Available online at http://iorajournal.org/indx.php/orics /index

Operations Research: International Conference Series

e-ISSN: 2722-0974
p-ISSN: 2723-1739

Vol. 3, No. 3, pp. 118-126, 2022

Performance Comparison of Covariance Function to Interpolate Unsampled Points with Simulation Data in Manado City

Claudya Soleman[1*] Winsy Weku[2], Deiby Salaki[3]

[1,2,3] *Departement of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Sam Ratulangi, Manado, Indonesia*

**Corresponding author winsy_weku@unsrat.ac.id*

Abstract

The covariance function measures the strength of statistical correlation as a function of distance. This follows Tobler's law which states that everything is usually related to all else but those which are near to each other are more related when compared to those that are further away. The correct weight of the basic covariance structure will produce the optimal kriging predictor. One interesting way to evaluate the strength of a kriging interpolation is to perform a simulation using a spatial structure. The simulation technique is executed in Manado City. The data is then applied to the variogram model using the spherical and matern covariance functions. The type of kriging method used in this simulation is ordinary kriging. The result shows that the suitable model to use is the matern model. Residual results from cross-validation show that the matern model has a lower biased estimation on both data. According to the RMSE and MAE criteria, the matern model outperforms the spherical model on data A and data B. The results of the interpolation are then visualized in the form of a map. Through this research, it can be concluded that the accuracy of the selection of the covariance function in the variogram model will provide a good estimate for the kriging method, and the most appropriate model for this case is the matern model.

Keywords: Covariance function, interpolation, simulation data

1. Introduction

Covariance modeling plays a key role in spatial data analysis because it provides information about the dependency structure of the underlying processes and determines the performance of spatial predictions. Various parametric models have been developed to accommodate the idiosyncratic features of a given dataset. However, parametric models can impose unjustified restrictions on the covariance structure and the procedure for selecting a particular model is often ad-hoc (Choi, 2014).

Covariance measures the assumption that things that are close tend to be more similar than things that are far apart. Covariance measures the strength of statistical correlation as a function of distance. The process of modeling the covariance function adjusts the curve to the available empirical data. It aims to reach the most suitable model which will then be used in the predictions (Esri, 2021).

The correct weight of the basic covariance structure will produce the optimal kriging predictor. An important, but not well-understood issue in kriging theory is the effect on the accuracy of the kriging predictor in substituting optimal weights for weights derived from the estimated covariance structure. In practice, the structure of this covariance is unknown and estimated from the data (Putter, 2021).

A comparison of the kriging interpolation variance model has been studied by (Marko, 2013) regarding the spatial distribution of groundwater quality, and the prediction of groundwater chemical parameters was carried out using geostatistical analysis on Geographic Information System (GIS) software using the Ordinary Kriging method. The best model for each parameter was assessed based on the root mean square error (RMSE) criteria. Rozalia et al. (2016) (Rozalia, 2014) have also researched estimating NO2 levels in the air in the city of Semarang using the Ordinary Kriging method, which then made a comparison between several variance models, namely spherical, exponential, and gaussian to get the best model to be used in the estimation. Based on this analysis, it was found that the best model is the spherical model which produces the highest estimation of Nitrogen Dioxide in Sub Gebangsari and the lowest Nitrogen Dioxide in Sub Patemon. (Sophal, 2014) is also one of the groups that examine the spatial distribution of soil

data in the Imba-Numa watershed, Japan. They used three geostatistical interpolation methods: Ordinary Kriging (OK), Universal Kriging (UK), and Inverse Distance Weighting (IDW), applied with 4 models, namely the Gaussian model, Exponential model, Pentaspherical model, and Hole-Effect model for each method. Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Error Percentage (MEP) were used to evaluate the performance of the method. Comparison Cross Validation is also used to compare and validate methods. The Matern semivariogram model was used in a study conducted by Pardo-Iguzquiza & (Chica-Olmo, 2008). They use Matern's model in modeling a natural phenomenon whose critical characteristic of spatial variability is random plane continuity.

Based on the problems of several previous studies, it can be seen that the covariance function has a significant role in variogram modeling and spatial interpolation. Therefore, a study that focuses on the comparison of interpolation with the covariance function of the Spherical and Matern models will be carried out using simulation data.

