Integrated GIS - Geostatistics System for Environmental Modelling

Nguyen Chi Quang¹ and Dinh Lan Anh²

Abstract

GIS and Geostatistics can provide powerful space analytical tools that result in quantitative characterization the extent and spatial distribution of environmental data. This paper provides an approach, which allows answering the question of "How one might interpolate in order to make predictions (or estimates) at points where measurements of environmental parameters are not available?". Additionally, this approach may help a researcher to determine a single representative value for an area that is represented by several measured or estimated values or both. Remediations and solutions for a more desirable environment therefore can be prioritized accordingly. The objectives of this paper are: to provide principal Integrated GIS - Geostatistical methodology for environmental modelling and to illustrate the methodology by case studies, which were actually carried out in coal mining zones in Quang Ninh Province, Vietnam.

1. Introduction

Analysis of environmental parameters contamination requires a reliable, quantitative understanding of natural conditions. Uncontrolled variability complicates the analysis of data on environmental contamination and increases the risk of incorrect data interpretation. As a consequence, modeling has become an increasingly important tool for evaluating complex environment problems.

2. Setting the problems

In 1999, a study group of the IT&E, Inc. collected dust, water, noise and soil sample within a coal mining zone of Cam Pha, Quang Ninh province, Vietnam. This zone is 5,749.30 km² in area, and located from 20°50'00" to 21°15'00" North latitude and from 106°50'00" to 107°25'00" East longitude. The area includes 15 open pit and 10 underground coal mines (Fig. 1).

¹ Ph.D, Institute for Environmental Science and Technology (INEST)
² B. Sc., Hanoi University of Technology (HUT)
These samples shown some significant environmental problems, especially for dust concentration. Subsequently, the IT&E, Inc. was assigned its services to help VINACOAL evaluate the data. Case study, which is presented in this section, is carried out to deal with the most serious environmental problem of dust pollution in Cam pha. Total 77 samples of dust were taken and shown on the map (Fig. 2).

3. Methods of Data Analysis

Geostatistics are statistical methods used to describe spatial relationships among sample data and to apply this analysis to the prediction of spatial and temporal phenomena. They are used to explain spatial patterns and to interpolate values at unsampled locations. Geostatistics have traditionally been used in the sphere of geosciences: meteorology, mining, soil science, forestry, fisheries, remote sensing, and cartography. Geostatistical techniques were originally developed by Vietnam scientists for geological data analysis. Geostatistical techniques were later successfully applied to mining and other disciplines.

Deterministic approaches to interpolation (e.g. trend surface, inverse distance weighting, triangulation, and splining) are based upon a priori mathematical models of spatial variation and can lead to smooth but inaccurate maps. This is because error is an inherent part of the sampling process. In practice, error can not be eliminated but only minimized. Therefore, in most cases one cannot produce a representative map of estimated values in unsampled locations with these techniques.

Geostatistical processes are comprised of the three major components of regionalized variable analysis: variographical analysis, prediction making, and error analysis. During structural analysis, spatial (auto)correlation can be analyzed using covariance and variogram. After structural analysis, predictions at unsampled locations are made using kriging (i.e. transposition of multiple linear regression into a spatial
context). For producing risk-qualified maps, the nonparametric geostatistical algorithms such as indicator kriging, indicator cokriging, and probability kriging are primarily recommended. They were developed to process data of highly variant phenomena without having to trim off important high-valued data.

In the article we used the following geostatistical approaches: simple, ordinary, indicator, probability and disjunctive kriging as well as ordinary and indicator kriging modifications for environmental data.

**Simple kriging** (SK) (in this method the estimation of the mean can be established *a priori* based upon a different data set from the data used for the present estimation; for example, declustering mean) and **ordinary** (with unknown mean) krigings (OK) (Gandin, 1963) can be applied successfully to data having, or close to, a multivariate Gaussian distribution. The assumption of the multivariate Gaussian distribution is rarely realized in practice. To check for the multivariate normal distribution of the data, one should first transform the initial data to a univariate normal distribution (i.e. performing linearization of the original data). Such transformation performed prior to the variogram/covariance analysis allows one to reduce the variability in the original data, making the variogram modeling more reliable and stable. The normal score transform function can be defined for any continuous cumulative distribution function. Transformation of the distribution of the initial data set into univariate Gaussian distribution does not guarantee that all of the initial data will be transformed into an exact multivariate Gaussian distribution. It is, however, relatively easy to check for bivariate normality of the transformation (Deutsch, Journel, 1998). If bivariate distribution is fulfilled, transformation could be adopted. Simple and ordinary kriging provide an optimal estimation in the class of linear models. Suggestion about normality of the transformed data allows one to calculate the confidence interval for the estimation and to present results of the estimation as a map with the desirable level of significance and as a map of the probability that selected critical level is exceeded or not, in addition to the map of estimation.

