximation, drilhole data, estimate model, coal calorie values

W górnictwie wykorzystuje się rozmaite modele estymacji do dokładnego określenia wielkości i rozkładu zawartości pierwiastka użytecznego w rudzie. Estymację położenia i właściwości skał w nieznanych obszarach z wykorzystaniem próbek losowych o znanym położeniu przeprowadzano na początku z wykorzystaniem przybliżenia wielomianowego. Pomimo tego, że rozwój technik komputerowych i statystycznych metod ewaluacji próbek losowych sprawiły, że po roku 1950 metody przybliżenia wielomianowego straciły na znaczeniu, nadal teoretyczna powierzchnia najlepszej zgodności przechodząca przez zmienne losowe wyrażana jest właśnie poprzez przybliżenie wielomianowe. W geofizyce, gdzie liczba próbek losowych jest zazwyczaj bardzo wysoka, wiarygodne rozwiązania uzyskać można jedynie przy wykorzystaniu wielomianów wyższych stopni. Określenie współczynników w tego typu wielomia-

* DOKUZ EYLUL UNIVERSITY, ENGINEERING FACULTY, DEPARTMENT OF MINING ENGINEERING, BUCA-IZMIR, TURKEY

nach jest skomplikowaną procedurą obliczeniową. W pracy tej poszukiwane współczynniki wielomianu wyższych stopni obliczono przy zastosowaniu metody uogólnionej macierzy odwrotnej. Opracowano odpowiedni algorytm komputerowy do obliczania stopnia wielomianu, zapewniający najlepszą regresję pomiędzy wartościami otrzymanymi z rozwiązań bazujących na wielomianach różnych stopni i losowymi danymi z obserwacji, o znanych wartościach. Rozwiązanie to przetestowano z użyciem danych uzyskanych z zastosowań praktycznych. W tym zastosowaniu użyto danych o wartości opałowej pochodzących z 83 odwiertów wykonanych w zagłębiu węglowym w południowo- zachodniej Turcji, wyniki obliczeń przedyskutowano w kontekście zagadnień uwzględnionych w niniejszej pracy.

**Słowa kluczowe:** przybliżenie wielomianowe, dane z otworów, model estymacji, wartości opałowa węgla

# 1. Introduction

Mathematical and statistical approximations are used to determine the size and grade distribution of ore bodies that are detected during preliminary surveys. In the mining industry, which involves more risks than other sectors, it is very important to accurately approximate the size and location of an ore body located below ground. At present, many estimation models exist to determine the nature and grade of an ore. In these models, the estimations are based on functions developed using random variables obtained in drilling tests. Among these methods, geostatistics and inverse distance weighting are commonly used. In the geostatistical method, the regional variance in the observational data is evaluated statistically, whereas in the inverse distance weighting method, the evaluation is based on the relationships among the positional properties of the observational data. On the other hand, Soltani and Hezerkhani (2009) studied on drillhole location optimization for ore body modelling and intended to increase geostatistic estimation accuracy.

Howarth (2001) reviewed the approximation methods used in geoscience in detail, giving an account of the historical development of this field and the methods employed. According to Howarth, data analysis on modern geosciences began with the work of Hutton in the 18th century and this analysis continues today. Early pioneering studies involved the estimation of the Earth's magnetic field lines and the methods used were an inspiration for later studies in a variety of disciplines. One of these studies was by Krumbein (1952, 1956), who used the method of polynomial approximation in the mapping of asymmetric regional variations. The aim of this method was to differentiate large-scale regional variations from small-scale localized deviations and thereby determine the directions and behaviors of both the general variations and local anomalies. Krumbein (1956) called his approximation method the "trend surface" method. In this method, the positional values (coordinates) of the random observations are expressed as the variables of a polynomial, whereas the observations are expressed as the values of random samples at a given position. Krumbein (1959) and Oldham and Sutherland (1955) defined the trend surface by using a functional description as a basis:

$$X = t + \epsilon \tag{1}$$

Here, X is the observed data for given spatial coordinates, i.e., north (x) and east (y) positional properties, and t is a polynomial function created by the use of coordinates as defined by the trend surface fitting method (Howarth, 2001):

$$t = A_{00} + A_{10}x + A_{01}y + A_{20}x^2 + A_{11}xy + \ldots + A_{nm}x^n y^m \tag{2}$$

where $\epsilon$ is a constant described for random error that is often neglected in estimation models.

It is thus necessary to determine the A coefficients to solve these approximation problems. Generalized inverse matrix operations can be used for the solution of unknown coefficients, but the size of the matrix and the inverse matrix operation requires excessive computer capacity. More specifically, these operations require the numbers associated with the coordinates to be raised to high-order powers, as well as the solution of the inverse matrix operation itself, which renders the task of determining these coefficients difficult (Karakus et al., 2011). Dwyer and Waugh (1953) were the first researchers to note this difficulty. Other authors have reported that the errors associated with these calculations result in approximation with low regression coefficients (Norcliffe, 1969; Tinkler, 1969; Chayes, 1970; Howarth, 2001,).

