Computing Uncertainty in the Mean with a Stochastic Trend Approach

Martha E. Villalba and Clayton V. Deutsch

The use of a trend equation relaxes the assumption of stationarity in the mean and defines the mean dependent on the sample location in the domain. The mean is defined by an equation that contains coefficients. The research proposes to randomize the regression coefficients to evaluate the uncertainty in the global mean. The technique is simple to apply because it requires few input parameters. The technique considers the size of the domain and accounts for the data at their locations. A FORTRAN program (uncregcoef) is presented to implement the proposed stochastic trend technique.

1. Introduction

Estimates are made under uncertainty because there are few data; a number of input parameters are required to compute the estimates. The uncertainty in the input parameters has a large influence on uncertainty at large scales, in particular, global uncertainty. Decision making partially depends on a measure of uncertainty in the estimates (Rose, 2001). The uncertainty of the estimates must be evaluated accurately. In general, the most important parameter of a statistical model is the mean because it has a first order effect on the resources and reserves.

Techniques like the bootstrap and spatial bootstrap are questionable because they do not directly account for the conditioning data or the size of the domain being considered. The conditional finite domain approach accounts for the data and domain, but is tricky to apply because of the sensitivity on the translation (see paper 115 in this volume). Moreover, all of these techniques assume the mean is constant over the domain. The mean in most study areas exhibit some dependence on location. The use of a trend equation relaxes the strong decision of first order stationarity and defines the mean dependent on the sample location in the domain, see Equation (1). There are *L+1* trend or drift terms, a_l represents the unknown regression coefficients and the f_l (**u**) terms specify the form of the trend. The form of the trend may be linear, quadratic, sine/cosine, drawn by hand or some other arbitrary, but known functions.

$$z(\mathbf{u}) = \sum_{l=0}^{L} a_l \, f_l(\mathbf{u}) \tag{1}$$

In practice, linear and quadratic equations are considered, different equations could be used if the available data require better fitting. A FORTRAN code called *uncregcoef.for,* which is a modification of the *correlate* program, was implemented to develop a stochastic trend technique. The mean of the geological variable is defined by an equation that contains coefficients, then, uncertainty in the regression coefficients are used to obtain uncertainty in the global mean.

2. Methodology

The coefficients of the trend are calculated based on linear regression theory. The sum of the squared differences between the mean function and the available data are as small as possible. This difference is not zero because the data values fluctuate about their expected values. The method of least square selects the regression coefficients with the criteria of minimizing the sum of squared differences (Johnson & Wichern, 2007). The multivariate (L+1)-variate distribution of uncertainty in the regression coefficients can be understood from the statistics literature (see below). Many realizations of the coefficients are sampled and the mean over the area of interest is calculated. This approach will be referred to as a stochastic trend (ST) approach to compute the uncertainty in the mean. The details of the stochastic trend (ST) approach are as follow:

- Define the number of parameter or term for the trend equation; it could be linear or quadratic.
- Calculate the regression coefficients for the trend using least square method.
- Define the covariance matrix and correlation matrix for the regression coefficients.
- Perform Cholesky decomposition of the correlation matrix $\rho = LU$

- Sample independent normal Gaussian score values $W=G^{-1}(p)$ and correlate them by the use of the lower decomposed correlation matrix L, $Y = L\,W$.

- Non-standardize these values Y by the use of their respective fitted regression coefficient a_a and their respective standard deviation σ_a.

$$a^{sim} = a_a + \sigma_a \times Y \tag{2}$$

- Evaluate the Z(**u**) at the locations required with the polynomial. This should use the set of coefficients to obtain the mean of the realization (m_z).

- Repeat the previous three step many times; then, assemble the distribution of m_z to model the uncertainty in the mean. The sets of sampled coefficients will reproduce the correlation matrix.

The coefficients of the trend and their variance are calculated by using the theory of multiple regression models. Where, the well known linear regression model is defined by the equation matrix (3), this equation has as independent variables, all the parameters that affect the dependent variable **Z**. These variables could be any spatially distributed variable.

