Application of universal kriging for prediction pollutant using GStat R

Annisa Nur Falah¹, Betty Subartini², and Budi Nurani Ruchjana²

¹ Master Student at Study Program of Mathematics
Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, Indonesia

² Department of Mathematics, Faculty of Mathematics and Natural Sciences
Universitas Padjadjaran, Indonesia

E-mail: annisanurfalah02@gmail.com

Abstract. In the universe, the air and water is a natural resource that is a very big function for living beings. The air is a gas mixture contained in a layer that surrounds the earth and the components of the gas mixture is not always constant. Also in river there is always a pollutant of chemistry concentration more than concentration limit. During the time a lot of air or water pollution caused by industrial waste, coal ash or chemistry pollution is an example of pollution that can pollute the environment and damage the health of humans. To solve this problem we need a method that is able to predict pollutant content in locations that are not observed. In geostatistics, we can use the universal kriging for prediction in a location that unobserved locations. Universal kriging is an interpolation method that has a tendency trend (drift) or a particular valuation method used to deal with non-stationary sample data. GStat R is a program based on open source R software that can be used to predict pollutant in a location that is not observed by the method of universal kriging. In this research, we predicted river pollutant content using trend (drift) equation of first order. GStat R application program in the prediction of river pollutants provides faster computation, more accurate, convenient and can be used as a recommendation for policy makers in the field of environment.

1. Introduction
Geostatistics is a blend of science mining, geology, mathematics, and statistics [1]. The data used in a geostatistics is spatial data based on the observation that the value of the location. One of main geostatistics tools is an interpolation process which initially is used to make predictions of mineral reserves, called as Kriging [6]. Kriging, or spatial Best Linear Unbiased Estimation (BLUE) is coined by Matheron (1962) in honour of D. G. Krige, a South African mining engineer [3]. Kriging is a valuation method that uses spatial data. Kriging is a method to estimate, in a spatial context, the value of a variable of interest at a location where it has not been measured, using data in the neighborhood [10]. Kriging calculation process can be distinguished several kinds, namely: Ordinary Point Kriging, Ordinary Block Kriging, Co-Kriging, Universal Kriging [8]. Universal Kriging is Kriging of data that have a particular tendency to trend [4]. Universal Kriging can also be interpreted as a valuation method used to deal with problems non-stationarity of sample data and can help the problems in real life. One example is the predicted levels of pollutant cadmium in Meuse river floodplain [5]. Thus it is necessary to predict the location that contains cadmium. However there is only few sample data about the cadmium concentration of Meuse river, hence the missing data in unobserved regions need to be generated.
In the calculation of prediction with Universal Kriging Method, it is required gstat and sp library in GStat-R [9]. The current condition of the river has a lot of waste contains pollutants, for example pollutant cadmium that can damage ecosystems and interfere with human health. Thus the need for a method that can predict pollutant of cadmium that pollute the river flood plains ie Universal Kriging Method. GStat R Program is a program for geostatistical model, prediction and simulation in one, two, or three dimensions comprising the sample variogram calculation, cross variogram and match the model. At the beginning of this stand-alone program, but in 2004 were made in the software package R as one package containing spatial data processing functions or command processing program geostatistical applications [2]. Thus, in this study to analyze the Universal Kriging Method can use a software based GStat R Program.

2. Literature Study
In this section, we briefly describe the main theoretical tools that will be used in the analysis of the Universal Kriging Method.

2.1. The Experimental Semivariogram for Universal Kriging
According to [1] develop hypotheses intrinsic meaning that the mean and variance of the shift \( Z(x + h) - Z(x) \) does not depend on the position of the location \( x \) or defined as:

\[
E[Z(x + h) - Z(x)] = 0 \quad (1)
\]

\[
Var[Z(x + h) - Z(x)] \quad (2)
\]

According the equation (1) and (2) so the experimental semivariogram is defined as:

\[
\hat{\gamma} = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [(Z(x_i + h) - m(x_i)) - (Z(x_i) - m(x_i))]^2 \quad (3)
\]

with :
- \( \hat{\gamma}(h) \) : experimental semivariogram value with distance \( h \)
- \( Z(x_i) \) : value of observations in \( x_i \)
- \( Z(x_i + h) \) : value of observations in \( x_i + h \)
- \( m(x) \) : trend (drift) equation
- \( N(h) \) : number of point pairs within \( h \)

2.2. The Theoretical Semivariogram Models
Generally the experimental semivariogram is computed by formula (3), we usually observe the distribution of experimental semivariogram and then identify a reasonable theoretical semivariogram model based on the experimental semivariogram distribution or prior knowledge. The most commonly used theoretical semivariogram models are exponential model, gaussian model and spherical model [11]. These theoretical models must be fitted to the experimental semivariogram by determining three parameters: sill \( c \), range \( a \) and distance \( (h) \)

Spherical Model

\[
\gamma(h) = \begin{cases} 
  c \left[ \frac{3}{2} \left( \frac{h}{a} \right) - \left( \frac{1}{2} \right) \left( \frac{h}{a} \right)^2 \right], & h < a; \\
  c, & h \geq a.
\end{cases} \quad (4)
\]