The fact that the averages of the values calculated by polynomial regression methods are fairly similar attracted the early attention of statisticians, and the concept of moving averages was then investigated (Krendall, 1946; Wold, 1949; Potter, 1955; Schlee, 1957; Pelletier, 1958). These studies were performed in diverse fields; perhaps the most interesting study was that of Krige (1960), whose research focused on gold reserves. Matheron (1962, 1963, 1965), who drew on the work completed by Krige and others, subsequently published the seminal principles of geostatistics, which entailed a group of statistical estimation methods. Wang and Zhang (1999) compared the kriging and trend surface analysis methods; they selected a fifth-degree polynomial to determine regional variations in heavy metal content and found that lateral variations could be detected using the kriging method.

The aim of this study was to determine the best-fitting degree for a polynomial function in a polynomial approximation method that uses random variables. In previous trend surface mapping studies, the best way to select the degree of the approximating polynomial has not always been clear; likewise, which coefficients best estimate the constructed trend surface is often unclear. The consequent ambiguity regarding estimation power in practical applications leads to the erroneous interpretation of the results of polynomial approximation methods. In this study, the mean values obtained from 83 drilling sites in a coal seam were used to perform surface fitting with polynomials of varying degrees, and the results of these approximations are discussed.

## 2. The theory of bivariate polynomial approximation

An approximation problem (function) with the goal of estimating a value at a point in a region of interest with minimum error using randomly positioned samples is referred to as the problem of finding the best-fitting function. The most commonly used approximation functions in engineering and scientific applications are polynomial functions. One reason for this popularity is the fact that every analytical function can be expanded as a Taylor series at a given point.

In this study, a finite number of observational data at known positions (north = $x$, east = $y$) were used to evaluate the solutions of bivariate polynomial approximations. The polynomial approximation was chosen here rather than the intermediate value theorem for polynomials because the number of data points is very high, and the data points are randomly distributed throughout the studied region. To illustrate, assume that we have a number $k$ of different sampling points $(x_i, y_i)$, $1 \leq i \leq k$ defined on the $x$-$y$ plane, all belonging to the unknown function $z = f(x,y)$. The function $z = f(x,y)$ thus defines a surface in the sampling space (in three dimensions). The purpose here is to determine the bivariate polynomial function that reproduces the data with

minimum error and/or estimates the value of $\hat{z} \approx f(\hat{x},\hat{y})$ with minimum error at a point $(\hat{x},\hat{y})$ that is not given in the dataset:

$$f(x,y) \approx p(x,y)$$

The bivariate polynomial, written in terms of powers of $x$ and $y$ ($x^m y^n$) and having a degree of $m + n$, is defined as

$$p(x,y) = \sum_{r=0}^{m}\sum_{s=0}^{n} a_{rs} x^r y^s \tag{1}$$

and has $(m + n)(n + 1)$ coefficients. In terms of matrices, the bivariate polynomial (1) is written as follows:

$$\boldsymbol{u}(x,y)\boldsymbol{a} = p(x,y) \tag{2}$$

Here, $1 \times (m + n)(n + 1)$

$$\boldsymbol{u}^T(x,y) = [1 \;\; y \;\; y^2 \;\; \cdots \;\; y^n \;\; x \;\; xy \;\; xy^2 \;\; \cdots \;\; x^m \;\; x^m y \;\; x^m y^2 \;\; \cdots \;\; x^m y^n]$$

and $(m + 1)(n + 1) \times 1$

$$\boldsymbol{a}^T = [a_{00} \;\; a_{01} \;\; a_{02} \;\; \cdots \;\; a_{0n} \;\; a_{10} \;\; a_{11} \;\; a_{12} \;\; \cdots \;\; a_{1n} \;\; \cdots \;\; a_{m0} \;\; a_{m1} \;\; a_{m2} \;\; \cdots \;\; a_{mn}]$$

are vectors. To obtain the $(m + 1)(n + 1)$ coefficients for $m + n$ polynomials, at least $(m + 1)(n + 1)$ data points are required. The number of data points is thus much larger than the number of coefficients to be obtained; that is, $k \gg (m + 1)(n + 1)$. For this reason, the approximation method is used.

The points $(x_i, y_i)$, $1 \le i \le k$ must satisfy the following equation:

$$p(x_i,y_i) = f(x_i,y_i) \tag{3}$$

Therefore, the linear system of equations below is obtained:

$$\boldsymbol{U}\boldsymbol{a} = \boldsymbol{b} \tag{4}$$

Here, the matrix is $\boldsymbol{U} = \begin{bmatrix} \boldsymbol{u}^T(x_1,y_1) \\ \boldsymbol{u}^T(x_2,y_2) \\ \boldsymbol{u}^T(x_3,y_3) \\ \vdots \\ \boldsymbol{u}^T(x_k,y_k) \end{bmatrix}$ $k \times (m+1)(n+1)$, and the vector is $\boldsymbol{b} = \begin{bmatrix} f(x_1,y_1) \\ f(x_2,y_2) \\ f(x_3,y_3) \\ \vdots \\ f(x_k,y_k) \end{bmatrix}$ $k \times 1$.