2. Description of the study area and data collection

The simulation technique is executed in Manado City. According to BPK RI data, the total population in Manado (January 2014) is approximately 430,790. The area of Manado City consists of land area and archipelagic area with a total area of 15,726 ha. The archipelago includes Bunaken island, Manado Tua island, and Siladen island. Administratively, Manado City is divided into 11 sub-districts and 87 districts. Manado City is located at the northern tip of Sulawesi Island and is the largest city in North Sulawesi as well as the capital city of North Sulawesi Province. Geographically, it is located between 1^0 25′ 88″ – 1^0 39′ 50″ North Latitude and 124^0 47′ 00″ – 124^0 56′ 00″ East Longitude.

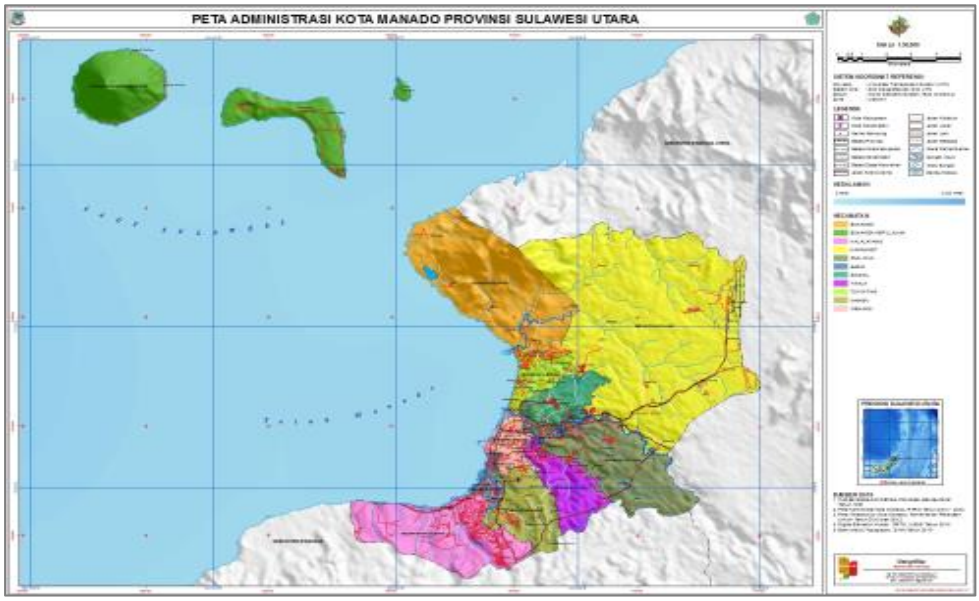
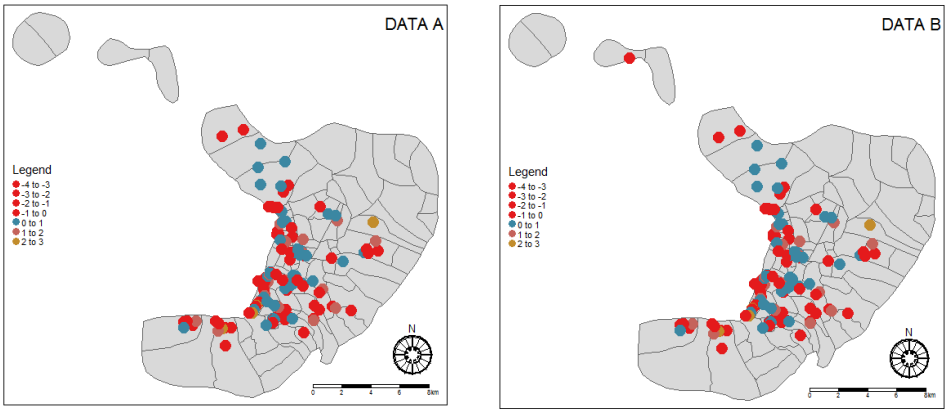


