The Korean Journal of Applied Statistics (2010)
23(3), 609–619

Data-Dependent Choice of Optimal Number of Lags in Variogram Estimation

Seungbae Choi[1] · Changwan Kang[2] · Jangsik Cho[3]

[1]Department of Data Information Science, Dongeui University
[2]Department of Data Information Science, Dongeui University
[3]Department of Information Statistics, Kyungsung University

(Received March 2010; accepted April 2010)

Abstract

Geostatistical data among spatial data is analyzed in three stages: (1) variogram estimation, (2) model fitting for the estimated variograms and (3) spatial prediction using the fitted variogram model. It is very important to estimate the variograms properly as the first stage(*i.e.*, variogram estimation) affects the next two stages. In general, the variogram is estimated with the moment estimator. To estimate the variogram, we have to decide the 'lag increment' or the 'number of lags'. However, there is no established rule for selecting the number of lags in estimating the variogram. The present paper proposes a method of choosing the optimal number of lags based on the PRESS statistic. To show the usefulness of the proposed method, we perform a small simulation study and show an empirical example with with air pollution data from Korea.

Keywords: Variogram, lag increment, number of lags, optimal lag, default lag.

1. Introduction

Geostatistical data among spatial data, also termed random field data, consists of observations measured at known specific locations or within specific regions. Because there are innumerable situations in which data are collected at various locations in space, application fields of spatial statistics are extensive. For example, the application fields include geology, soil science, image processing, epidemiology, crop science, ecology, forestry, astronomy, atmospheric science and environmental science. For these fields, practically, many studies have been carried out. A representative example of how to use geostatistics in environmental problems is given by Journel (1984). Istok and Cooper (1988) demonstrated how to predict ground contaminant concentrations using geostatistics and Myers (1984) implemented it to assess the movement of a multi-pollutant plume. Furthermore, Webster (1985) investigates soil characteristics and Piazza *et al.* (1981) analyse gene frequencies.

The analysis of such data is carried out in three stages: (1) estimating variograms, (2) fitting variogram models to the estimated variograms and (3) predicting the value at a specified location

This work was supported by National Research Foundation of Korea Grant funded by the Korean Government(KRF-2007-313-C00136).

[1] Corresponding author: Associate Professor, Department of Data Information Science, Dongeui University, Busan 614-714, Korea. E-mail: csb4851@deu.ac.kr

using the fitted variogram model. It is very important to estimate the variograms well as the first stage(i.e., variogram estimation) affects the next two stages. In general, the variogram is estimated with the moment estimator (Matheron, 1963), and we have to decide the ‘lag increment’ or the ‘number of lags’ as it is estimated. Practically in real data analysis, data analyst estimates the variogram by using several numbers of lags, and then selects the most proper number of lags among them. This can lead to the subjectivity of the analyst and a preposterous result in estimating the variogram.

This paper proposes a method of choosing the optimal number of lags(or the optimal lag increment) in estimating the variograms based on given spatial data. To show the usefulness of the proposed method, we perform a small simulation study and show an empirical example with the air pollution data in Korea.

For simplicity, we assume that the underlying process of the observed spatial data is stationary and isotropic. For computations, the software S−Plus is used with its spatial module S+SpatialStats.

The article is organized as follows. Section 2 briefly outlines the basics of spatial statistics. In Section 3 introduces a method selecting the optimal number of lags in variogram estimation. A small simulation study is performed in Section 4 to illustrate the usefulness of the proposed method. In Section 5 presents an empirical example and discuss its results. Concluding remarks are made in the last Section.

2. Outline of Spatial Statistics

Spatial data can be considered as a realization of a stochastic process $Z(s)$, *i.e.*,

$$\{z(s) : s \in D \subset R^r\}, \tag{2.1}$$

where s indicates a location in D and R^r is a r-dimensional Euclidean space. Most often r is 1, 2 or 3. The basic form of spatial data can be expressed as (z_i, s_i), $i = 1, \ldots, n$, where z_i is the i^{th} observation of a phenomenon of interest at location s_i. In spatial data analysis, it is assumed that the observed data have the following structure

$$z(s) = m(s) + e(s), \tag{2.2}$$

where $m(s)$ denotes a large-scale variation called drift or trend and $e(s)$ a small-scale variation. The latter term is a fluctuating random component with zero expectation and a covariance structure(or model) containing some parameters like random variation or measurement error. In most cases, a spatial data set represents a single realization of a random process. As such, some degree of stationarity must be assumed in order to make inferences about the data. Stationarity refers to some form of ‘location invariance’ of the data. It implies that the relationships within any subset of points remain the same no matter where the points reside in space (Mathsoft, 1996). In particular, when the mean and variance of the first difference of the stochastic process $Z(s)$ satisfy

$$\begin{aligned} E(Z(\boldsymbol{s}+\boldsymbol{d}) - Z(\boldsymbol{s})) &= 0, \\ \mathrm{Var}(Z(\boldsymbol{s}+\boldsymbol{d}) - Z(\boldsymbol{s})) &= 2\gamma(\boldsymbol{d}), \quad \boldsymbol{s},\ \boldsymbol{s}+\boldsymbol{d} \in D. \end{aligned} \tag{2.3}$$