The rank of the linear equation (4) is $(m + 1)(n + 1)$, and the system (4) cannot be solved because the data differ from each other. The best approximated solution to the unsolvable linear equation system (4) is obtained by determining the vector $\boldsymbol{a}$ and using the method of least squares. The equation normal to (4) is obtained as follows:

$$\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{a} = \boldsymbol{U}^T\boldsymbol{b} \tag{5}$$

Equation (5) has $(m + 1)(n + 1)$ unknowns. $\boldsymbol{U}^T\boldsymbol{U}$ is a square and symmetric matrix and has a $rank(\boldsymbol{U}^T\boldsymbol{U}) = (m + 1)(n + 1)$.The only solution to the normal equation (5) is obtained as

$$\boldsymbol{a} = (\boldsymbol{U}^T\boldsymbol{U})^{-1}\,\boldsymbol{U}^T\boldsymbol{b} \tag{6}$$

Here, the matrix $(\boldsymbol{U}^T\boldsymbol{U})^{-1}\,\boldsymbol{U}^T$ is the Moore-Penrose *g*-inverse matrix of $U$ and is denoted $\boldsymbol{U}^+ = (\boldsymbol{U}^T\boldsymbol{U})^{-1}\,\boldsymbol{U}^T$. The coefficients to be obtained from (6) represent the optimal approximated solution to (4) and are the coefficients of the polynomial (1). The bivariate polynomial verifying the observational data with minimum error is calculated.

## 3. Methods

The coefficients of the polynomial were calculated using the generalized inverse matrix solution defined above. A computer program was written in Visual Basic for this solution. The general algorithm of this program is illustrated in Figure 1. First, the random observational values were read from a data file, and the position of each drilling site (the *x*-*y* coordinates) and values to be modeled (calorie and ash) were assigned to variables. Then, using the position variables of the drilling sites and setting the degree of the polynomial equal to the maximum number of drilling points, the mathematical problem was constructed. For points in the data file with known values, the estimation was performed using the coefficient vector obtained after solving the problem. A linear regression was performed between the original and estimated values of $n$ drilling points, and the best polynomial degree was determined. This operation was repeated from the smallest to the highest possible $(n - 1)$ polynomial degree, and the degree with the maximum linear regression value was displayed on the screen.

## 4. Analysis

We tested the developed solution algorithm using data from 83 drilling points in the Afsin-Elbistan coal basin in Turkey. The drilling sites were randomly located in an area 2,050 m long on the east-west axis and 2,315 m long on the north-south axis (Fig. 2). The problem defined above was constructed and solved for the average calorie values of the drills that were used as coal cutters.

Using the computer program we developed, the approximation surface coefficients of the calorie values of the drilling sites were solved for various polynomial degrees. The general statistical evaluations of the raw calorie values and the calorie values estimated with different polynomial degrees are listed in Table 1. Here, the mean calorie values remained constant as the polynomial degree decreased, but the normal distributions and standard deviations progressively deviated from the raw calorie values (Fig. 3). In this case, as the polynomial degree decreased, the surface obtained did not include regional anomalies and thus led to surfaces that can be interpreted differently. To allow comparison with previous studies, the polynomial degrees selected were 2, 3, 4, and 8 (where the m and n values are the same). Additionally, we evaluated the powers $m = 5$ and $n = 12$, which represented the closest surface to the original calorie values determined using the developed algorithm solution that resulted from our analysis.

