MALAYSIAN JOURNAL OF SCIENCE

Vol. 38 • Special Issue (2) • 2019

International Seminar on Mathematics in Industry & International Conference on Theoretical and Applied Statistics 2018 (ISMI-ICTAS18)

JURNAL SAINS MALAYSIA

MJS

ISSN 1394 - 3065

The Malaysian Journal of Science is indexed or cited in the following Scientific Databases: Elsevier Bibliographic Databases (ACI, Scopus, EMBASE, Compendex, GEOBASE, EMBiology, Elsevier BIOBASE, FLUIDEX and World Textiles); CAB Abstracts and Chemical Abstracts Service Database

http://www.myjurnal.my/public/browse-journal-view.php?id=194
www.mjs.um.edu.my

Selected papers from the

International Seminar on Mathematics in Industry & International Conference on Theoretical and Applied Statistics 2018 (ISMI-ICTAS18)

Universiti Teknologi Malaysia Kuala Lumpur, Malaysia
4th - 6th September 2018

International Seminar on Mathematics in Industry & International Conference on Theoretical and Applied Statistics 2018 (ISMI-ICTAS18) was a joint conference that focusses on all areas of mathematics applications. The conference was held at UTM Kuala Lumpur, Malaysia from Sept 4 through Sept 6, 2018. ISMI-ICTAS18 was a collaborative effort between the main organiser, UTM Centre for Industrial and Applied Mathematics (UTM-CIAM) and UTM Department of Mathematical Sciences, ITS Department of Statistics, ITS Department of Mathematics, Oxford Centre for Industrial and Applied Mathematics (OCIAM) and Asia Pacific Consortium of Mathematics for Industry (APCMfI).

The theme of the conference was "STEM Harnesses 4th Industrial Revolution (4IR) Challenges & Opportunities". This event included lectures, research paper presentations and the Malaysian Mathematics In Industry Workshop (MMIW2018). The scope of conference consists of research and development in Mathematics for Industrial and Real Applications in the broad areas of Applied Mathematics, Theoretical Statistics, Applied Statistics, Operational Research and Computational Mathematics. Selection of conference papers from Theoretical Statistics and Applied Statistics were presented in this issue which covers a range of applications in various industries including Industrial Network System, Biotechnology, Meteorology, Hydrology and Environment.

Temperature and Humidity Forecast via Univariate Partial Least Square and Principal Component Analysis

Sutikno[1a*], Zahrotun Nisaa'[1b], Kartika Nur 'Anisa'[1c]

[1]Department of Statistics, Faculty of Mathematics Computing and Data Science, Institut Teknologi Sepuluh Nopember, Kampus Sukolilo, Surabaya 60111, INDONESIA. E-mail: sutikno@statistika.its.ac.id[a]; zahrotunnisaa11@gmail.com[b]; kartika.nuranisa9@gmail.com[c]
* Corresponding Author: sutikno@statistika.its.ac.id[a]
Received: 21st April 2019 Revised : 6th August 2019 Published: 30th September 2019
DOI : https://doi.org/10.22452/mjs.sp2019no2.1

ABSTRACT Indonesian Meteorology, Climatology, and Geophysics Agency (BMKG) uses Numerical Weather Prediction (NWP) for short-term weather forecast but it gives biased result. Therefore, this study implements Univariate Partial Least Square (PLS) as Model Output Statistics (MOS) for temperature and humidity forecast. This study uses the maximum temperature (Tmax), minimum temperature (Tmin), and relative humidity (RH) which are called response variables and NWP as predictor variable. The results show that the performance of the model based on Root Mean Square Error of Prediction (RMSEP) are considered to be good and intermediate. The RMSEP for Tmax in all stations is intermediate (0.9-1.2), Tmin in three stations is good (0.5-0.8), and humidity in three stations is also good (2.6-5.0). The prediction result from the PLS is more accurate than the NWP model and able to correct an 89.94% of the biased NWP for Tmin forecasting.

Keywords: MOS, NWP, PCA, PLS, Temperature and Humidity.

1. INTRODUCTION

Indonesia is one of the archipelago states with a tropical climate, having a dynamic and complex weather and atmospheric system. The atmosphere also has a significant role in the global weather and climate systems (Tjasyono, 2004). Weather is considered to be the part that cannot be separated from human activity and influences the various areas of life. Dealing with it, an efficient method is needed for weather forecasting, especially in the short-term forecasting (Wardani, 2010). Indonesian Meteorology, Climatology, and Geophysics Agency (BMKG) has forecasted a short-term weather by comparing and observing a weather pattern and condition that happened the day before, and generally, the accuracy of forecasting will vary since it largely depends on the forecaster's experience.

Information about weather forecasts has been published by BMKG including maximum temperature (Tmax), minimum temperature (Tmin), and the relative humidity (RH). Since 2004, BMKG has been doing a study for a short-term weather forecasting using Numerical Weather Prediction (NWP) data, but the result of the NWP forecasting was biased for a location that had complex high-resolution topography and vegetation. Thus, Clark et al. (2001) used the Model Output Statistics (MOS) to optimize the utilization of NWP output to produce more accurate weather forecasts.

MOS is a method for modeling of the relation between the weather observation result and the NWP output based on a regression method. MOS will determine the statistical relationship between the predictor variable and the NWP model response variable for a certain time projection (Glahn and Lowry, 1972). In this study, we use Univariate Partial Least

Square (PLS) as the MOS method, PLS utilizes a univariate response and only has a single objective function and a single response variable.

The response variable is weather observation data, while the predictor variable is the output data of the Numerical Weather Prediction Conformal Cubic Atmospheric Model (NWP CCAM). The NWP data is taken from 9 measurement grids for every variable so that the complexity will be high and the multicollinearity potentially occurs. This high complexity can be tolerated using PCA (Principal Component Analysis) process to reduce the dimension of the grid. The result from this dimension reduction will be used as the predictor variable for the PLS. Then, the PLS result through the PCA as its pre-processing stage will be compared with the actual data and the NWP model by looking at RMSEP (Root-Mean-Square-Error Prediction) and %IM (percentage improval) criteria.

We describe the Principal Component Analysis (PCA) method, MOS Modeling using PLS, variables used, and model evaluation in section 2. In section 3, we apply the method to forecast temperature and humidity, also show the results of our analysis. Finally, section 4 presents the conclusion of this study. In this study, we use statistical approach to explain about temperature and humidity forecast.

2. METHODS AND MATERIALS

2.1 Principal Component Analysis

Principal component analysis (PCA) is to reduce multicollinearity and the dimension of data. The result will be a new data with reduced variable but still able to explain the variability of data (Joliffe, 1986). If a random Vector $\mathbf{X}^T = [X_1, X_2, ..., X_p]$ has a covariance matrix of $\mathbf{\Sigma}$ with the eigenvalue of $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$, then the linear combination will be in (1).

$$\begin{aligned} PC_1 &= \mathbf{e}_1^T \mathbf{X} = e_{11}X_1 + e_{12}X_2 + ... + e_{1p}X_p \\ &\vdots \\ PC_p &= \mathbf{e}_p^T \mathbf{X} = e_{p1}X_1 + e_{p2}X_2 + ... + e_{pp}X_p \end{aligned} \tag{1}$$

PC_p = the p^{th} linear combination, the p^{th} biggest variance

X_p = the p^{th} origin variable

$\mathbf{e}_p$ = the p^{th} eigenvector

The i^{th} linear combination can be generally written as follows in (2).

$$PC_i = \mathbf{e}_i^T \mathbf{X},\ i = 1, 2, ..., p \tag{2}$$

So that, $Cov(PC_i, PC_k) = \mathbf{e}_i^T \mathbf{\Sigma} \mathbf{e}_k$, $i, k = 1, 2, ..., p$. The principal components do not have any correlation among each of them and have the same variance with eigenvalue from Σ, so as in (3).

$$\sigma_{11} + \sigma_{22} + ... + \sigma_{pp} = \sum_{i=1}^{p} Var(X_i) = \lambda_1 + \lambda_2 + ... + \lambda_p = \sum_{i=1}^{p} Var(PC_i) \tag{3}$$

The number of principal components is k where $k < p$ and the proportion of total variance that can be explained by the k^{th} principal component as follows:

$$\text{Variance Proportion for } k^{th}\ PC = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \ldots + \lambda_p} \quad . \tag{4}$$

According to Johnson and Winchern (2007), there are several points to determine the amount of PC:

1. Observing the scree plot, as it shows the amount of eigenvalue λ_i. If the line created at the scree plot has a certain big range, then the PC on this line will be taken.
2. The amount of the PC taken is chosen according to the amount of eigenvalue that is greater than 1 (if the PC is obtained from the correlation matrix).
3. The amount of PC taken should have a cumulative variance percentage of 80% to 90%. It means that the PC should be able to explain data variability of at least 80%.

2.2 MOS Modeling using PLS

MOS is a modeling between the weather observation result and the output of NWP based on regression. According to Wilks (2006), the general mathematical model of MOS is shown in (5).

$$\hat{Y}_t = f_{MOS}\left(X_t\right) \tag{5}$$

$\hat{Y}_t$ = weather forecast at the time-*t*

X_t = output variables of NWP at the time-*t*

PLS (Partial Least Square) is an efficient statistical method for predicting a small data sample with a lot of variables that might be correlated with each other. By doing a computer calculation, PLS becomes easier to be implemented for a great amount of data without the need to provide assumption (Wilks, 2006). In PLS, the dimensional reduction and the regression process are done simultaneously. Then $\mathbf{T}$ is denoted as the latent variable or score, which is obtained from random sample variable matrix decomposition $n \times c$. $\mathbf{P}$ is called the X-loadings $p \times c$ and $\mathbf{Q}$ is called Y-loadings $q \times c$. The PLS is based on the latent component decomposition from (6)

$$\begin{aligned} \mathbf{X} &= \mathbf{T}\mathbf{P}^T + \mathbf{E} \\ \mathbf{Y} &= \mathbf{T}\mathbf{Q}^T + \mathbf{F} \end{aligned} \quad . \tag{6}$$

Hence, the $\mathbf{X}$ matrix is $n \times p$ and $\mathbf{Y}$ is $n \times q$. $\mathbf{E}$ and $\mathbf{F}$ are residual matrices that are each of which are $n \times p$ and $n \times q$.