$Z(\cdot)$ is said to be intrinsically stationary where $2\gamma(\boldsymbol{d})$ and $\gamma(\boldsymbol{d})$ are called the variogram and semivariogram, respectively. Furthermore, if $2\gamma(\boldsymbol{d}) = 2\gamma(\| \boldsymbol{d} \|)$, the variogram is said to be isotropic. If $2\gamma(\boldsymbol{d})$ depends on the direction of $\boldsymbol{d}$ as well as its distance $||\boldsymbol{d}||$, it is anisotropic. For simplicity, we use the term variogram instead of semivariogram where there is no confusion.

The first step of spatial analysis is to estimate the variogram $\gamma(\boldsymbol{d})$ using the observed data. When we can assume that the variogram is isotropic, an estimator for the variogram called sample variogram (Matheron, 1963) can be computed by

$$\hat{\gamma}(d) = \frac{1}{2N_d} \sum_{N(d)} (z(\boldsymbol{s}_i) - z(\boldsymbol{s}_j))^2. \tag{2.4}$$

Here $N(d)$ is the set of all pairs with Euclidean distances d, N_d is the number of distinct pairs in $N(d)$, and $z(\boldsymbol{s}_i)$ and $z(\boldsymbol{s}_j)$ are data values at spatial locations $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$, respectively. When the variogram is an isotropic, the directional sample variogram is computed using the same formula by replacing d by vector $\boldsymbol{d}$. Practically, in order to calculate the variogram values using Equation (2.4), we select first the lag distances 'd', then calculate the variogram values by regarding pairs with distance within 'd±lag.tol' as the pairs in $N(d)$. Symbol 'lag.tol' means lag tolerance which establishes distance bins for the lag increments, to accommodate for unevenly spaced observations, and if it is half the lag increment, no overlap of the distance classes is allowed. The 'lag increment(d±lag.tol)' defines the distances at which the variogram is calculated and the number of lags in conjunction with the size of the lag increment will define the total distance over which the variogram is calculated. To estimate the variogram, we have to decide the 'lag increment' or the 'number of lags'. They are in the reciprocal relationship, *i.e.*, 'the number of lags = maxdist / lag increment', where maxdist is the maximum distance over the field of data. A specific number of lags' as k is denoted throughout this paper.

The next stage of spatial analysis is to fit a variogram model that explains best the dependence(autocorrelation structure) of the underlying stochastic process. Most variogram models contain three parameters; namely, range, sill and nugget(or nugget effect). As the separation distance between pairs increase, the corresponding variogram value will also generally increase. Eventually, however, an increase in the separation distance between pairs of values and the variogram reaches a plateau. The distance at which the variogram reaches this plateau is called the range. The sill is the plateau the variogram reaches at the range. When we conduct the simulation, the partial sill is estimated as the sill of the sample variogram minus any nugget is used (Isaaks and Srivastava, 1989; Mathsoft, 1996). The nugget is the vertical jump from the value of 0 at the origin to the value of the variogram at extremely small separation distances. Also, $\gamma(0, \theta) = 0$ but $\lim_{d\to 0} \gamma(d, \theta) = c_0$ for any constance c_0 that is called the nugget because it is believed that microscale variation is causing a discontinuity at the origin (Cressie, 1991). It is estimated from the sample variogram as the value of $\gamma(d)$ for $d = 0$.

There have been proposed so far several variogram models(functions) according to its form; for example, spherical, gaussian, exponential as bounded variogram models, and power and linear models as unbounded variogram models. Here, we introduce only three models which are used for our study. They are given as follows (Cressie, 1991).