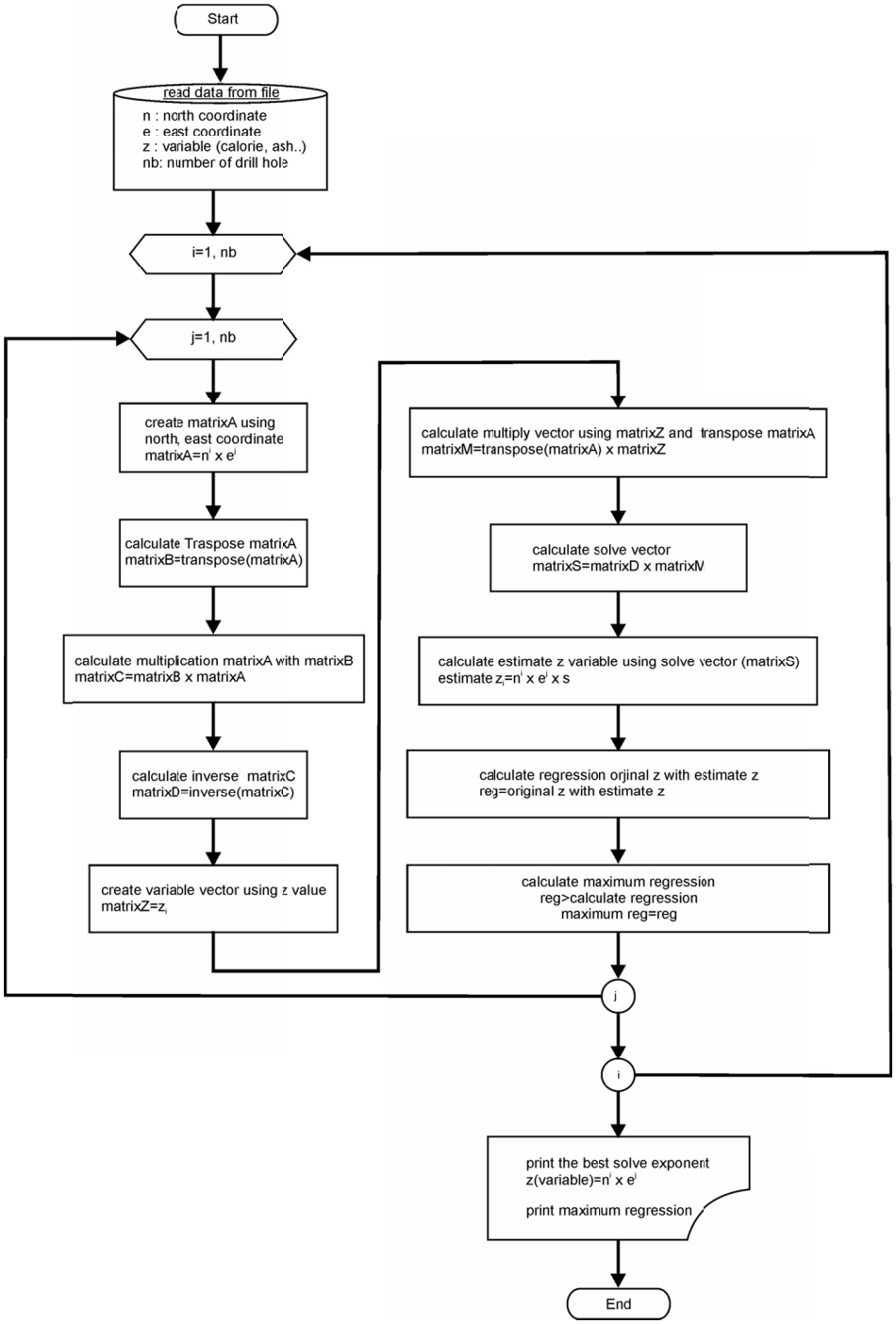


Fig. 1. The computer algorithm to determine the highest polynomial approximation degree