The PLS is just like the principal component regression that is a method that forms the latent component matrix $\mathbf{T}$ as the linear transformation from $\mathbf{X}$,

$$\mathbf{T} = \mathbf{X}\mathbf{W}^* \tag{7}$$

$\mathbf{W}^*$ is the weighting matrix sized $p \times c$ with c is the number of latent components. The $\mathbf{W}^*$ can be obtained using (8).

$$\mathbf{W}^* = \mathbf{W}\left(\mathbf{P}^T\mathbf{W}\right)^{-1} \tag{8}$$

The latent component is used to predict **Y**, substituting the origin variable, **X**. When **T** is formed, we can then obtain $\mathbf{Q}^T$ from the smallest quadratic method as in (9).

$$\hat{\mathbf{Q}}^T = \left(\mathbf{T}^T\mathbf{T}\right)^{-1}\mathbf{T}^T\mathbf{Y} \tag{9}$$

From equation (6), $\mathbf{Y} = \mathbf{TQ}^T + \mathbf{F}$ and the matrix **B** is a regression coefficient matrix for the model $\mathbf{Y} = \mathbf{XB} + \mathbf{F}$, then the equation (10) is obtained.

$$\begin{aligned} \mathbf{XB} &= \mathbf{TQ}^T \\ \mathbf{XB} &= \mathbf{XW}*\mathbf{Q}^T \\ \mathbf{B} &= \mathbf{W}*\mathbf{Q}^T \end{aligned} \tag{10}$$

The estimator of $\mathbf{B}$ is $\hat{\mathbf{B}} = \mathbf{W}*\left(\mathbf{T}^T\mathbf{T}\right)^{-1}\mathbf{T}^T\mathbf{Y}$. So that we can obtain a conjecture for $\mathbf{Y}$ as in (11).

$$\begin{aligned} \hat{\mathbf{Y}} &= \mathbf{X}\hat{\mathbf{B}} \\ \hat{\mathbf{Y}} &= \left[\mathbf{TW}^{-1}\mathbf{W}\left(\mathbf{T}^T\mathbf{T}\right)^{-1}\mathbf{T}^T\mathbf{Y}\right] \\ \hat{\mathbf{Y}} &= \left[\mathbf{TI}\left(\mathbf{T}^T\mathbf{T}\right)^{-1}\mathbf{T}^T\mathbf{Y}\right] \\ \hat{\mathbf{Y}} &= \left[\mathbf{T}\left(\mathbf{T}^T\mathbf{T}\right)^{-1}\mathbf{T}^T\mathbf{Y}\right] \end{aligned} \tag{11}$$

The PLS can be used for both univariate response and multivariate response. This study is utilizing the PLS for the univariate response with the intent to obtain each modeling result from the response variable separately. The amount of latent variable is determined by a statistic assessing the accuracy of estimation, Prediction Residual Sum of Square *(PRESS)*. The *PRESS* value for the univariate response is shown in (12).

$$PRESS = \sum_{t=1}^{n}\left(y_t - \hat{y}_{t-1}\right)^2 \tag{12}$$

The modeling using PLS is done when the response variable is to be analyzed separately so that Y is a response matrix variable $n \times 1$. For a certain weight amount $\mathbf{w_i} = \left(w_{11},...,w_{pi}\right)^T$, the covariance between the response variable Y and the random variable $T_i = w_{1i}X_1 + w_{2i}X_2 ... + w_{pi}X_p$ can be obtained using (13)

$$COV\left(Y,T_i\right) = \frac{1}{n}\mathbf{w}_i^T\mathbf{X}^T\mathbf{Y}. \tag{13}$$

Covariance between T_i and T_j for $i \neq j;\ j = 1,2,...,c$

$$COV\left(Y,T_i\right) = \frac{1}{n}\mathbf{w}_i^T\mathbf{X}^T\mathbf{w}_j = \frac{1}{n}\mathbf{t}_i^T\mathbf{t}_j \tag{14}$$

$\mathbf{w}$ is defined to be the square of the covariance between Y and the latent component, $\mathbf{w}$ is maximized when each of the latent components does not have any correlation. Generally, the PLS only has one objective function. This objective function that is maximized on PLS for $i = 1, 2, ..., c$ will produce a weighting vector using (15)

$$\mathbf{w}_i = \arg\max_i \mathbf{w}^{\mathrm{T}} \mathbf{X}^{\mathrm{T}} \mathbf{Y} \mathbf{Y}^{\mathrm{T}} \mathbf{X} \mathbf{w} \tag{15}$$

as long as: $\mathbf{w}_i^T \mathbf{w}_i = 1$; $\mathbf{w}_i^T \mathbf{X}^T \mathbf{X} \mathbf{w}_j = \mathbf{t}_i^T \mathbf{t}_j = 0$, for $j = 1, 2, ..., i-1$.

We can see from the formula that the latent component formed on PLS has maximum covariance with the response variable so that the prediction is very good (Clark et al., 2001). *PLS Algorithm* (Boulesteix et al., 2006)

a. First iteration h=1, Maximum iteration $h_{\max} = p$
b. Determine $\mathbf{w} = \mathbf{X}^T \mathbf{y} / \mathbf{y}^T \mathbf{y}$
c. Calculate $\mathbf{t} = \mathbf{X}\mathbf{w}$
d. Calculate the loading $\mathbf{Y}$, $\mathbf{q} = \mathbf{y}^T \mathbf{t} / \left(\mathbf{t}^T \mathbf{t}\right)$
e. Renew $\mathbf{X}$ and $\mathbf{Y}$, as in (16)

$$\begin{aligned} \mathbf{p} &= \mathbf{X}^T \mathbf{t} / (\mathbf{t}^T \mathbf{t}) \\ \mathbf{X} &= \mathbf{X} - \mathbf{t}\mathbf{p}^T \\ \mathbf{Y} &= \mathbf{Y} - \mathbf{t}\mathbf{q}^T \end{aligned} \tag{16}$$

The value for measuring the goodness of the model's prediction is the determination of coefficient value (R^2) that can be calculated using (17)

$$R^2 = 1 - \frac{\sum_{t=1}^{n} \left(y_t - \hat{y}_t\right)^2}{\sum_{t=1}^{n} \left(y_t - \bar{y}\right)^2}. \tag{17}$$

2.3 Model Validation

One of the measurements that can be used to know the quality of forecasting result is Root Mean Square Error of Prediction (RMSEP) (Wold et al., 2001). The formula we can use to obtain the RMSEP value from the univariate modeling is as (18).

$$RMSEP = \sqrt{\frac{\sum_{t=1}^{n_{pred}} \left(Y_t - \hat{Y}_t\right)^2}{n_{pred}}} \tag{18}$$

The smaller the RMSEP value, the better the forecasting model. The criteria of RMSEP value can be used as a base for model validation which is shown in Table 1.

Table 1: RMSEP value criteria (Source: BMKG).

Criterion	RMSEP	
	Temperature	**Humidity**
Very good	0.0 - 0.4	0.0 - 2.5
Good	0.5 - 0.8	2.6 - 5.0
Intermediate	0.9 - 1.2	5.1 - 7.5
Bad	1.3 - 1.6	7.6 - 10.00
Very bad	> 1.6	> 10.00

2.4 Bias Corrector Measurement

The percentage improvement of MOS model against the NWP is shown by the Percentage Improval (%IM) that can be calculated using formulas as (19)

$$\%IM = \frac{RMSEP_{NWP} - RMSEP_{MOS}}{RMSEP_{NWP}} \times 100\%. \quad (19)$$

The value of %IM is from 0% to 100%. The higher value of %IM means the MOS model has a better correction of the NWP's biased forecasting result.

2.5 Data and Variables

The data used in this study is a secondary data from BMKG, i.e. the output of the daily NWP CCAM from 1 January 2009 to 31 December 2010. Four observation stations that are used in this study are Citeko, Kemayoran, Pondok Bentung, and Tangerang. The response variable is the surface's weather observation data that consist of Tmax, Tmin, and RH measured directly in every station. The predictor variable is the output of the NWP CCAM model. Meanwhile, the NWP CCAM parameter used is taken from the previous study's parameter by a meteorologist, shown in Table 2 for the MOS model.

The used parameters from the NWP CCAM for every observation station are 18 parameters. The 11 parameters are measured on the surface level (with a height of ±2 meters above the sea level), while the other 7 parameters are measured on a three level of different air pressures, where level 1 is 100 millibar pressure, level 2 is 950 millibar pressure, and level 4 is 850 millibar pressure. Therefore, the total parameters are 32 parameters. Each parameter is measured on 9 measurement grids (3×3) in the nearest location from the place of observation station.

Table 2: NWP CCAM parameters.

No.	Variable	Level
1	Surface Pressure Tendency dpsdt)	Surface
2	Water Mixing Ratio (mix)	1, 2, 4
3	Vertical Velocity (omega)	1, 2, 4
4	PBL depth (pblh)	Surface
5	Surface Pressure (ps)	Surface
6	Mean Sea Level Pressure (psl)	Surface
7	Screen Mixing Ratio (qgscm)	Surface
8	Relative Humidity (rh)	1, 2, 4
9	Precipitation (rnd)	Surface
10	Temperature	1, 2, 4
11	Maximum Screen Temperature (tmaxcr)	Surface
12	Minimum Screen Temperature (tmincr)	Surface
13	Pan Temperature (tpan)	Surface
14	Screen Temperature (tscrn)	Surface
15	Zonal Wind (u)	1, 2, 4
16	Friction Velocity (ustar)	Surface
17	Meridional Wind (v)	1, 2, 4
18	Geopotential Height (zg)	1, 2, 4

3. RESULTS AND DISCUSSION

The analysis and evaluation steps for Tangerang Station will be explained in detail, while the rest of the stations will be just a slight summary since the occurrence analysis steps are actually the same.

3.1 Pre-Processing the NWP Data using PCA Method

Each NWP variable is measured on 9 measurement grids. Hence, there are 162 (18×9) predictor variables will increase the complexity of the model. To solve it, this study used a dimensional reduction i.e. PCA. The amount of principal components is determined by choosing which have an eigenvalue larger than one. The principal component for the NWP variable in Tangerang Station is shown in Table 3.

Table 3: NWP variable's principal components in Tangerang station.

Variable	PC	Eigen Value	Var.	Variable	PC	Eigen Value	Var.
Dpsdt	1	9.2904	99.9857	temp2	1	8.3576	97.1532
mixr1	1	8.7048	92.4047	temp4	1	8.7006	99.0090
mixr2	1	8.9565	96.2157	Tmaxscr	1	8.5420	98.0922
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
temp1	1	8.4996	95.8768	zg4	1	8.5784	97.5735

Table 3 shows that in Tangerang station, each NWP variables produces 1 component, except for the *zg level 1* variable that is 3 components, and *zg level 2* variable that is 2 components. Therefore, the total amount of the principal components that are

formed in Tangerang Station is 35 components, 39 components in Citeko Station, 35 components in Kemayoran and Pondok Betung Station. The variability of NWP variables explained by the principal components varies from 92.40% until almost 100%. The principal components will be used as the predictor variables on the MOS modeling using PLS.

3.2 Prediction Modeling of Tmax, Tmin, and RH using PLS Method

The first step of PLS modeling in Tangerang Station is to determine the optimum amount of component of each model using a cross validation.

Table 4: The amount of the optimal components in four stations.

Station	Variable	Amount of components	Smallest PRESS value
Citeko	Tmax	11	0.7317
	Tmin	7	0.8627
	RH	29	0.7554
Kemayoran	Tmax	9	0.7035
	Tmin	6	0.8627
	RH	6	0.8653
Pondok Betung	Tmax	22	0.7154
	Tmin	5	0.9079
	RH	5	0.9084
Tangerang	Tmax	6	0.7081
	Tmin	3	0.9478
	RH	2	0.9476

On the cross-validation process, every iteration will produce a *PRESS* value. Model with the smallest *PRESS* value will be the model that holds the optimum amount of components. The optimal component from the PLS in the four stations is shown in Table 4.

The optimal amount of component in each station is then used for the predictive modeling process of Tmax, Tmin, and RH. The modeling process will be explained according to the steps of the PLS modeling which have been described previously.

1. *Calculating PLS Weighting in Tangerang Station*

The weighting matrix (**W**) is obtained from a merge of an every weighting vector extracted according to the amount of optimal component that has already been determined before. The component of **W** matrix in Tangerang Station (i.e. Tmax, Tmin, and RH) is shown in Table 5, Table 6, and Table 7 respectively.

Table 5: The weight value of X used for Tmax of PLS modeling in Tangerang station.

Variable	**w1**	**w2**	**w3**	...	**w6**
PC.dpsdt	0.0417	0.1248	-0.1882	...	-0.2782
PC.mixr1	-0.0130	0.0309	-0.2308	...	-0.2094
PC.mixr2	0.0714	0.1292	-0.2660	...	0.0641
⋮	⋮	⋮	⋮	⋱	⋮
PC.zg4	-0.2603	0.1695	-0.0784	...	-0.1689

Table 6: The weight value of X used for Tmin of PLS modeling in Tangerang station.

Variable	**w1**	**w2**	**w3**
PC.dpsdt	-0.0318	0.1803	0.1639
PC.mixr1	-0.3385	-0.1831	0.1406
PC.mixr2	-0.2958	-0.0666	0.1023
⋮	⋮	⋮	⋮
PC.zg4	-0.2933	0.0598	-0.1046

Table 7: The weight value of X used for RH of PLS modeling in Tangerang station.

