Choice of Optimal Number of Lags in Variogram Estimation

Seungbae Choi
Dongeui University, Statistical consulting Center
24, Kaya-Dong, PusanJin-Gu
Pusan 614-714, Korea
statcst@hyomin.dongeui.ac.kr

Changwan Kang
Dongeui University, Department of Computer Science and Statistics
24, Kaya-Dong, PusanJin-Gu
Pusan 614-714, Korea
cwkang@hyomin.dongeui.ac.kr

Yutaka Tanaka
Okayama University, Department of Environmental and Mathematical Sciences
3-1-1, Tsushima Naka
Okayama 700-8530, Japan
tanaka@stat.ems.okayama-u.ac.jp

1. Introduction

Spatial data are analyzed in three stages; variogram estimation, model fitting for the estimated variograms, and the spatial prediction using the fitted variogram model. As the first stage, i.e., variogram estimation, affects the next two stages, it is very important to estimate the variograms well. In general, the variogram is estimated with the moment estimator(Matheron, 1965) as follows;

$$\hat{g}(d) = \sum_{N(d)} \{z(s_i) - z(s_j)\}^2 / (2N_d),$$

where $N(d)$ is the set of all pairs with Euclidean distance d, N_d is the number of pairs in set $N(d)$, and $z(s_j)$ is the observed value at location s_j.

To estimate the variogram, we have to input the "lag increment" or the "number of lags" in most packages. There is no established rule for selecting the number of lags in estimating the variogram. For example, the default number of lags is 20.

The present paper proposes a method of choosing the optimal number of lags based on a set of given spatial data. Especially, to show the validity of the proposed method, we perform a small simulation study and show a numerical example. Here, we assume that the underlying process of the observed spatial data is stationary and consider three models i.e., spherical, exponential, gaussian, for the simulation study in Chapter 2.

2. The optimal lag

We propose to use the k^* as the optimal number of lags in variogram estimation, where k^* is

the number of lags which minimizes the PRESS(Predicted Error Sum of Square) using the given dataset.

2. 1 Simulation study

To show the validity of optimal number k^*, a simulation study is conducted as follows; 1) Fix the models(spherical, exponential gaussian) and parameters(sill, nugget, range). 2) Generate 100 datasets with the fixed models and parameters. Each dataset is composed of 400 locations and observed values. 3) Select the optimal number of lags k^* on the basis of the PRESS. 4) Fit a variogram model using k^* and estimate the parameters in the variogram model. 5) Calculate the MSE(Mean Square Error) for the estimated parameters for 100 datasets. Here, we summarize the MSEs the three parameters into the Euclidean norm. 6) Calculate the MSE for the default k for 100 datasets. 7) Compare the results between 5) and 6).

2. 2 Simulation results

We show only the result in the case of spherical model for the sake of the limitation of space. The performance of optimal k^* (MSE : 44.45) is better than the default k (MSE : 47.43).

3. Numerical example

A numerical example is given to show that the obtained k^* is of better performance than the default k. As a numerical example, we use 116 carbon monoxide(Co) data taken from the ministry of environment of Korea. This is conducted as follows; 1) determine the optimal k^* by the proposed method. 2) fit a variogram model using the selected number of lags. 3) predict the values in the deleted locations using the fitted variogram model, and calculate the PRESS. 4) obtain the PRESS using the default k. 5) compare the PRESSs obtained by 3) and 4). The results show that the performance in the case of k^* (PRESS : 7.79) is better than the default k (PRESS : 7.84).

REFERENCE

Eulogio, P. I.(1999). VARFIT: a fortran-77 program for fitting variogram models by weighted least squares, *Computer & Geosciences*, 25, 251-261.

Matheron, G.(1965). *Les variables regionalizees and leur estimation*, Masson, Paris, 305 pp.

MathSoft Inc.(1996). *S+SPATIALSTATS User's manual*, MathSoft Inc., Seattle, Washington.

SUMMARY

A method of selecting the number of lags is proposed in estimating the variograms in spatial data analysis. The validity or usefulness of the proposed method is established through a simulation study with three models. A numerical example is given to illustrate that it is better to use the proposed method than to use the default number of lags in variogram estimation.