Pakistan Journal of Statistics and Operation Research

Spatial Prediction Simulation with Nonlinear Multicovariate

Geneveve Parreño-Lachica[1,2*]

* Corresponding Author

1. Department of Mathematics, College of Arts and Sciences, West Visayas State University, Iloilo City, 5000, Philippines; Email: gmparreno@wvsu.edu.ph
2. Center for Research and Innovations in Science, Mathematics, and Education, West Visayas State University, Iloilo City, 5000, Philippines

Abstract

Cokriging is a multivariate spatial method used to predict the observed value for a primary variable in an unknown location with the help of a spatially correlated secondary variable. The existence of two or more nonlinear secondary variables in predicting spatial data usually arises, especially in cokriging. Therefore, a method that can improve the model's predictive power by adding the interaction of variables is proposed. The proposed method can be effectively used, especially when the primary and secondary variables have a nonlinear relationship. By transforming the nonlinear variables, a higher correlation can be attained. This study used principal component analysis with interaction (PCAI) method among secondary variables to reduce two or more secondary variables into one dimension as a secondary variable in the cokriging technique. The proposed method was tested and verified through simulation and real data using the 2015 South Korea Air Pollution dataset, a dataset known for its complex spatial patterns and high variability, to prove its validity and usefulness. The predicted residual error sum of squares (PRESS) statistic was used for cross-validation. Computations were done using the R Project for Statistical Computing software. PCAI as a secondary variable gives the lowest PRESS value compared to only one secondary variable or principal component analysis (PCA). Considering the criterion, the lowest value of PRESS indicates the best model. Thus, PCAI cokriging outperformed PCA cokriging. Using PCAI as a secondary variable may be a better method than PCA for cokriging with nonlinear multicovariates.

Key Words: Cokriging, Nonlinear Multicovariate, Simulation, Spatial Prediction, Spatial Statistics.

1. Introduction

This study focuses on spatial statistics, which encompasses three main forms of data collection: geostatistical data, lattice data, and spatial point pattern data. For the purposes of our research, we specifically consider geostatistical data, which consists of observations measured at known specific locations or within particular regions (Choi et al., 2010). The application of spatial statistics is wide-ranging, as it is often necessary to collect data at various locations in space (Acevedo, 2012; O'Sullivan & Unwin, 2014). Our research aims to contribute to this field by proposing a new method to enhance the predictive power of cokriging in spatial statistics, particularly when dealing with nonlinear multicovariates.

The geostatistical data analysis procedure was conducted in stages. These are the following: (1) Estimating the sample variograms; (2) Fitting theoretical variograms; and (3) Kriging or predicting the value at a specified location (Ripley, 1981; Isaaks & Strivastava, 1989; Dale & Fortin, 2014).

Cokriging is a multivariate spatial method that estimates spatially correlated variables as an extension of the kriging method in geostatistics (Schabenberger & Gotway, 2017). In geostatistics, the dataset structure of cokriging consists of locations and observed values for two variables (the primary and secondary variables) in the locations (Wackernagel, 1998; Wackernagel, 2013). The cokriging method helps predict the observed value for a primary

variable in an unknown location with the help of a secondary variable (Chica-Olmo, 2007; Usman et al., 2013). Generally, cokriging is conducted using one primary and one secondary variable.

Principal Component Analysis (PCA) is widely used as a dimensionality reduction technique that transforms correlated variables into a smaller number of uncorrelated components while preserving most of the variability present in the original dataset, thereby facilitating the identification of underlying patterns and structures in multivariate data (Ahmed & Siddiqui, 2014).

Kang et al. (2008) used Principal Component Analysis (PCA) to improve cokriging and obtain better results than kriging when the secondary variables are two or more. Parreño et al. (2017b) studied a method using the PCA with interaction ($PCAI$) among secondary variables to reduce two or more variables into one dimension as a secondary variable for the cokriging technique. Comparing Kriging, PCA, and $PCAI$ with Root Mean Square Error ($RMSE$) as a criterion, $PCAI$ cokriging is superior to kriging and PCA cokriging, respectively. Parreño et al. (2017a) studied a method that carries out cokriging after transforming the variable in case the secondary variable has a lower correlation with a primary variable. A simulation method was performed to show the validation of the method using the location, primary variable, and two secondary variables. One secondary variable has a low correlation value, while the other has a high correlation value. As per the cross-validation criteria of the Predicted Residual Error Sum of Squares (PRESS) statistic, cokriging with a higher correlation yields better simulation results than cokriging with a lower correlation. However, in various research conducted using Nonlinear Multivariate Analysis (e.g., Perada et al., 2005), spatial prediction using nonlinear multicovariate analysis has not received much attention.

Recent developments further support the need to revisit PCA-based dimension reduction within spatial statistics, particularly under nonlinear and spatially dependent settings. More recent studies have emphasized that PCA and its functional extensions remain central tools for dimension reduction, but that conventional linear formulations may be inadequate when nonlinear dependence and spatial structure are both present (Li et al, 2022). For instance, spatial effects have been shown to be important in PCA for geographically distributed data, since ignoring spatial dependence may lead to incomplete or less appropriate interpretation of multivariate patterns (Cartone & Postiglione, 2021).

In parallel, newer studies on nonlinear principal component methods have shown that linear functional principal component analysis (FPCA) may fail to capture nonlinear structure adequately, motivating nonlinear extensions that preserve both linear and nonlinear variation while accounting for complex spatial structure (Zhong & Song, 2025). Likewise, recent multivariate spatial functional PCA approaches have demonstrated that incorporating spatial dependence directly into dimension reduction can improve coherence and predictive performance relative to methods that ignore spatial context (Si-Ahmed et al, 2025; Li, Huang, & Härdle, 2019).

According to several studies, interaction terms are beneficial for polynomial regression models because they increase their predictive power, especially when dealing with nonlinearity in the connection between the input and target variables. However, the study of the effect of correlation coefficients in collocated cokriging describes a situation in which estimates primarily depend on secondary data when there is a strong correlation between the primary and secondary variables (Legendre, 1993; Kang et al., 2006; Rocha et al., 2012). Within these premises, this study proposes a method in which cokriging can still be performed when two or more nonlinear secondary variables exist. That is by transforming the nonlinear variables and reduce into one dimension using $PCAI$ (principal component analysis with interaction among secondary variables) so that the prediction value can be attained. Furthermore, adding variables' interaction improves the model's predictive power. This is effectively used, especially when the primary and secondary variables have a nonlinear relationship.

R is powerful in analyzing and visualizing spatial data (Maindonald & Braun, 2010; Bivand et al., 2013). Using the R Project for Statistical Computing software, a simulation was performed, and the $PCAI$ method was applied to a real-world data set (air pollution data) to prove the validity and usefulness of the proposed method. $PRESS$ (Predicted residual error sum of squares) criterion was used for cross-validation. Allen (1971) described the PRESS statistic as a leave-one-out refitting and prediction method. The low PRESS value indicated the best model (Lloyd, 2010).

2. Methodology

2.1 Construction of the Principal Component Analysis with Interaction (PCAI) Method

Let $Z(s)$ denote the primary spatial random field observed at location $s \in D \subset \mathbb{R}^2$, where $s = (longitude, latitude)$. Suppose that, at the same set of locations, we observe p transformed secondary variables denoted by

$$X_1(s), X_2(s), \dots, X_p(s).$$

These secondary variables may exhibit nonlinear relationships with the primary variable $Z(s)$ and among themselves.