Variable	**w1**	**w2**
PC.dpsdt	-0.0651	-0.0237
PC.mixr1	-0.1994	-0.0207
PC.mixr2	-0.2408	-0.0206
⋮	⋮	⋮
PC.zg4	0.0579	-0.3081

2. X-Scores Formation

The obtained X-scores will be the **T** matrix consisted of a vector **t** component. The **X** matrix is the predictor matrix from the result of PCA operation, while **w** is the weighting value that is obtained previously. The X-scores for Tmax is shown in Table 8, and Table 9 for the Tmin and RH.

Table 8: The X-*scores* for Tmax of PLS modeling in Tangerang station.

N	**t1**	**t2**	...	**t6**
1	-4.2611	0.1455	...	-1.0065
2	-0.6483	1.0485	...	-0.5370
3	-1.0081	2.1319	...	-2.1069
⋮	⋮	⋮	⋱	⋮
637	0.9811	-0.2294	...	0.2630

Table 9: The X-*scores* for Tmin and RH of PLS modeling in Tangerang station.

N	Tmin			RH	
	t1	**t2**	**t3**	**t1**	**t2**
1	-1.1287	1.7879	-1.7765	2.9974	-2.8224
2	-1.4827	0.5584	1.5362	-1.3023	-2.7302
3	-3.9506	0.8837	3.6885	-2.6981	-3.8481
⋮	⋮	⋮	⋮	⋮	⋮
637	0.1494	-0.9879	-0.5842	-0.5341	0.2010

3. Loading Factor Matrix Formation for Y

Y-loading is a loading related to the response variable. The loadings factor is obtained from the combination of loadings Y factor of each component. The loadings factor matrix for Y is shown in Table 10.

Table 10: Loadings Y factor of PLS modeling in Tangerang station.

Q	Tmax	Q	Tmin	q	RH
q1	0.2356	q1	0.1354	q1	0.2150
q2	0.1921	q2	0.0815	q2	0.1577
q3	0.1332	q3	0.0532		
q4	0.0821				
q5	0.0874				
q6	0.0543				

4. Calculating Regression Coefficient

The PLS coefficient (**B**) can be obtained after matrix **W**, **Q**, and **T**. The component of the PLS coefficient matrix on Tangerang Station is shown in Table 11.

Table 11: PLS coefficient in Tangerang station.

Variable	Tmax	Tmin	RH
PC.dpsdt	0.0012	0.0243	-0.0212
PC.mixr1	-0.0320	-0.0707	-0.0566
PC.mixr2	0.0436	-0.0520	-0.0677
⋮	⋮	⋮	⋮
PC.zg4	-0.0553	-0.0479	-0.0331

5. PLS Preparation

The preparation of PLS is done by the regression coefficient taken from Table 11 with the predictor variable that is obtained from PCA. When the PLS is formed, the conjectured value of the Tmax, Tmin, and RH can also be obtained. Those conjecture value, especially the conjectured value of the training data can be used to test how good the formed model with R^2 is, as the higher the R^2 value, the better the model is. The obtained R^2 value from the PLS in the four stations is shown in Table 12.

Table 12: The value of R^2 from PLS in four stations.

Station	**Variable**	$\boldsymbol{R^2}$**(%)**
Citeko	Tmax	52.75
	Tmin	41.99
	RH	50.89
Kemayoran	Tmax	54.57
	Tmin	31.26
	RH	46.33
Pondok Betung	Tmax	53.57
	Tmin	24.44
	RH	47.03
Tangerang	Tmax	55.33
	Tmin	14.93
	RH	31.96

Table 12 shows that the average value of R^2 obtained from the PLS is generally not that good, despite that the R^2 value for the Tmax modeling (maximum temperature as response) alone is good, ranging from 52.75% to 55.33%, because the R^2 value from the Tmin modeling (minimum temperature as response) is quite small with a range of 14.93% - 41.99% and the R^2 value from the RH modeling that ranges from 31.95% to 50.89%. The R^2 value for the Tmax modeling in Tangerang Station is 55.33%, means that there is 55.33% Tmax variance that can be explained by the formed model.

3.3 PLS Validation

The model validation aims to know the accuracy and the goodness of the formed model. The PLS validation is done by testing data with the observation data so that we can obtain the RMSEP value. The RMSEP value in four stations is shown in Table 13.

Generally, the RMSEP value of the Tmax modeling using PLS has an intermediate result according to the BMKG criterion. In the other side, the RMSEP value of the Tmin in Citeko, Kemayoran, and Pondok Betung Station has a good result of 1.0857. This PLS modeling is also has a good criterion if used for RH modeling in Citeko, Kemayoran, and Pondok Betung Station, while the RH modeling in Tangerang Station has an intermediate criterion because it holds the RMSEP value of 5.7314. The result of this PLS modeling is then regarded as the MOS model.

Table 13: RMSEP value for PLS in four stations.

Station	Variable	RMSEP	RMSEP Criterion
Citeko	Tmax	1.1261	Intermediate
	Tmin	0.5183	Good
	RH	4.9307	Good
Kemayoran	Tmax	0.9698	Intermediate
	Tmin	0.7502	Good
	RH	4.3629	Good
Pondok Betung	Tmax	1.0479	Intermediate
	Tmin	0.8563	Good
	RH	4.6994	Good

	Tmax	0.9485	Intermediate
Tangerang	Tmin	1.0857	Intermediate
	RH	5.7314	Intermediate

3.4 Comparison of Accuracy between NWP Prediction Result and MOS Model Result

The NWP model produces a biased forecast so that it needs a post-processing using the MOS method, i.e. PLS. The percentage of the amount of biased NWP data that can be corrected by the MOS model is shown by the Percentage Improval (%IM), where the $RMSEP_{NWP}$ is obtained based on the comparison of the NWP data on the fifth grid (the nearest grid from the observation station) and the observation data for the Tmax, Tmin, and RH variables. The amount of biased data that can be corrected by the MOS model with the PLS in the four stations are shown in Table 14.

Table 14: The value of $RMSEP_{NWP}$, $RMSEP_{MOS}$, dan %IM.

Station	**Variable**	**$RMSEP_{MOS}$**	**$RMSEP_{NWP}$**	**%IM**
	Tmax	1.1261	4.1572	72.9121
Citeko	Tmin	0.5183	5.1505	89.9369
	RH	4.9307	13.0509	62.2195
	Tmax	0.9698	2.9491	67.1154
Kemayoran	Tmin	0.7502	1.9110	60.7431
	RH	4.3629	7.1804	39.2388
	Tmax	1.0479	3.3227	68.4624
Pondok Betung	Tmin	0.8563	1.0812	20.8010
	RH	4.6994	7.5821	38.0198
	Tmax	0.9485	3.1089	69.4908
Tangerang	Tmin	1.0857	1.3400	18.9776
	RH	5.7314	6.5589	12.6164

Table 14 shows that the RMSEP that is obtained from the NWP model is greater than the RMSEP from the MOS model, which means that the MOS model is consistently better to be used to predict the Tmax, Tmin, and RH rather than the NWP model. The MOS model is able to correct from 18.9776% to 89.9369% of the biased NWP for forecasting the Tmin. The same table also shows that the $RMSEP_{NWP}$ in the Citeko Station is the greatest among the other four stations so that Citeko Station has a %IM that holds the greatest bias corrector. This is because the Citeko Station is located in the mountain area that holds a complex vegetation, therefore producing a big amount of bias for the NWP model.

4. CONCLUSION

Most of the principal components that are formed by the result of the NWP variables reduction within the 9 measurement grids are exactly one component. The validation result of the PLS with the RMSEP criterion shows that the Tmax belongs to intermediate for all stations, Tmin has a good criterion in three stations (i.e. Citeko, Kemayoran and Pondok Bentung), and RH has a good criterion in three stations (i.e. Citeko, Kemayoran and Pondok Bentung). The prediction results from the PLS is more accurate than the NWP model and able to correct an 89.94% of the biased NWP for Tmin forecasting (response as a result of PLS modeling). Therefore, we can conclude that the

PLS can solve the NWP problem regarding the relation function and dimension reduction.

The modeling result from this study is recommended to be used for BMKG in forecasting the temperature and humidity because this model is capable to produce a smaller bias compared to the NWP model from the BMKG itself. It must be noted that a method comparison should be done in each station to obtain the best method due to a potential spatial effect that may occur.

5. ACKNOWLEDGEMENT

The entire data used in this study were supported by the Meteorology, Climatology and Geophysics Agency (BMKG) of Indonesia. The fund for this study is supported by the Ministry of Research, Technology and Higher Education of Indonesia for the grant of the National Strategic Research 2018.

6. REFERENCES

BMKG. (2006). *Uji Operasional dan Validasi Model Output Statistik (MOS).* Jakarta: BMKG.

Boulesteix, Anne-Laure, and Strimmer, K. (2006). Partial Least Squares: A Versatile Tool for the Analysis of High-Dimensional Genomic Data. *Journal of Briefings in Bioinformatics, 8*: 32-44.

Clark, M. P., Hay, L. E., and Whitaker, J. S. (2001). Development of operational hydrologic forecasting capabilities. *American Geophysical Union, Fall Meeting*.

Glahn, H. R., and Lowry, D. A. (1972). The Use of a Model Output Statistics (MOS) in Objective Weather Forecasting. *Applied Meteorology*, 1203-1211.

Johnson, R. A., and Winchern, D. W. (2007). *Applied Multivariate Statistical Analysis 6th Edition*. United States: Pearson Education.

Joliffe, I. T. (1986). *Principal Component Analysis* (2nd ed.). New York: Springer-Verlag.

Tjasyono, B. (2004). *Klimatologi 2nd Edition*. Bandung: ITB.

Wardani, I. K. (2010). *Manfaat Prediksi Cuaca Jangka -Pendek Berdasarkan Data Radiosonde dan Numerical Weather Prediction (NWP) untuk Pertanian Daerah.*

Wilks, D. S. (2006). *Statistical Methods in the Atmospheric Sciences 2nd Edition*. Boston: Elvesier.

Wold, S., Sjostrom, M., and Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Journal of Chemometrics and Intelligent Systems, 58*: 109-130.

Malaysian Journal of Science 38 (Special Issue 2): 14- 22 (2019)

**THE INTERNATIONAL SEMINAR ON MATHEMATICS IN INDUSTRY (ISMI)
AND THE INTERNATIONAL CONFERENCE ON THEORETICAL AND APPLIED STATISTICS (ICTAS)
ISMI-ICTAS18 [4-6 SEPTEMBER 2018]**

Assessing Dynamic-Time-Warping Dissimilarity Measures in Regionalization of River Discharges

Nur Syazwin Mansor[1a], Norhaiza Ahmad[1b]*, Arien Heryansyah[2c]

[1] Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia, 81310 Johor Bahru, MALAYSIA. E-mail: nsyazwin3@live.utm.my[a] ; norhaiza@utm.my[b]
[2] Faculty of Engineering, Universitas Ibn Khaldun (UIKA), Bogor. Jalan KH Sholeh Iskandar KM.2, Kedung Badak, Tanah Sereal, Kota Bogor, Jawa Barat 16162, INDONESIA. Email: arengga@gmail.com[c]
* Corresponding Author: norhaiza@utm.my[b]
Received: 21st April 2019 Revised : 6th August 2019 Published: 30th September 2019
DOI : https://doi.org/10.22452/mjs.sp2019no2.2

ABSTRACT Regionalization of river discharges is a process of transferring hydrological information to generalize hydrological information from one river to another. One approach to regionalize river discharge is to use a distance-based regional analysis by employing a Dynamic Time Warping (DTW) dissimilarity measure to cluster homogeneous river discharge patterns based on sequenced of time series discharge data. However, clustering homogeneous river discharge patterns can be sensitive to the choice of distance metric measures used due to out of phase behavior in the discharge time series. In this study, we assess three types of Dynamic Time Warping (DTW) measures specifically conventional DTW, a feature based DTW and a weighted based DTW on four annual discharge time series from six rivers in the state of Johor, Malaysia. A comparison of eight different clustering validation indices to determine the optimal number of rivers clusters with similar discharge patterns. These indices are used to measure the internal and external strength of the identified clusters. The results indicate that weighted based DTW outperform the conventional DTW and feature based DTW with 75% of the clustering indices agree that there are three optimal clusters of river discharge. By using weight as a function in DTW, it helps to cater the out of phase behavior in river discharge time series with the highest agreement of clustering indices compared to other types of DTW measures. We also found that three of the rivers (Sayong, Bekok, and Segamat) have similar river discharge patterns and could be used together in the generalization process. Meanwhile, the other rivers (Johor, Kahang, and Muar) varies in their time series patterns.