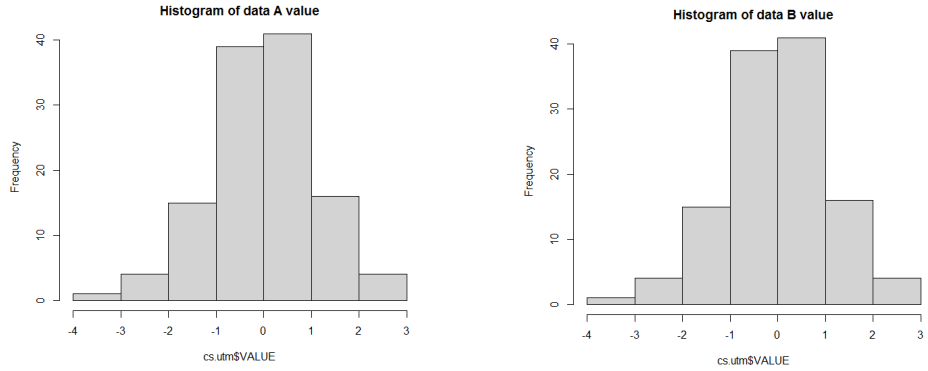
FIGURE 1. Study area (source : https://petatematikindo.wordpress.com)



(a) (b)

Figure 2: (a) Data A location points/without outlier, (b) Data B location points/with outlier

This simulation is done using 2 data. Data A consist of 120 random location points on the main island. Data B consist of 119 location points which are the same as data A but have a different location point that is located on one of the small islands which are considered as an outlier. The values of these points are generated randomly with a normal distribution (with $\mu=0$, $\sigma=1$). Data A and B have the same value. It means to see whether the prediction results with the same covariance function model will change due to the location point outlier or not. Dark purple color represent lower concentration and green color represent high concentration.



(a) (b)

Figure 3: (a) Histogram Data A, (b) Histogram Data B

2.1. Methods of Analysis

The data interpolation process usually considers the normality of the data. However, by conducting simulations, it is intended that the data used can be fully controlled. Data A and data B values have been generated with a normal distribution since the beginning. Therefore data normalization process is not necessary. This can be seen in Figure 3.

The geostatistical analysis is used to model the spatial distribution of the simulation data. It relates to autocorrelation spatial data that has a basic spatial structure or pattern that can be realized in (semi)variogram analysis. (Semi)variogram is a characterization of the spatial correlation of the variables studied. The selection of the theoretical variogram model is appropriate by fitting the theoretical variogram model is under the experimental variogram. Non-monotone variograms with periodicity will converge to a constant when lag increases, or is called sill. The oscillation that occurs can cause the curve to go through sill for several times (Weku, 2019). The semivariogram shows the relationship between the lag distance on the horizontal axis and the semivariogram value on the vertical axis.

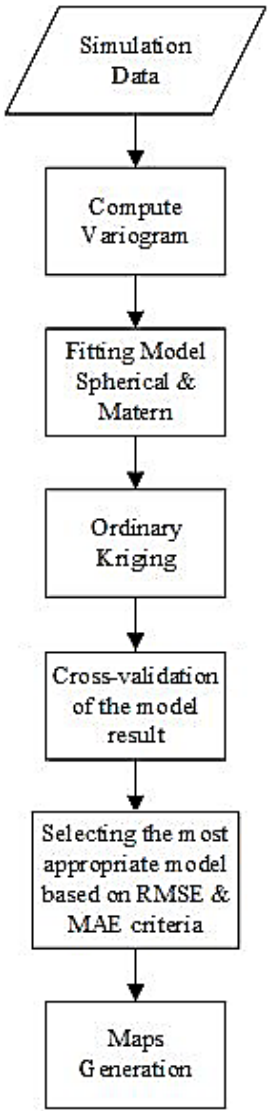


Figure 4: Flow chart of the steps followed for the geostatistical analysis

Theoretically, to calculate the semivariogram, the following formula is usually used:

$$\gamma(\mathrm{h}) = \frac{1}{2\mathrm{n(h)}}\sum_{\mathrm{i=1}}^{\mathrm{m(h)}}[\mathrm{Z(x_i)} - \mathrm{Z(x_i + h)}]^2 \quad (1)$$

where *γ(h)* is the semivariogram value for the lag distance *(h)*, *n(h)* is the number of pairs of variables separated by the lag distance *(h)*, and *Z(x)* is the variable value [4].