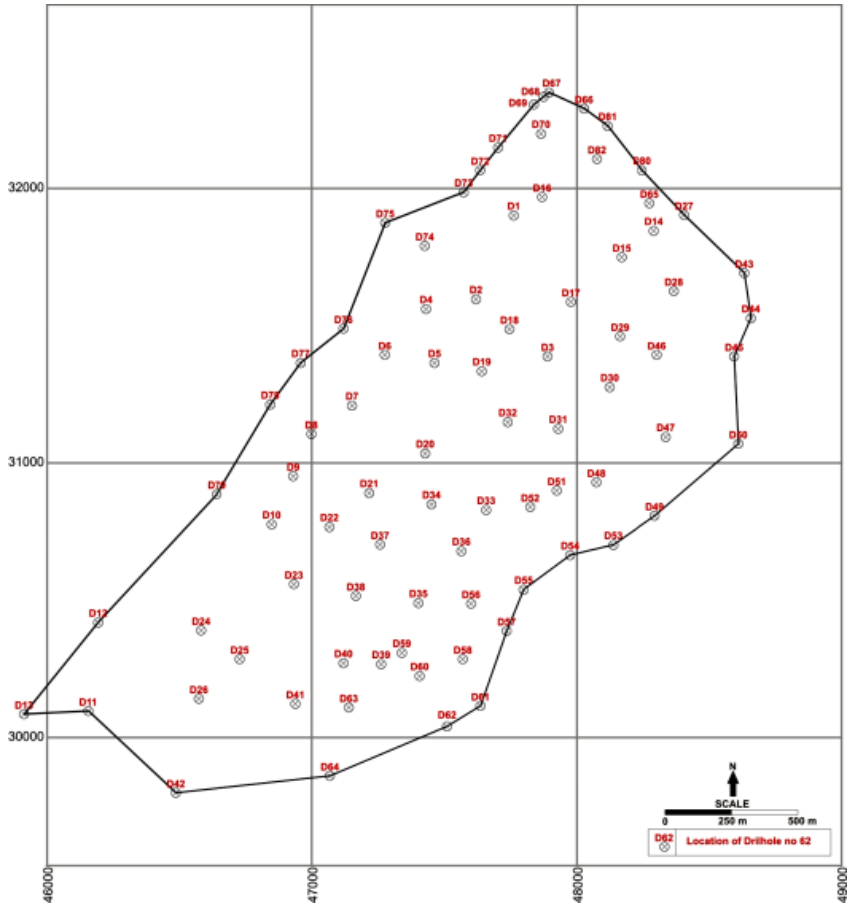


Fig. 2. Locations of the drilling tests used to provide the random data

TABLE 1

General statistical evaluation

| | **Raw Calorie Values** | **Estimated values $m = 2, n = 2$** | **Estimated values $m = 3, n = 3$** | **Estimated values $m = 4, n = 4$** | **Estimated values $m = 8, n = 8$** | **Estimated values* $m = 5, n = 12$** |
|---|---|---|---|---|---|---|
| N | 83 | 83 | 83 | 83 | 83 | 83 |
| Mean | 1,196.2 | 1,196.2 | 1,196.2 | 1,196.2 | 1,196.2 | **1,196.2** |
| Standard deviation | 138 | 92.5 | 101.8 | 105.5 | 129.1 | **130.5** |
| Minimum | 765 | 884.6 | 883.7 | 859.6 | 786.7 | **749.4** |
| Maximum | 1,427 | 1,346.2 | 1,382.7 | 1,404.9 | 1,426.6 | **1,423.2** |
| Skewness | -0.81 | -0.89 | -0.92 | -0.89 | -1.06 | **-1.07** |
| Kurtosis | 1.08 | 1.09 | 1.61 | 1.30 | 1.74 | **1.87** |

* The powers of the best-fitting approximation surfaces.

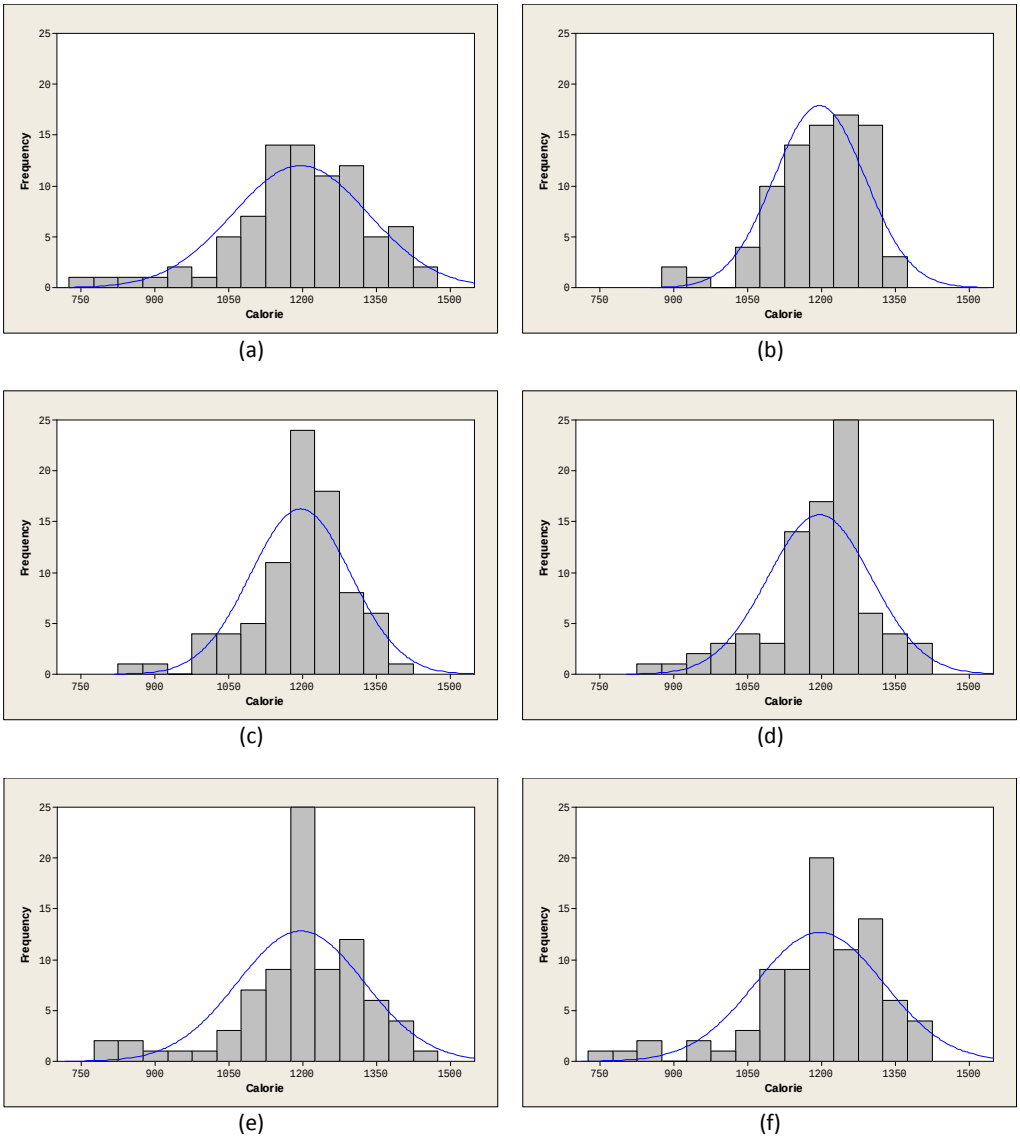


Fig. 3. General statistical evaluation of the calorie values obtained in 83 drilling tests. (a) Raw calorie value; (b) $m = 2$, $n = 2$; (c) $m = 3$, $n = 3$; (d) $m = 4$, $n = 4$; (e) $m = 8$, $n = 8$; (f) $m = 5$, $n = 12$; the best-fitting approximation surface was obtained at this degree