Keywords: Dynamic Time Warping, Clustering, Dissimilarity measure

1. INTRODUCTION

Regionalization refers to a process of transferring hydrological information to generalize hydrological information from one river to another (Razavi & Coulibaly, 2012; Snelder et al., 2005). It is essential not only to ensure the transferability of information when applying regionalization methods, but can also provide valuable indications to improve the understanding of the dominant physical phenomena in the different groups (Toth, 2013). For instance, similar river discharge can be classified according to their annual maximum flows, coefficient of variation and skewness of annual maximum discharges, latitude and longitude of selected stations (Agarwal et al., 2016; Corduas, 2011; Dikbas et al., 2013; Elesbon et al., 2015). Such information can be used to estimate low discharge magnitude and frequency in river streams for the purpose of river management

(Kahya et al., 2007; Laaha & Bloschl, 2006).

One approach of regionalization is through a time-based cluster analysis. It is used to cluster homogeneous river discharge patterns based on sequence of time series discharge data. This approach classifies a set of individual time series which possess similar patterns, shapes, or changes through selected interval period of time (Ouyang et al., 2010). Each data points of the time series are objects of ordered time sequences. This time-based clustering takes into account of the whole sequence of river discharge time series and is analyzed to identify the similarity across different sets of time series discharge data.

In time-based clustering, a nonlinear dissimilarity measure is commonly considered. A common nonlinear mapping dissimilarity measure used by hydrologists for ordered time sequence data is Dynamic Time Warping (DTW) dissimilarity measure (Gertseema et al., 2016; Gupta & Chaturvedi, 2013; Mishra et al., 2015). This nonlinear dissimilarity measure aligns each data points elastically which allows similar shapes to match when the points are out of phase in a time series sequence. It does not only take into account the smallest distance between data points of each discharge sequences but also the phase-difference of the two sequences. This method allows similar shape to be matched and the variability of the river discharge is considered by warping the time of discharge occurrences.

In the literature, there are several variations of DTW. However, as far as we are concerned only the conventional DTW has been applied to analyze the river discharge data to date. Therefore, in this study, we assess three types of Dynamic Time Warping (DTW) measures specifically conventional DTW, a feature based DTW and a weighted based DTW in the regionalization of river discharge data.

2. RIVER DISCHARGE DATA

The river discharge data used in the analysis were obtained from the Malaysian Drainage and Irrigation Department which consist of the daily observed of six rivers (Johor, Sayong, Bekok, Kahang, Muar, and Segamat) over 34 years from 1980 to 2014. However, after the missing records are removed, only 28 years remain in the dataset from 1980 until 2008. In this study, only four annual discharge data (1981, 1982, 1983, and 1987) for each river were used.

River discharge time series of each river for 1981 are plotted in Figure 1. At a glance, it can be seen that there are several homogeneous groups of river. First, based on the river discharge pattern, it can be observed that Sungai Johor and Sungai Kahang share the same pattern of river discharge which consistently fluctuate throughout the year and have a very high peak at the end of the year. Sungai Sayong and Sungai Bekok appear to share the same pattern of river discharge where there is a small peak during the month of May-June and moderately high in December. However, this assumption cannot be verified only by visualization. It needs to be statistically proven. The process of identifying homogenous river discharge pattern using a time-based cluster analysis is described in the following section.

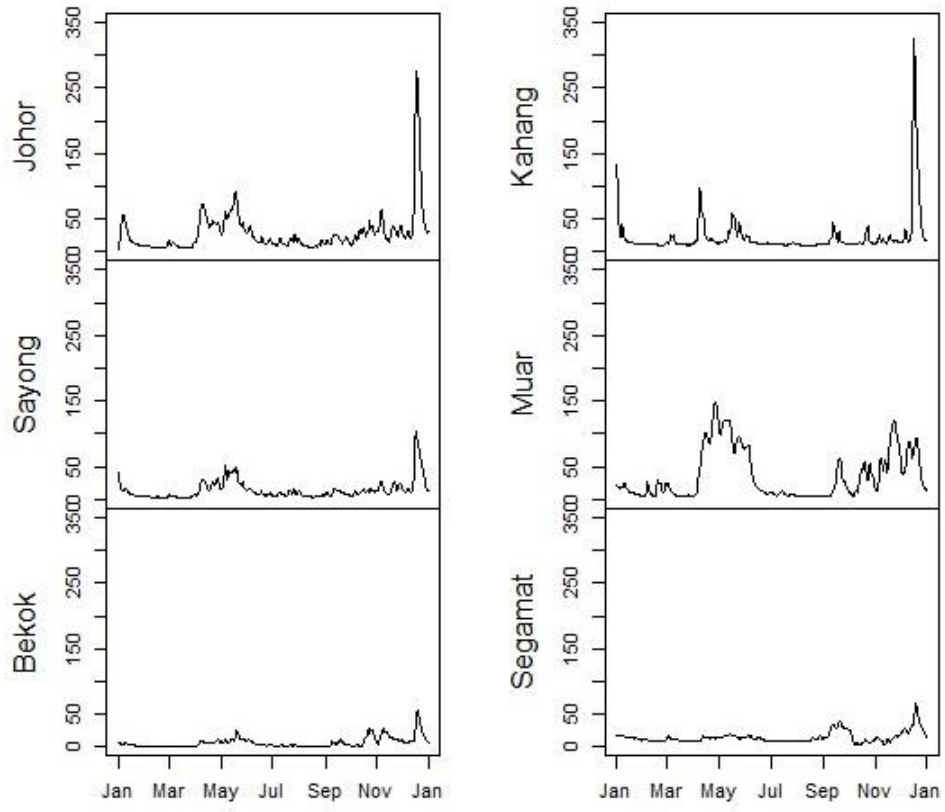


Figure 1: River discharge of six rivers in 1981.

3. METHODOLOGY

In general, the process of identifying the temporal grouping of river discharge process using cluster analysis comprises of four major steps: transformation of raw data, dissimilarity measure between annual river discharge, clustering algorithm to identify the membership of cluster, and validation index to validate the cluster membership.

In this study, transformation is not required since the measurement scale of river discharge are the same for each year. The homogeneous annual discharge processes are then identified by comparing three different types of Dynamic Time Warping (DTW) measures specifically conventional DTW, a feature based DTW and a weighted based DTW. Identification of regional grouping of river discharge is done using K-medoid clustering algorithm and C-Index measure to validate the membership of clusters.

3.1 Dynamic Time Warping (DTW)

Dynamic Time Warping (DTW) is an algorithm for measuring optimal similarity between two river discharge time sequences. The time series data vary not only on the time amplitudes but also in terms of time progression (Mishra et al., 2015). The non-linear alignment produces a similar measure, allowing similar shapes to match even if they are out of phase in the time axis. Sequences are "warped" non-linearly in the time dimension to determine their measure of similarity independence for certain non-linear variations in the time dimension. In respect of river discharge time series data, the DTW algorithm for dissimilarity measure is as follow:

$$Y_{1j} = y_{11}, y_{12}, ..., y_{1m} \quad (1)$$

$$Y_{2j} = y_{21}, y_{22}, ..., y_{2m} \quad (2)$$

An $m \times n$ matrix is constructed using DTW aligning for these two sequences. Each element in the matrix contains the dissimilarity between two points Y_{1j} and Y_{2j}, called Euclidean distance, D_{EUC}. The distance is defined as:

$$D_{EUC}(Y_i, Y_j) = \sqrt{\sum_{k=1}^{p} \left(Y_{ik} - Y_{jk}\right)^2} \quad (3)$$

where Y_{ik} and Y_{jk} are respectively the k^{th} river discharge value of the p-dimensional monsoon period for individuals i and j. A warping path, W is a contiguous set of matrix elements that defines a mapping between Y_{1j} and Y_{2j}. The kth element of W is defined as $w_k = (i, j)_k$, so we have:

$$W = w_1, w_2, ..., w_k, ..., w_K \text{ where } (m, n) \le K < (m + n - 1) \quad (4)$$

The warping path is typically subject to several constraints:

i. Boundary conditions: $w_1 = (1,1)$ and $w_k = (m, n)$. The warping path needs to start and finish diagonally opposite corner cells of the matrix.

ii. Continuity: Given $w_k = (a, b)$ then $w_{k-1} = (a', b')$, where $a - a' \le 1$ and $b - b' \ge 1$. The allowable steps in the warping path are restricted to adjacent cells (including diagonally adjacent cells).

iii. Monotonicity: Given $w_k = (a, b)$ then $w_{k-1} = (a', b')$, where $a - a' \ge 0$ and $b - b' \le 0$. The points in W need to be monotonically spaced in time.

Many warping paths satisfy the constraints, but only one path is chosen which minimizes the warping cost taken by:

$$D_{DTW}(X_1, X_2) = \min\left(\frac{1}{k}\sum_{k=1} w_k\right) \quad (5)$$

where k in the denominator used to compensate the fact that warping paths may have different lengths.

3.2 Derivative Dynamic Time Warping (DDTW)

Derivative Dynamic Time Warping (DDTW) utilize information on the shape of the time series by considering the first derivative of the sequences. Previously, the conventional DTW construct $m \times n$ path matrix where the matrix contains the dissimilarity between two points Y_{1j} and Y_{2j}, called Euclidean distance, D_{EUC}. With DDTW the dissimilarity measure is not Euclidean but rather the square of the estimated derivatives of Y_{1j} and Y_{2j}. Estimate derivatives of each data points are:

$$D_{DER}[Y] = \frac{(y_i - y_{i-1}) + ((y_{i+1} - y_{i-1})/2)}{2} \quad (6)$$

This estimate is simply the average of the slope of the line through the point in question and its left neighbor, and the slope of the line through the left neighbor and the right neighbor. Empirically this estimate is more robust to outliers than any estimate considering only two data sequences. Note the estimate is not defined for the first and last elements of the sequence. Instead, we use the estimates of the second and penultimate elements respectively.

3.3 Weighted Dynamic Time Warping (WDTW)

Weighted Dynamic Time Warping (WDTW) is the modified version of the conventional DTW as described in 3.1 which calculates the distance of all pairwise points with equal weight of each point regardless of phase difference, WDTW penalizes the points according to the phase difference between discharge of the two river discharge time series (Jeong et al., 2011). In WDTW algorithm, when creating an $m \times n$ path matrix, the distance between the two points y_i and y_j is calculated as:

$$D_W(y_i, y_j) = \left\| w_{|i-j|}(y_i - y_j) \right\| \quad (7)$$

where $w_{|i-j|}$ is the positive weight value between the two points y_i and y_j. A Modified Logistic Weight Function (MLWF) is used to assign weight and the weight value $w_{(i)}$ is defined as:

$$w_{(i)} = \left[\frac{w_{\max}}{1 + \exp(-g(i - m_c))} \right] \quad (8)$$

where $i = 1, \ldots, m$, m is the length of a sequence and m is the midpoint of a sequence. $w_{\max}$ is the desired upper bound for the weight parameter, and g is an empirical constant that controls the curvature (slope) of the function; that is, g controls the level of penalization for the points with larger phase difference. The value of $w_{\max}$ is set to 1 and g is set to 0.6 as recommended by Jeong et al. (2011).

