A spatial model for a stream networks of Citarik River with the environmental variables: potential of hydrogen (PH) and temperature

A Bachrudin¹, N B Mohamed², S Supian³, Sukono⁴*, Y Hidayat⁵

¹,⁵Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, INDONESIA
²School of Informatics and Applied Mathematics, Universiti Malaysia Terengganu, MALAYSIA
³,⁴Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, INDONESIA

*Corresponding author e-mail: sukono@unpad.ac.id

Abstract. Application of existing geostatistical theory of stream networks provides a number of interesting and challenging problems. Most of statistical tools in the traditional geostatistics have been based on a Euclidean distance such as autocovariance functions, but for stream data is not permissible since it deals with a stream distance. To overcome this autocovariance developed a model based on the distance the flow with using convolution kernel approach (moving average construction). Spatial model for a stream networks is widely used to monitor environmental on a river networks. In a case study of a river in province of West Java, the objective of this paper is to analyze a capability of a predictive on two environmental variables, potential of hydrogen (PH) and temperature using ordinary kriging. Several the empirical results show: (1) The best fit of autocovariance functions for temperature and potential hydrogen (ph) of Citarik River is linear which also yields the smallest root mean squared prediction error (RMSPE), (2) the spatial correlation values between the locations on upstream and on downstream of Citarik river exhibit decreasingly

Keywords: Stream networks, spatial dependence, autocovariance function, kernel convolution, kriging.

1. Introduction

A Classical geostatistics is widely used in environmental science and ecology which are normally applied the statistical instruments such as variogram, semivariogram, covariance functions and almost all of the statistical instruments are based upon a Euclidean distance framework (e.g. see Cressie, [2], Recently, there was a new class of geostatistics that is in the basis of non-Euclidean distance [9], Curriero, [4; 5], Rathbun, [12], and Higdon, [7]. Barry & Ver Hoef, [1]). Curriero [4], pointed out that conditions of a metric are not sufficient proof of the validity of distance to yield positive definite functions. Such distances cannot be used without proof in covariance and semivariogram models. Covariance \( \text{cov}(d(s_i, s_j)) \) calculated using metric \( d(s_i, s_j) \) must satisfy the non-negative definiteness
property. Briefly, a metric distance that must meet the geometric properties and covariance function based on the metric distance must satisfy the non-negative definiteness property.

Ver Hoef et al.[14] and Cressie et al. [3] developed a geostatistics for non-Euclidean distance which are applied for a stream networks. The non-Euclidean distance is called as a stream or hydrological distance. They built several autocovariance functions which were permissible for stream distance met the both properties. Using moving-average construction (kernel convolution), they extended a spatial model for stream networks. Ver Hoef et al. [14] developed the spatial model based on both stream distance and flow direction, namely tail-down and tail-up and this paper will discuss tail-up model. A process of this model moves from downstream to upstream which is the opposite of flow direction. In the tail-up model, it is possible that flow volume from one upstream segment is larger than the flow volume for the other upstream segment, then it might want to weight according to flow volume. The weighting processes can be used a flow volume, a stream order, or a proxy variable such as a watershed area, or other measures.

The objective of this paper is to analyze a capability of a predictive on two environmental variables, potential of hydrogen (ph) and temperature using ordinary kriging.

2. A Spatial model for a stream networks

As has mentioned previously that Ver Hoef et al. [14] developed a new class of geostatistics using a non-Euclidean distance for a stream networks, and this distance is well known as a stream or a hydrological distance. However, for studies related to hydrology, using Euclidean distance is not appropriate. Since the movement of materials in the water only occurs within stream networks, it is more appropriate to use distance defined only along the steam. Stream distance between any two points is obtained by summing up the length of segments which connect them [8]. The different stream and Euclidean distance is described in Figure 1. One problem arises is that how to develop a permissible autocovariance function using stream distance. An answer of this problem is described as follows.