Regression analysis was performed to assess the statistical relationship between the estimated values obtained using the solution vector and the raw data. As shown in Figure 4, the degree of the polynomial with the largest coefficients was $n = 8$ and $m = 8$ for the case where $(m + 1)(n + 1) < 83$, which defined the upper limits for m and n using drilling data from 83 points. Theoretically, the surface most accurately representing the raw data should be obtained with this

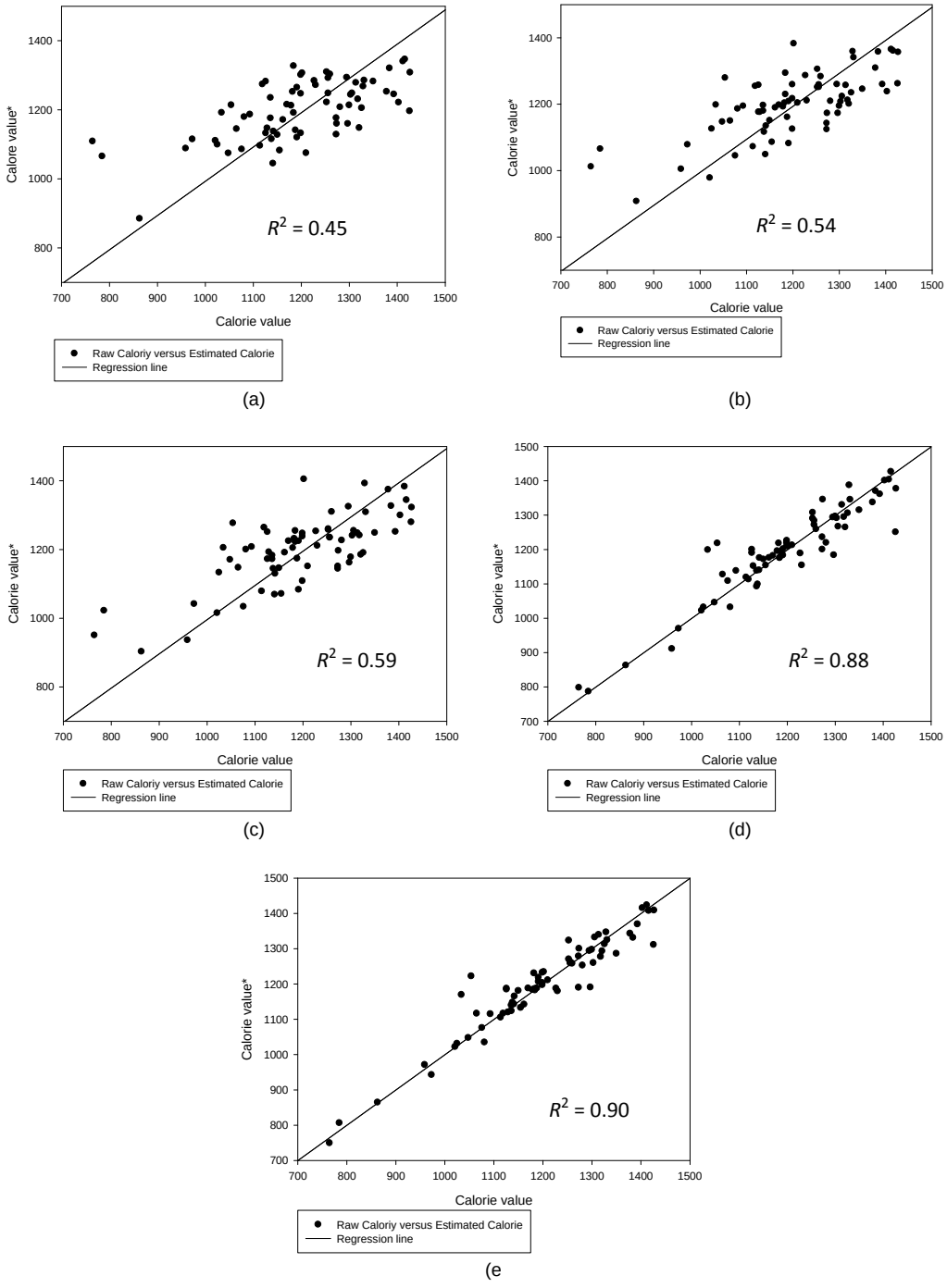


Fig. 4. Regression analyses. (a) $m = 2$, $n = 2$; (b) $m = 3$, $n = 3$; (c) $m = 4$, $n = 4$; (d) $m = 8$, $n = 8$; (e) $m = 5$, $n = 12$; the best-fitting approximation surface was obtained at this degree ($R^2 = 0.91$)

polynomial degree. However, the results of our analysis revealed that the closest surface to the raw data approached the degrees of $n = 5$ and $m = 12$.