Spherical model:

$$\gamma_s(d; \boldsymbol{\theta}) = \begin{cases} 0, & d = 0, \\ \theta_n + \theta_s \left[\left(\frac{3}{2}\right) \frac{d}{\theta_r} - \left(\frac{1}{2} \left(\frac{d}{\theta_r} \right)^3 \right) \right], & 0 < d \leq \theta_r, \\ \theta_n + \theta_s, & d \geq \theta_r. \end{cases} \tag{2.5}$$

Gaussian model:

$$\gamma_s(d;\boldsymbol{\theta}) = \begin{cases} 0, & d = 0, \\ \theta_n + \theta_s \left[1 - \exp\left(- \left(\dfrac{d}{\theta_r} \right)^2 \right) \right], & d \neq 0, \end{cases} \tag{2.6}$$

Exponential model:

$$\gamma_r(d;\boldsymbol{\theta}) = \begin{cases} 0, & d = 0, \\ \theta_n + \theta_s \left[1 - \exp\left(- \dfrac{d}{\theta_r} \right) \right], & d \neq 0, \end{cases} \tag{2.7}$$

where $\boldsymbol{\theta} = (\theta_n, \theta_s, \theta_r)$, θ_n is nugget effect, $\theta_s = (\sigma^2 - \theta_n)$ is sill, and θ_r is range. Strictly speaking, θ_s is partial sill and σ^2 is sill. The above variogram models can be fit by several methods; Maximum Likelihood(ML), Restricated Maximum Likelihood(REML), Least Squares, Weighted Least Squares method, and so on. Zimmerman and Zimmerman (1991) compare several estimation methods and conclude that ordinary nonlinear least squares or some form of weighted nonlinear least squares is usually as good as many of the more complicated and computationally intensive methods (Mathsoft, 1996). Thus, we used the weighted nonlinear least squares estimator in this paper.

An important problem in spatial statistics is to predict the unobserved value $z(\boldsymbol{s}_0)$ at a specified location $\boldsymbol{s}_0$ based on the information of n observations $z(\boldsymbol{s}_i)$, $i = 1, \ldots, n$. This kinds of spatial prediction problem is known as kriging in spatial statistics, and there are some types of kriging such as simple, ordinary, and universal kriging. Here we will consider only ordinary kriging that is often associated with the B.L.U.E.(best linear unbiased estimator) because it is used in our simulation. We will look at how the ordinary kriging weights are calculated. In order to solve above a question, we consider a linear predictor as

$$\hat{Z}(\boldsymbol{s}_0) = \sum_{i=1}^{n} w_i Z(\boldsymbol{s}_i). \tag{2.8}$$

It is a best linear unbiased prediction method under the assumption that the stochastic process underlying the observations is second-order stationary, *i.e.*,

$$E\left(Z\left(\boldsymbol{s}_i\right)\right) = 0, \qquad \mathrm{Var}\left(Z\left(\boldsymbol{s}_i\right)\right) = \sigma^2, \qquad \mathrm{Cov}\left(Z\left(\boldsymbol{s}_i\right), Z(\boldsymbol{s}_j)\right) = C(\boldsymbol{s}_i - \boldsymbol{s}_j), \tag{2.9}$$

for any i, j. For unbiasedness, the weights must satisfy $\sum_{i=1}^{n} w_{ij} = 1$. By minimizing the prediction variance $E(Z(\boldsymbol{s}_0) - \hat{Z}(\boldsymbol{s}_0))^2$ under the equality constraint on the weights, we obtain the following system of equations:

$$-\sum_{j=1}^{n} w_j C(\boldsymbol{s}_i - \boldsymbol{s}_j) + C(\boldsymbol{s}_0 - \boldsymbol{s}_j) - \lambda = 0, \quad (i = 1, \ldots, n), \ \sum_{i=1}^{n} w_i = 1, \tag{2.10}$$

where λ is a Lagrange multiplier and the covariance function is related to the variogram as $C(\boldsymbol{s}_i - \boldsymbol{s}_j) = \sigma^2 - \gamma(\boldsymbol{s}_i - \boldsymbol{s}_j)$. If the covariogram $C(\boldsymbol{s}_i - \boldsymbol{s}_j)$ and $C(\boldsymbol{s}_0 - \boldsymbol{s}_i)$ are known, the optimal weights $\{w_i\}$ can be obtained by solving the above equations. In practical data analysis we usually do not know these variograms. Therefore, before applying this kriging method we have to estimate them. More precisely, at first we calculate sample variograms, omnidirectional or directional variograms depending on the structure of spatial correlation, and then find a theoretical variogram model which fits best the sample variograms. There are two reasons why the sample variogram cannot be used

Table 3.1. Estimated parameters in an exponential model for the variograms estimated by assigning three numbers of lags

	Sill	Nugget	Range
Lag = 10	0.139	0.034	10.65
Lag = 15	0.103	0.003	13.76
Lag = 20	0.095	0.012	15.52

directly in the ordinary kriging system. First, there are often situations in which the distance from the point being estimated to a particular sample is smaller than the distance between any pair of available samples. Since the sample data set cannot provide any pairs for these small distances, we must rely on a function that provides variogram values for all distances and directions, even those that are not available from sample data. Second, the use of the sample variogram does not guarantee the existence and uniqueness of the solution to the ordinary kriging system (Isaaks and Srivastava, 1989).