3.4 K-medoid Clustering

Instead of using the mean point as the center of a cluster, K-medoids uses an actual point in the cluster to represent it. Medoid is the most centrally located object of the cluster, with minimum sum of distances to other points. First, we select K representative points to form initial clusters, and they are repeatedly moved to better cluster representatives. All possible combinations of representative and non-representative points are analyzed, and the quality of the resulting clustering is calculated for each pair. An original representative point is replaced with the new point which causes the greatest reduction in distortion function. At each iteration, the set of best points for each cluster form the new respective medoids.

3.5 Internal Validity Measure

Internal validity measures are evaluated based on the compactness and separation. Compactness refers on how close the annual

river discharges within a cluster. Meanwhile, separation refers to how distinct a cluster is from other clusters of river discharges. In this study, we use eight different internal validity measures which are Calinski-Harabasz (CH), Silhouette, Dunn, C, Davies-Bouldin (DB), McClain, Krzanowski-Lai (KL), and S_Dbw indices (Desgraupes, 2013; Zhao, 2012).

4. RESULTS AND DISCUSSION

We use four years of annual discharge for each six rivers in Johor as described in Section 2. This dataset is reflected by a 365×24 data matrix from the corresponding vectors. Three dissimilarity measures (DTW, DDTW, and WDTW) are applied to the dataset to produce different 24×24 dissimilarity matrices. Each dissimilarity matrix is used in the K-medoid clustering procedure to extract different cluster solution from the river discharge. In order to find the optimal number of clusters, we tested different number of cluster solutions ranging from three to five clusters based on its internal clustering indices.

Table 1 shows the internal indices value for three to five cluster solution. The first three indices (CH, Silhouette, and Dunn) indicate strong similar river discharge pattern when the values are large. The larger the value, the more accurate closeness within each cluster. In contrast, the next five indices (C, DB, McClain, KL, and SDBW) indicate strong similar river discharge pattern when the values are small. The smaller the value, the more accurate closeness within each cluster.

We summarize the agreement between all the indices in the last row of Table 1. WDTW clearly outperforms the other two dissimilarity measures with 75% of the clustering indices agree that there are three optimal cluster solutions for the river discharge. Meanwhile, only 25% of the indices agree that there are three or five optimal numbers of clusters using DTW and DDTW respectively.

Table 1: Internal clustering indices for 3 to 5 cluster solution.

Index/No. of Clusters	**3**			**4**			**5**		
	DTW	DDTW	WDTW	DTW	DDTW	WDTW	DTW	DDTW	WDTW
CH	6.51	6.16	**6.97**	5.64	5.27	5.96	5.79	5.49	5.44
Silhouette	0.29	0.29	**0.47**	0.30	0.30	0.36	0.30	0.30	0.36
Dunn	0.29	0.29	**0.45**	0.37	0.36	0.33	0.43	0.43	0.41
C	**0.29**	0.30	0.43	0.34	0.34	0.39	0.48	0.49	0.42
DB	1.63	1.67	**0.99**	1.63	1.68	1.62	1.46	1.51	1.60
McClain	0.94	0.96	**0.26**	0.99	1.01	0.93	1.02	1.04	1.09
KL	2.37	2.46	2.09	0.82	**0.75**	1.18	2.15	2.21	1.26
SDBW	0.98	1.01	**0.49**	0.93	0.94	0.53	0.66	0.67	0.65
Percentage of agreement (%)	12.5%	-	75%	-	12.5%	-	-	-	-

In the following Table 2, we detailed out the membership of the three cluster solution using WDTW dissimilarity measure. Cluster 1 consist of 11 annual river discharge, Cluster 2 consist of 8 annual river discharge, and Cluster 3 consist of only 5 annual river discharge. In particular, Sayong, Bekok, and Segamat river are categorized in Cluster 1. These three rivers are located at the North of Johor. This implies that they have similar river discharge patterns and could be used together in the generalization process. Meanwhile, the other rivers (Johor, Kahang, and Muar) in the other clusters vary in their time series patterns.

Table 2: Membership of clusters using WDTW dissimilarity measure.

Rivers	**Cluster 1**	**Cluster 2**	**Cluster 3**
Johor river discharge 1981		X	
Johor river discharge 1982			X
Johor river discharge 1985			X
Johor river discharge 1987			X
Sayong river discharge 1981	**X**		
Sayong river discharge 1982		X	
Sayong river discharge 1985	**X**		
Sayong river discharge 1987	**X**		
Bekok river discharge 1981	**X**		
Bekok river discharge 1982	**X**		
Bekok river discharge 1985	**X**		
Bekok river discharge 1987	**X**		
Kahang river discharge 1981		X	
Kahang river discharge 1982		X	
Kahang river discharge 1985	**X**		
Kahang river discharge 1987		X	
Muar river discharge 1981		X	
Muar river discharge 1982		X	
Muar river discharge 1985			X
Muar river discharge 1987			X
Segamat river discharge 1981	**X**		
Segamat river discharge 1982		X	
Segamat river discharge 1985	**X**		
Segamat river discharge 1987	**X**		
Total	**11**	8	5

5. CONCLUSION

In this study, we assess three different types of Dynamic Time Warping (DTW) dissimilarity measures in regionalization of river discharge data. Weighted DTW (WDTW) was found to be superior compared to the other two types of DTW (conventional DTW and feature based DTW). Although the six rivers are located in the same state, they have different fluctuation pattern throughout the year. This may result to unaligned mapping of river discharge time series due to slightly higher/lower feature between sequences when similarities are sought. Hence, it could lead to singularities problem where a single point on one-time series maps onto a large subsection of another time series (Jeong et al., 2011).

Thus, weighted DTW helps to re-align the higher or lower feature between sequences

by penalizing the points according to the phase difference between river discharge time series. This caters the issue of small time shift between discharge peaks in the main river and in its tributaries as highlighted by Gertseema et al. (2016). However, the types of DTW dissimilarity measure compared in this study used raw river discharge data or only local feature of discharge data that represent relationship with two adjacent neighboring points. Further research will be conducted on studying a better feature to be included in DTW dissimilarity measure that is able to include the overall significant or global features that occur in the sequence of river discharge data.

6. REFERENCES

Agarwal, A.; Maheswaran, R.; Sehgal, V.; Khosa, R.; Sivakumar, B. & Bernhofer, C. (2016). Hydrologic regionalization using wavelet-based multiscale entropy method. *Journal of Hydrology,* 538: 22-32.

Corduas, M. (2011). Clustering streamflow time series for regional classification. *Journal of Hydrology,* 407: 73-80.

Desgraupes, B. (2013). Clustering indices. *University of Paris Ouest-Lab Modal'X,* 1: 34.

Dikbas, F.; Firat, M.; Koc, A. C. and Gungor, M. (2013). Defining homogeneous regions for streamflow processes in Turkey using a K-means clustering method. *Arabian Journal for Science and Engineering,* 38: 1313-1319.

Elesbon, A. A.; Silva, D. D. d.; Sediyama, G. C.; Guedes, H. A.; Ribeiro, C. A. and Ribeiro, C. B. d. M. (2015). Multivariate statistical analysis to support the minimum streamflow regionalization. *Engenharia Agricola, SciELO Brasil,* 35: 838-851.

Geertsema, T.; Teuling, A.; Uijlenhoet, R.; Torfs, P.; Hoitink, A. and Weerts, A. (2016). Simultaneous occurrence of discharge peaks in a large river and its lowland tributaries. In *River Flow - Proceedings of the International Conference on Fluvial Hydraulics,* St. Louis, 11-14 July 2017, pp. 1626-1630.

Gupta, A. and Chaturvedi, S. K. (2013). Real Time Prediction System of Discharge of the Rivers using Clustering Technique of Data Mining. *International Journal of Engineering Research and Development,* 9: 12-24.

Jeong, Y. S., Jeong, M. K., and Omitaomu, O. A. (2011). Weighted dynamic time warping for time series classification. *Pattern Recognition,* 44(9), 2231-2240.

Kahya, E.; Demirel, M. Cü. and Piechota, T. C. (2007). Spatial grouping of annual streamflow patterns in Turkey. *27th AGU Hydrology Days,* 169-176

Laaha, G. and Blöschl, Gü. (2006). A comparison of low flow regionalisation methods: catchment grouping. *Journal of Hydrology, Elsevier,*323: 193-214.

Mishra, S.; Saravanan, C.; Dwivedi, V. and Pathak, K. (2015). Discovering flood rising pattern in hydrological time series data mining during the pre monsoon period. *Indian Journal of Geo-Marine Sciences,* 44: 3.

Ouyang, R.; Ren, L.; Cheng, W. and Zhou, C. (2010). Similarity search and pattern discovery in hydrological time series data mining. *Hydrological Processes,* 24(9): 1198-1210.

Razavi, T. & Coulibaly, P. (2012). Streamflow

prediction in ungauged basins: review of regionalization methods. *Journal of Hydrologic Engineering, American Society of Civil Engineers,*18: 958-975.

Snelder, T. H.; Biggs, B. J. & Woods, R. A. (2005). Improved eco-hydrological classification of rivers. *River Research and Applications*, 21(6): 609-628.

Toth, E. (2013). Catchment classification based on characterisation of streamflow and precipitation time series. *Hydrology and Earth System Sciences*, 17: 1149-1159.

Zhao, Q. (2012). *Cluster Validity in Clustering Methods*. Dissertations in Forestry and Natural Sciences, University of Eastern Finland.

Multivariate T^2 Control Chart Based on James-Stein and Successive Difference Covariance Matrix Estimators for Intrusion Detection

Muhammad Ahsan[1a], Muhammad Mashuri[1b*], Heri Kuswanto[1c], Dedy Dwi Prastyo[1d] and Hidayatul Khusna[1e]

[1] Department of Statistics, Institut Teknologi Sepuluh Nopember, Surabaya 60111, INDONESIA. E-mail: ahsan4th@gmail.com[a] ; m_mashuri@statistika.its.ac.id[b] ; heri_k@statistika.its.ac.id[c] ; dedy-dp@statistika.its.ac.id[d] ; khusna16@mhs.statistika.its.ac.id[e]
* Corresponding Author: m_mashuri@statistika.its.ac.id
Received: 21st April 2019 Revised : 19th September 2019 Published: 30th September 2019
DOI : https://doi.org/10.22452/mjs.sp2019no2.3

ABSTRACT The intrusion detection is a process to monitor the events taking place in a computer system or network and analyse the monitoring results to find signs of intrusion. The multivariate control chart, which is often used in the intrusion detection system, is Hotelling's T^2. In this study, the Hotelling's T^2 chart performance for intrusion detection is improved using the successive difference covariance matrix to estimate the covariance matrix and James-Stein estimator to estimate the mean vector. The control limits of the proposed chart are calculated using kernel density estimation. The performance of the proposed method, using T^2 based on kernel density estimation control limit, outperforms the other control chart approaches in both training and testing dataset.

Keywords: T^2 control chart, James-Stein estimator, successive difference covariance matrix, kernel density estimation, intrusion detection.

1. INTRODUCTION

The network security systems are needed to be improved because of the rapid development of information exchange processes. In this case, a security system that can prevent and warn of the attacks on the network is needed. One security mechanism that can be used in preventing attacks is the intrusion detection system (IDS). The intrusion detection is a process to monitor the events, which are taking place in a computer system or network and to analyse the monitoring results to find the signs of anomalies in the network (Bace & Mell, 2001). The IDS is an essential tool in modern computing infrastructure to monitor and identify unwanted and suspicious network traffic, related to the unauthorised system access or poorly configured systems (Shenfield, Day, & Ayesh, 2018). The IDS has become an important component in computer network architecture because it has similar purposes to the burglar alarm that gives warning to every malicious event in the network.