Gaussian Model

\[
\gamma(h) = \begin{cases} 
  c \left[ 1 - \exp \left( -\frac{3h^2}{a^2} \right) \right], & h < a; \\
  c, & h \geq a.
\end{cases} \quad (5)
\]

Exponential Model

\[
\gamma(h) = \begin{cases} 
  c \left[ 1 - \exp \left( -\frac{3h}{a} \right) \right], & h < a; \\
  c, & h \geq a.
\end{cases} \quad (6)
\]
2.3. **Universal Kriging Method**

Universal Kriging Method is one of the Kriging Method which assumes that the average (mean) is not constant. If the equation $m(x)$ is the trend (drift), with the following equation [4]:

$$m(x) = \sum_{i=0}^{n} a_i f_i(x)$$

Because $m(x)$ not constant, then the estimator of universal kriging as follows [8]:

$$\hat{Z}(x_0) - m(x) = \sum_{i=1}^{k} \lambda_i [Z(x_i) - m(x)]$$

where

$$\hat{Z}(x_0) = m(x) + \sum_{i=1}^{k} \lambda_i [Z(x_i) - m(x)]$$

3. **Main Result**

3.1. **Research Data**

The data used in this paper is secondary data in the program GStat R. Dataset drawn named Data Meuse is a data set that consists of four heavy metals (Cadmium, Zinc, Lead, Copper) as measured from the top layer of soil along the Meuse River flood plain located around Meers and Maasband, Limburg, Netherlands [2].

| Table 1. Table of Cadmium Data |
|-------------------------------|
| Locations x(m) y(m) Cadmium (ppm) |
| 1 | 181072 | 333611 | 11.7 |
| 2 | 181025 | 333558 | 8.6 |
| ... | ... | ... | ... |
| 164 | 180627 | 330190 | 2.7 |

| Table 2. Table of Descriptive Statistics |
|------------------------------------------|
| x | y | Cadmium |
| Min. | 178605 | 330179 | 0.7 |
| 1st Qu. | 179442 | 331065 | 1.8 |
| Median | 180283 | 332213 | 2.9 |
| Mean | 180156 | 332022 | 4.689 |
| 3rd Qu. | 180935 | 332778 | 7.05 |
| Max. | 181390 | 333611 | 18.1 |

From the Table 2 above minimum values obtained coordinate $x$ 178605 meters and a maximum of 181390 meters, the minimum coordinate $y$ value of 329714 meters and a maximum of 333611 meters, while the content of cadmium ie 0.7 ppm to 18.1 ppm and the minimum value to the maximum value.

3.2. **Stationarity**

In universal kriging necessary assumption that the data is a second order not stationary, while the second order stationary universal kriging not apply, so all the data is done plotting the data in order to determine whether the obtained data is stationary or not. Non-stationary from sample data will look after points in the plot and the content will be visible if the data has a tendency to trend or not. Here is a plot of the distribution of the data based on Table I.
Based from Fig. 1. and Fig. 2. can be seen that the data of pollutant cadmium is not stationary, because there are some outliers or a trend (drift). Therefore the universal kriging method is suitable to use to predict the data of the cadmium.

3.3. Table of The Experimental Semivariogram

The experimental semivariogram is one tool used to get the best of the theoretical semivariogram models. To be able to know the theoretical semivariogram models are suitable for use, first look for the average of the value semivariogram, with the help of R.3.1.1 program average values obtained semivariogram from a pair of data shown in the following table:

Based on the Table 3 the average value obtained table experimental semivariogram as many as 15 locations. To determine the best of the theoretical semivariogram model, can be done by the plot of data from the experimental semivariogram.
Table 3. The Experimental Semivariogram

| No | Number of point pairs within $h$ | Distance | Experimental Semivariogram |
|----|---------------------------------|----------|-----------------------------|
| 1  | 57                              | 79.29244 | 4.927280                    |
| 2  | 299                             | 163.97367| 6.470288                    |
| 3  | 419                              | 267.36483| 6.603915                    |
| 4  | 457                              | 372.73542| 8.581387                    |
| 5  | 547                              | 478.47670| 9.326972                    |
| 6  | 533                              | 585.34058| 10.494006                   |
| 7  | 574                              | 693.14526| 10.009457                   |
| 8  | 564                              | 796.18365| 10.610774                   |
| 9  | 589                              | 903.14650| 11.097955                   |
| 10 | 543                              | 1011.29177| 12.220606                  |
| 11 | 500                              | 1117.86235| 14.103238                  |
| 12 | 477                              | 1221.32810| 13.323896                  |
| 13 | 452                              | 1329.16407| 15.156499                  |
| 14 | 457                              | 1437.25620| 10.608018                  |
| 15 | 415                              | 1543.20248| 11.153523                  |

Figure 3. The Experimental Semivariogram

3.4. Fitting The Experimental Semivariogram with The Theoretical Semivariogram Models
Based on the plot of the results of the experimental semivariogram value calculation, the following is the result fitting of the three theoretical semivariogram models.