To verify the regression analysis, contour maps were produced from the raw and estimated calorie values using the triangulation method. Figure 5 shows the contour curves drawn with the values estimated using the same position coordinates; the surface closest to the raw data is situated at the polynomial degrees $m = 5$ and $n = 12$, where the highest regression coefficient was found. However, in the estimation of intermediate values using the solution coefficients, errors may arise, especially outside the region defined by the drilling locations. Furthermore, there were significant differences between the surfaces obtained at lower polynomial degrees and those obtained using the raw data (Figure 4, a–c). Therefore, lower-degree polynomial approximations should not be used, especially for the analysis of large datasets.

## 5. Results and Discussion

Mathematical estimation methods have a long history in modern geoscience research, and the shortcomings of polynomial approximation methods paved the way for pioneering work in geostatistical methods. The rising popularity of modern geostatistics has limited research on polynomial approximation methods and their applications. In this study, we investigated the solution to an estimation problem constructed from the positional properties of random variables using a polynomial approximation method. To determine the solution coefficients, a Moore-Penrose generalized inverse matrix solution was used, and a computer program was written to test the results obtained with different polynomial degrees. Drilling data from a coal basin were used in this test case, and the calorie values of this coal seam were selected as the variables. Based on our analysis, the following results were obtained:

i) The Moore-Penrose generalized inverse matrix solution can be used in approximation methods for nonlinear systems. However, this method can be difficult because the solution of high-ranking matrix operations is limited due to the large size of the dataset. Another limitation is that the degree of the polynomial used in the estimation process is determined by the number of data points. Additionally, in studies with an insufficient number of samples, the polynomial degree is small, and the solution for the resulting approximation surface may be misleading. Generalized inverse matrix methods are commonly used in the numerical analysis of nonlinear systems with polynomial approximations. However, no method has been developed to determine the degree of the polynomial that is most accurate in these approximation solutions. In this study, a computer program was developed to evaluate every possible degree of the approximating polynomial to yield the best approximation, and the program was tested in the analysis of drilling data. To determine the closest approximation model, linear regression was performed between the solution coefficients calculated as a function of polynomial degree and the raw data.

ii) In theory, the surface most closely representing a domain of random variables is expected to occur at the polynomial degree equal to the number of data points. However, due to the structural irregularities encountered in geoscience applications, the solution providing the best approximation may be found at a different polynomial degree. This study found that, in the analysis of calorie values from the drilling data, the most accurate regression of the approximation surface was obtained at the polynomial degrees $n = 5$, $m = 12$. This result