The variable **Z** represents a vector (*n,1*) of the available values, called dependent variable, where *n* is the number of observations or number of data available, *L= l+1* is the number of parameters, explanatory variables or predictor variabless that define the fitted equation of the model equation. **X** is an independent variable and is defined as the matrix of the number of observations and number of parameters, which dimension is (*n,*(*l+1*)). The geological data are located in three dimensions, the columns two, three and four of the matrix **X** correspond to their coordinates. **a** corresponds to a vector of the fitted equation coefficients or vector of regression parameters, with dimension ((*l+1*),*n*). ε corresponds to an error vector with dimension (*n,1*), this error vector is assumed to be a random part of the model equation that have a distribution of mean zero and variance s^2.

$$\underset{\mathbf{Z}}{\begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{bmatrix}} = \underset{\mathbf{X}}{\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1l} \\ 1 & x_{21} & x_{22} & \cdots & x_{2l} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nl} \end{bmatrix}} \underset{\mathbf{a}}{\begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_l \end{bmatrix}} + \underset{\boldsymbol{\varepsilon}}{\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}} \tag{3}$$

The regression analysis specifies how to predict the response **Z**. The location of the available data is translated to the **X** matrix according determined function polynomial; this function may reproduce the values corresponding to the available data. The least square fitting estimates the coefficient vector by:

$$\hat{\mathbf{a}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Z} \tag{4}$$

The fitted equation is given by the operation of the matrix *X* (5), which is the same as the first term of the general trend equation.

$$\hat{\mathbf{Z}} = \mathbf{X}\hat{\mathbf{a}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Z} \tag{5}$$

The residual vector (6) is calculated by the difference between the value of the available data and the expected value calculated based on fitted model equation. The variance of residuals (7) is the sum product of residual divided by the residual degree of freedom defined as the number of available data minus the number of fitted coefficients. Notice that coefficients $\hat{\mathbf{a}}$ and $\hat{\boldsymbol{\varepsilon}}$ are not correlated and the number of samples should be greater than number of parameters plus one.

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{Z} - \hat{\mathbf{Z}} \tag{6}$$

$$s^2 = \frac{\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}}}{n-(l+1)} \tag{7}$$

$$E\left\{s^2\right\} = \sigma^2 \tag{8}$$

The least square estimator (regression coefficients) has a multivariate normal distribution with mean vector and covariance matrix.

$$E\{\hat{\mathbf{a}}\} = \mathbf{a} \tag{9}$$

$$Cov(\hat{\mathbf{a}}) = \sigma^2(\mathbf{X}\mathbf{X}^T)^{-1} \tag{10}$$

Each location of the matrix (11) corresponds to the variance between two regressions coefficients; the diagonal of the matrix corresponds to the variances of the coefficients.

$$Cov(\hat{\mathbf{a}}) = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1L} \\ c_{21} & c_{22} & \cdots & c_{2L} \\ \vdots & \vdots & \ddots & \vdots \\ c_{L1} & c_{L2} & \cdots & c_{LL} \end{bmatrix} = \begin{bmatrix} \sigma^2{}_{11} & c_{12} & \cdots & c_{1L} \\ c_{21} & \sigma^2{}_{22} & \cdots & c_{2L} \\ \vdots & \vdots & \ddots & \vdots \\ c_{L1} & c_{L2} & \cdots & \sigma^2{}_{LL} \end{bmatrix} \tag{11}$$

The correlation matrix of regression coefficients is defined by dividing each term of the covariance matrix by their respective standard deviation of each coefficient.

