Computers & Geosciences 40 (2012) 146–152

Contents lists available at SciVerse ScienceDirect

Computers & Geosciences

journal homepage: www.elsevier.com/locate/cageo

Mapping an uncertainty zone between interpolated types of a categorical variable

J.K. Yamamoto[a], X.M. Mao[b,*], K. Koike[c], A.P. Crosta[d], P.M.B. Landim[e], H.Z. Hu[b], C.Y. Wang[b], L.Q. Yao[b]

[a] *Department of Environmental and Sedimentary Geology, University of Sao Paulo, Sao Paulo, Brazil*
[b] *College of Water Conservancy and Civil Engineering, China Agricultural University, No. 17 Qinghua East Road, Beijing 100083, China*
[c] *Graduate School of Engineering, Kyoto University, Kyoto, Japan*
[d] *Department of Geology and Natural Resources, University of Campinas, Campinas, Brazil*
[e] *Department of Applied Geology, State University of Sao Paulo, Rio Claro, Brazil*

ARTICLE INFO

Article history:
Received 11 December 2010
Received in revised form
10 September 2011
Accepted 12 September 2011
Available online 22 September 2011

Keywords:
Categorical variable
Indicator function
Indicator kriging
Interpolation variance

ABSTRACT

Categorical data cannot be interpolated directly because they are outcomes of discrete random variables. Thus, types of categorical variables are transformed into indicator functions that can be handled by interpolation methods. Interpolated indicator values are then backtransformed to the original types of categorical variables. However, aspects such as variability and uncertainty of interpolated values of categorical data have never been considered. In this paper we show that the interpolation variance can be used to map an uncertainty zone around boundaries between types of categorical variables. Moreover, it is shown that the interpolation variance is a component of the total variance of the categorical variables, as measured by the coefficient of unalikeability.

1. Introduction

Collecting field samples at certain locations, and then using the information gathered in a field campaign to produce maps showing the spatial distribution of analyzed variables is a common procedure in Earth sciences. Quite often the nature of the variables being analyzed is categorical. For example, a soil survey usually comprises the collection of soil samples and the assignment of soil types to the samples. In this case, soil types are categorical variables or, in statistical terms, discrete random variables.

In order to produce maps based on categorical variables it is necessary to interpolate values for unsampled locations in between collection points. Even when we code categorical data as numbers, varying from 1 to the number of categories or types, we cannot use this information to interpolate a value for unsampled locations. This problem requires coding available information as indicator functions that can be interpolated and backtransformed to the original types of categorical variables (Koike and Matsuda, 2005; Teng and Koike, 2007; Leuangthong et al., 2008).

After transforming categorical data into indicator functions, the indicator kriging approach (Journel, 1983) can be used to estimate values for unsampled locations. This approach requires calculating and modeling a number of experimental semivariograms equal to the number of types of categorical variables (Leuangthong et al., 2008). However, when there are types of variables presenting low proportions this approach is almost impossible in practical terms. The indicator semivariograms in these cases are based on a few possible pairs and they might present statistical fluctuations.

Furthermore, after the interpolation and the generation of maps showing the distribution of the categorical types, the uncertainty zones that represent the boundaries between adjacent zones of different categorical types need to be properly established. This is usually done by defining buffer zones. However, these buffer zones are defined arbitrarily, using a constant or a variable distance, in a procedure known in geographic information systems (GIS) as "proximity analysis" (Star and Estes, 1990).

In this paper we propose the use of multiquadric equations (Hardy, 1971) that do not depend on semivariogram models. Thus, indicator functions will be interpolated for unsampled locations using multiquadric equations. Interpolated indicator values are backtransformed into original and mutually exclusive types of categorical variables (Teng and Koike, 2007). In this way

* Corresponding author. Tel.: +86 10 82373119; fax: +86 10 62737796.
E-mail addresses: jkyamamo@usp.br (J.K. Yamamoto), maoxiaomin@tsinghua.org.cn (X.M. Mao), koike.katsuaki.5x@kyoto-u.ac.jp (K. Koike), alvaro@ige.unicamp.br (A.P. Crosta), plandim@rc.unesp.br (P.M.B. Landim), haizhuhu@163.com (H.Z. Hu), wangchunying1987@yahoo.com (C.Y. Wang), liqiang85327@gmail.com (L.Q. Yao).

0098-3004/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cageo.2011.09.005

we are able to produce a map of interpolated types of the categorical variable. There is an uncertainty zone around these boundaries because the resulting interpolated map is based on sample data points. We propose the use of the interpolation variance (Yamamoto, 2000) for types of categorical variables for mapping uncertainty for the transition zones between interpolated types. By incorporating this concept into the resulting map the limitations of using arbitrary zones that result from spatial buffering can be overcome.

2. Mapping an uncertainty zone between interpolated types of a categorical variable

2.1. Categorical variables

Categorical variables come from observations in which certain qualitative characteristics are recognized such as color, texture, and pattern. Variables measured on a nominal scale (Stevens, 1946) and on ordinal scale (Stevens, 1946) are called categorical variables. Ordinal scales assign numbers representing the rank order of certain characteristics (Stevens, 1946). For example, sediments can be described as fine, medium, and coarse, depending on the grain size. Variables measured on an ordinal scale can be analyzed as categorical variables as well.