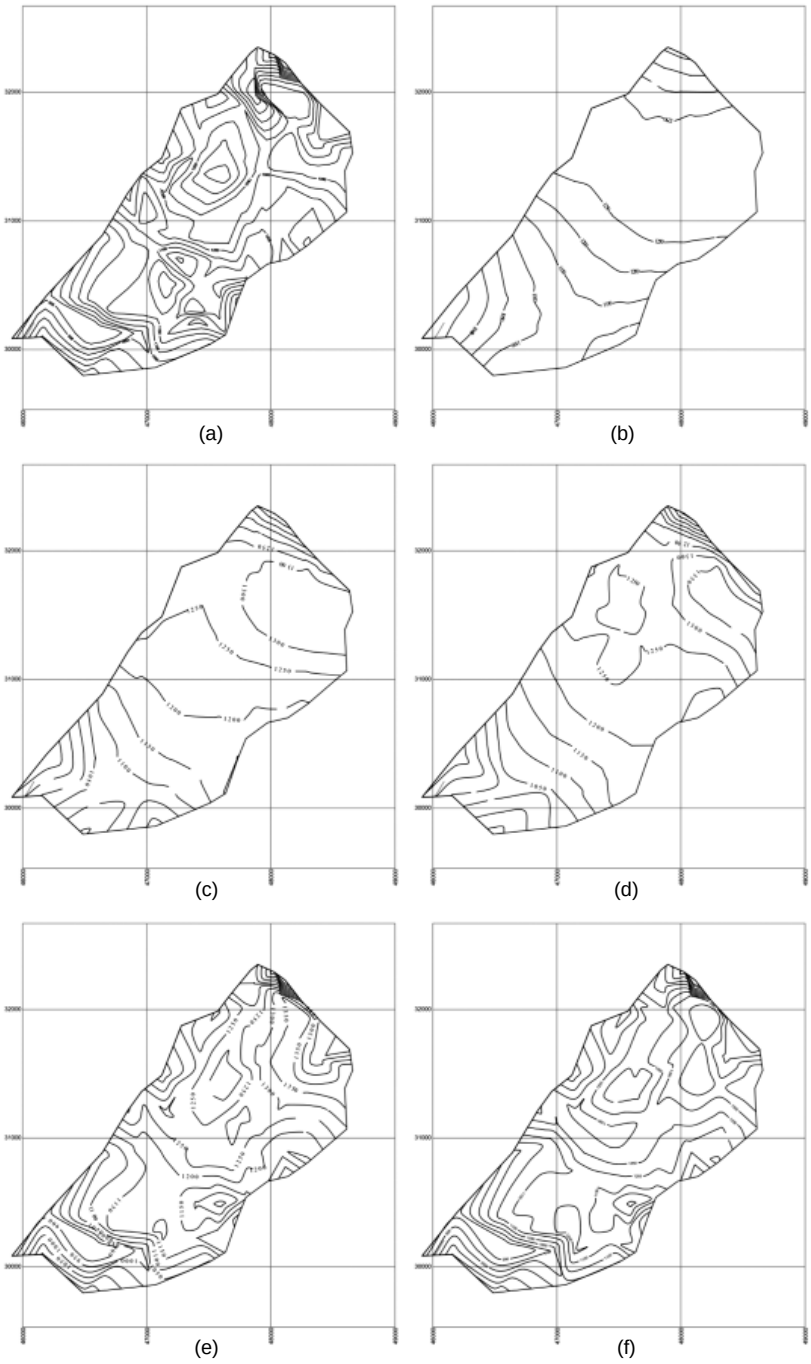


Fig. 5. Triangulated contour maps. (a) Raw calorie values; (b) $m = 2$, $n = 2$; (c) $m = 3$, $n = 3$; (d) $m = 4$, $n = 4$; (e) $m = 8$, $n = 8$; (f) $m = 5$, $n = 12$; the best-fitting approximation surface was obtained at this degree ($R^2 = 0.91$)

suggests that, as is often the case in geoscience, the location-dependent directional changes of random samples can determine the degree of the polynomial in the application of a polynomial approximation method.

## References

Campbell S.L., Meyer C.D., 1979. *Generalized inverses of linear transformations*. Pitmann Publishing, London.

Chayes F., 1970. *On deciding whether trend surfaces of progressively higher order are meaningful*. Geol. Soc. America Bull. 81,4: 1,273-1,278.

Dwyer P.S., Waugh F.V., 1953. *On errors in matrix inversion*. Jour. Am. Stat. Assoc., 48-262:289-319.

Howarth R.J., 2001. *A History of regression and related model-fitting in the earth sciences 1,636?-2,000*. Natural Resources Research 10-4: 241-286.

Karakus D., Safak S., Onur A.H., 2011. *A New Approximation Method For The Trend Hypersurface Analysis For Elevation Of Drilhole Data*. Arch. Min. Sci., Vol. 56, No. 1, p. 47-58.

Kendall M.G., 1946. *The advanced theory of statistics*. Griffin, London.

Krige D.G., 1960. *On the departure of ore value distributions from the lognormal model in South African gold mines*. Jour. South African Inst. Mining and Metallurgy, 64-4: 231-244.

Krumbein W.C., 1952. *Principles of facies map interpretation*. Jour. Sedimentary Petrology, 22-4: 200-211.

Krumbein W.C., 1956. *Regional and local components in facies maps*. Am. Assoc. Petroleum Geologists Bull. 40-9: 2,163-2,194.

Krumbein W.C., 1959. *The sorting of geological variables illustrated by regression analysis of the factors controlling beach firmness*. Jour. Sedimentary Petrology, 29-4: 575-587.

Matheron G., 1962. *Trait ´e de g´eostatistique appliqu´ ee*. 2 vols. Technip, Paris.

Matheron G., 1963. *Principles of geostatistics*. Econ. Geology, 58: 1246-1266.

Matheron G., 1965. *Les variables régionalisées et leur estimation: Une application de la théorie des fonctions aléatoires aux sciences de la nature*. Masson, Paris.

Norcliffe G.B., 1969. *On the use and limitations of trend surface analysis*. Can. Geographer, 13-4: 338-348.

Oldham C.H.G., Sutherland D.B., 1955. *Orthogonal polynomials: their use in estimating the regional effect*. Geophysics, 20-2: 295-306.

Pelletier B.C., 1958. *Pocono paleocurrents in Pennsylvania and Maryland*. Geol. Soc. America Bull., 69-8: 1,033-1,064.

Potter P.E., 1955. *The petrology and origin of the Lafayette Gravel: Part I.* Mineralogy and Petrology: Jour. Geology 63-1: 1-38.

Rao C.R., Mitra S.K., 1971. *Generalized inverses of matrices and their applications*. John Wiley, New York.

Schlee J., 1957. *Upland gravels of southern Maryland*. Geol. Soc. America Bull. 68-10: 1,371-1,410.

Soltani S., Hezarkhani A., 2009. *Additional exploratory boreholes optimization based on three-dimensional model of ore deposit*. Arch. Min. Sci., Vol. 54, No. 3, p. 495-506.

Tinkler K.J., 1969. *Trend surfaces with low "explanations;" the assessment of their significance*. Am. Jour. Science 267-1: 114-123.

Wang X.J., Zhang Z.P., 1999. *A comparison of conditional simulation, kriging and trend surface analysis for soil heavy metal pollution pattern analysis*. Journal of Environmental Science and Health, 34-1: 73-89

Wold H.O.A., 1949. *A large-sample test for moving averages*. Jour. Royal Stat. Society B11: 297-305.

*Received: 02 April 2012*