(Cressie, 1993) recommends an experimental variogram using a robust estimator to handle the presence of outliers, which is written as follows:

$$\gamma(\mathrm{h}) = \frac{1}{2}\left(\frac{\frac{1}{|N(h)|}\sum_{|N(h)|}\left(\left|Z(s_i)-Z(s_j)\right|^{1/2}\right)^4}{(0.457+0.494/|N(h)|)}\right) \quad (2)$$

The Kriging method is a geostatistical analysis method used to interpolate a content value based on sample data taken in irregular places (Rozalia, 2014). Ordinary Kriging (OK) is one of the most basic kriging methods. At the no-sampled location x_0, Z is estimated by:

$$\mathrm{Z(x_0)} = \sum_{\mathrm{i=1}}^{\mathrm{n}} \lambda_\mathrm{i} \mathrm{Z(x_i)} \quad (3)$$

where $Z(x_0)$ is the estimated value of the random variable *(RV) Z* at the unsampled location x_0 and λ_i is the n weight assigned to the observation point $Z(x_i)$. The weight of λ_i is one to ensure unbiased conditions and is found by minimizing the variance of the estimate. *RV Z(x)* can be decomposed into trend component *m(x)* and residual component *R(x)*:

$$Z(x) = m(x) + R(x) \tag{4}$$

OK assumes mean stationarity and assumes *m(x)* to be a constant, but unknown value. Nonstationary conditions are accounted for by confining the stationary domain to the local environment and moving it across the study area. The residual component *R(x)* is modeled as a stationary *RV* with a mean of zero and with an intrinsic stationary assumption (Sophal, 2014).

For the cross-validation that was applied, N-fold cross-validation makes partitions the data set in N parts. For all observations in a part, predictions are made based on the remaining N-1 parts; this is repeated for each of the N parts (Pebesma, 2021). RMSE and MAE are used to investigate the most suitable model by comparing the values. The smallest RMSE and MAE values indicate the model that is most compatible with the data.

$$\boldsymbol{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\mathcal{E}_i^2} \tag{5}$$

$$\boldsymbol{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\mathcal{E}^2| \tag{6}$$

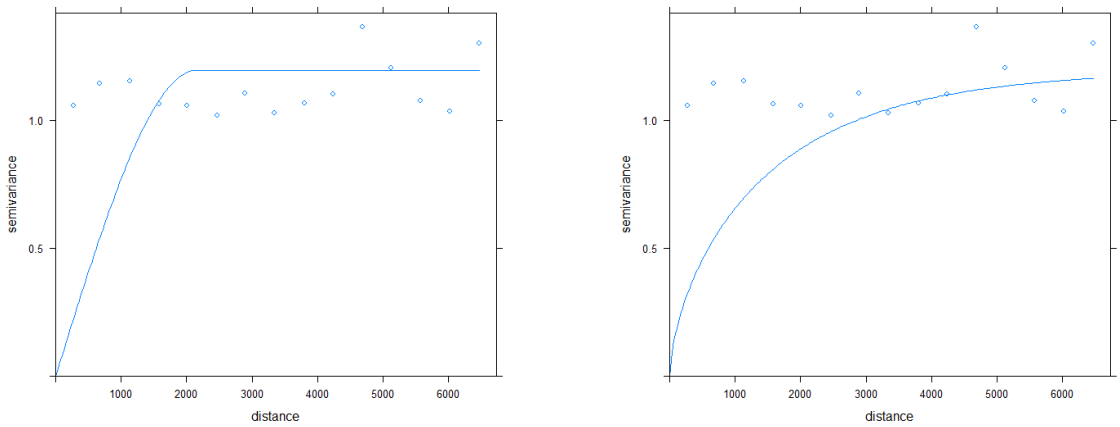
where $\mathcal{E}_i = Z_p - Z_o$ with Z_p and Z_o represent the predicted and observed values of the respective parameters, and N is the number of observation points from the data set, $N = 120$.

In this study, the procedure for applying the methodology for the spatial analysis of simulation data is illustrated from the picture (Marko, 2013) and the following steps were followed:

a. The value of the spatial distribution of simulation data is generated using Microsoft Excel software.
b. Spatial interpolation and geostatistical analysis for the simulation data were performed using the R software.