The statistical process control (SPC) approach can be performed not only in the industrial field, but also be adopted in network intrusion detection. The SPC is a quality control method, which utilises the statistical methods to monitor and control the process. The control charts, which is one of the favourite tools used in SPC, is a graph employed to study the stability of the process over time. Based on the type of the quality characteristics monitored, the control charts are classified into attribute (Ahsan, Mashuri, & Khusna, 2017; Wibawati et al., 2016; Wibawati et al., 2018) and variable (Page, 1961; Roberts, 1959; Shewhart, 1924) control charts. Meanwhile, based on the number of quality characteristics, the control charts are classified into univariate and multivariate control charts. The univariate control chart only considers one characteristic. On the other

hand, the multivariate control chart is used to control production process with more than one characteristics. The recent development for the multivariate control chart includes Hotelling's T^2 control chart (Abu-Shawiesh, Kibria, & George, 2014; Ahsan et al., 2018a; Alkindi, Mashuri, & Prastyo, 2016), MCUSUM control chart (Arkat, Niaki, & Abbasi, 2007; Issam & Mohamed, 2008; Khusna et al., 2018a) and MEWMA control chart (Khusna et al., 2018b; Pirhooshyaran & Niaki, 2015).

The SPC can be used as a capable technique, which can guarantee the system security and stability in network monitoring and intrusion detection process (Bersimis, Sgora, & Psarakis, 2016). The superiority of applying this method to monitor the anomalies in the network does not require the knowledge of information from the prior attacks. This advantage makes the SPC based IDS to be easily executed in the online detection system (Catania & Garino, 2012). There are many studies on SPC, that have been implemented in IDS, for both the univariate and multivariate cases (Park, 2005). Based on the previous research, it can be known that the Hotelling's T^2 chart is the most commonly used control chart for the intrusion detection. The Hotelling's T^2, which uses the conventional mean and covariance matrix, is reactive to the outlier. As a result, the conventional method is not effective to use for multiple outliers' case, because of the masking effect (Alfaro & Ortega, 2009). The masking effect in the monitoring process happens as a result of the outlier, which cannot be detected by the control chart. To overcome the problem that arises, several robust methods have been proposed to reduce the effect of multiple outliers by substituting the existing estimators with the more robust ones. Moreover, the performance of Hotelling's T^2 control chart, in detecting the shift of mean vector, is increasing when the robust covariance matrix estimator is implemented (Williams et al., 2006). The successive difference covariance matrix (SDCM) is one of the robust covariance matrix estimators, that can be used in this instance. The Hotelling's T^2 control chart, based on SDCM, is powerful in detecting the shift of mean vector (Sullivan & Woodall, 1996; Vargas, 2003). Moreover, SDCM can also be exploited for auto-correlated processes, such as T^2 control chart based on the SDCM for multivariate process, using residuals of vector autoregressive (VAR) model (Wororomi et al., 2014).

Many researchers have proved the effectiveness of Hotelling's T^2 control chart based on SDCM. However, the exact distribution for this control chart has not been discovered. Sullivan and Woodall (1996) and Williams et al. (2006) suggested the approximate distribution for Hotelling's T^2 control chart based on SDCM. In addition, some studies have been conducted to improve Hotelling's T^2 chart control limit, by using nonparametric approaches, to overcome the limited research of Hotelling's T^2 based on SDCM distribution. Those studies have been conducted to improve the control limit of Hotelling's T^2 using kernel density estimation (KDE) technique (Chou, Mason, & Young, 1999; Phaladiganon et al., 2011; Phaladiganon et al., 2013). Using the same approach, Ahsan et al. (2018b) implemented the idea of KDE into the Hotelling's T^2 based on SDCM to monitor the anomalies in the network. The multivariate Hotelling's T^2 based on SDCM has a good performance to monitor the anomalies in the network when it is applied to the bootstrap resampling method to calculate the control limit (Ahsan, Mashuri, & Khusna, 2018).

The improvement not only can be done on covariance matrix as stated before, but also can be done on the mean vector. The shrinkage estimators, which have smaller mean squared errors than the traditional estimators, can be practised in this issue (Lehmann & Casella, 2006; Stein, 1956). The James-Stein estimator (James & Stein, 1961), which is the improved estimator of the mean vector, can be employed

to get the better result in estimating the mean vector of the Hotelling's T^2 control chart. Wang, Huwang and Yu (2015) found that the performance of the multivariate control chart with the James-Stein estimator is better than that of the existing control charts. Therefore, the aim of this study is to propose Hotelling's T^2 control chart based on the hybrid James-Stein and SDCM using the KDE approach. The James-Stein estimator is employed to estimate the mean vector, while the SDCM is adopted to estimate the robust covariance matrix. Theoretically, applying the proposed control chart into IDS can refine the performance in network monitoring system. To prove this statement, the performance of the proposed method is compared with the other methods.

This paper is organised as follows: section 2 describes T^2 control chart based on James-Stein and SDCM, while the control limit of Hotelling's T^2 control chart using KDE is explained in section 3; section 4 presents the dataset and methodology that were used in this research; the performance comparisons of the proposed IDS with the other control charts are presented in section 5; and finally, section 6 summarises the obtained results and presents a future research.

2. HOTELLING'S T^2 CONTROL CHART BASED ON JAMES-STEIN AND SDCM ESTIMATORS

In this section, the proposed Hotelling's T^2 control chart based on James-Stein and SDCM estimators is presented. The Hotelling's T^2 is one of multivariate control charts that can be used to monitor the mean of a multivariate process (Montgomery, 2009). Let $\mathbf{X} = \left[\mathbf{x}_1', \mathbf{x}_2', \ldots, \mathbf{x}_n'\right]'$, where $i = 1, 2, \ldots, n$ number of observations are identic and independent random vectors, which follow multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Using $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$ and $\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)'$, T^2 statistic can be calculated as follows (Hotelling, 1974):

$$T_i^2 = \left(\mathbf{x}_i - \bar{\mathbf{x}}\right)' \mathbf{S}^{-1} \left(\mathbf{x}_i - \bar{\mathbf{x}}\right) \tag{1}$$

By assuming the data to follow multivariate normal distribution, the control limit of Hotelling's T^2 can be obtained with the following equation:

$$CL = \frac{p(n+1)(n-1)}{n^2 - np} F_{(\alpha, p, n-p)}, \tag{2}$$

where n is number of observations, p is number of variables and α is false alarm rate. The process is concluded as in-control, if T^2 statistic in Equation (1) is lower than the control limit, CL formulated in Equation (2).

The SDCM is another alternative method to estimate the covariance matrix that was first introduced by Hawkins and Merriam (1974) as well as Holmes & Mergen (1993). This method is constructed by changing the estimated covariance matrix $\mathbf{S}$ with the SDCM. Under in-control condition, the SDCM, $\mathbf{S}_D$ is an unbiased estimator for covariance matrix $\boldsymbol{\Sigma}$ (Sullivan & Woodall, 1996). The T^2 based on SDCM can be calculated as follows:

$$T_{D,i}^2 = \left(\mathbf{x}_i - \bar{\mathbf{x}}\right)' \mathbf{S}_D^{-1} \left(\mathbf{x}_i - \bar{\mathbf{x}}\right), \tag{3}$$

where

$$\mathbf{S}_D = \frac{1}{(n-1)} \sum_{i=2}^{n} \left(\mathbf{x}_i - \mathbf{x}_{i-1}\right)\left(\mathbf{x}_i - \mathbf{x}_{i-1}\right)'. \tag{4}$$

There are some approaches in constructing the control limit of T_D^2 statistics that follow the multivariate normal distribution, such as the control limit based on Sullivan and Woodall (CL_{SW}) (Sullivan & Woodall, 1996), control limit based on Mason and Young (CL_{MY}) (Mason & Young, 2002) and control limit based on Chi-square distribution (CL_{χ^2}). Those control limits can be calculated using Equations (5) to Equation (7).

$$CL_{SW} = \frac{(n-1)^2}{n} BETA_{(1-\alpha),\frac{p}{2},\frac{(g-p-1)}{2}}, \tag{5}$$

$$CL_{MY} = \frac{(f-1)^2}{f} BETA_{(1-\alpha),\frac{p}{2},\frac{(g-p-1)}{2}} \tag{6}$$

$$CL_{\chi^2} = \chi^2_{(1-\alpha),u} \tag{7}$$

where $BETA_{(1-\alpha),p,g}$ $BETA_{(1-\alpha),p,g}$ is $[1-\alpha]$-th quantile of Beta distribution, p is number of quality characteristics and g is the shape parameter, $\chi^2_{(1-\alpha),u}$ is $[1-\alpha]$-th quantile of Chi-square distribution with u degree of freedom and $g = \frac{2(n-1)^2}{3n-4}$.

In this study, the James-Stein estimator is used to construct better Hotelling's T^2 control charts. The basic form of the James-Stein estimator is formulated as follows:

$$\bar{\mathbf{x}}_0^{JS} = \left(1 - \frac{p-2}{n\left(\bar{\mathbf{x}} - \mathbf{v}\right)^T \boldsymbol{\Sigma}^{-1} \left(\bar{\mathbf{x}} - \mathbf{v}\right)}\right)\left(\bar{\mathbf{x}} - \mathbf{v}\right) + \mathbf{v}, \tag{8}$$

where $\mathbf{v}$ is a fixed vector, that contains the target value, by which $\bar{\mathbf{x}}$ will be shrunk. According to Lehmann & Casella (2006), $\mathbf{v}$ can be determined as any p-dimensional vector. The James-Stein estimator can be improved to have smaller mean square error (MSE) than the standard James-Stein estimator, as in Equation (9), by Wang et al. (2015). The improved James-Stein estimator can be calculated as follows:

$$\bar{\mathbf{x}}^{JS} = \left(1 - \frac{p-2}{n\left(\bar{\mathbf{x}} - \mathbf{v}\right) \boldsymbol{\Sigma}^{-1} \left(\bar{\mathbf{x}} - \mathbf{v}\right)}\right)\left(\bar{\mathbf{x}} - \mathbf{v}\right) + \mathbf{v}, \tag{9}$$

where the notation $f(x)^+$ is defined as:

$$f(x)^+ = \begin{cases} f(x), \text{ if } f(x) > 0 \\ 0, \text{ otherwise.} \end{cases}$$

The improved James-Stein estimator, as in Equation (9) is utilised to construct the Hotelling's T^2 control chart based on James-Stein estimator, which is constructed by replacing the $\bar{\mathbf{x}}$ in Equation (1) with $\bar{\mathbf{x}}^{JS}$. The statistics of the proposed chart is formulated as follows:

$$T^2_{JSD,i} = \left(\mathbf{x}_i - \bar{\mathbf{x}}^{JS}\right)' \mathbf{S}_D^{-1} \left(\mathbf{x}_i - \bar{\mathbf{x}}^{JS}\right). \tag{10}$$

The improved James-Stein estimator $\bar{\mathbf{x}}^{JS}$ in Equation (9) can be improved further by replacing the covariance matrix $\boldsymbol{\Sigma}$ with SDCM, $\mathbf{S}_D$ in Equation (4), such that:

$$\bar{\mathbf{x}}_D^{JS} = \left(1 - \frac{p-2}{n\left(\bar{\mathbf{x}} - \mathbf{v}\right)^T \mathbf{S}_D^{-1}\left(\bar{\mathbf{x}} - \mathbf{v}\right)}\right)^+ \left(\bar{\mathbf{x}} - \mathbf{v}\right) + \mathbf{v} \tag{11}$$

The T^2 control chart based on James-Stein estimator in Equation (10) is enhanced using the new James Stein estimator $\bar{\mathbf{x}}_D^{JS}$ in Equation (11) as follows:

$$\tilde{T}^2_{JSD,i} = \left(\mathbf{x}_i - \bar{\mathbf{x}}_D^{JS}\right)' \mathbf{S}_D^{-1} \left(\mathbf{x}_i - \bar{\mathbf{x}}_D^{JS}\right). \tag{12}$$

Because the distribution of the proposed chart is still unknown, its control limits are calculated using KDE.