The proposed Principal Component Analysis with Interaction ($PCAI$) method constructs a composite secondary variable that incorporates both the main effects and interaction effects of the secondary variables prior to dimensional reduction. The procedure is described as follows.

Step 1. Construction of Interaction Terms
For each pair of secondary variables $X_i(s)$ and $X_j(s)$, where $i < j$, define the two-way interaction term as

$$X_{ij}(s) = X_i(s)X_j(s). \tag{1}$$

Thus, the augmented secondary variable set becomes

$$\mathcal{X}^*(s) = \{X_1(s), \dots, X_p(s), X_{12}(s), X_{13}(s), \dots, X_{(p-1)p}(s)\}. \tag{2}$$

The total number of variables in the augmented set is

$$p + \binom{p}{2}.$$

This step allows nonlinear dependence structures among secondary variables to be explicitly represented.

Step 2. Scaling of Variables
All secondary variables and their interaction $\mathcal{X}^*(s)$ terms were standardized to zero mean and unit variance prior to PCA computation.

$$\tilde{X}_k(s) = \frac{X_k(s) - \bar{X}_k}{\sigma_k}, \tag{3}$$

where $\bar{X}_k$ and σ_k denote the sample mean and standard deviation, respectively.

Standardization ensures that all variables contribute equally to the principal component extraction by removing scale effects. Consequently, PCA was performed on the correlation matrix rather than the covariance matrix.

Step 3. Principal Component Extraction
Principal Component Analysis was applied to the correlation matrix of the standardized augmented variables. Eigenvalues and eigenvectors were computed, and components were ordered according to explained variance.
Let the first principal component be defined as

$$PCAI(s) = \sum_{i=1}^{p} \alpha_i \tilde{X}_i(s) + \sum_{i<j} \beta_{ij} \tilde{X}_{ij}(s), \tag{4}$$

where α_i and β_{ij} are elements of the eigenvector corresponding to the largest eigenvalue of the covariance matrix.
In this study, only the first principal component was retained, as it accounts for the largest proportion of total variance and serves as a one-dimensional composite secondary variable suitable for cokriging.

Step 4. Integration into the Cokriging Framework
The resulting $PCAI$ score, denoted by $PCAI(s)$, is then treated as a single composite secondary variable in the cokriging model for predicting $Z(s_0)$ at an unsampled location s_0.

Thus, the cokriging predictor becomes

$$\hat{Z}(s_0) = \sum_{i=1}^{n} \lambda_i Z(s_i) + \sum_{i=1}^{n} \mu_i \,\mathrm{PCAI}(s_i), \tag{5}$$

where $\lambda_i$and $\mu_i$are cokriging weights determined from the auto- and cross-variogram models.

By incorporating interaction terms before dimensional reduction, $PCAI$ captures nonlinear interdependencies among secondary variables that standard PCA cannot represent, thereby improving predictive performance in nonlinear multicovariate spatial settings. The same scaling and component selection procedures were applied to both PCA and $PCAI$ to ensure comparability of results.

2.2 Simulation

The flowchart provides an easy understanding of how the simulation procedure is being done. In this study, Figure 1 is the flowchart of the simulation procedure. For clarity and reproducibility, the simulation procedure is presented below as an ordered sequence of steps, specifying (i) the distributions used for random generation, (ii) the variogram-based spatial dependence imposed on simulated random fields, and (iii) how these components enter the construction of primary and secondary variables. The simulation was conducted in the following stages.

First, the location variables latitude and longitude were randomly generated from a uniform distribution with one (1) as the minimum and ten as the maximum values for the range of locations.

Second, spatially correlated random vectors $X$and Ywere generated as Gaussian geostatistical realizations. Specifically, independent innovations were first generated from a standard normal distribution, i.e., $\varepsilon \sim N(0,1)$, and were then used to simulate two zero-mean Gaussian random fields whose spatial dependence structure follows the selected theoretical variogram model (Spherical, Gaussian, or Exponential). In the simulation study, variogram parameters (range = 2, sill = 0.5, nugget = 0) were predetermined to ensure controlled comparison across methods and consistent evaluation of predictive performance under identical spatial dependence structures. For each model, the same conventional parameter values were used, producing two spatial random vectors $X = \big(X(s_1), \dots, X(s_n)\big)$and $Y = \big(Y(s_1), \dots, Y(s_n)\big)$ over the n generated locations. These simulated spatial vectors serve as the stochastic components embedded in the construction of the primary and secondary variables in Equations (6) and (7).

Third, the primary variable was listed as Equation 6:

$$A = X. \tag{6}$$

For the secondary variables, vectors B, C, D, PCA, and $PCAI$ were made through Equation 7:

$$V = \exp(\rho X + \sqrt{1-\rho^2}Y) \tag{7}$$

where $V = B$ when $\rho = 0.65$, $V = C$ when $\rho = 0.70$ and $V = D$ when $\rho = 0.75$. Different values of ρ were used to achieve the purpose of the study. In other words, different ρ value gives a different correlation result. Next, the principal component analysis was executed using the variables B, C, and D, and the result was named PCA. Then, using variables $B, C, D, B \times C, B \times D$, and $C \times D$, the principal components analysis was performed, and the result was named $PCAI$.

Fourth, the following three theoretical variogram spatial models combined the latitude and longitude location data simulated by a uniform distribution and simulated geostatistical data B, C, D, PCA, and $PCAI$. The data sets were generated with sizes $n = 100, 300,$ and 500. Each dataset has a size of 100 simulated values. That is, random vectors B, C, D, PCA, and $PCAI$ have n data sets where $n = 100,\ 300,\ 500$, and $m = 100$ observations in each data set. These data sets were constructed into one dataset with a size $(n \times 8)$ for the analyses.

Fifth, the cokriging was executed one by one. Variable A was used as a primary variable, and variables B, C, D, PCA, and $PCAI$ as secondary variables. Each cokriging method is subject to cross-validation and $PRESS$ comparison.