**Indicator kriging** (IK) (Journel, 1988) is an example of the techniques in which the uncertainty model depends only on the information available. In indicator kriging a 0-1 transformation is introduced, which will be the new variable for which we will compute variograms and carry out kriging methods. Among the shortcomings of indicator kriging is the loss of information after coding data through indicator functions. For example, if the data values are in the range of 1 to 100 and the indicator value (cutoff) selected is 20, then the data points with values 21 and 99 will be interpreted as being equivalent. It is possible, therefore, in such situations to use a soft indicator function. This indicator function can be prepared by the user based upon knowledge of the data sets and processing by ordinary kriging. One of the possible solutions is to remove data near the threshold from consideration.

**Probability kriging** (PK) (Journel, 1988) kriging is considered an improvement over the indicator kriging method in the sense that the data is used more completely
for the task of estimating conditional probabilities that given thresholds are exceeded. This nonparametric approach uses the original data set in addition to indicators.

**Disjunctive kriging (DK)** (Rivoirard, 1994) is also based on a specific nonlinear transformation of the original data. The first step in this type of estimator is the normal score transformation of the data. The next step is to express the normalized function as an expansion of hermitical polynomials. It should be noted that a sufficient number of hermitical polynomials depends on searching neighborhoods and to receive reliable results of estimation we should use a different number of hermitical polynomials for different locations. It is also assumed that the condition of bivariate standard normal distribution is fulfilled. It should be noted that bivariate normal distribution is weaker than the multi-Gaussian condition. As mentioned above, it is possible to check the bivariate normality of the transformation. If bivariate normality exists, disjunctive kriging can produce better results than other geostatistical methods. If it is not so, it is desirable to use another type of kriging. Validation and cross-validation techniques can be used to confirm this.

**Environmental applications** Geostatistical methods were initially developed in the 1960’s for the specific problem of estimating resources in the mining industry. However the very idea on which they are founded means that they are equally applicable in other earth science industries like the environment. The correlation between sample grades used in mining geostatistics to estimate ore resources is equivalent to using the correlation between soil pollutant concentrations in the environment to obtain the best pollutant map from which to evaluate remediation costs.

The Integrated GIS - geostatistical System includes space estimation and simulation algorithms. It thus provides all the tools needed to obtain an accurate image of the state of the environment from which the most appropriate strategy can be adopted. Typical problems to be solved include:

- defining optimal sampling strategies based on 2 or 3D variographic analysis
- mapping pollutants using linear kriging
- estimating volumes of material exceeding a critical threshold using non-linear techniques
- performing risk analysis/assessment studies, based on conditional simulations.

This study shows that geostatistics can be used as an important aid when defining the environmental management strategy for a mine site in Vietnam.

3. **Statistical assessment**

From comprehensive database of VINACOAL, data on dust ware generated into set for the first step of statistical testing prior to modeling procedure. This step was carried out to define basic statistical parameters such as number of samples, mean, min, max, variance, standard error, rejected sample, etc and graphs and distribution func-
tions of dust concentrations. The assessment was rejected one sample of dust because its concentration value is unspecific (extremely high in this case) (Figure 3).

Distribution function of dust [equation (1)]:

\[
y = 72 \times 100 \times e^{\frac{x-a}{b}} \times e^{-e^{\frac{x-a}{b}}} \\
- \infty < x < \infty; \quad a = 45.52; \quad b = 78.86
\]

\[y = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(x-c)^2}{2}} \, dx\]

\[x \rightarrow \infty, \quad c = 100, \quad \sigma = 100\]

Figure 3: Statistical analysis of dust (Number of samples: 72, Mean: 102.51, Min: 0.58, Max: 650, Variance: 22830.89, Standard error: 17.81, Rejected sample (unspecific value): 1 (C = 1340 mg/m³))