Outcomes of discrete random variables cannot be combined directly to give the values at unsampled locations. However, some functions of these categorical variables can be used to estimate linearly the value at a given location or domain (Rivoirard, 1994). These functions, known as indicator functions, are used to indicate a type present within a categorical variable.

2.2. The indicator function

Given a categorical variable with K types, the indicator function for the kth type is defined as

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

The indicator variable is also known as an all-or-nothing variable (Journel and Huijbregts, 1978), because within K types of a categorical variable just one type k will have a value equal to one and all other equal to zero.

The mean of the indicator variable can be calculated as

$$E\left[I(x;k)\right]=\frac{f_k}{N}=p_k, \quad (2)$$

where p_k is the proportion of type k and $N=\sum_k f_k$ is the total number within the domain.

The variance of the indicator variable is

$$Var[I(x;k)]=E[I^2(x;k)]-(E[I(x;k)])^2=p_k-p_k^2=p_k(1-p_k). \quad (3)$$

Note that $E[I^2(x;k)]=E[I(x;k)]$. The variance is therefore the proportion of type k times the proportion of types different to k.

A single indicator variable that has two possible outcomes, one or zero, follows the Bernoulli distribution (Kader and Perry, 2007). Let p_1 be the proportion of ones and p_2 the proportion of zeroes then we have the mean equal to p_1 and the variance equal to p_1p_2 (Kader and Perry, 2007).

When we have K types within a categorical variable, then we have K indicator variables that follow a categorical distribution as a generalization of the Bernoulli distribution. Since indicator variables are mutually exclusive and exhaustive $\sum_k^K i(x;k)=1$ (Leuangthong et al., 2008) the categorical distribution is a special case of the multinomial distribution.

In expression (3) we can calculate the variance for kth indicator function. The global variance for all K types is given by the coefficient of unalikeability proposed by Kader and Perry (2007)

$$\mu_2=\sum_k^K p_k(1-p_k). \quad (4)$$

According to Kader and Perry (2007), the coefficient of unalikeability gives the proportion of possible comparisons that are unalike.

Now K indicator variables replace a categorical variable with K types. Indicator variables can be combined linearly to obtain estimated values at unsampled locations.

2.3. Indicator kriging

Indicator kriging is the most common interpolator to estimate every category type k at an unsampled location x_o as follows:

$$i_{IK}^*(x_o;k)=\sum_{i=1}^{n}\lambda_i i(x_i;k). \quad (5)$$

For example, if k=A it means that we are estimating the probability that the category type is A at location x_o

$$i_{IK}^*(x_o;k)=P(x_o;k=\text{A}).$$

We can also calculate the uncertainty associated with the indicator kriging estimate (4) as

$$s_o^2(x_o;k)=\sum_{i=1}^{n}\lambda_i[i(x_i;k)-i_{IK}^*(x_o;k)]^2. \quad (6)$$

This is none other than the interpolation variance proposed by Yamamoto (2000). Rewriting this expression we obtain

$$s_o^2(x_o;k)=\sum_{i=1}^{n}\lambda_i i^2(x_i;k)-(i_{IK}^*(x_o;k))^2.$$

Since $\sum_{i=1}^n\lambda_i i^2(x_i;k)=\sum_{i=1}^n\lambda_i i(x_i;k)$ the interpolation variance can be written as

$$s_o^2=i_{IK}^*(x_o;k)-(i_{IK}^*(x_o;k))^2=i_{IK}^*(x_o;k)(1-i_{IK}^*(x_o;k)).$$

For example, for k=A the interpolation variance is

$$s_o^2(x_o;k=\text{A})=P(x_o;k=\text{A})P(x_o;k\neq\text{A}), \quad (7)$$

and, therefore, the variance is equal to the product of probabilities that the category type at location x_o is A and that the category type is different than A.

The indicator kriging approach requires K indicator semivariogram models (Leuangthong et al., 2008). This is very difficult because some types can present just few data points and consequently few pairs presenting large statistical fluctuations. Thus, instead of indicator kriging we can apply multiquadric equations for interpolation of indicator variables at unsampled locations.

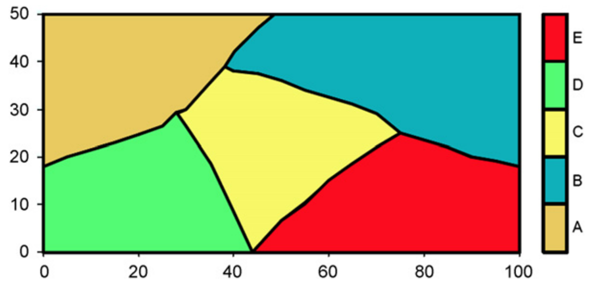

Fig. 1. Exhaustive data set showing a categorical variable with 5 types.