Barry and Ver Hoef [1] showed that large class of autocovariance functions can be developed by creating random variables as the integration of a moving average function over white noise random process:

\[ Z(s) = \int_{-\infty}^{\infty} g(x-s|\theta[W(x)]dx \]  

where \( W(x) \) the white noise is random process and \( g(x-s|\theta) \) is the moving average function. Based on the moving average construction, the autocovariance of \( Z(s) \) and \( Z(s+h) \) is given as follows (Ver Hoef et al. [14]):

\[ C(h|\theta) = Cov(Z(s),Z(s+h)) = \int_{-\infty}^{\infty} g(x-s|\theta)g(x|\theta) \]  

Figure 1. Stream and Euclidean distance
Equation (2) is a permissible autocovariance for stream networks and \( h \) is a stream distance. Proof of Equation (2) can be found in appendix. From Equation (2) can be obtained a variety of valid covariance functions which is given in Table 1 as follows:

| Name          | Autocovariance function                                                                 |
|---------------|----------------------------------------------------------------------------------------|
| Linear with sill | \( C(h|0) = \begin{cases} \theta_0 + \theta_1 & h = 0 \\ \theta_0 \left( 1 - \frac{h}{\theta_2} \right) & 0 < h < \theta_2 \\ 0 & h \geq \theta_2 \end{cases} \) |
| Spherical     | \( C(h|0) = \begin{cases} \theta_0 + \theta_1 & h = 0 \\ \theta_0 \left( 1 - \frac{h}{2\theta_2} + \frac{1}{2} \left( \frac{h}{\theta_2} \right)^3 \right) & 0 < h < \theta_2 \end{cases} \) |
| Mariah        | \( C(h|0) = \begin{cases} \theta_0 & h = 0 \\ \theta_1 \ln \left( \frac{\theta_2}{h/h_2} + 1 \right) & 0 < h \end{cases} \) |
| Exponential   | \( C(h|0) = \theta_0 \exp \left( -\frac{h}{\theta_2} \right), h > 0 \) |

Equation (2) and the autocovariance functions in Table are described in detail in Table 1. In a stream networks, there will be a finite number of stream segments, and it gives index them arbitrarily with \( i = 1, 2, 3 \) as Figure 2. One of stream networks that is called as a tail up model, the moving average function begins from downstream (segment \( i = 3 \) in Figure 2), then the moving average into two parts, with one part going up segment \( i = 2 \), and the other going up segment \( i = 3 \). If it uses no weighting at all, so variance of the random variable is \( \int_{-\infty}^{\infty} \left( g(x|0) \right)^2 dx + \nu^2 \) from Figure 2 for \( r_3 \) would be greater than for \( r_2 \). When the moving average split into parts, then the variance of each parts could be a different. To keep the variances are constant, sum of each weight segment must be one. For example, let \( \omega_1 \) and \( \omega_2 \) be a weight for segment \( i = 2 \) and \( i = 3 \) in Figure 2, so to keep the variances are constant that must be \( \omega_1 + \omega_2 = 1 \). This constraint that stream networks is to keep a stationary process.

![Figure 2. Moving average constructions](image-url)
Autocovariance function in Equation (2) incorporates a weighting $\omega$ of a segment of stream networks is given as follows:

$$ C(s,t|\theta) = \begin{cases} 
0 & \text{if } s \text{ and } t \text{ are not flow-connected,} \\
C(0) + v_j^2 & \text{if } s = t, \\
\prod_{k \in B_{s,t}} \sqrt{\omega_k} C(d(s,t_j)) & \text{otherwise} 
\end{cases} $$

(3)

where $v_j^2$ is a discontinuity at $h=0$ which is the nugget effect in geostatistical terminology, $d(s,t_j)$ is the stream networks, $C(d(s,t_j))$ can be found in Table 1, and $B_{s,t,j}$ is the set of stream segments between two locations, including the segment for the upstream location, but excluding the segment for the downstream location. If the weighting is applied in a river, so it can incorporate flow volume. In the absence of any flow volume characteristics, it can simply weight each split in the moving average function by $\frac{1}{2}\sqrt{2}$ [14].

The ultimate objective of this paper is to measure a predictive ability for stream networks which is measured by root mean square prediction error (RMSPE):

$$ \text{RMSPE} = \sqrt{\frac{\sum_{i=1}^{n} (\hat{z}_i - \hat{z}_i) ^2}{n}} $$

(4)

It is estimated by leave-one-out cross-validation after estimating the covariance parameters. Where $\hat{z}_i$ is the $i$th site kriged using all the other $n-1$ sites.