4. Modeling the local uncertainty

Objectives of modeling the local uncertainty are to estimate values of dust or noise at unsampled points and estimate the probability of being no greater than any \( z \) at \( u \). To achieve these objectives, appropriate formulas were used [equation (2)]:

Estimated value = linear combination of neighboring data; that mean:

\[z \ast K(u) = \sum_{a=1}^{n} \lambda_{a} \ast z(u_{a})\]

Of which, weights \( \lambda_{a} \) are such that:

- Estimator is unbiased [equation (2.1)]:
  \[E\{Z \ast K(u) - Z(u)\} = 0\]  \hspace{1cm} (2.1)

- Estimation variance is minimum [equation (2.2)]:
  \[\sigma_{K}^{2}(u) = \text{Var}\{Z \ast K(u) - Z(u)\}\]  \hspace{1cm} (2.2)

Model the conditional cumulative distribution function (ccdf) by equation (3):
\[ F(u; z(n)) = \text{Prob}\{Z(u) \leq z(n)\} \forall n \] (3)

Using ordinary Kriging approach, interpolated dust and noise contamination maps were established. The ordinary Kriging formula used is equation (4):

\[ F(u; z(n)) = E\{(u); z(n)\} \] (4)

These maps provide clear pictures about current situations of dust in Campha coal mining zone. They show where dust concentrations (in mg/m³) are high and where are low at a certain time (Figs. 4). However, these pictures sometimes not fulfill the needs of researcher because they only provide numerical estimated value of the interest parameters at unsampled points at a certain time. The question of "Is U polluted?" is answered already but the other one of "Should we take remediation at U?" is still remained.

Figure 4a: Variogram analysis of dust

Figure 4b: Kriging model of dust contamination
Figure 4c: Statistical analysis of dust interpolated data (No of samples: 2641, Mean: 129.13, Min: 0.64, Max: 739.00, Variance: 27768.65, Standard error: 3.24)

Figure 4: Geostatistical Interpretation and modelling

Indicator Kriging approach is well suited for the second question. Instead of estimating value of the parameters, we estimate their probabilities. It means that indicator Kriging formula is used [equation (5)]: Probability = expected value of an indicator:

\[
F(u;z(n)) = E\{I(u);z(n)\}
\]

With \( I(u;z) = 1 \) if \( Z(u) \leq z \), 0 otherwise

Details of indicator Kriging approach practice were carried out by steps. Firstly, we discretize the range of \( Z \) using \( K \) thresholds \( Z_k \) as follows:

Code each observation \( z(u) \) into an indicator of exceedence of \( Z_k \). Note that no uncertainty at sampled locations, e.g. \( z(u_1) = 0.5 \) mg/m\(^3\) \( \Rightarrow \) \( i(u_1;0.8) = 1; \)
\( z(u_2) = 1.1 \) mg/m\(^3\) \( \Rightarrow \) \( i(u_2;0.8) = 0. \)

- Compute a variogram from indicator data and model it by equation (6):

\[
\gamma(h; z_k) = \frac{1}{2N(h)} \sum_{\alpha=1}^{N(h)} [I(u_{\alpha}; z_k) - i(u_{\alpha} + h; z_k)]^2
\]

- Estimate the local probability at \( u \) by Kriging of indicator values by equation (7):

\[
\text{Prob}\{Z(u) \leq z_k \} = \sum_{\alpha=1}^{n} \lambda_{\alpha} \times i(u_{\alpha}; z_k)
\]

Secondly, decisions were made in the face of uncertainties. In principle, there are many ways to give decision. For example, we can decide based on the exceedence
of a probability threshold. It means \( U \) is classified as contaminated if equation (8) is fulfilled:

\[
\Pr \{ Z(u) > z_c \mid (n) \} > p_c
\]  

(8)

Where the choice of \( p_c \) ensures that: (1) \( p_c \) is marginal probability of contamination and (2) threshold may vary with the current land use.