$$\rho(\hat{\mathbf{a}}) = \begin{bmatrix} 1/\sigma_{11} & 0 & 0 & 0 \\ 0 & 1/\sigma_{22} & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 1/\sigma_{LL} \end{bmatrix} \times \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1L} \\ c_{21} & c_{22} & \cdots & c_{2L} \\ \vdots & \vdots & \ddots & \vdots \\ c_{L1} & c_{L2} & \cdots & c_{LL} \end{bmatrix} \times \begin{bmatrix} 1/\sigma_{11} & 0 & 0 & 0 \\ 0 & 1/\sigma_{22} & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 1/\sigma_{LL} \end{bmatrix}$$

$$\rho(\hat{\mathbf{a}}) = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1L} \\ \rho_{21} & 1 & \cdots & \rho_{2L} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{L1} & \rho_{L2} & \cdots & 1 \end{bmatrix} \tag{12}$$

The correlation matrix of the coefficients is important because the regression coefficients are necessarily correlated and cannot be drawn independently. The covariance matrix of the sample locations (C11) is used to condition the sampling in technique as in the spatial bootstrap. Both matrices are symmetric and are positive definite.

3. Application and challenges

A set of 11 synthetic data are located in along 100 meters in 1-D. The mean of the values is 2.4 and the variance 0.695. The trend of the data is fitted with a linear trend that considers two coefficients. The matrix **X** is the location of the data and has a dimension (11,2). The values of the data are the matrix **Z** with dimension (11,1). The least square method gives the coefficients through the following operation of matrices. Where, the first coefficient a_0 is 1.2403 and second coefficient a_1 is 0.0247.

$$\hat{\mathbf{a}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Z} = \begin{bmatrix} 1.2403 \\ 0.0247 \end{bmatrix}$$

The equation of the trend is defined by:

$$Z(\mathbf{u}) = 1.2403 + 0.0247x$$

The data location is replaced in that previous equation to calculate the residual and σ^2.

$$s^2 = \frac{\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}}}{n-(l+1)} = \frac{1.21}{11-(1+1)} = 0.134$$

The covariance is defined by:

$$Cov(\hat{\mathbf{a}}) = \sigma^2(\mathbf{X}\mathbf{X}^T)^{-1} = \begin{bmatrix} 4.19E-02 & -6.22E-04 \\ -6.22E-04 & 1.32E-05 \end{bmatrix}$$

The diagonal of the covariance matrix is the variance of each coefficient, then, each term is divided by their respective standard deviations to obtain the correlation matrix.

$$\rho = \begin{bmatrix} \rho_{11} & \rho_{12} \\ \rho_{21} & \rho_{22} \end{bmatrix} = \begin{bmatrix} 1 & -0.837 \\ -0.837 & 1 \end{bmatrix}$$

Note that the regression coefficients are negatively correlated – as they should be. When the intercept is high, then the slope must be lower and vice versa. The correlation matrix is symmetric and is positive-definite, then, the Cholesky decomposition is possible. The lower triangular matrix helps to correlate w_1 and w_2.

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \underbrace{\begin{bmatrix} \rho_{11} & 0 \\ \rho_{21} & \sqrt{1-\rho_{21}^{\,2}} \end{bmatrix}}_{\text{Lower matrix}} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -0.837 & \sqrt{1-(-0.837)^2} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

The realizations of the Gaussian coefficients Y_1 and Y_2 are non-standardized to get a_0 and a_1 in original units.

$$a_0 = a_{a_0} + \sigma_{a_0} Y_1 = 1.2403 + 0.2048 \times Y_1$$

$$a_1 = a_{a_1} + \sigma_{a_1} Y_2 = 0.0247 + 0.0036 \times Y_2$$

One hundred realizations are sampled the regression coefficients a_0 and a_1, those values show a correlation of -0.820 close to the input correlation of -0.837.

The one hundred realizations of the coefficients are used in the trend model to obtain values in $Z(\mathbf{u})$. The fluctuations of the trend gives one hundred means that are in function of their locations.

The stochastic trends give an uncertainty of the mean equal to 0.114 and the mean of 100 realizations is 2.414 close to the input mean of the data 2.409. The variance of the 11 values is 0.695, then; the uncertainty that consider independence in the sampling is 0.251. This value represents two times the uncertainty that is given by the stochastic trend approach.