A kriging method is normally applied in geostatistics to predict $z(s)$ at point $s_0$ using a model:

$$ Z(s_0) = \sum_{i=1}^{n} \lambda_i Z(s_i) $$

(5)

Using the unbiased constraint $\sum_{i=1}^{n} \lambda_i = 1$, and minimizing variance of predictor error, $\lambda_i$ can be computed the following expression as follows:

$$ \lambda' = (\lambda_1, \ldots, \lambda_n) = \left( \textbf{c} + 1 \frac{(1-\textbf{1}' \Sigma^{-1} \textbf{1}) \textbf{c}}{\textbf{1}' \Sigma^{-1} \textbf{1}} \right) \Sigma^{-1} \textbf{c} $$

(6)

where $\Sigma(\theta)$ a covariance matrix is whose elements is in Equation (3), $\textbf{c} = (C(s_0 - s_1), \ldots, C(s_0 - s_n))^T$. The interpolation $\hat{z}(s_0)$ gives a best linear unbiased predictor [2].

3. Methodology

Many applications of this model are widely used on a river (e.g. see Gareta et al. [6], Peterson et al. [10], [11]). A case study of this paper is the basin of the Citakik River which is a small watershed of the Citakik River. This river is located in Bandung regency. A study area is described by Figure 3. The collected dataset has two environmental variables (ph and temperature) sampled at 25 sites. These variables are potential of hydrogen (PH) and temperature. Addition to these variables, to measure a spatial dependence is a volume discharge.
After computing a matrix of stream distance, then compute a matrix of covariance \( \Sigma(\theta) \) whose elements are in Equation (3) for each autocovariance function is in Table 1. Maximum likelihood is applied to Estimate of the parameters \( \theta = (\theta_0, \theta_1, \theta_2) \) by assuming that \( Z(s) = (Z(s_1), \ldots, Z(s_n))^T \) is normally multivariate distributed, or \( Z(s) \sim N(0, \Sigma(\theta)) \). Finally, compute RMSPE using Equation (4). Its computation is used R programming and a mapping of the prediction results is employed GIS software.

4. Empirical Results
Figure 4 shows the histogram of potential hydrogen (PH) and temperature indicated a crude distribution of their variables. The histograms plays an important role in a diagnostic of the probability distribution. Besides the histogram, some descriptive statistics of the two variables are computed to see some statistical measures shown in Table 2.

![Histogram of pH and Temp](image)
Table 2 Descriptive statistics for PH and temperatures

|          | PH          | Temperature |
|----------|-------------|-------------|
| Mean     | 7.44        | 25.04       |
| Standard Error | 0.135277   | 0.567098    |
| Median   | 7.3         | 25          |
| Standard Deviation | 0.676387   | 2.835489    |
| Kurtosis | 2.473181    | -0.77294    |
| Skewness | 1.106285    | 0.003662    |
| Range    | 3.2         | 10          |
| Minimum  | 6.3         | 20          |
| Maximum  | 9.5         | 30          |
| Sum      | 186         | 626         |
| Count    | 25          | 25          |

Following Figure 4 and Table 2 show that the histogram shape of PH variable tends leptokurtic distribution and has heavier tails distribution supported by the estimate kurtosis is a positive value. In addition, this distribution exhibits a non-symmetric shape. For temperature variable shows that the distribution of data is fairly symmetric since the estimate of skewness is almost zero. Because of the negative kurtosis value is less than zero, then the distribution is light tails and is called a platykurtic.

Table 3. Model selection of autocovariance functions using aic statistic and the parameter estimates of the selected model

| Environmental Variables | AIC Statistic | The Estimated Parameters |
|-------------------------|---------------|--------------------------|
|                         | Exponential   | Spherical                | Linear | $\theta_0$ | $\theta_1$ | $\theta_2$ |
| Temperature             | 111.93        | 124.01                   | 102.30 | 0           | 256.96     | 1194.36     |
| PH                      | 195.47        | 233.25                   | 198.92 | 3.89        | 23.81      | 19999.96    |

Table 3. shows the least values of AIC statistics are respectively linear and Exponential autocovariance function for the variables of temperature and PH, but the values of AIC statistic for PH are not significantly different, so it is concluded the best fit of auto covariance for both of the variables is linear. For Linear that the estimated parameters are $\hat{\theta}_0 = 256.96$, and $\hat{\theta}_1 = 1194.36$ and for ph that the estimated parameters are $\hat{\theta}_0 = 3.89$, $\hat{\theta}_1 = 23.81$, and $\hat{\theta}_2 = 19999.96$. Of this result indicates that dependency of ph is longer than temperature to the quality of water along on the Citarik River, but the degree of dependence is higher.