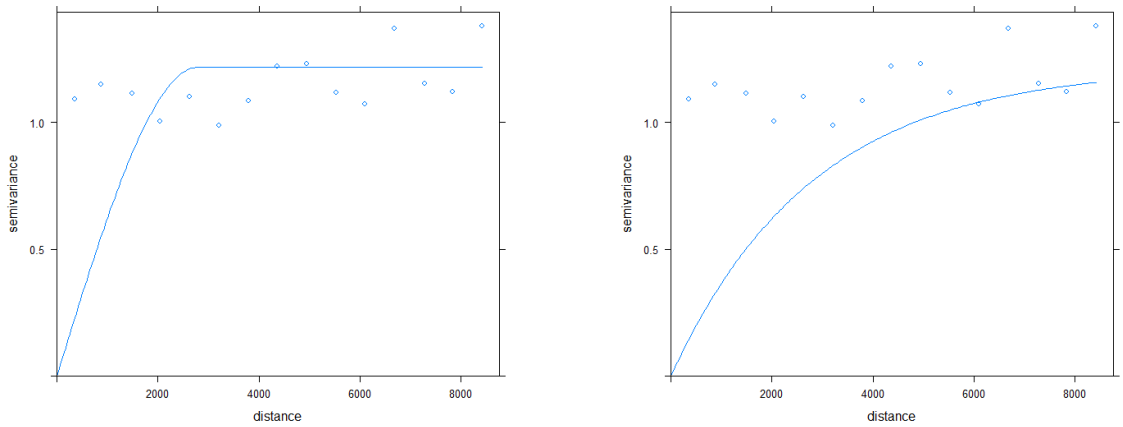
3. Results and Discussion

3.1 DATA A



(a) (b)

DATA B



(c) (d)

Figure 5: (a) Spherical Data A, (b) Matern Data A, (c) Spherical Data B, (d) Matern Data B

TABLE 1. The models characteristics

	Model	**Nugget**	**Sill**	**Range**	**Kappa**
Data A	Spherical	0.0	1.195458	2159.03	-
Data A	Matern	0.0	1.195458	2159.03	0.3
Data B	Spherical	0.0	1.218914	2808.433	-
Data B	Matern	0.0	1.218914	2808.433	0.3

The covariance function modeling was carried out on the fixed model obtained from the gstat package in the R software. The only parameter that changed was kappa which is a special parameter in the matern model whose default value was 0.5 then changed to 0.3. It is because the matern model whose kappa is 0.5 is equal to the exponential model. In this analysis, the sill, nugget, and range values were kept constant to get a fair comparison between the two different models. The results can be seen in Figure 5. There are three spatial classifications seen from the ratio of the nugget to sill variance. The former is strong if the ratio is less than 25%, moderate if the ratio is between 25% and 75%, and the last, the spatial dependence is weak if the ratio is more than 75% (Marko, 2013). As a result of the fixed model that comes from the package that has been available by the R software, the spatial dependence of all models carried out is a strong spatial dependence because the ratio of all models is less than 25%.

After the model is installed, cross-validation is used to ensure how well the interpolation of the model is used. The cross-validation technique removes the data, one at a time, predicts groundwater chemistry, and then compares them with values for each site. In cross-validation, the statistics used to diagnose whether the model and/or its associated parameter values are reasonable (Marko, 2013).

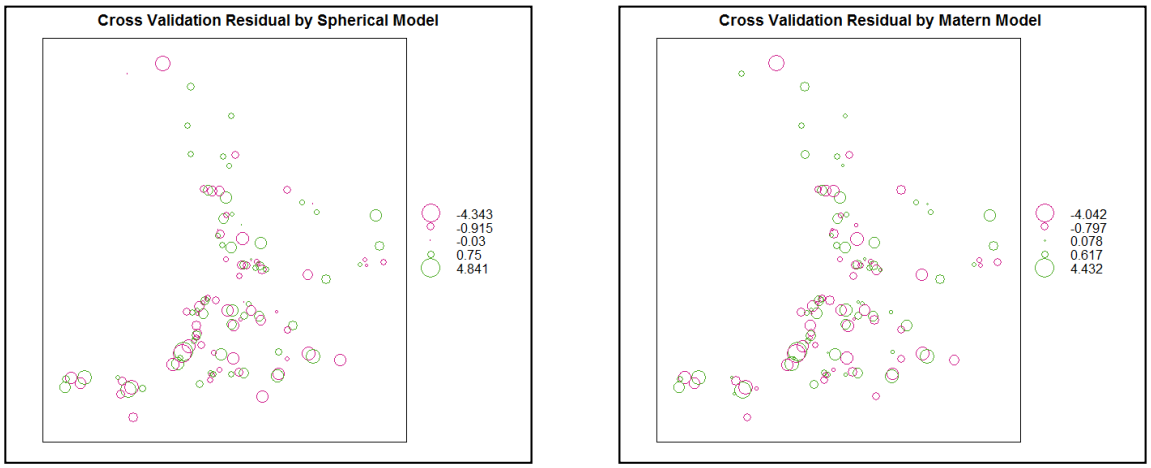
Table 2: Cross-validation residual summary