3. The Optimal Number of Lags

We consider the following example before introducing the optimal number of lags.

Example 3.1. We consider the effect on variogram model fitting of the assigned numbers of lags. The exponential model is fitted to Korean air pollution data(Co), which will be analyzed in Section 4. Table 3.1 shows that the estimated parameters are somewhat different depending on the assigned number of lags.

This implies that the choice of the number of lags is important for model fitting. In the above example, we can recognize that we need the criterion for selecting the number of lags.

Journel and Huijbregts (1978) suggest the following two practical rules in choosing the lag increment and number of lags: (1) the experimental variogram should only be considered for distances d for which the number of pairs is greater than 30. (2) The distance of reliability for an experimental variogram is $d < 1/2$, where D is the maximum distance over the field of data. However, these rules are ambiguous in choosing practically the number of lags or lag increment. The present paper proposes a criterion for selecting the optimal number of lags.

Here, we focus on the number of lags denoted by symbol k because the above two factors are mutually reciprocal. Now, our main interest is in finding the optimal number of lags among possible values of k.

We define the Predicted Residual Sum of Square(PRESS) for finding the optimal number of lags as

$$\text{PRESS}_k = \sum_{i=1}^{n} \left(z_i - \hat{z}_{(i)}\right)^2, \quad \text{for a fixed number of lags } k, \tag{3.1}$$

where z_i is an observation at the i^{th} location and $\hat{z}_{(i)}$ is the predicted value at the i^{th} location using the observed values excepting the i^{th} one. We define the optimal number of lags as follows.

Definition 3.1. *The optimal number of lags $\hat{k}^*$ is the value of k which minimizes the PRESS statistic.*

In some statistical packages, the number of lags is set to 10 or 20 as a default number of lags. Here, we put the default numbers of lags k as 10(the default value in SAS package, SAS Institute Inc.

(1998)) and 20(the default value in S−PLUS package), and compare the performance with that of optimal $\hat{k}^*$. From now on, we represent 'the optimal numbers of lags' as 'the optimal lag' or 'optimal $\hat{k}^*$' and 'the default number of lags' as 'the default lag'.

4. Simulation Study

We carry out a small simulation to investigate the performance of the optimal $\hat{k}^*$ compared to the default $k = 10$ and 20. In this study, we consider three models(spherical, exponential and Gaussian), all of which contain three parameters(sill, nugget and range) and we restrict the scope of the number of lags from 6 to 25 in selecting the optimal $\hat{k}^*$. The range is related only to the observation positions(latitude and longitude or east and west) and sill and nugget are also to the observation values. For this reason, this simulation is conducted with varying only the range parameter in each model.

Also, to compare performance of $\hat{k}^*$ with $k = 10$ and 20, we use the root mean squared errors RMSE_1 and RMSE_2.

$$\text{RMSE}_1 = \sqrt{\sum_{i=1}^{n} \left|\left|\hat{\boldsymbol{\theta}}_i^* - \boldsymbol{\theta}_i\right|\right|^2}, \qquad \text{RMSE}_2 = \sqrt{\sum_{i=1}^{n} \left|\left|\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i\right|\right|^2}, \tag{4.1}$$

where $\boldsymbol{\theta} = (\theta_n, \theta_s, \theta_r)(\theta_n$: nugget, θ_s: sill, θ_r: range) is the true parameter vector, and $\hat{\boldsymbol{\theta}}^*$ and $\hat{\boldsymbol{\theta}}$ are the estimated parameter vectors for each of the 100 datasets generated by using the optimal $\hat{k}^*$ and the default $k = 10$ and 20, respectively, symbol $||\cdot||$ indicating the Euclidean norm.

To show the usefulness of the optimal $\hat{k}^*$ obtained by our method, a simulation study is conducted in the following steps; (1) Fix the models(spherical, exponential, and Gaussian) and parameters(sill, nugget, and range), and generate m(100400) lattice positions. In case of sample size 100 they are generated by a 10×10 grid while for 400 observations a 20×20 grid is used. (2) Generate m observed values at the m lattice positions with the fixed models and parameters, *i.e.*, $\theta_s = 0.7$, $\theta_n = 0.1$, $\boldsymbol{\theta}_r = (2, 20)$ for the three models. m observations with the specified model and parameter sat the m lattice positions can be generated using the function 'rfsim' in spatial module S+SpatialStats of S−Plus. These observations are autocorrelated with errors of specified covariance model. (3) Generate 100 datasets, each of which is composed of m positions and the corresponding m observations. (4) Select the optimal $\hat{k}^*$ on the basis of the PRESS statistic. (5) Estimate the parameters of the variogram model by using the optimal $\hat{k}^*$ obtained for each dataset. (6) Calculate the RMSE_1 for the parameters estimated by the optimal $\hat{k}^*$ in 100 datasets. (7) Calculate the RMSE_2 for the default $k = 10$ and 20 in 100 datasets. (8) Compare the results between (6) and (7).