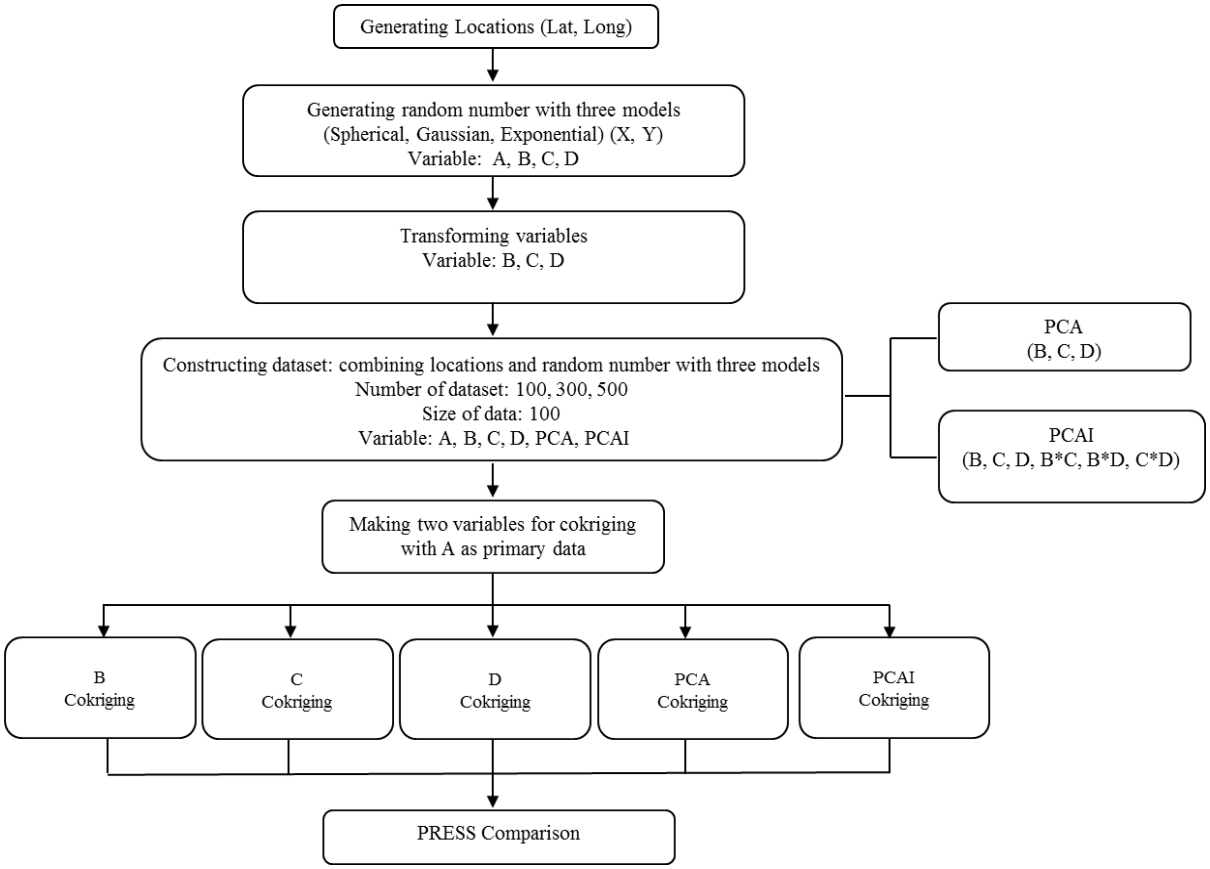


Figure 1: Flowchart of the simulation procedure for generating spatial data and evaluating the PCA-interaction cokriging method

Lastly, model performance was evaluated using the Predicted Residual Error Sum of Squares ($PRESS$) statistic, which is equivalent to leave-one-out cross-validation (LOOCV). $PRESS$ computes prediction errors by systematically removing one observation at a time and refitting the model. Similarly, leave-one-out cross-validation provides a robust framework for assessing predictive accuracy by systematically removing individual observations during model estimation and evaluating prediction errors on the omitted observations (Kuh et al., 2023)

For a dataset with n spatial observations, the $PRESS$ statistic is defined as:

$$\text{PRESS} = \sum_{i=1}^{n} \left(Z(s_i) - \hat{Z}_{-i}(s_i)\right)^2, \tag{8}$$

where $Z(s_i)$ is the observed value at location s_i, and $\hat{Z}_{-i}(s_i)$ is the predicted value at $s_i$obtained by fitting the model using all observations except the i-th observation.

2.3 Empirical Example

A total of 253 observations of the year 2015 South Korea Air Pollution data were used in this study. The raw data (Table 1) of this study has its respective location (latitude and longitude) that included five different air pollution data such as carbon monoxide (CO), particulate matter (PM_{10}), ozone (O_3), sulfur dioxide (SO_2), and nitrogen dioxide (NO_2). To further understand the context and applicability of this study, we consider utilizing secondary data from another country since the researcher's home country needs more data.

Table 1: Raw air-pollution measurements and spatial coordinates (South Korea, 2015)

Longitude	Latitude	CO	PM_{10}	O_3	SO_2	NO_2
129.18	35.33	0.36	30.9691	0.0124	0.0041	0.0123

128.62	36.81	0.76	55.2110	0.0898	0.0149	0.0438
128.11	37.36	0.77	60.1394	0.0872	0.0136	0.0406
127.44	36.34	0.55	44.6179	0.0195	0.0054	0.0187
126.98	37.57	0.43	34.1306	0.0139	0.0025	0.0159
⋮	⋮	⋮	⋮	⋮	⋮	⋮
128.12	35.18	0.57	55.9560	0.0843	0.0104	0.0365
126.72	37.45	0.82	60.5120	0.0898	0.0147	0.0448
128.60	35.88	0.4	32.5874	0.0148	0.0044	0.0121
127.13	37.46	0.71	57.1914	0.0885	0.0144	0.0403
129.36	36.03	0.59	53.3090	0.0817	0.0084	0.0283

The procedure in Figure 1 was followed to achieve the purpose of this study. A step-by-step analysis procedure using real air pollution data was used.

Step 1. Let CO be the chosen primary variable. On the other hand, the four remaining air pollutants: PM_{10}, O_3, SO_2, and NO_2, served as secondary variables.

Step 2. The correlation between the primary variable and all secondary variables was checked. Figure 2 shows the correlation matrix and scatter plot of the raw data. In the figure, the variables are written on a diagonal. The bivariate scatter plots with fitted lines are displayed at the bottom of the diagonal. The correlation value plus the significance level (*) is displayed on the diagonal top. It can be seen in the figure that the primary variable CO has a nonlinear correlation to all secondary variables PM_{10}, O_3, SO_2, and NO_2. Each is significant at $p < 0.001$ and is associated with a symbol (***).

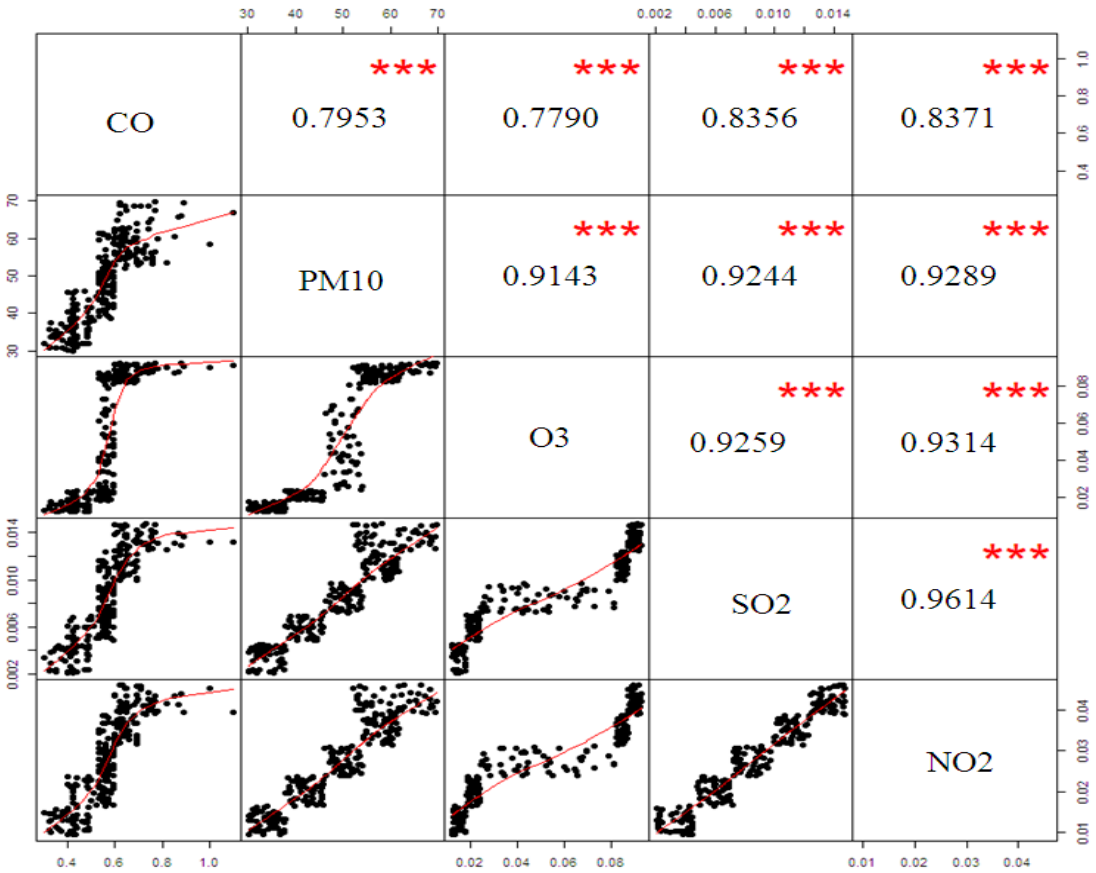


Figure 2: Correlation matrix and scatter plots of the raw air-pollution variables