RESIDUAL SUMMARY DATA A						
Spherical Model	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	-4.3432	-0.9150	-0.0304	-0.0300	0.7495	4.8409
Matern Model	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	-4.04164	-0.79739	0.07824	-0.02006	0.61715	4.43181
RESIDUAL SUMMARY DATA B						
Spherical Model	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	-4.36402	-0.92928	-0.07813	-0.03871	0.71010	4.79688
Matern Model	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	-4.355860	-0.883701	-0.005579	-0.038906	0.785462	4.852858

For the model generated by interpolation to be reliable, it is preferable if the residue has two main characteristics. The first is that there is no recognizable cluster spatial pattern to indicate a random spatial distribution. This is so that the over predictions and underpredictions reflect the distribution of the random process as much as possible. The second is normally distributed, or as close as possible to the normal distribution. This implies that the average value (average) that is owned is or is close to zero. When these two conditions are met, it can be said that the model that is owned is not affected by any bias, that is, it does not systematically overpredict or underpredict the value of the variable being modeled (Exprodat, 2010).

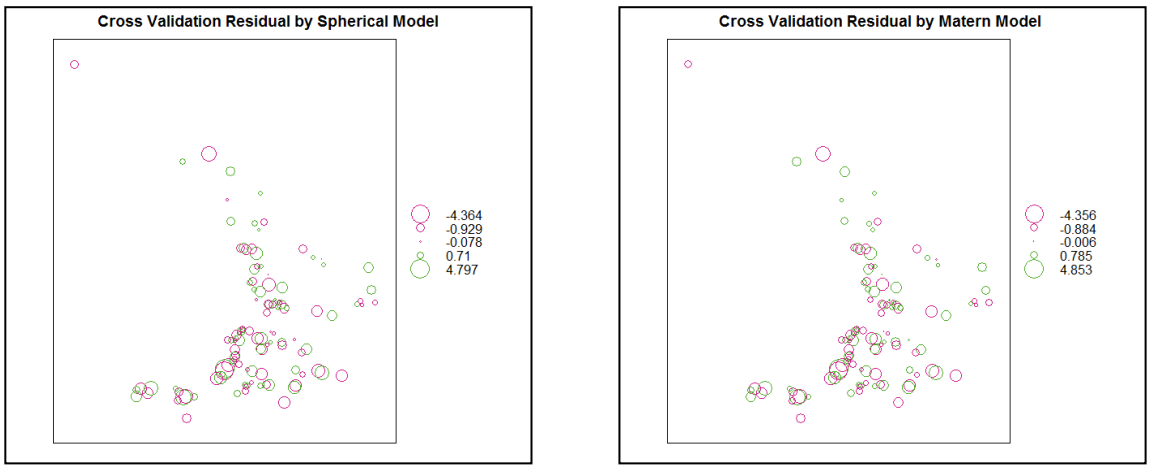
The residual results from the cross-validation can be seen in Table 2. In data A the average value closest to zero is the mean value of the matern model, which means that the model produces unbiased predictions compared to the spherical model. For data B, the matern model also generates a more unbiased estimation compared to the spherical model. The results of the visualization of the residuals can be seen in Figure 6 with the value of underpredictions (negative errors) represented by pink color, and the value of over predictions (positive errors) represented by green color.

DATA A



(a) (b)