The selection of the optimal $\hat{k}^*$ can be explained as below.

Step 1. For a fixed lag k, estimate the variogram using $n-1$ observations excepting the i^{th} one and obtain the predicted value $\hat{z}_{(i)}$ at the i^{th} location based on the estimated variogram.

Step 2. For every $i(i = 1, \ldots, m)$, calculate $(z_i - \hat{z}_{(i)})$ by repeating step 1 m times.

Step 3. For the fixed lag k, calculate the PRESS statistic $\sum_{i=1}^{m}(z_i - \hat{z}_{(i)})^2$.

Step 4. Calculate the PRESS for every $k(6 \le k \le 25)$, and select the optimal $\hat{k}^*$ which minimizes the PRESS.

The results of the simulation study are given in from Table 4.1 to 4.4. They show a result of in sample size 100 and 400 according to the range. We can find that the performance of the optimal

Table 4.1. Simulation results(Sample size: 100, Range: 2)

Model	$\boldsymbol{\theta}$	Optimal($\hat{k}^*$)			Default($k = 10$)			Default($k = 20$)		
		$E(\hat{\theta}^*)$	S.E.	RMSE$_1$	$E(\hat{\theta})$	S.E.	RMSE$_2$	$E(\hat{\theta})$	S.E.	RMSE$_2$
Sph	$\theta_s = 0.7$	0.73	0.023	0.76	0.73	0.024	0.84	0.74	0.023	0.80
	$\theta_n = 0.1$	0.10	0.008		0.11	0.009		0.10	0.009	
	$\theta_r = 2$	2.30	0.065		2.35	0.073		2.33	0.069	
Gau	$\theta_s = 0.7$	0.71	0.025	0.83	0.70	0.026	0.91	0.71	0.026	0.89
	$\theta_n = 0.1$	0.14	0.005		0.16	0.006		0.15	0.005	
	$\theta_r = 2$	2.47	0.064		2.56	0.068		2.51	0.068	
Exp	$\theta_s = 0.7$	0.80	0.032	1.41	0.82	0.034	1.57	0.80	0.033	1.55
	$\theta_n = 0.1$	0.07	0.014		0.07	0.015		0.08	0.013	
	$\theta_r = 2$	2.48	0.128		2.52	0.144		2.54	0.141	

Sph: spherical, Gau: Gaussian, Exp: exponential

Table 4.2. Simulation results(Sample size: 100, Range: 20)

Model	$\boldsymbol{\theta}$	Optimal($\hat{k}^*$)			Default($k = 10$)			Default($k = 20$)		
		$E(\hat{\theta}^*)$	S.E.	RMSE$_1$	$E(\hat{\theta})$	S.E.	RMSE$_2$	$E(\hat{\theta})$	S.E.	RMSE$_2$
Sph	$\theta_s = 0.7$	0.77	0.026	5.57	0.79	0.025	6.46	0.78	0.026	6.29
	$\theta_n = 0.1$	0.10	0.010		0.07	0.012		0.08	0.011	
	$\theta_r = 20$	22.4	0.501		21.9	0.617		22.1	0.595	
Gau	$\theta_s = 0.7$	0.71	0.025	7.22	0.70	0.026	8.10	0.71	0.026	7.78
	$\theta_n = 0.1$	0.14	0.005		0.15	0.006		0.14	0.005	
	$\theta_r = 20$	24.4	0.570		25.3	0.616		24.8	0.611	
Exp	$\theta_s = 0.7$	0.85	0.036	24.1	0.87	0.038	27.2	0.85	0.037	25.3
	$\theta_n = 0.1$	0.07	0.015		0.06	0.017		0.08	0.014	
	$\theta_r = 20$	29.2	2.242		30.5	2.523		30.2	2.326	

Table 4.3. Simulation results(Sample size: 400, Range: 2)

