13th International Symposium on Mineral Exploration

Toward Sustainable Society with Natural Resources – Frontiers in Earth Resources Technologies and Environmental Conservation

Hanoi, 22-24 September 2014

VIETNAM NATIONAL UNIVERSITY PRESS, HANOI

Proceedings of
13th International Symposium on Mineral Exploration
(ISME-XIII)

Toward Sustainable Society with Natural Resources
- Frontiers in Earth Resources Technologies and
Environmental Conservation -

2679527

Proceedings of 13th International Symposium on Mineral Exploration
Hanoi, Vietnam, September 22 - 24, 2014

A Survey into Multiquadric Interpolation of Categorical Data

J. K. Yamamoto[1], **K. Koike**[2]

[1]*Institute of Geosciences, University of Sao Paulo, Sao Paulo, Brazil.*
[2]*Graduate School of Engineering, Kyoto University, Kyoto, Japan.*

Corresponding author: jkyamamo@usp.br

ABSTRACT: Interpolation parameters for mulquadric interpolation are investigated in this paper. Both the number of neighbor data and the shape parameter C have a strong influence on interpolation results. Actually, for multiquadric interpolation of categorical data we have to use small number of neighbor data and a tiny shape parameter as well.

Key words: categorical data, multiquadric equations, smoothing

1. Introduction

Categorical data represent an important source of geological information. It is commonly used for interpolation of geological models. Categorical variables cannot be handled directly by numerical methods, but only after transformation into indicator functions as follows. Given a categorical variable composed of K types, the indicator function for the kth type is:

$$I(x_i;k)=\begin{cases}1 \text{ if type } k \text{ is present at location } x_i \\ 0 \text{ if type } k \text{ is not present at location } x_i\end{cases}$$

2. Multiquadric interpolation

As the most suitable interpolation method, Yamamoto et al. (2012, p. 148) proposed to use multiquadric equations (Hardy, 1971, p. 1907) instead of indicator kriging. The indicator function for the k^{th} type can be interpolated at an unsampled location x_o based on n neighboring indicator values as (Yamamoto et al. 2012, p. 148):

$$I^*(x_o;k)=\sum_{i=1}^{n} w_i I(x_i;k) \qquad (1)$$

The weights $\{w_i,\ i=1,\ldots,n\}$ come from the solution of a system of linear equations (Yamamoto et al. 2012, p. 148):

$$\begin{cases}\sum_{j=1}^{n} w_j\phi(x_j-x_i)+\mu=\phi(x_o-x_i) \text{ for } i=1,n \\ \sum_{j=1}^{n} w_j=1\end{cases}$$

For interpolation of indicator functions the multiquadric kernel gives good results. The multiquadric kernel is $\phi(x)=\sqrt{|x|^2+C}$, where $|x|$ is the norm of a vector in $\mathbf{R}^n$ and C is a positive constant. This constant is known as shape parameter and the accuracy of multiquadric interpolation depends on this parameter (Bayona et al., 2011, p. 7384-7385).

The variance associated with the k^{th} interpolated type can be computed as (Yamamoto et al. 2012, p. 149):

$$S_o^2(x_o;k)=\sum_{i=1}^{n} w_i\left[I(x_i;k)-I^*(x_o;k)\right]^2 \quad (2)$$

For every unsampled location x_o we have K interpolated indicator functions (1) and K variances (2). The most likely type $k_{\max}$ at x_o is given by (Teng & Koike, 2007, p. 533):

$$I^*(x_o;k_{\max})=\max\left(I^*(x_o;k),k=1,\ldots,K\right) \quad (3)$$

Therefore, the variance associated with the most likely type is also the maximum variance:

$$S_o^2(x_o;k_{\max})=I^*(x_o;k_{\max})\left(1-I^*(x_o;k_{\max})\right) \quad (4)$$

In order to visualize uncertainties on the interpolated model, Yamamoto et al. (2012, p. 151) defined the uncertainty zone in which variances are greater than 0.20 (Figure 1). Actually, Yamamoto et al. (2012, p. 151) proposed overlapping the uncertainty zone with most likely types. Locations presenting variances greater than 0.20 are considered uncertain and as shown by Yamamoto et al. (2012, p. 151) they are close to geologic contacts.

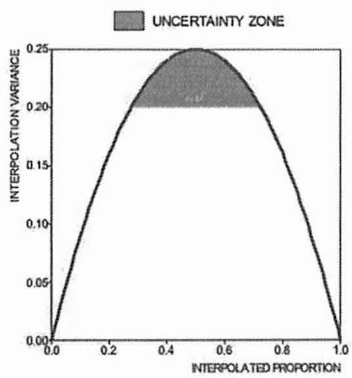

Fig 1. Uncertainty zones defined for interpolation variances greater than 0.20.

3. Materials and methods

As we know there are two important parameters for multiquadric interpolation: the constant C and the number of neighbor data. In order to study the effect of both parameters on interpolation results we considered using a stratified random sample with 40 data points (Figure 2).

The sample set in Figure 2B was used to study the effect of both neighbor data points and the constant C in multiquadric equations. Since we have 40 data points, the number of nearest neighbor data points varies from 4 to 20 in steps of 4.