Statistical and geostatistical considerations guided the choice of transformation functions. In cokriging, improving the linear association between primary and secondary variables enhances cross-variogram stability and predictive accuracy. Square root and exponential transformations were applied to address skewness, heteroscedasticity, and nonlinear dependence. Variables with larger, right-skewed distributions were adjusted to stabilize variance, while smaller-magnitude variables were square-root transformed to reduce dispersion. These transformations improve correlation structure and approximate linear relationships prior to PCA and $PCAI$ construction.

Step 3. All the raw data of secondary variables were transformed to attain a much higher correlation. For variable transformation, Equations 9 and 10 were used:

$$\exp(\rho Z(s) + \exp(1-\rho^2)\, X_i(s)) \tag{9}$$

$$\sqrt{\rho Z(s) + \sqrt{1-\rho^2}\, X_i(s)} \tag{10}$$

where $Z(s)$ is the primary variable and $X_i(s)$ is the second variable. Here, $\rho = 0.60$. Exponential was used for a variable containing large data values, such as PM_{10}. In comparison, the square root was used for variables containing small data values, such as O_3, SO_2, and NO_2.

Figure 3 shows the transformed data correlation matrix and scatter plot. Notice that the correlation values between the primary variable CO and secondary PM_{10}, O_3, SO_2, and NO_2 became higher compared to the raw data in Figure 2. Also, all secondary variables were statistically highly correlated, associated with a symbol (***) to the primary variable at a 0.001 significance level.

Step 4. Using the PCA, the four secondary variables PM_{10}, O_3, SO_2, and NO_2 were reduced into one dimension and named PCA. Next, an interaction of all the secondary variables PM_{10}, O_3, SO_2, NO_2, $O_3\ x\ PM_{10}$, $SO_2\ x\ PM_{10}$, $NO_2\ x\ PM_{10}$, $O_3\ x\ SO_2$, $NO_2\ x\ O_3$, and $NO_2\ x\ SO_2$, was conducted and then reduced into one dimension and named $PCAI$.

The result of the correlation analysis among variables and the correlation coefficient value used for the $PCAI$ is given in Table 2. All variables are highly significant at a 0.001 significance level.

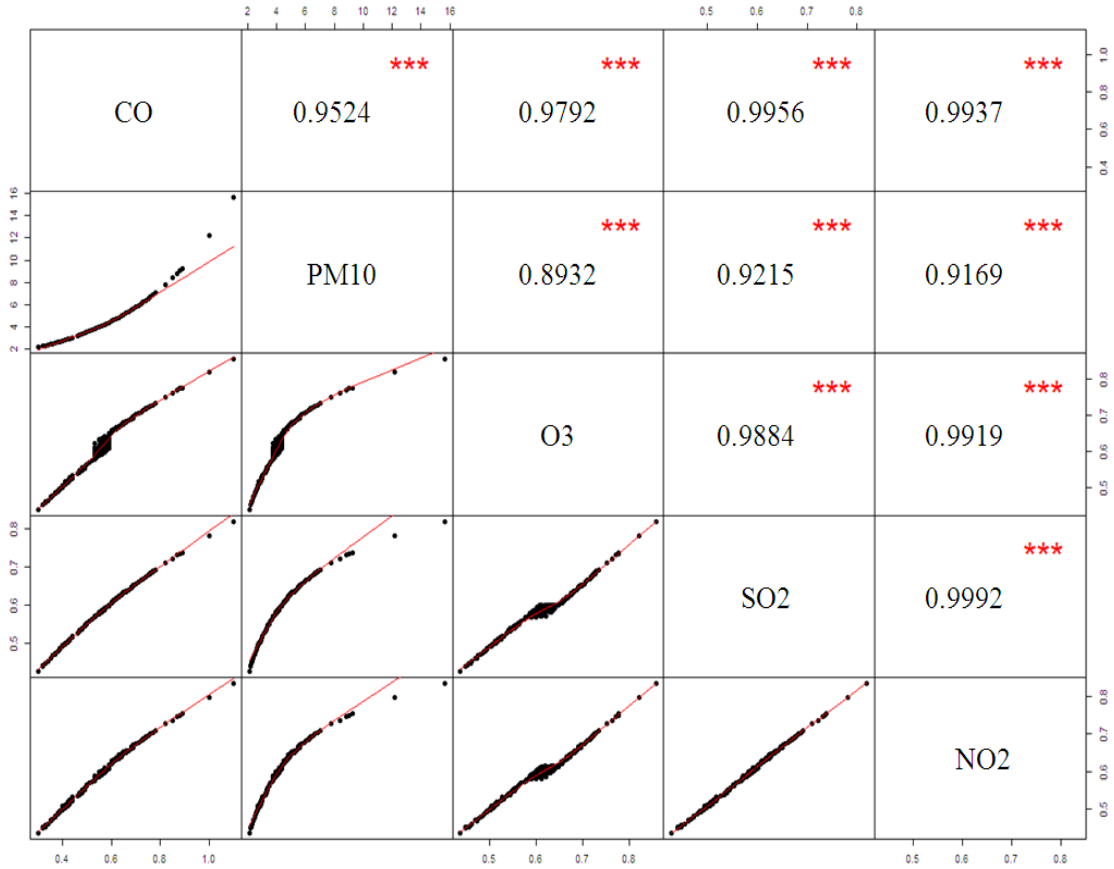


Figure 3: Correlation matrix and scatter plots of transformed secondary variables

Table 2: Correlation coefficients among transformed secondary variables used in PCA and $PCAI$

	PM_{10}	O_3	SO_2	NO_2	O_3xPM_{10}	SO_2xPM_{10}	NO_2xPM_{10}	O_3xSO_3	NO_2xO_3	NO_2xSO_3
PM_{10}	1.00									
O_3	0.8932***	1.00								
SO_2	0.9215***	0.9884***	1.00							
NO_2	0.9169***	0.9919***	0.9992***	1.00						
O_3xPM_{10}	0.9988***	0.8847***	0.9088***	0.9050***	1.00					
SO_2xPM_{10}	0.9987***	0.8713***	0.9011***	0.8962***	0.9992***	1.00				
NO_2xPM_{10}	0.9989***	0.8741***	0.9029***	0.8984***	0.9994***	0.9999***	1.00			
O_3xSO_3	0.9343***	0.9930***	0.9943***	0.9954***	0.9293***	0.9194***	0.9216***	1.00		
NO_2xO_3	0.9373***	0.9940***	0.9933***	0.9952***	0.9266***	0.9162***	0.9185***	0.9998***	1.00	
NO_2xSO_3	0.9475***	0.9847***	0.9964***	0.9960***	0.9377***	0.9304***	0.9322***	0.9976***	0.9967***	1.00

***p<0.001

Correspondingly, for the subsequent analyses, the earned principal component analyzed data sets served as secondary data. Table 3 is the analyzed principal component data with the location variables such as longitude and latitude.