Model	$\boldsymbol{\theta}$	Optimal($\hat{k}^*$)			Default($k = 10$)			Default($k = 20$)		
		$E(\hat{\theta}^*)$	S.E.	RMSE$_1$	$E(\hat{\theta})$	S.E.	RMSE$_2$	$E(\hat{\theta})$	S.E.	RMSE$_2$
Sph	$\theta_s = 0.7$	0.68	0.013	0.89	0.69	0.019	1.27	0.68	0.013	0.94
	$\theta_n = 0.1$	0.12	0.009		0.12	0.017		0.12	0.011	
	$\theta_r = 2$	2.19	0.086		2.32	0.116		2.20	0.090	
Gau	$\theta_s = 0.7$	0.72	0.027	0.58	0.73	0.028	0.65	0.73	0.029	0.60
	$\theta_n = 0.1$	0.10	0.003		0.11	0.004		0.10	0.003	
	$\theta_r = 2$	2.10	0.050		2.15	0.059		2.11	0.055	
Exp	$\theta_s = 0.7$	0.79	0.046	2.94	0.84	0.050	3.55	0.82	0.045	3.14
	$\theta_n = 0.1$	0.10	0.006		0.06	0.018		0.08	0.009	
	$\theta_r = 2$	2.78	0.281		2.89	0.346		2.79	0.304	

$\hat{k}^*$ is better than those of the two default values $k = 10$ and 20 in all cases of three models of tables in the terms of the RMSEs. For example, the RMSE$_1$ of the optimal $\hat{k}^*$ is 0.76, while the RMSE$_2$s of the default $k = 10$ and 20 are 0.84 and 0.80, respectively, for spherical model with sample size 100 and range 2. The rest show the similar results.

In tables, $E(\hat{\theta}^*)$ indicates mean of the optimal individual parameters $\hat{\theta}^*$s and $E(\hat{\theta})$ mean of the default individual parameters $\hat{\theta}$s(cases of 10 and 20) estimated from 100 datasets for each model, and θ is true parameters. The S.E.s represent the empirical standard error of the optimal individual parameters $\hat{\theta}^*$s and of the default individual parameters $\hat{\theta}$s(cases of 10 and 20).

If we judge with the $|E(\hat{\theta}^*) - \theta|$ that is the difference between real and the mean of fitted parameters

Table 4.4. Simulation results(Sample size: 400, Range: 20)

Model	$\boldsymbol{\theta}$	Optimal($\hat{k}^*$)			Default($k=10$)			Default($k=20$)		
		$E(\hat{\theta}^*)$	S.E.	RMSE$_1$	$E(\hat{\theta})$	S.E.	RMSE$_2$	$E(\hat{\theta})$	S.E.	RMSE$_2$
Sph	$\theta_s = 0.7$	0.69	0.015	3.92	0.69	0.018	9.84	0.69	0.015	4.07
	$\theta_n = 0.1$	0.11	0.006		0.12	0.014		0.11	0.007	
	$\theta_r = 20$	20.7	0.388		22.5	0.957		20.6	0.403	
Gau	$\theta_s = 0.7$	0.72	0.026	3.68	0.72	0.026	4.66	0.73	0.026	4.15
	$\theta_n = 0.1$	0.10	0.002		0.11	0.003		0.10	0.003	
	$\theta_r = 20$	20.9	0.357		21.3	0.448		21.0	0.405	
Exp	$\theta_s = 0.7$	0.86	0.053	27.1	0.90	0.056	33.7	0.89	0.059	33.3
	$\theta_n = 0.1$	0.12	0.006		0.11	0.013		0.12	0.007	
	$\theta_r = 20$	30.8	2.490		35.6	3.006		33.9	3.045	

in the case of the optimal and $|E(\hat{\theta}) - \theta|$ that is the result of default case of both 10 and 20, we find that values of $|E(\hat{\theta}^*) - \theta|$ are smaller than those of $|E(\hat{\theta}) - \theta|$ in most cases. In this result, we find that values of $|E(\hat{\theta}^*) - \theta|$ are smaller than those of $|E(\hat{\theta}) - \theta|$ in most cases. This also shows that the performance of the optimal $\hat{k}^*$ is better than those of the default $k = 10$ and 20 because parameters obtained by the optimal $\hat{k}^*$ are closer to the true parameter values than those by the default k. From the parameter ranges view, we can note that case of range = 2 is fitter than that of range = 20 for all models. Also, we find values of the S.E. in the case of the optimal are smaller than those of both the default $k = 10$ and 20 in most cases. For reference, the computational costs of the proposed method take about one minute to find the optimal $\hat{k}^*$ in case sample size equal to 100 and about two minutes in case sample size equal to 400.

5. Empirical Example

In this section, we apply the proposed method to an actual dataset. First we show that the optimal $\hat{k}^*$ is of better performance than the default $k = 10$ and 20, then spatial data analysis is conducted with the obtained optimal $\hat{k}^*$. This dataset consists of 116 measurements of carbon monoxide(Co), which are taken from the Korean Ministry of the Environment in 2003. The measurement sites are displayed in Figure 5.1. These data are measured by ppm(parts per million) unit with TM(Transverse Mercator) coordinates which are generally used in Korean Map.