3. CONTROL LIMIT OF THE PROPOSED CHART BASED ON KDE

The kernel density estimation (KDE) method is a non-parametric method to estimate the probability density function of a random variable. This method was first introduced by Rosenblatt (1956) and Parzen (1962)—the so-called Rosenblatt-Parzen kernel density estimator. Chou, Mason and Young (2001) proposed KDE to estimate the distribution of T^2 statistic. Using the same procedure, $\tilde{T}^2_{JSD,i}$ obtained under in-control condition, can be estimated by KDE. The empirical distribution of $\tilde{T}^2_{JSD,i}$ statistic can be calculated using the following kernel function:

$$\hat{f}_h(\tilde{T}^2_{JSD}) = \frac{1}{n}\sum_{i=1}^{n} K\left[\frac{\left(\tilde{T}^2_{JSD} - \tilde{T}^2_{JSD,i}\right)}{h}\right], \tag{13}$$

where K and h define kernel function and smoothing parameter, respectively. The most used kernel is Gaussian kernel; therefore, in this paper, it is used in the analysis. Furthermore, the cumulative distribution function of $\hat{f}_h(\tilde{T}^2_{JSD})$ can be written as:

$$\hat{F}_h(t) = \int_0^t \hat{f}_h(\tilde{T}^2_{JSD}) d\tilde{T}^2_{JSD} \tag{14}$$

The control limit of $\tilde{T}^2_{JSD,i}$ based on KDE could be estimated by taking the percentile of kernel distribution. Hence, the control limit of T^2 based on KDE, equal to $\left[100(1-\alpha)\right]$-th percentile of $\tilde{T}^2_{JSD}$ distribution, can be calculated as follows:

$$CL_{KDE} = \hat{F}_h(t)^{-1}(1-\alpha). \tag{15}$$

The cumulative distribution function $\hat{F}_h(t)$ in Equation (14) can be calculated using tables of integral in the closed form distribution. However, the control limit might be inefficient to be calculated, if the distribution is not closed form. To overcome such problem, the kernel control limit is solved by employing trapezoidal rule to calculate the integral. The trapezoidal rule is one of the numerical integration methods to approximate definite value of the integral equation.

4. METHODOLOGY

4.1 IDS based T^2 James-Stein and SDCM Control Chart

In this section, the procedures of the proposed IDS based on T^2 James-Stein and SDCM control chart are described. The algorithm for the proposed IDS, using KDE control limit, can be divided into two phases, as follows:

Phase I: Building Normal Profile

The mean vector, the covariance matrix and KDE control limit are calculated from the normal profile of the dataset in this phase. The estimated values are then used in the next phase to monitor the new connection. The procedures for building the normal profile phase are defined as follows:

Step 1 Form matrix $\mathbf{x}_{normal}$, which is the normal connection data.

Step 2 Calculate vector $\bar{\mathbf{x}}_{normal}$, which is the mean of each column of the normal connection data $\mathbf{x}_{normal}$.

Step 3 Calculate the matrix of $\mathbf{S}_{DN}$ as in Equation (4), which is the estimated covariance matrix of the normal connection data $\mathbf{x}_{normal}$.

Step 4 Calculate vector $\bar{\mathbf{x}}^{JS}_{normal}$, which is the estimated mean vector from James-Stein estimator of normal connection data $\mathbf{x}_{normal}$ using Equation (11).

Step 5 Calculate statistics $\tilde{T}^2_{JSDN,i}$, as in Equation (12), using the normal connection data $\mathbf{x}_{normal}$.

Step 6 Determine α and calculate the KDE control limit CL_{KDE}, as in Equation (15).

Phase II: Detection

The estimated value of $\bar{\mathbf{x}}^{JS}_{normal}$, $\mathbf{S}_{DN}$ and CL_{KDE} from Phase I are then used in this phase. The following steps explain the procedures of the detection phase:

Step 1 Form matrix $\mathbf{x}_{test}$, which is the new connection data.

Step 2 Calculate statistics $T^2_{JSDT,i}$ from new connection data $\mathbf{x}_{test}$ as follows:

$$T^2_{JSDT,i} = \left(\mathbf{x}_i - \bar{\mathbf{x}}^{JS}_{normal}\right)' \mathbf{S}^{-1}_{DN} \left(\mathbf{x}_i - \bar{\mathbf{x}}^{JS}_{normal}\right),$$

where $\bar{\mathbf{x}}^{JS}_{normal}$ and $\mathbf{S}_{DN}$ are taken from the normal connection data in phase I.

Step 3 If $T^2_{JSDT,i} > CL_{KDE}$, then the connection is an intrusion and for $T^2_{JSDT,i} < CL_{KDE}$, the connection is normal.

4.2 Performance Evaluation

The NSL-KDD is the dataset that was used in this study. This dataset was first proposed by Tavallaee et al. (2009), as a solution for outdated KDD-99 dataset (Stolfo, 1999), which has been accessible for more than 15 years. The NSL-KDD dataset consists of 41 variables, with 34 quantitative and 7 qualitative variables. Nevertheless, this study only uses 32 quantitative variables, because the value of the rest of the quantitative variables is equal to zero.

In this study, the NSL-KDD dataset is monitored by using Hotelling's T^2 based on James-Stein and SDCM with KDE control limit. The performance of the proposed control chart will be compared with the conventional Hotelling's T^2 chart and T^2 based on SDCM chart, using various control limits, as stated in Ahsan et al. (2018b). Those various control limits include F distribution control limit according to Equation (2); Sullivan and Woodall control limit approach is based on Equation (5); Mason and Young control limit approach is according to Equation (6); and Chi-square control limit is based on Equation (7). In addition, the performance of each IDS approach is evaluated using a confusion matrix as shown in Table 1 (Ahsan, Mashuri, & Khusna, 2018). The occurrence of false positive (FP) in the network causes a false alarm that can disturb the system. Meanwhile, false negative (FN), taking place in the network, will allow an intrusion in the system.

Table 1: Intrusion detection confusion matrix.

	Prediction	
	Intrusion	Normal
Intrusion	True Positive (TP)	False Negative (FN)
Normal	False Positive (FP)	True Negative (TN)

The level of accuracy used is the hit rate that can be calculated as follows:

$$\text{Hit Rate} = \frac{TP + TN}{TP + TN + FP + FN}$$

Based on the type of error, the level of error in intrusion detection can be divided into two types, namely FP rate and FN rate. The FP and FN rate formulas are calculated as follows:

$$FP \text{ Rate} = \frac{FP}{TN + FP}$$

$$FN \text{ Rate} = \frac{FN}{TP + FN}.$$

5. RESULTS AND DISCUSSIONS

The performance comparisons of the proposed IDS, with the other approaches, are presented and discussed in this section. The performance of the proposed IDS (JS-$SDCM_{KDE}$) is compared with the conventional Hotelling's T^2 (T^2) and Hotelling's T^2 based on SDCM, using several control limit approaches, as has been stated in Ahsan et al. (2018b). Those control limits are F distribution control limit ($SDCM_F$), Sullivan and Woodall approach ($SDCM_{SW}$), Mason and Young approach ($SDCM_{MY}$) and chi-square control limit ($SDCM_{CH}$).

5.1 Results

The performance of the proposed chart, using KDE control limit, is compared with the other control chart methods, such as conventional Hotelling's T^2 control chart and Hotelling's T^2 control chart based on SDCM with various control limits. Table 2 reports the performance comparison between the proposed method and the other control charts for the training dataset. The values of hit rate from Table 4 are presented in a single graphic to simplify the interpretation process. According to Figure 1a, it can be observed that the proposed chart with KDE control limit has the highest hit rate compared to the other approaches. The proposed chart also has the better result in term of FN rate, which is depicted in Figure 2a. However, the proposed chart with KDE control limit has a similar value of FP rate, with Hotelling's T^2 control chart based on SDCM with Sullivan and Woodall ($SDCM_{SW}$) control limit. Thus, the proposed chart with KDE control limit has better accuracy to detect anomaly in the network than the other control charts' approach, for the training dataset based on hit rate and FN rate criteria.

Table 2: Performance of various IDS for training data.

IDS	Hit Rate	FN	FP	FN Rate	FP Rate
T^2	0.91330	5428	5494	0.0806	0.0937
$SDCM_F$	0.91338	5417	5495	0.0804	0.0937
$SDCM_{SW}$	0.91705	4280	6170	0.0636	0.1052
$SDCM_{MY}$	0.91331	5429	5492	0.0806	0.0937
$SDCM_{CH}$	0.91332	5427	5492	0.0806	0.0937
JS-$SDCM_{KDE}$	0.91751	4115	6277	0.0611	0.1071

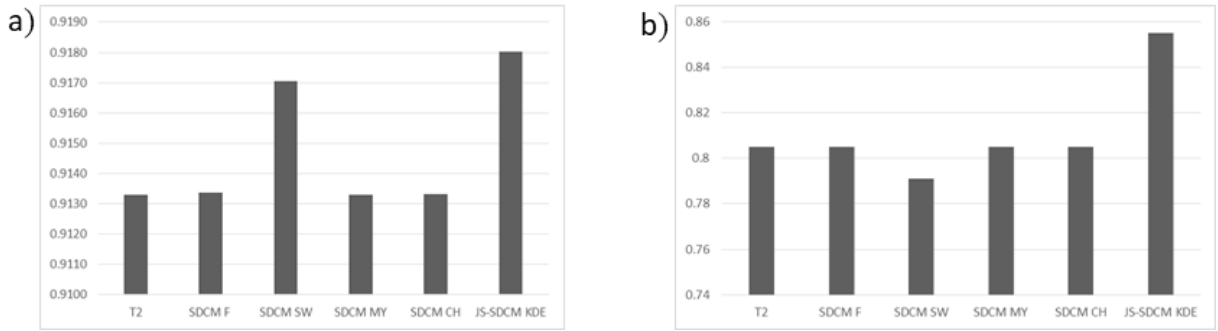


Figure 1: Hit rate comparison of various control charts for: a) training dataset b) testing dataset.

Table 3 exhibits the performance comparison between the proposed chart and the other control chart methods for the testing dataset. Similar to the previous result, the performance of the proposed chart with KDE control limit is better than the other approaches based on the hit rate criteria, as shown in Figure 1b. Although the proposed chart with KDE control limit has the higher FN rate compared to the other approaches, its performance outperforms the other methods based on the FP rate criteria, as depicted in Figure 2b. It is noteworthy that the difference between the FN rate of the proposed chart with KDE and the other approaches is not significant.

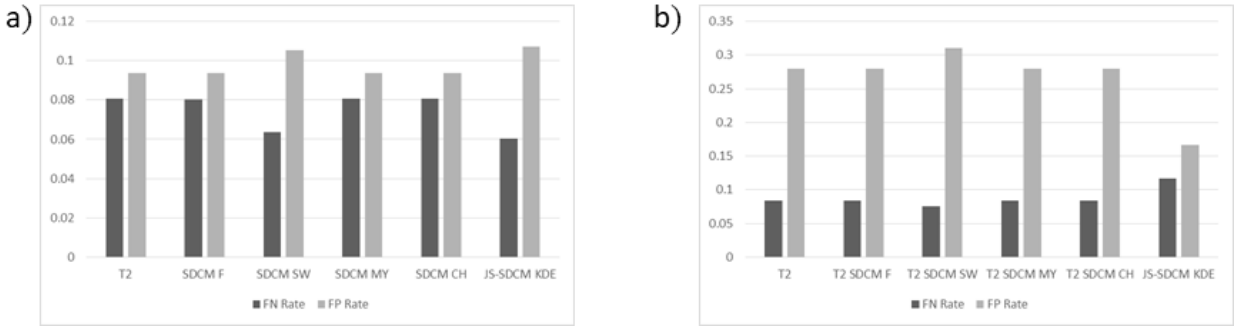


Figure 2: FN and FP rate comparison of various control charts for: a) training dataset b) testing dataset.

Table 3: Performance of various IDS for testing data.