Table 4. MSE Semivariogram

| MSE Of The Theoretical Semivariogram Models | Spherical | Gaussian | Eksponensial |
|--------------------------------------------|-----------|----------|--------------|
| 3.455941                                   | 4.195471  | 18.24626 |

Based on the Fig. 4. and Table 4, the theoretical semivariogram model is obtained it is best spherical semivariogram model, because it has the most minimum mean square error between gaussian semivariogram model and exponential semivariogram model.
3.5. The Trend (drift)

The universal kriging is kriging of data that have a tendency to trend (drift). In this paper the order of the equation trend (drift) bounded is order one, because the shape is more simple than the other order. Using R.3.1.1 program can be sought from the equation coefficients trend (drift) as follows:

![Figure 4. Fitting Theoretical with Experimental Semivariogram](image)

**Table 5. Coefficient of the Trend Equation**

| Coefficients | Intercept | x         | y         |
|--------------|-----------|-----------|-----------|
|              | -326.65228| -0.00294  | 0.00259   |

3.6. Predictions at Unobserved Locations

**Table 6. Prediction at Unobserved Locations**

| Locations | x    | y    | Prediction | Error Variance |
|-----------|------|------|------------|----------------|
| 1         | 181071 | 333612 | 8.92048    | 6.081463       |
| 2         | 180111 | 331811 | 0.8250955  | 6.076976       |
| 3         | 181190 | 333114 | 1.611239   | 5.655549       |
| 4         | 180553 | 332706 | 8.903248   | 5.769673       |
| 5         | 180915 | 332754 | 1.727731   | 5.641941       |
| 6         | 180384 | 332475 | 9.630817   | 5.863041       |
| 7         | 180198 | 331590 | 0.9197342  | 5.727752       |
| 8         | 179472 | 331124 | 4.242874   | 5.987837       |
| 9         | 178813 | 330665 | 3.679977   | 5.90064        |
| 10        | 179302 | 330178 | 0.9584527  | 5.813183       |

Based on the Table 5, this means that when the value of the absis x is equal to zero, and the value of the ordinate y is equal to zero, then the content of the pollutant cadmium will be equal to intercept, the value would be decrease by 326.65228. If the value of the absis x shifted one unit then the level will be decrease by 0.00294 whereas if the ordinate y shifted one unit then the content of pollutant cadmium would be increase by 0.00259.
According to the Table 6, it can be seen that the predicted results and the variance of pollutant cadmium in 5 locations at unobserved locations is has the variation value, is not likely to go increase or not likely to go decrease, this is because the type of data that is predictable is the data that is non-stationary or have a tendency trend (drift) so that the results of predictions is very variation.

4. Conclusion
GStat R procedure in calculating prediction at unobserved locations with universal kriging method include the following stages:

a. Testing stationarity by way of plot distribution data.

b. The process of calculating the value of the experimental semivariogram.

c. Fitting the theoretical semivariogram models that best by finding the minimum sum of squared errors.

d. Searching for the coefficients of the equation trend (drift) of the order one.

e. The prediction at unobserved locations using universal kriging method.

Acknowledgment
The work reported in this paper is supported by the Directorate General of Higher Education, Ministry of Research-Technology and Higher Education, Indonesia, through Penelitian Kerjasama Luar Negeri dan Publikasi Internasional 2016.

References
[1] Armstrong M 1998 Basic Linear Geostatistics. New York: Springer
[2] Bivand R, Pebesma E, and Rubio V 2013 Applied Spatial Data Analysis with R (Second Edition) New York: Springer
[3] Cressie N A C 1993 Statistics For Spatial Data. New York: John Wiley and Sons, Inc.
[4] Kambhammettu B, Allena P, and King J 2011 Application and evaluation of universal kriging for optimal contouring of groundwater levels. J.Eart Syst. Sci.
[5] Middelkoop H 2000 Heavy-metal pollution of the river Rhine and Meuse floodplains in the Netherlands. Netherlands Journal of Geosciences 79(4) pp 411–428
[6] Montero J M, Aviles G F, and Mateu J 2015 Spatial and Spatio-Temporal Geostatistical Modeling and Kriging New York: John Wiley and Sons.
[7] Nerini D, Monestiez P, and Mante C 2010 Cokriging for spatial functional data. Journal of Multivariate Analysis pp 409–418
[8] Olea R A 1999 Geostatistics for engineers and earth scientists. United States of America: Kluwer Academic Publishers
[9] Pebesma E 2004 Multivariable geostatistics in S the gstat package Computers and Geosciences 30(7) pp 683–691
[10] Wackernagel H 1994 Cokriging versus kriging in regionalized multivariate data analysis Elsevier Science B V pp 83–92
[11] Youkuo C, Yongguo Y, and Wangwen W 2015 Coal Seam Thickness Prediction based on Least Squares Support Vector Machines and Kriging Method EJGE 20 Bund. 1