DATA B



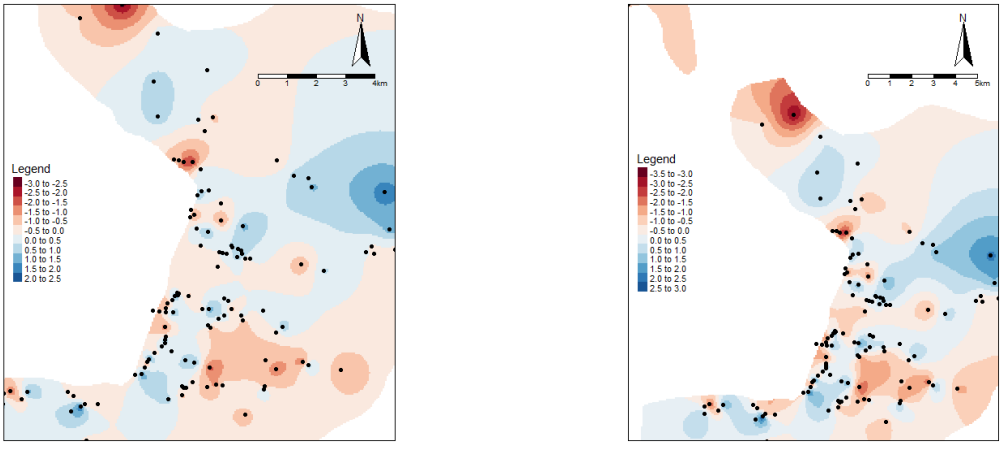
(c) (d)

Figure 6: (a) Data A Cross-validation Residual Spherical Model, (b) Data A Cross-validation Residual Matern Model, (c) Data B Cross-validation Residual Spherical Model, (d) Matern Data B Cross-validation Residual Matern Model.

Table 3: RMSE and MAE comparison

	Model	RMSE	MAE
Data A	Spherical	1.381903	1.077259
Data A	Matern	1.297191	0.9967506
Data B	Spherical	1.387596	1.077814
Data B	Matern	1.381952	1.073306

The results of the comparison of the RMSE and MAE values indicate that the appropriate model to be used in the interpolation is the same for the two simulation data used. In data A, the smallest RMSE value is owned by the matern model, which means that data A will have a better estimate if the prediction uses the model. The results of data B also show that the matern model has a smaller error value than the spherical model. Therefore, using the best model, namely matern for both data, the kriging interpolation is then executed. The visualization of the spatial prediction map of the simulation data is shown in Figure 7. The dark blue color represents a high concentration of predicted value, while the dark pink color represents a low concentration.



(a) (b)

Figure 7: (a) Data A prediction map, (b) Data B prediction map.

4. Conclusion

Through the research, it concluded that the matern model outperforms the spherical model in this simulation. Residual results from cross-validation also show that the matern model has an unbiased estimate on both data compared to the spherical model. According to the RMSE and MAE criteria, data A and B have better predictions using the matern model. It proves that the selection of the covariance function on the variogram model is important because it is very influential in providing a suitable estimate for the kriging method.

References

Choi, I. K. (2014). Modeling spatial covariance functions, Ph.D dissertation, Purdue University, West Lafayette, 2014. Accessed on: November 4, 2021. [Online]. Available:https://docs.lib.purdue.edu/open_access_dissertations.

Cressie, N. A. C. (1993). *Statistics for Spatial Data Revised Edition*. New York, A Wiley-Interscience Publication.

Esri, Semivariogram and covariance functions. Accessed on: November 15, 2021. [Online]. Available: https://desktop.arcgis.com/en/arcmap/latest/extensions/geostatistical-analyst/semivariogram-and-covariance-functions.htm

Marko, K., Al-Amri, N. S. & Elfeki, A. M. M. (2013). Geostatistical analysis using GIS for mapping groundwater quality: case study in the recharge area of Wadi Usfan,western Saudi Arabia. *Springer).*

Pardo-Iguzquiza, E. & Chica-Olmo, M. (2008). Geostatistics with the Matern semivariogram model: A library of computer programs for inference, kriging and simulation, *Computers & Geosciences.* pp. 1073-1079.

Pebesma, E. (co)kriging cross validation, n-fold or leave-one-out, (R Documentation). Accessed on: December 1, 2021. [Online].

Putter, H. & Young, G. A. (2001). On the effect of covariance function estimation on the accuracy of kriging predictors, *Bernoulli*, pp. 421-438.

Rozalia, G., Yasin, H. & Ispriyanti, D. (2014). PENERAPAN METODE ORDINARY KRIGING PADA PENDUGAAN KADAR NO2 DI UDARA (Studi Kasus: Pencemaran Udara di Kota Semarang), *Jurnal Gaussian*, *5*(1), pp. 113-121.

Sarann, T. c, L. & Kato, T. (2014). Assessment of Geostatistical Interpolation Method for Spatial Soil Mapping in Imba-Numa watershed, Japan, *Techno-Science Research Journal*, pp. 61-70.