A second example is evaluated to show the influence of domains in the uncertainty and show a complex model equation that describes the mean as a function of their locations in three dimensions. A lognormal gold data is used to evaluate the uncertainty in the mean. Forty nine lognormal values were assembled with coordinates at East, North and Elevation; values of gold are located in the grid of 200 × 200 meters in the horizontal direction and 20 meters in the vertical direction. The mean of the value is 1.10 g/t and the standard deviation is 1.4. The evaluations consider three different domains:

- Pessimistic criteria, some data locate close to the limits of the domain and the distance from the data to the limits of the domain is equivalent to values less than the half distance between samples.

- Average criteria, where the distance from the data to the limits of the domain is equivalent to the half distance between samples.

- Optimistic criteria, where the distance from the data to the limits of the domain is equivalent to almost twice or more than the distance between samples.

The uncertainty of the gold values at this example is calculated by a quadratic trend of 10 coefficients. The evaluation also considers a lineal trend of 4 coefficients to observe the sensitivity of the number of coefficients in the uncertainty.

$$Z(\mathbf{u}) = a_0 + a_1 x + a_2 y + a_3 z + a_4 x^2 + a_5 y^2 + a_6 z^2 + a_7 xy + a_8 xz + a_9 yz$$

$$Z(\mathbf{u}) = a_0 + a_1 x + a_2 y + a_3 z$$

The variables *x*, *y* and *z* are the coordinates of the vector location **u.** These values are transferred to the matrix **X** to obtain the regression coefficients, $\mathbf{a} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Z}$. The vector matrix **Z** contains the 49 values. These coefficients are evaluated for both function polynomials that define the trend:

$$Z(\mathbf{u}) = 8.554 - 0.0029x + 0.0685y - 0.7056z + 0.0001x^2 + 0.0003y^2 + 0.0092z^2 - 0.0008xy + 0.0024xz - 0.0017yz$$

$$Z(\mathbf{u}) = 5.403 - 0.0008x + 0.004y - 0.153z$$

The covariance matrix and the correlation matrix of the coefficients are calculated in each case. These correlation matrices are decomposed to obtain the lower matrix. Lower matrices of (10×10) and (4×4) are used to correlate 100 times the random Gaussian values. One hundred realizations of the coefficients are non-standardized to get values in original units. These values are replaced in the function polynomials to estimate the grade in each node of the respective domain. The evaluations show that quadratic trend reproduce the tails of the input values, then, the uncertainty with quadratic trend is greater than linear trend.

Evaluation of uncertainty of gold values in a pessimistic domain gives an uncertainty of 0.227 (std) using lineal polynomial and 0.515 (std) using quadratic polynomial. Those values are greater than the uncertainty $1.4/\sqrt{49} = 0.20$ that assume independence of the data. The uncertainty increases as the domain increase. The optimistic domain obtains an uncertainty of 0.234 using lineal trend and 0.721 using quadratic trend.

4. Conclusions

The proposed technique is simple to apply because it does not require covariance model of the data and needs few input parameters. This technique considers the size of the domain and assumes that the mean is dependent on location; however, the form of the mean equation could be a challenge. Uncertainty in the mean with this technique is very sensitive to the choice of the form of the mean.

References

Deutsch, C. V. (2002). *Geostatistical Reservoir Modeling.* New York: Oxford University Press.

Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation.* New York: Oxford University Press.

Johnson, R. A., & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis.* New Jersey: Person Education, Inc.

Rose, P. R. (2001). Risk Analysis and Management of Petroleum Exploration Ventures. *AAPG Methods in Exploration Series* , 5-10.

Sathe, S. T., & Vinod, H. D. (1974). Bounds on the Variance of Regressions coefficients due to Heteroscedastic or Autoregressive Errors. *The Econometric Society , 42*, 33-340.