Other way, decision could be given based on the exceedence of a physical threshold. It means \( U \) is classified as contaminated if equation (9) is fulfilled:

\[
Z^*_{\text{L}}(u) > Z_c
\]  

(9)

For this case, we must care of two things:

1. **Choice of concentration estimate \( Z^*_{\text{L}}(u) \):**
   - The first criterion (equation 9.1) is consideration of \( p - \) quantile of the ccdf:
     \[
     Z^*_{\text{L}}(u) = F^{-1}(u; p_{\mid (n)})
     \]  
     (9.1)
   - The second criterion (eq. 9.2) is consideration of expected value of the ccdf:
     \[
     Z^*_{\text{L}}(u) = Z^*_E(u) \quad (E\text{-Type estimate})
     \]  
     (9.2)

2. **Associated risks of misclassification:**
   - Risk \( \zeta(u) \) of classifying wrongly \( u \) as hazardous (equation 9.3):
     \[
     \zeta(u) = \Pr \{ Z(u) \leq Z_c \mid \{ Z^*_{\text{L}}(u) > Z_c(n) \} \} = F[u; Z_c \mid (n)]
     \]  
     (9.3)
   - Risk \( \zeta(u) \) of classifying wrongly \( u \) as safe (equation 9.4):
     \[
     \zeta(u) = \Pr \{ Z(u) > Z_c \mid \{ Z^*_{\text{L}}(u) \geq Z_c(n) \} \} = 1 - F[u; Z_c \mid (n)]
     \]  
     (9.4)

Also, minimization of an expected cost could be used as a base for decision-making. For this case, the decision that minimizes the costs based on the evaluation of the economical impacts of alternatives will be taken. Requirements for this case concerned with probability distribution and economical functions, which measuring the costs on principles of: (1) overestimation of concentrations of dust and noise is unnecessary cleaning, and (2) underestimation of concentrations of dust and noise is possible ill health.
Figure 5: Indicator spatial distribution of dust. In this paper, this step of making-decision was illustrated by the first kind - decide based on the exceedence of a probability threshold (Figs 5, 6).

Threshold 427.6 mg/m³  Threshold 10 mg/m³

Threshold 150 mg/m³  Threshold 30 mg/m³

Threshold 250 mg/m³  Threshold 100 mg/m³

Figure 6: Indicator Variogram analysis of dust.

5. Indicator simulation

This is the application of the general sequential simulation method to the case of an indicator function, or more generally of several nested indicators. The method can easily be extent to nested indicators with increasing thresholds of an RF Z(x)

\[ I_i(x) = 1 \text{z}(x)<z_i \quad \text{where } z_1 < z_2 < \ldots < z_m \]

Threshold 100 mg/m³

\[ I(x) = 1 \Rightarrow I_j(x) = 1 \quad \forall j < 1 \]

Nested indicators satisfy the characteristic property:

(10)

In the case where the indicators are defined by (11), Y is simply a discrete version of Z: Y(x) = i means that \( z_i \leq Z(x) < z_{i+1} \) (with the convention that \( z_0 = -\infty \) and \( z_{m+1} = +\infty \)). If Z is continuous, Y varies by unit jumps, like a discrete diffusive random function. In the other cases Y can vary by larger jumps.
The local information can be the value taken on by an auxiliary variable at this point \( x_a \). The spatial distribution of the simulations generated by the sequential indicator algorithm are not yet known, except in special cases. In the case of conditional simulations, it depends on the spatial distribution of the conditioning RF and has no reason to be identical to it.

Sequential indicator simulation is a very flexible method. Its use requires some approximations. A shortcoming, mainly when nested indicators are simulated, is that a result is obtained even if the model defined by the indicator direct and cross-covariances is consistent (Figs. 7).

The estimates, obtained directly from the kriging algorithm, are in the form of distributions. These distributions provide considerable information about the unknown
dust concentration. Moreover, the expected value, any quantile, any confidence interval or any cumulative probability of the unknown dust can be readily estimated from these distributions. The about presented results have been selected from the data set to illustrate the calculations and the main features of the various estimators. In addition, the various types of estimates mapped for the entire mining site are presented also.

6. Conclusion

Uncertainty due to complexity of natural phenomena is critical to assess and incorporate this uncertainty in decision-making processes.

Integrated GIS - Geostatistics System allows one to:

- Characterize spatial patterns of quantitative/qualitative attributes
- Investigate scale-depend correlations between attributes
- Predict unsampled attributes using various types of information (hard and soft) defined on different supports
- Assess the uncertainty at unsampled locations
- Incorporate the uncertainty in decision-making processes
- Simulate the spatial distribution of attributes to evaluate impact of different scenarios

Space interpolation is a successful method for interpolating environmental data. However Space Interpolation does not scale well which further motivates the need for optimized solutions, which make the best use of the available resources.

References