Furthermore, the secondary variable PCA was obtained from the 253 x 4 fastened points, and $PCAI$ was obtained from 253 x 10 fastened points. Step 4 served as a center stage to put forward the proposed method in the context of this study.

Table 3: Principal component scores of transformed air-pollution variables and their spatial locations

Longitude	Latitude	PCA (Principal Component Score)	$PCAI$ (Principal Component Interaction Score)
129.18	35.33	-3.3472	-4.7951
128.62	36.81	3.1249	5.0049
128.11	37.36	3.2445	5.2291
127.44	36.34	3.2445	5.2291
126.98	37.57	-2.2350	-3.3443
⋮	⋮	⋮	⋮
128.12	35.18	0.3012	0.2802
126.72	37.45	4.0366	6.6957
128.60	35.88	-2.6996	-3.9614
127.13	37.46	2.3657	3.6622
129.36	36.03	0.5249	0.6387

Step 5. One at a time, the cokriging was executed. All 253 observations were analyzed. CO is the primary variable, and PM_{10}, O_3, SO_2, NO_2, PCA, and $PCAI$ as secondary variables. Here, six cokriging results were expected for result comparison.

For the empirical analysis, variogram parameters (range, sill, nugget) were estimated from the data using experimental variogram fitting procedures. The theoretical model (Exponential) and corresponding parameters were selected based on visual fit and least-squares minimization between the empirical and theoretical variograms.

Step 6. The sample variogram was estimated using the primary variable CO and the secondary variable PM_{10} with parameters range=0.8, sill=0.045, and nugget=0.00. The performed cokriging resulted in a fitted model as Exponential, range=0.8 ($\simeq 88.8\,km$), sill=0.144394189, and nugget=0.071230276. The CO variogram has an Exponential model with range=0.8, sill=0.012120766, and nugget=0.005981423 while PM_{10} has an Exponential model with range=0.8, sill=1.862518661, and nugget=0.917905866. Figure 4 shows the auto- and cross-variograms for CO and PM_{10} fitted as an Exponential model in Equation 11.

Exponential model (CO, PM_{10}):

$$\gamma_r(d;\theta) = \begin{cases} 0\,, & d = 0 \\ 0.071230276 + 0.144394189\left[1 - exp\left(-\frac{d}{0.8}\right)\right], & d \neq 0 \end{cases} \tag{11}$$

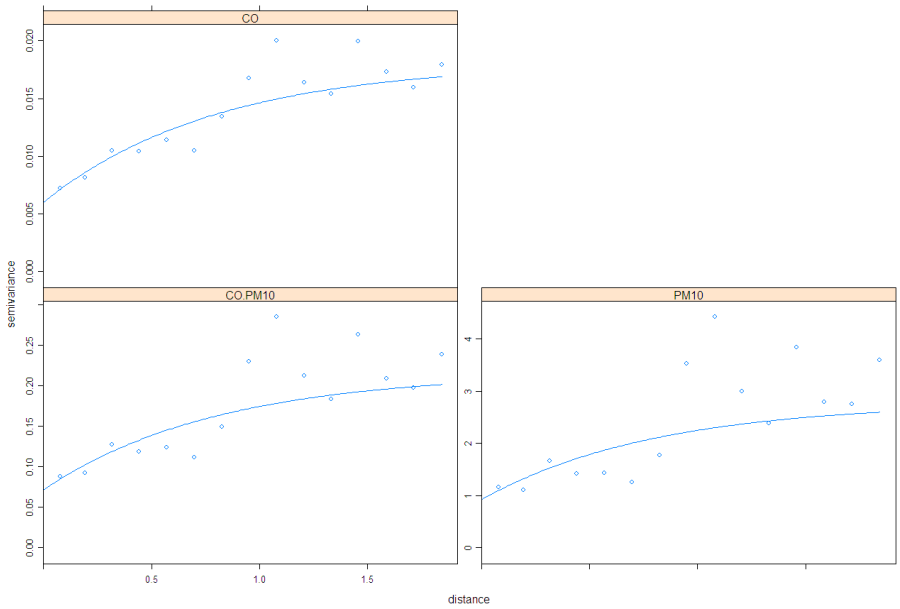


Figure 4: Auto and cross variogram plot for CO and PM_{10}

Step 7. The sample variogram was estimated using the primary variable CO and secondary variable O_3 with parameters range=0.8, sill=0.009, and nugget=0.004. The performed cokriging resulted from a fitted model as Exponential, range=0.8 ($\simeq 88.8\,km$), sill=0.007657712, and nugget=0.003749383. The CO variogram has an Exponential model with range=0.8, sill=0.012120766, and nugget=0.005981423 while O_3 has an Exponential model with range=0.8, sill=0.004933472, and nugget=0.002518936. Figure 5 shows the auto- and cross-variograms for CO and O_3 fitted with the Exponential model in Equation 12.

Exponential model (CO, O_3):

$$\gamma_r(d;\theta) = \begin{cases} 0\,, & d = 0 \\ 0.003749383 + 0.007657712\left[1 - exp\left(-\frac{d}{0.8}\right)\right], & d \neq 0 \end{cases} \quad (12)$$

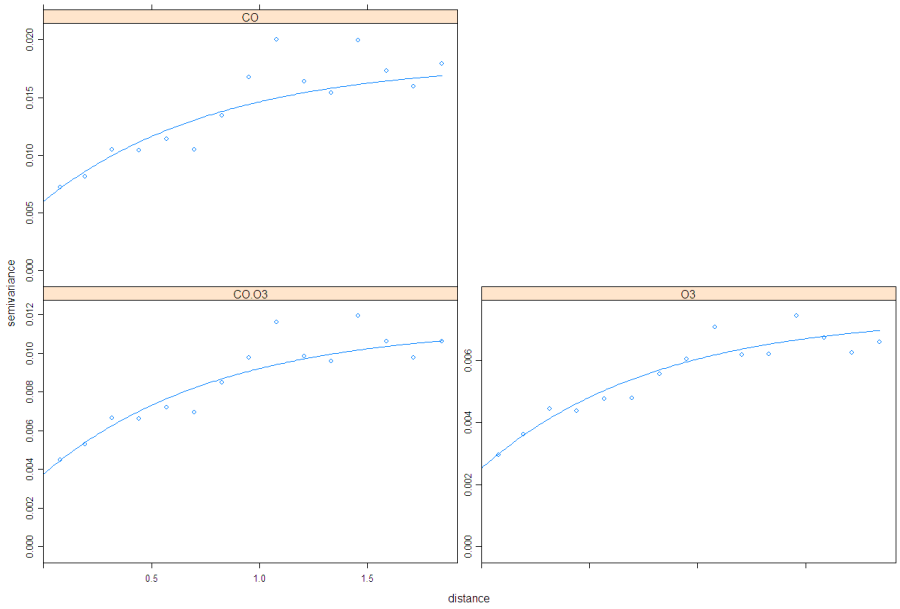


Figure 5: Auto and cross variogram plot for CO and O_3

Step 8. The sample variogram was estimated using the primary variable CO and the secondary variable SO_2 with parameters range=0.8, sill=0.045, and nugget=0.00. The cokriging resulted from a fitted model as exponential, range=0.8 ($\simeq 88.8\,km$), sill=0.006446285, and nugget=0.003095572. The CO variogram has an exponential model with range=0.8, sill=0.012120766, and nugget=0.005981423, while SO_2 has an exponential model with range=0.8, sill=0.003449920, and nugget=0.001616220. Figure 6 shows the auto- and cross-variograms for CO and SO_2 fitted to the exponential model in Equation 13.