The Table 4 and 5 indicate a spatial correlation between the locations of the Citarik River. The first column In Table 3 shows that the spatial correlation between the Location 37 and the locations 33, 36, 35, 34, 31, 30, 29, 28, 64, 63, and 69 are 0.234, 0.211, 0.192, 0.179, 0.079, 0.075, 0.071, 0.067, 0.05, 0.023, and 0.023, respectively. The values of the spatial correlations equal to zero in Table 3 and 4 which indicate unconnected flows between the locations. Following this result can be concluded that spatial correlation values between the locations on upstream and on downstream of Citarik river exhibit decreasingly, as well as spatial correlation in Table 4. The spatial correlation of temperature is generally higher than PH.
Based on the Table 6 shows that the smallest value of RMSPE for pH and temperature for linear, respectively that the results are consistent to using AIC statistic.

Table 6. Performances of models and predictions using RMSPE

| Autocovariance Function | RMSPE | PH   | Temperature |
|--------------------------|-------|------|-------------|
| Linear                   | 0.62  | 0.66 |
| Spherical                | 0.70  | 0.69 |
| Exponential              | 0.73  | 1.89 |

It is usually in forecasting analysis to compare the predicted result with the actual or real data shown in Figures 5 and 6. For temperature, based on Figure 5 shows that there is a positive correlation between actual and prediction whose determination coefficient is 57%. For PH, Figure 6 shows that there is weakly correlated and determination coefficient equals only 16%. Comparing the predicted result and real data in the Citarik River is shown in Figures 7 and 8. Figures 7 and 8 shows that prediction values of PH and temperature make more “alkali” and become a high temperature,
respectively. These mappings are appropriate to analyzing by means of quantitative results. Anyway, the predictions are almost the same result as (6) and [10] that means RMPE for PH is less than 1, and for temperature is less than 3, respectively.

Figure 5. Plot actual and prediction for temperature

Figure 6. Plot actual and prediction for PH

Figure 7. A mapping actual and prediction in study area for PH
Figure 8. A mapping actual and prediction in study area for temperature

Conclusion
In this paper has been discussed about a spatial model for a stream networks of Citarik River with the environmental variables: potential of hydrogen (PH) and temperature. The following conclusions is based on the empirical results: The best fit of autocovariance functions for temperature and potential hydrogen (PH) of Citarik river is linear which also yields the smallest root mean squared prediction error (RMSPE). The spatial correlation values between the locations on upstream and on downstream of Citarik river exhibit decreasingly.

References
[1] Barry, R.P. and Ver Hoef, J.M. 1996. Blackbox kriging: Spatial prediction without specifying the variogram. J. Agricult. Biol. Environ Stat. 1, 297-322.
[2] Cressie, N.A.C 1993. Statistics for Spatial data, revised edition, John Wiley & Sons, Inc., New York.
[3] Cressie, Frey, B., Frey, B.H. and Smith, M. 2006. Spatial prediction on a river network. Journal of Agricultural Biological and Environmental Statistics, 11, 127-150.
[4] Curriero, F.C. 2006. On the use of non-Euclidean distance measures in geostatistics. Math. Geology, 38, 907-925.
[5] Curriero, F. C. 1997. The use of non-Euclidean distances in geostatistics, PhD Dissertation, Department of Statistics, Kansas State University.
[6] Garreta, V., Moniestier, P, and Ver Hoef, J.M. 2010. Spatial modelling and prediction on river network: up model, down model or hybrid?. Environmetrics, 21, 439-456.
[7] Higdon, D. 1998. A process-convolution approach to modelling temperatures in the North Atlantic Ocean, Journal of Environmental and Ecological Statistics 5, 173-190.
[8] Li, Jie 2009. Spatial Multivariate design in the plane and stream networks, Dissertation, University of Iowa.
[9] Little LS, Edwards D, Porter D.E. 1997. Kriging in estuaries: as the crow flies or as the fish swims? Journal of experimental marine biology and ecology, 213, pp. 1-11.
[10] Peterson, E.E., Theobald, D.M., and Ver Hoef, J.M. 2007. Geostatistical modelling on stream networks: developing valid covariance matrices based on hydrologic distance and stream flow. Freshwater Biology. 32.267-279.
[11] Peterson, E.E. and Ver Hoef J.M. 2010. A mixed model moving average approach to geostatistical modelling in stream network. Ecology, 91, 644-655.
[12] Rathbun, S. 1998. Spatial modelling in irregularly shaped regions: kriging estuaries. Environmetrics, 9, 109-129.