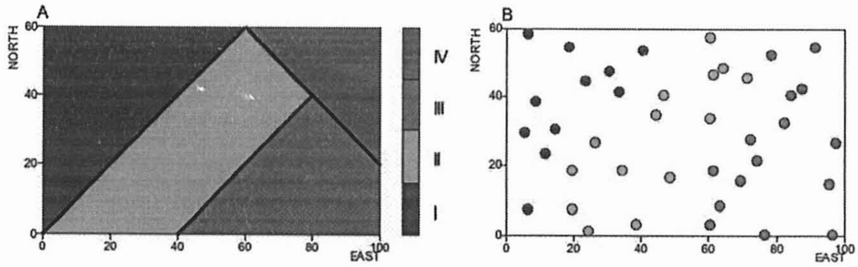

Fig 2. Synthetic model composed of 100 by 60 points presenting a categorical variable with four types (A) and stratified random sample (B).

4. Results and discussion

For categorical data interpolation we must work with a minimum number of neighbor data points. The effect of value C is investigated by changing its value from 10 to 1000 and varying the number of neighbor data points as mentioned before. Results are presented in Figure 3.

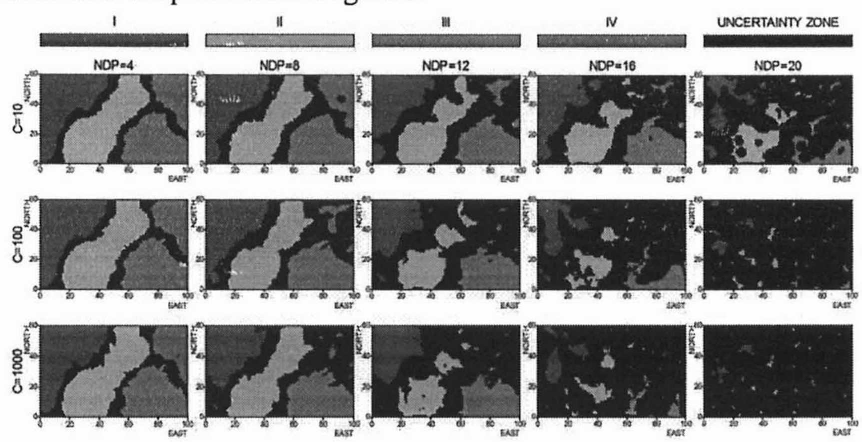

Fig 3. Resulting images of most likely types and uncertain types after multiquadric interpolation with varying C and number of neighbor data points from 4 to 20. NDP means number of neighbor data points.

Increasing the number of neighbor data points means decreasing estimated probabilities. However, the constant C has a strong negative effect on resultant multiquadric interpolated images. A number of neighbor data points greater than or equal to 12 results in deteriorated images. For example, the type IV starts vanishing with 12 neighbor data points and increasing constant C. It becomes worse with increasing the number of neighbor data points.

We know that increasing the number of neighbor data is the cause of smoothing for continuous data interpolation. However, smoothing for categorical variables is different from continuous data because when we take the most likely type (expression 3), the smoothing vanishes. For categorical variables we have to examine the smoothing effect under the view of global variances. The global variance for all K types can be computed as (Kader and Perry, 2007):

$$\mu_2 = \sum_k^K p_k(1-p_k)$$

where proportion (p_k) is simply $p_k = \dfrac{f_k}{M}$ and M is the number of interpolated locations.

We computed the global variances for varying number of neighbor data points and for different constants, according to results displayed in Figure 4.

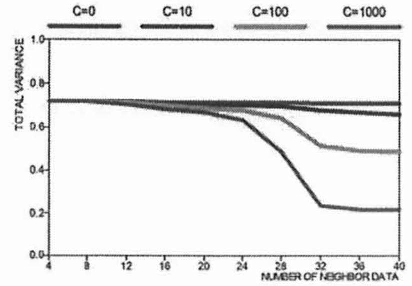

Fig 4. Total variance according to the number of neighbor data and constant C

Figure 4 shows that for a C=0 there is no variance reduction even with increasing the number of neighbor data. Nevertheless, with C greater than zero smoothing effect becomes conspicuous with variance. Actually, the higher C the lower is the global variance. Moreover, the smoothing in multiquadric interpolation is caused by the combination of larger number of neighbor data and larger constants.

5. Conclusions

In this paper we presented the results of a survey into multiquadric interpolation for categorical data interpolation. The main conclusion of this survey is related to the use of a small number of neighbor data and a minimum constant C.

Acknowlegments: We want to acknowledge CNPq – Conselho Nacional de Desenvolvimento Científico e Tecnológico – for funding this research.

REFERENCES

Bayona, V.; Moscoso, M.; Kindelan, M. 2011. Optimal constant shape parameter for multiquadric based RBF-FD method. J. Comput. Physics, 230, p. 7384-7399.

Hardy, R.L. 1971. Multiquadric equations of topography and other irregular surfaces. J. *Geophysical Research* (76) 1905-1915.

Kader, G.D.; Perry, M. 2007. Variability for categorical data. *J. of Statistics Education*, v. 15 (no. 2) downloaded from http://www.amstat.org/publications/jse/v15n2/kader.html on June 22nd, 2010.

Teng, Y.; Koike, K. 2007. Three-dimensional imaging of a geothermal system using temperature and geological models from a well-log data set. *Geothermics*, 36: 518-538.

Yamamoto, J.K.; Mao, X.M.; Koike, K.; Crosta, A.P. Landim, P.M.B.; Hu, H.Z.; Wang, C.Y.; Yao, L.Q. 2012. Mapping an uncertainty zone between interpolated types of a categorical variable. *Comput. & Geosci*, 40: 146-152.