Exponential model (CO, SO_2):

$$\gamma_r(d;\theta) = \begin{cases} 0\,, & d = 0 \\ 0.003095572 + 0.006446285\left[1 - exp\left(-\frac{d}{0.8}\right)\right], & d \neq 0 \end{cases} \quad (13)$$

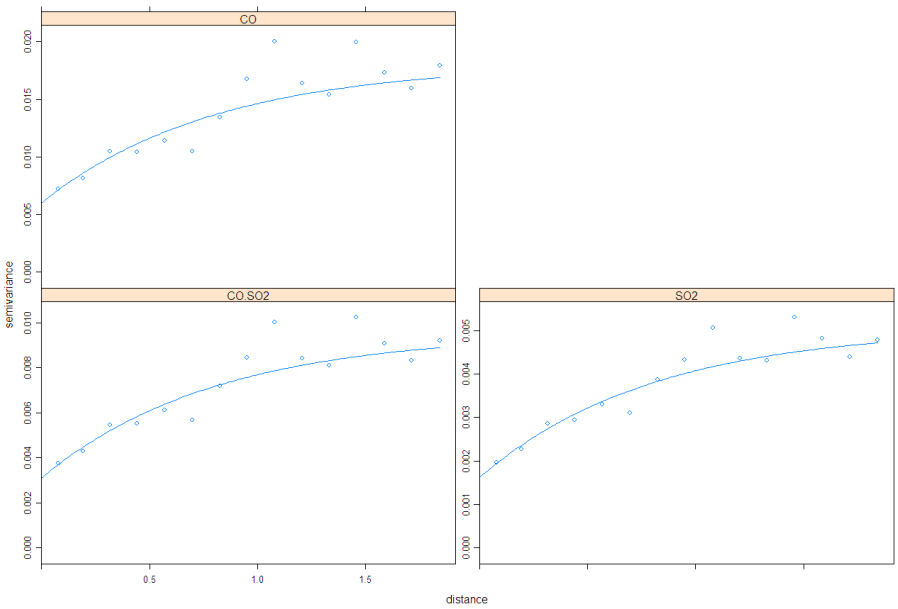


Figure 6: Auto and cross variogram plot for CO and SO_2

Step 9. The sample variogram was estimated using the primary variable CO and secondary variable NO_2 with parameters range=0.8, sill=0.045, and nugget=0.00. The cokriging resulted from a fitted model as Exponential, range=0.8 ($\simeq 88.8\,km$), sill=0.006750542, and nugget=0.003216978. The CO variogram has an Exponential model with range=0.8, sill=0.012120766, and nugget=0.005981423, while NO_2 has an Exponential model with range=0.8, sill=0.003794747, and nugget=0.001753470. Figure 7 shows the auto- and cross-variograms for CO and NO_2 fitted with the Exponential model in Equation 14.

Exponential model (CO, NO_2):

$$\gamma_r(d;\theta) = \begin{cases} 0\,, & d = 0 \\ 0.003216978 + 0.006750542\left[1 - exp\left(-\frac{d}{0.8}\right)\right], & d \neq 0 \end{cases} \quad (14)$$

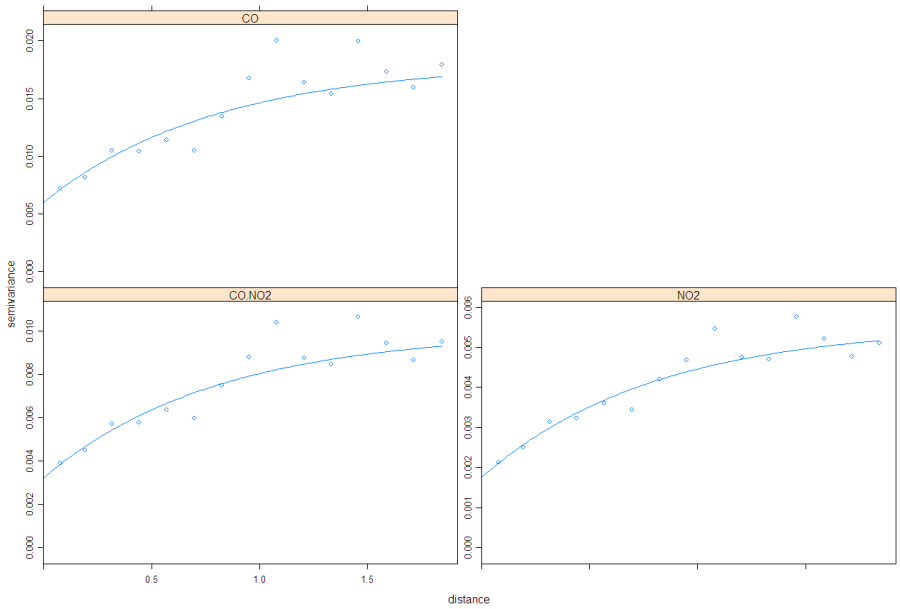


Figure 7: Auto and cross variogram plot for CO and NO_2

Step 10. The sample variogram was estimated using the primary variable CO and the secondary variable PCA with parameters range=0.8, sill=0.045, and nugget=0.00. The cokriging resulted from a fitted model as Exponential, range=0.8 ($\simeq 88.8\,km$), sill=0.200291310, and nugget=0.097075082. The CO variogram has an Exponential model with range=0.8, sill=0.012120766, and nugget=0.005981423, while PCA has an Exponential model with range=0.8, sill=3.313273269, and nugget=1.585483030. Figure 8 shows the auto- and cross-variograms for CO and PCA fitted with the Exponential model in Equation 15.

Exponential model (CO, PCA):

$$\gamma_r(d;\theta) = \begin{cases} 0\,, & d = 0 \\ 0.097075082 + 0.200291310\left[1 - exp\left(-\frac{d}{0.8}\right)\right], & d \neq 0 \end{cases} \quad (15)$$