[13] Ver Hoef, J.M. and Peterson, E.E. 2010. A moving average approach for spatial statistical models of stream networks. *Journal of the American Statistical Association*, 105, 6-18.

[14] Ver Hoef, Peterson, E.E, and Theobald, D. 2006. Spatial statistical models that use flow and stream distance. *Environ Ecol Stat*. 13, 449-464.

[15] Ver Hoef, Peterson, E.E., Clifford, D., and Shah, R. 2014. SSN: An R package for spatial statistical modelling on stream networks. *Journal of Statistical Software*, 56, 1-45.
Appendix

This proof can be found in [8].

Where $W(x)$ is a white noise random process, with $E(W(x)) = 0$ and $\text{var}(W(x)) = 1$, and $g(x|\theta)$ is the moving average function. The integral in Equation (1) is ordinary Riemann integral, but it should be really written as $Z(s, \omega) = \int_{-\infty}^{\infty} g(\|x-s\|\theta)W(x, \omega)\,dx = \int_{-\infty}^{\infty} g(\|x-s\|\theta)dW(x, \omega)$ which is a Lebesgue integral.

$$c(h|\theta) = \text{cov}(Z(s), Z(s+h))$$

$$= E\left[\int_{-\infty}^{\infty} g(\|x-s\|\theta)W(x, \omega)\,dx\int_{-\infty}^{\infty} g(\|y-s-h\|\theta)W(y, \omega)\,dy\right]$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} E\left[g(\|x-s\|\theta)g(\|y-s-h\|\theta)W(x, \omega)W(y, \omega)\right]dxdy$$

(Fubini’s Theorem)

$$= \int_{-\infty}^{\infty} g(\|x-s\|\theta)g(\|x-s-h\|\theta)\,dx$$

$$= \int_{-\infty}^{\infty} g(\|x\|\theta)g(\|x-h\|\theta)\,dx,$$

In order to Fubini’s Theorem can be used that must be met:

$$E\left[\left|\int_{-\infty}^{\infty} g(\|x-s\|\theta)W(x, \omega)\,dx\right|\int_{-\infty}^{\infty} g(\|y-s-h\|\theta)W(y, \omega)\,dy\right]$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} E\left|g(\|x-s\|\theta)g(\|y-s-h\|\theta)W(x, \omega)W(y, \omega)\right|dxdy$$

(Tonelli’s Theorem)

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |g(\|x-s\|\theta)g(\|y-s-h\|\theta)|E\left|W(x, \omega)W(y, \omega)\right|dxdy$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |g(\|x-s\|\theta)g(\|y-s-h\|\theta)|\sqrt{EW^2(x)EW^2(y)}\,dxdy,$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |g(\|x-s\|\theta)g(\|y-s-h\|\theta)|\,dxdy$$

$$< \infty$$

The Equation (1) can be as a function of distance $h$ which is notated by $\rho(h)$. The moving average construction allows a valid autocovariance to be expressed by:
\[
C(h|0) = \begin{cases} 
\int_{-\infty}^{\infty} (g(x|0))^2 dx + v^2, & h = 0 \\
\int_{-\infty}^{\infty} g(x|0) g(x-h|0) dx, & h > 0 
\end{cases}
\] (2)

Relationship between (1) and (2) can be formulated by [14]:

\[
C(x|0) = \begin{cases} 
\theta_0 + \theta_1, & h = 0 \\
\theta_1 \rho(h/\theta_2), & h > 0 
\end{cases}
\] (3)

For example, \( g(x) = \frac{1}{x+1}, x \geq 0 \), using (2) and (3) is obtained:

\[
C(x|0) = \begin{cases} 
\theta_0 + \theta_1, & h = 0 \\
\ln\left(\frac{h}{\theta_2} + 1\right), & h > 0 
\end{cases}
\] (4)

The Equation (4) is a valid autocovariance function for the moving average function, \( g(x) = \frac{1}{x+1}, x \geq 0 \). It can be derived for other functions to construct a valid autocovariance function.