We divide the dataset into two subsets, *i.e.* the training dataset containing 100 observations and the validation dataset containing 16 observations. Here, we randomly selected 16(6^{th}, 9^{th}, 11^{st}, 16^{th}, 19^{th}, 23^{rd}, 31^{st}, 32^{nd}, 76^{th}, 85^{th}, 86^{th}, 91^{st}, 93^{rd}, 111^{st}, 114^{th}, 116^{th}) observations as the validation dataset and use the remaining 100 observations as training dataset. We analyze based on for the purpose of this study and also show that the superiority of the proposed method remain in case of different training and validation datasets. In Figure 5.1, symbol '·' represents the positions of the training data and '+' positions of the validation data. To detect an appropriate variogram model, we consider all three models described above as candidates, then select a model that minimizes the PRESS.

Analysis is conducted with a given model as follows; (1) Determine the optimal $\hat{k}^*$ based on the PRESS using the training dataset. (2) Fit the variogram model using the selected optimal $\hat{k}^*$. (3) Predict the values at the locations of observations in the randomly selected validation dataset using the fitted variogram model, and calculate the MSE(mean squared error) using the optimal $\hat{k}^*$. (4) Obtain the MSE using the default $k = 10$ and 20. (5) Compare the MSEs obtained by (3) and (4).

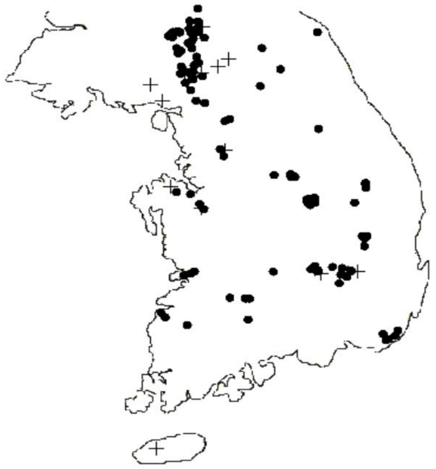

Figure 5.1. 116 measured locations for carbon monoxide(Co) data in Korea

Table 5.1. PRESSs for default and optimal in exponential model

	Optimal($\hat{k}^*$)	Default	
		$k = 10$	$k = 20$
PRESS	7.22(17)	7.49	7.24

To decide the optimal $\hat{k}^*$, we calculate the PRESS by applying the proposed method to the training dataset. As the results, the PRESS of the optimal $\hat{k}^*(7.22)$ is smaller than that of the default $k = 10(7.49)$ and $20(7.24)$ as shown in Table 5.1. From Table 5.1, in terms of PRESS, we can select 17 as the optimal $\hat{k}^*$ for given exponential model.

Now, as we know the optimal $\hat{k}^*$, we can fit the variogram model by using it. For comparison, we also consider the default $k = 10$ and 20 with the exponential model. Plots of the fitted variogram models for the three numbers of lags($\hat{k}^* = 17$, $k = 10$, 20) are given in Figure 5.2, and the fitted models for the three cases are as follows;

$$\hat{\gamma}_{10}(d;\boldsymbol{\theta};k=10) = \begin{cases} 0.0149 + 0.1015\left[1-\exp\left(-\dfrac{d}{21.8255}\right)\right], & d \neq 0, \\ 0, & d = 0, \end{cases} \tag{5.1}$$

$$\hat{\gamma}_{20}(d;\boldsymbol{\theta};k=20) = \begin{cases} 0.0296 + 0.0905\left[1-\exp\left(-\dfrac{d}{28.7212}\right)\right], & d \neq 0, \\ 0, & d = 0, \end{cases} \tag{5.2}$$

$$\hat{\gamma}^{*}(d;\boldsymbol{\theta};k^{*}=17) = \begin{cases} 0.0291 + 0.0826\left[1-\exp\left(-\dfrac{d}{16.1157}\right)\right], & d \neq 0, \\ 0, & d = 0. \end{cases} \tag{5.3}$$

The sum of the squared residuals indicate the goodness of fit of the variogram models are 0.0083 for the optimal $\hat{k}^*$, 0.0039 for the default $k = 10$, and 0.0128 for the default $k = 20$. The MSEs for predicting the validation data are 0.0453 for the optimal $\hat{k}^*$, 0.0460 for the default $k = 10$ and 0.0461 for the default $k = 20$. Here we can note that the optimal $\hat{k}^*$ is of better performance than the rest from the MSE viewpoint, although the case of $k = 10$ has the smallest value of the sum of