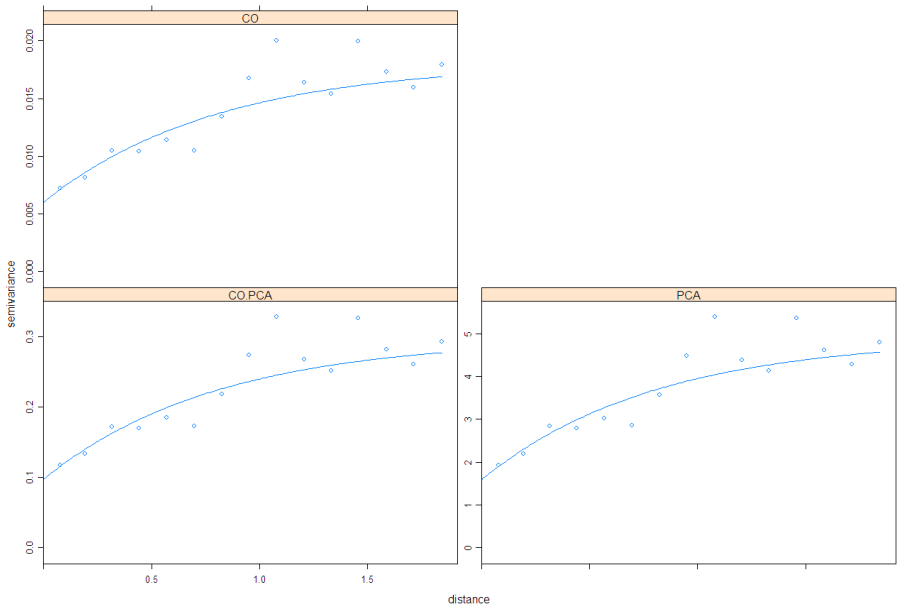


Figure 8: Auto and cross variogram plot for CO and PCA

Step 11. This time, the sample variogram was estimated using the primary variable CO and the secondary variable $PCAI$ with parameters range=0.8, sill=0.10, and nugget=0.00. The cokriging resulted from a fitted model as Exponential, range=0.8 ($\simeq 88.8\,km$), sill=0.312922908, and nugget=0.153143267. The CO variogram has an Exponential model with range=0.8, sill=0.012120766, and nugget=0.005981423, while $PCAI$ has an Exponential model with range=0.8, sill=8.138403629, and nugget=3.969611314. Figure 9 shows the auto- and cross-variogram for CO and $PCAI$ fitted with the Exponential model in Equation 16.

Exponential model (CO, $PCAI$):

$$\gamma_r(d;\theta) = \begin{cases} 0\,, & d = 0 \\ 0.153143267 + 0.312922908\left[1 - exp\left(-\frac{d}{0.8}\right)\right], & d \neq 0 \end{cases} \tag{16}$$

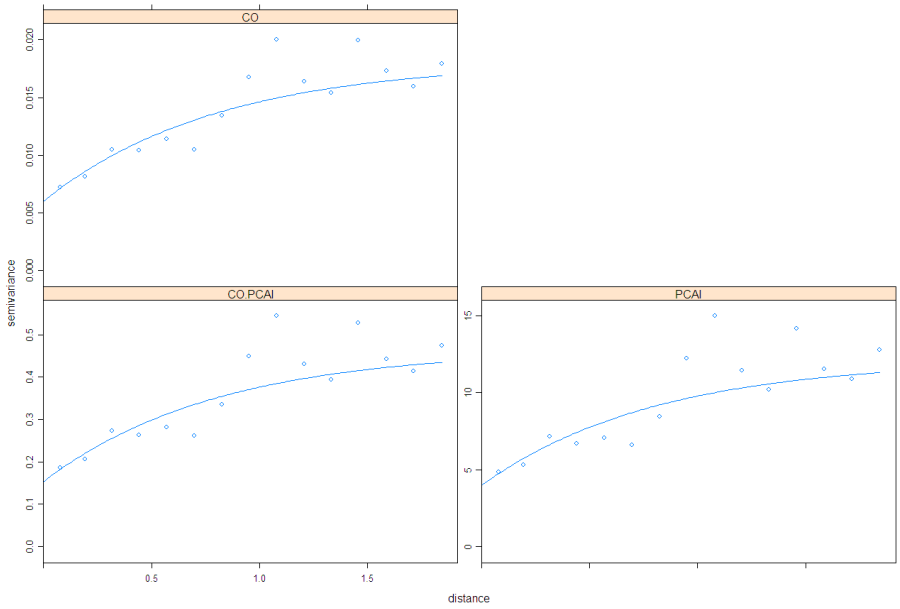


Figure 9: Auto and cross variogram plot for CO and $PCAI$

Step 12. The summary of cross variograms for CO and the secondary variables fitted to the models are presented in Table 4. Observe that in the cross-variogram results, the range=0.8($\simeq 88.8\,km$). All secondary variables are the same for the reason that these data came from the same locations. The obtained predicted results of cokriging, PCA, and $PCAI$ cokriging executed from [Step 1] to [Step 11] were then compared. Here, all the data within the set were used for cross-validation.

Table 4: Summary of cross variograms for CO and secondary variables fitted to the model

Secondary Variable	Model	Sill	Nugget	Range
PM_{10}	Exponential	0.1444	0.0712	0.8
O_3	Exponential	0.0076	0.0037	0.8
SO_2	Exponential	0.0064	0.0030	0.8
NO_2	Exponential	0.0068	0.0032	0.8
PCA	Exponential	0.0971	0.2003	0.8
$PCAI$	Exponential	0.3129	0.1531	0.8

3. Results and Discussion

3.1 Simulation

As the simulation procedure in Figure 1 was followed, the results are summarized in Table 5. Given that A is the primary variable in all cokriging computations, we call the $PRESS$ results B Cokriging for using B as a secondary variable. Also, C Cokriging for using C and D Cokriging for using D as a secondary variable, respectively. Moreover, PCA Cokriging for the $PRESS$ result uses PCA as a secondary variable and $PCAI$ Cokriging for using $PCAI$ as a secondary variable correspondingly.

The result was obtained by generating, transforming, and simulating 100, 300, and 500 data sets, each with 100 data points. With these data sets, the cokriging was performed with three theoretical variogram models: Spherical, Gaussian, and Exponential. Then, the results were compared using the $PRESS$ statistic and used as a criterion for the cross-validation of the proposed method. Only isotopic data were used in this study. Wackernagel (1998) emphasized that data is isotopic when the primary data shares the exact location as the secondary data.

All reported values correspond to the $PRESS$ statistic obtained under leave-one-out cross-validation; Mean Square Error (MSE) was not used as the evaluation criterion. The simulated mean $PRESS$ result in Table 5 shows that using $PCAI$ as a secondary variable, gives the lowest $PRESS$ value compared to only one secondary variable and PCA in all three models with 100, 300, and 500 data sets. Considering the criterion, the lowest value of $PRESS$ indicates the best model. Thus, $PCAI$ cokriging outperformed PCA cokriging. In other words, using $PCAI$ as a secondary variable is a better method compared to PCA for cokriging with nonlinear multicovariate.

Table 5: Comparison of simulated mean $PRESS$ results using three models

Dataset	Model	B Cokriging	C Cokriging	D Cokriging	PCA Cokriging	$PCAI$ Cokriging
100	Spherical	0.16914	0.16138	0.15182	0.12223	0.11287
	Gaussian	1.28484	1.220938	1.16067	1.21940	1.07211
	Exponential	0.08087	0.77349	0.07395	0.06133	0.04541
300	Spherical	0.12832	0.12016	0.11066	0.09178	0.06866
	Gaussian	1.08555	1.03045	0.97757	1.02914	0.90649
	Exponential	0.12250	0.11502	0.10633	0.08433	0.06176
500	Spherical	0.12858	0.12046	0.11113	0.09344	0.06714
	Gaussian	1.04174	0.99144	0.94138	0.98995	0.88606
	Exponential	0.09729	0.08881	0.08003	0.06265	0.04676

3.2 Empirical Example

$PRESS$ (Predicted error sum of squares) was used as a criterion to compare the results between Cokriging, PCA cokriging, and $PCAI$ cokriging. Given that CO was the primary variable in all cokriging computations, the $PRESS$ result was called PM_{10} Cokriging for using PM_{10} as a secondary variable. Also, O_3 Cokriging for using O_3, SO_2 Cokriging for using SO_2, and NO_2 Cokriging for using NO_2 as a secondary variable. Moreover, PCA Cokriging for the $PRESS$ result uses PCA as a secondary variable and $PCAI$ Cokriging for using $PCAI$ as a secondary variable correspondingly.