Watson, G. S. (1955). Serail Correlation in Regression Analysis. *Biometrika , 42*, 327-341.

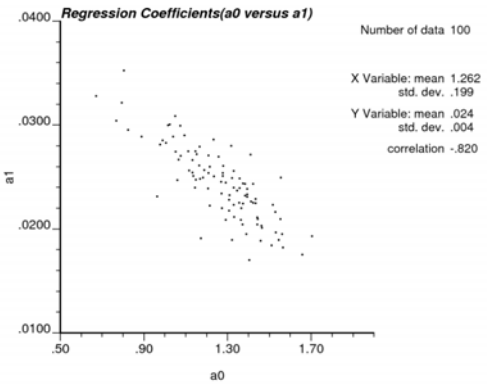

Figure 1: The correlation of the realizations of the regression coefficients a0 and a1.

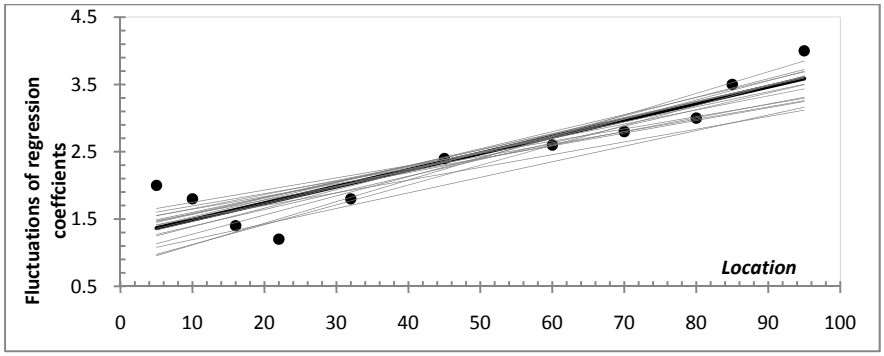

Figure 2: Realizations of the trend model of the section, where the black points are the data; the solid black line is the linear trend model with two parameters; and the grey lines are the stochastic trends.

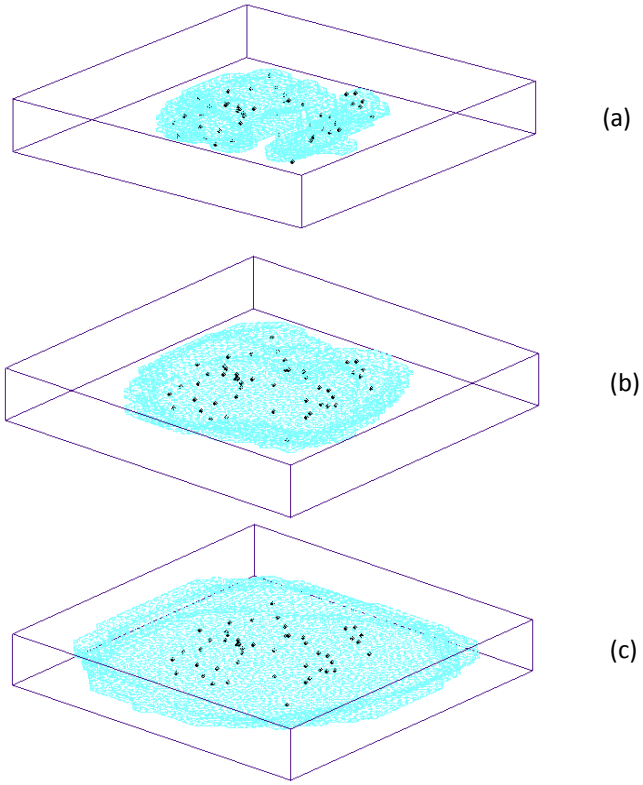

Figure 6.3: The data are illustrated with black points and the domain in 3D solid, where (a) shows the pessimistic criteria used to model the domain, (b) shows the average criteria and (c) shows the optimistic criteria.