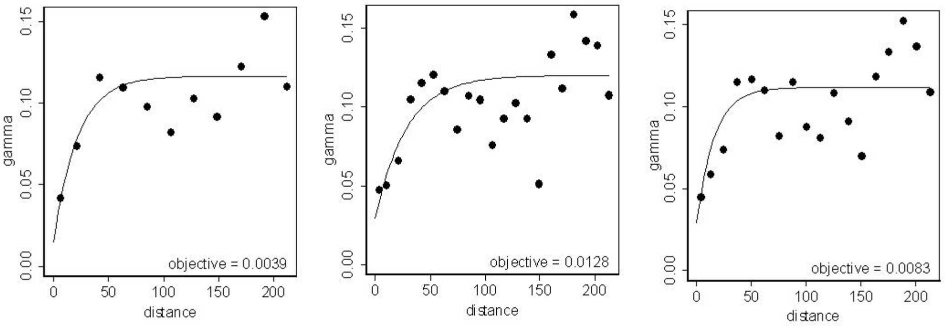

Figure 5.2. Plot of fitted variogram models(left: default = 10, middle: default = 20, right: optimal)

Table 5.2. Average of 10 MSEs for the different training and validation datasets

	Optimal($\hat{k}^*$)	Default	
		$k = 10$	$k = 20$
Average MSE	0.0442	0.0460	0.0470

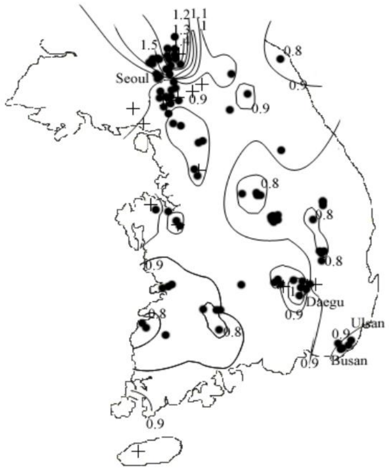

Figure 5.3. Contour plot of kriging prediction

squared residuals in training data.

To show that the superiority of the proposed method remain in case of different training and validation datasets, we tried the above same analysis with the 10 different training and validation datasets selected randomly among the same dataset with 116 observations. Table 5.2 shows the MSEs in case of each dataset and averages of those 10 MSEs obtained for the three numbers of lags($\hat{k}^*$, $k = 10$, 20). We can find that all of 10 MSEs in the optimal lag smaller than in the default lag $k = 10$ and 20. Therefore, we can note that the superiority of the proposed method remain in case of different training and validation datasets in Table 5.2.

Contour plot of the prediction surface based on the exponential model with $\hat{k}^* = 17$ is given in Figure 5.3. In this figure, note that the predicted value for the air pollution(Co) is high in Seoul, Busan, Daegu and Ulsan. We guess that the automobile and industrial plants cause this result.

6. Concluding Remarks

We propose a method for selecting the optimal number of lags in estimating the variograms in spatial data analysis. The usefulness of the proposed method is established through a simulation study. We applied the proposed method in an empirical example and then conducted further analysis with obtained optimal lag. The present paper, we proposed a method for finding the optimal lag using the PRESS assuming that variogram model is given however a method for selecting optimal variogram model remains to be studied.

References

Cressie, N. A. C. (1991). *Statistics for Spatial*, Wiley, New York.

Isaaks, E. H. and Srivastava, R. M. (1989). *An Introduction to Applied Geostatistics*, Oxford University Press, New York.

Istok, J. D. and Cooper, R. M. (1988). Geostatistics applied to ground water pollution III: Global estimates, *Journal of Environmental Engineering*, **114**, 915–928.

Journel, A. G. (1984). New ways of assessing spatial distribution of pollutants, *In Environmental Sampling for Hazardous Waters*, G. Schweitzer(Ed.), 109–118. American Chemical Society, Washington.

Journel, A. G. and Huijbregts, C. J. (1978). *Mining Geostatistics*, Academic Press, London.

Matheron, G. (1963). *Principles of Geostatistics, Economic Geostatistics*, Academic Press, London.

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

Myers, D. E. (1984). *Borden field data and multivariate geostatistics In Hydraulic Engineering*, M. A. Ports(Ed.), 795–800. American Society of Civil Engineering, NewYork.

Piazza, A., Menozzi, P. and Cavalli-Sforza, L. (1981). The making and testing of geographic gene frequency maps, *Biometrics*, **37**, 635–659.

SAS Institute Inc. (1998). SAS/STAT Technical Report: Spatial Prediction Using the SAS System, SAS Institute Inc..

Webster, R. (1985). Quantitative spatial analysis of soil in the field, In *Advances in Soil Science*, **3**, B. A. Stewart(ed.), 1–70. Springer-Verlag, New York.

Zimmerman, D. L. and Zimmerman, M. B. (1991). A comparison of spatial semivariogram estimators and corresponding ordinary kriging predictors, *Technometrics*, **33**, 77–92.