Table 6: $PRESS$ Comparison result

Cokriging	$PRESS$
PM_{10} Cokriging	0.00745564
O_3 Cokriging	0.00254374
SO_2 Cokriging	0.00980486
NO_2 Cokriging	0.01109226
PCA Cokriging	0.00505586
$PCAI$ Cokriging	0.00041369

The result in Table 6 shows that $PCAI$ Cokriging ($PRESS$=0.00041369) is superior compared to PM_{10} Cokriging ($PRESS$=0.00745564), O_3 Cokriging ($PRESS$=0.00254374), SO_2 Cokriging ($PRESS$=0.00980486), NO_2 Cokriging ($PRESS$=0.01109226), and PCA Cokriging ($PRESS$=0.00254374) respectively. The $PCAI$ Cokriging outperformed the other cokrigings. This indicates that the $PCAI$ Cokriging as a secondary variable provides better prediction performance. Based on this outcome, the study yields good results, confirming the validity of the proposed method.

The improved performance of $PCAI$ cokriging demonstrates that incorporating interaction-derived variance structures prior to dimensionality reduction enhances spatial prediction accuracy under nonlinear multicovariate conditions

4. Conclusion

The cokriging technique in this study is superior to kriging for spatial prediction, given the use of multivariate spatial data. The benefits of cokriging increase as secondary information becomes more abundant than primary information. However, in cokriging, a secondary variable is used to estimate the location value.

The analysis shows that the proposed method is reasonable. On the other hand, reducing multivariate secondary data to one dimension (PCAI) is better than directly reducing it to one dimension (PCA). The validity of the proposed method was proved through the criterion process.

Data Availability: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgements: This study was partly supported by a West Visayas State University - University Research and Development Center grant. In addition, Dr. Seung Bae Choi and Dr. Chang Wan Kang of Dong-Eui University, Busan, South Korea, provided the data used in this study.

References

1. Acevedo, M. F. (2012). Data analysis and statistics for geography, environmental science, and engineering. CRC Press.
2. Ahmed, S. A., & Siddiqui, J. S. (2014). Principal component analysis to explore climatic variability and dengue outbreak in Lahore. Pakistan Journal of Statistics and Operation Research, 10(2), 247. https://doi.org/10.18187/pjsor.v10i2.686
3. Allen, D. M. (1971). Mean Square error of prediction as a criterion for selecting variables. Technometrics, 13(3), 469–475. https://doi.org/10.1080/00401706.1971.10488811
4. Bivand, R. S., Pebesma, E. J., & Gómez-Rubio, V. (2013). Applied spatial data analysis with R. Springer Science & Business Media.

5. Cartone, A., & Postiglione, P. (2021). Principal component analysis for geographical data: the role of spatial effects in the definition of composite indicators. Spatial Economic Analysis, 16(2), 126-147. DOI: 10.1080/17421772.2020.1775876
6. Chica-Olmo, J. (2007). Prediction of housing location price by a multivariate spatial method: Cokriging. Journal of Real Estate Research, 29(1), 91-114. https://doi.org/10.1080/10835547.2007.12091188
7. Choi, S., Kang, C., & Cho, J. (2010). Data-dependent choice of optimal number of lags in Variogram estimation. Korean Journal of Applied Statistics, 23(3), 609-619. https://doi.org/10.5351/kjas.2010.23.3.609
8. Dale, M. R., & Fortin, M. (2014). Spatial analysis: A guide for ecologists. Cambridge University Press.
9. Isaaks, E. H., & Srivastava, R. M. (1989). An introduction to applied geostatistics. Oxford University Press.
10. Kang, C. W., Choi, S. B., & Cho, J. S. (2008). Study on Cokriging Method for Multivariate Spatial Data. Journal of the Korean Data Analysis Society, 10(5), 2661-2668.
11. Kang, H. C., Han, S. T., Choi, J. H., Lee, S. K., Kim, E. S., Eum, I. H., & Kim, M. K. (2006). Methodology of Data Mining. Freedom Academy Press: Paju, Korea.
12. Kuh, S., Kennedy, L., Chen, Q., & Gelman, A. (2023). Using leave‐one‐out cross validation (LOO) in a multilevel regression and poststratification (MRP) workflow: A cautionary tale. Statistics in Medicine, 43(5), 953-982. https://doi.org/10.1002/sim.9964
13. Legendre, P. (1993). Spatial Autocorrelation: Trouble or New Paradigm? Ecology, 74(6), 1659–1673. https://doi.org/10.2307/1939924
14. Li, Y., Huang, C., & Härdle, W. K. (2019). Spatial functional principal component analysis with applications to brain image data. Journal of Multivariate Analysis, 170, 263-274. https://doi.org/10.1016/j.jmva.2018.11.004
15. Li, Y., Qiu, Y., & Xu, Y. (2022). From multivariate to functional data analysis: Fundamentals, recent developments, and emerging areas. Journal of Multivariate Analysis, 188, 104806. https://doi.org/10.1016/j.jmva.2021.104806
16. Lloyd, C. D. (2010). Local models for spatial analysis (2nd ed.). CRC Press.
17. Maindonald, J., & Braun, J. (2010). Data analysis and graphics using R: An example-based approach (3rd ed.). Cambridge University Press.
18. O'Sullivan, D., & Unwin, D. (2014). Geographic information analysis. John Wiley & Sons.
19. Parreño, G., Kim, K. K., Kang, C., & Choi, S. (2017a). Prediction simulation study for Cokriging via variable transformation. The Korean Data Analysis Society, 19(5), 2311-2321. https://doi.org/10.37727/jkdas.2017.19.5.2311
20. Parreño, G., Kim, K. K., Kang, C., & Choi, S. (2017b). A study on prediction comparison of Kriging and Cokriging using PCA. The Korean Data Analysis Society, 19(4), 1721-1732. https://doi.org/10.37727/jkdas.2017.19.4.1721
21. Pereda, E., Quiroga, R. Q., & Bhattacharya, J. (2005). Nonlinear multivariate analysis of neurophysiological signals. Progress in Neurobiology, 77(1-2), pp. 1–37. https://doi.org/10.1016/j.pneurobio.2005.10.003
22. Ripley, R. D. (1981). Spatial Statistics. John Wiley and Sons, New York.
23. Rocha, M. M., Yamamoto, J. K., Watanabe, J., & Fonseca, P. P. (2012). Studying the influence of a secondary variable in Collocated Cokriging estimates. Annals of the Brazilian Academy of Sciences, 84(2), 335-346.
24. Schabenberger, O., & Gotway, C. A. (2017). Statistical methods for spatial data analysis. CRC Press.
25. Si-ahmed, I., Hamdad, L., Agonkoui, C. J., Kande, Y., & Dabo-Niang, S. (2025). Principal component analysis of multivariate spatial functional data. Big Data Research, 39, 100504. https://doi.org/10.1016/j.bdr.2024.100504
26. Usman, U., A., Yelwa, S., Gulumbe, S., & Danbaba, A. (2013). An assessment of the changing climate in Northern Nigeria using Cokriging. American Journal of Applied Mathematics and Statistics, 1(5), 90–98. https://doi.org/10.12691/ajams-1-5-3
27. Wackernagel H. (1998). Isotopic Cokriging. In: Multivariate Geostatistics. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-03550-4_25
28. Wackernagel, H. (2013). Multivariate Geostatistics: An introduction with applications. Springer Science & Business Media.
29. Zhong, Q., & Song, X. (2025). Functional nonlinear principal component analysis. Computational Statistics and Data Analysis, 209. https://doi.org/10.1016/j.csda.2025.108169