IDS	Hit Rate	FN	FP	FN Rate	FP Rate
T^2	0.8049	814	3584	0.0838	0.2793
$SDCM_F$	0.8049	814	3585	0.0838	0.2794
$SDCM_{SW}$	0.7911	731	3978	0.0753	0.3100
$SDCM_{MY}$	0.8049	814	3584	0.0838	0.2793
$SDCM_{CH}$	0.8049	814	3584	0.0838	0.2793
$JS\text{-}SDCM_{KDE}$	0.8554	1127	2134	0.1160	0.1663

5.2 Discussions

Based on the performance evaluation from the previous section, it could be shown that the T^2 based on James-Stein and SDCM with KDE control limit, has the highest hit rate for the training and testing dataset. For the training dataset, the misdetection happens because of the high value of FP rate produced by the proposed chart. The high value of FP rate from training dataset happens as a result of the control chart oversensitivity to detect an attack, while the attack is not actually happened in the network. However, the proposed IDS is superior to the other approaches in terms of the low value of FN rate criteria. Consequently, IDS constructed by this approach will successfully detect an actual attack in the network, but it has a higher false alarm.

For the testing dataset, similar to the training dataset, the misdetection happens because of the high value of FP rate. These findings indicate that the IDS constructed by using T^2, based on James-Stein and SDCM, will produce more false alarm. However, compared to the other approaches, the proposed chart has the lowest FP rate. This

proposed method also will not let the attack occur in the network without any warning shown by the low level of FN rate. Thus, by considering the performance from the training and testing dataset, IDS constructed by the proposed chart with KDE, outperforms the other approaches in terms of detecting the actual attacks in the network.

6. CONCLUSION

In this paper, the multivariate Hotelling's T^2 control chart is improved by James-Stein and SDCM estimators, while its control limit is calculated using KDE method before it is applied into IDS. The performance of proposed IDS is evaluated and compared with the other control limits by hit rate, FN rate and FP rate criteria. Furthermore, the performance of proposed IDS is also compared with some existing control charts. The performance evaluation results reveal that the proposed IDS using T^2, based on James-Stein and SDCM with KDE control limit, outperforms the other approaches in both in training and testing dataset. The proposed method is effective to be applied in IDS, based on its ability to detect anomaly in the network, confirmed by the low value of FN rate. The multiclass detection for each type of attack with an incremental algorithm can be considered as future research.

7. ACKNOWLEDGEMENT

This work was supported by Research, Technology, and Higher Education Ministry, the Republic of Indonesia through PMDSU scheme under Grant 128/SP2H//PTNBH/DRPM/2018.

8. REFERENCES

Abu-Shawiesh, M. O. A., Kibria, G., and George, F. (2014). A robust bivariate control chart alternative to the Hotelling's T2 control chart. *Quality and Reliability Engineering International*, 30(1): 25–35.

Ahsan, M., Mashuri, M., and Khusna, H. (2017). Evaluation of Laney p' chart performance. *International Journal of Applied Engineering Research*, 12(24): 14208–14217.

Ahsan, M., Mashuri, M., and Khusna, H. (2018). Intrusion detection system using bootstrap resampling approach of T2 control chart based on successive difference covariance matrix. *Journal of Theoretical and Applied Information Technology*, 96(8): 2128–2138.

Ahsan, M., Mashuri, M., Kuswanto, H., Prastyo, D. D., and Khusna, H. (2018a). Multivariate control chart based on PCA mix for variable and attribute quality characteristics. *Production & Manufacturing Research*, 6(1): 364–384. https://doi.org/10.1080/21693277.2018.1517055

Ahsan, M., Mashuri, M., Kuswanto, H., Prastyo, D. D., and Khusna, H. (2018b). T2 control chart based on successive difference covariance matrix for intrusion detection system. In *Journal of Physics: Conference Series*, 1028: 12220.

Alfaro, J. L., and Ortega, J. F. (2009). A comparison of robust alternatives to Hotelling's T2 control chart. *Journal of Applied Statistics*, 36(12): 1385–1396. https://doi.org/10.1080/02664760902810813

Alkindi, Mashuri, M., and Prastyo, D. D. (2016). T2 hotelling fuzzy and W2 control chart with application to wheat flour production process. In *AIP Conference Proceedings*, 1746. https://doi.org/10.1063/1.4953977

Arkat, J., Niaki, S. T. A., and Abbasi, B. (2007). Artificial neural networks in applying MCUSUM residuals charts for AR(1) processes. *Applied Mathematics and Computation*, 189(2): 1889–1901.

https://doi.org/10.1016/j.amc.2006.12.08

Bace, R., and Mell, P. (2001). NIST special publication on intrusion detection systems. *Special Publication (NIST SP) - 800-31*. https://doi.org/10.1016/S1361-3723(01)00614-5

Bersimis, S., Sgora, A., and Psarakis, S. (2016). The application of multivariate statistical process monitoring in non-industrial processes. *Quality Technology and Quantitative Management*, 3703(September): 1–24. https://doi.org/10.1080/16843703.2016.1226711

Catania, C. A., and Garino, C. G. (2012). Automatic network intrusion detection: Current techniques and open issues. *Computers & Electrical Engineering*, 38(5): 1062–1072. https://doi.org/10.1016/j.compeleceng.2012.05.013

Chou, Y.-M., Mason, R., and Young, J. (2001). The control chart for individual observations from a multivariate non-normal distribution. *Communications in Statistics: Theory & Methods*, 30(8-9): 1937-1949. https://doi.org/10.1081/STA-100105706

Chou, Y., Mason, R. L., and Young, J. C. (1999). Power comparisons for a hotelling's t 2 STATISTIC. *Communications in Statistics - Simulation and Computation*, 28(4): 1031–1050. https://doi.org/10.1080/03610919908813591

Hawkins, D. M., and Merriam, D. F. (1974). Zonation of multivariate sequences of digitized geologic data. *Journal of the International Association for Mathematical Geology*, 6(3): 263–269. https://doi.org/10.1007/BF02082892

Holmes, D. S., and Mergen, A. E. (1993). Improving the performance of the T2 control chart. *Quality Engineering*, 5(4): 619–625. https://doi.org/10.1080/08982119308919004

Hotelling, H. (1974). Multivariate quality control. In *Techniques of Statistical Analysis*. New York: McGraw-Hill.

Issam, B. K., and Mohamed, L. (2008). Support vector regression based residual MCUSUM control chart for autocorrelated process. *Applied Mathematics and Computation*, 201(1–2): 565–574. https://doi.org/10.1016/j.amc.2007.12.059

James, W., and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1: 361–379.

Khusna, H., Mashuri, M., Ahsan, M., Suhartono, S., and Prastyo, D. D. (2018a). Bootstrap based maximum multivariate CUSUM control chart. *Quality Technology & Quantitative Management*. https://doi.org/10.1080/16843703.2018.1535765

Khusna, H., Mashuri, M., Suhartono, Prastyo, D. D., and Ahsan, M. (2018b). Multioutput least square SVR based multivariate EWMA control chart. In *Journal of Physics: Conference Series*, 1028(1): 12221. Retrieved from http://stacks.iop.org/1742-6596/1028/i=1/a=012221

Lehmann, E. L., and Casella, G. (2006). *Theory of Point Estimation*. Springer Science & Business Media.

Mason, R. L., and Young, J. C. (2002). *Multivariate Statistical Process Control with Industrial Applications*. Society for Industrial and Applied Mathematics. Retrieved from http://epubs.siam.org/doi/book/10.1137/1.9780898718461

Montgomery, D. (2009). *Introduction to Statistical Quality Control*. New York: John Wiley & Sons Inc. https://doi.org/10.1002/1521-3773(20010316)40:6<9823::AID-ANIE9823>3.3.CO;2-C

Murray Rosenblatt. (1956). Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics*, 27: 832–837. https://doi.org/10.1214/aoms/1177728190

Page, E. S. (1961). Cumulative Sum Charts. *Technometrics*, 3(1): 1–9. https://doi.org/10.1080/00401706.1961.10489922

Park, Y. (2005). *A Statistical Process Control Approach for Network Intrusion Detection*. Georgia Insitute of Technology.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3): 1065–1076. https://doi.org/10.1214/aoms/1177704472

Phaladiganon, P., Kim, S. B., Chen, V. C. P., Baek, J.-G., and Park, S.-K. (2011). Bootstrap-based T2 multivariate control charts. *Communications in Statistics - Simulation and Computation*, 40(5): 645–662. https://doi.org/10.1080/03610918.2010.549989

Phaladiganon, P., Kim, S. B., Chen, V. C. P., and Jiang, W. (2013). Principal component analysis-based control charts for multivariate nonnormal distributions. *Expert Systems with Applications*, 40(8): 3044–3054. https://doi.org/10.1016/j.eswa.2012.12.020

Pirhooshyaran, M., and Niaki, S. T. A. (2015). A double-max MEWMA scheme for simultaneous monitoring and fault isolation of multivariate multistage auto-correlated processes based on novel reduced-dimension statistics. *Journal of Process Control*, 29: 11–22. https://doi.org/10.1016/j.jprocont.2015.03.008

Roberts, S. W. (1959). Control Chart Tests Based on Geometric Moving Averages. *Technometrics*, 1(3): 239–250. https://doi.org/10.1080/00401706.1959.10489860

Shenfield, A., Day, D., and Ayesh, A. (2018). Intelligent intrusion detection systems using artificial neural networks. *ICT Express*, 4(2): 95-99.

Shewhart, W. A. (1924). Some applications of statistical methods to the analysis of physical and engineering data. *Bell Labs Technical Journal*, 3(1): 43–87.

Stein, C. (1956). *Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution*. United States: Stanford University Stanford.

Stolfo, S. J. (1999). KDD cup 1999 dataset. *UCI KDD Repository. Http://Kdd.Ics.Uci.Edu*, 0.

Sullivan, J. H., and Woodall, W. H. (1996). A comparison of multivariate control charts for individual observations. *Journal of Quality Technology*, 28(4): 398–408.

Tavallaee, M., Bagheri, E., Lu, W., and Ghorbani, A. A. (2009). A detailed analysis of the KDD CUP 99 data set. In *IEEE Symposium on Computational Intelligence for Security and Defense Applications, CISDA 2009*. https://doi.org/10.1109/CISDA.2009.5356528

Vargas, N. J. (2003). Robust estimation in multivariate control charts for individual observations. *Journal of Quality Technology*, 35(4): 367–376.

Wang, H., Huwang, L., and Yu, J. H. (2015). Multivariate control charts based on the James–Stein estimator. *European Journal of Operational Research*, 246(1): 119–127.

Wibawati, Mashuri, M., Purhadi, and Irhamah. (2016). Fuzzy multinomial control chart and its application. In *AIP Conference Proceedings*, 1718(1): 110004. https://doi.org/10.1063/1.4943351

Wibawati, Mashuri, M., Purhadi, Irhamah, and Ahsan, M. (2018). Performance fuzzy multinomial control chart. In *Journal of Physics: Conference Series*, 1028(1): 12120. Retrieved from http://stacks.iop.org/1742-6596/1028/i=1/a=012120

Williams, J. D., Woodall, W. H., Birch, J. B., and Sullivan, J. O. E. H. (2006). On the distribution of Hotelling's T2 statistic based on the successive differences covariance matrix estimator. *Journal of Quality Technology*, 38: 217–229.

Wororomi, J. K., Mashuri, M., Irhamah, and Arifin, A. Z. (2014). On monitoring shift in the mean processes with vector autoregressive residual control charts of individual observation. *Applied Mathematical Sciences*, 8: 3491–3499. https://doi.org/10.12988/ams.2014.44298

Malaysian Journal of Science 38 (Special Issue 2): 36 - 45 (2019)

**THE INTERNATIONAL SEMINAR ON MATHEMATICS IN INDUSTRY (ISMI)
AND THE INTERNATIONAL CONFERENCE ON THEORETICAL AND APPLIED STATISTICS (ICTAS)
ISMI-ICTAS18 [4-6 SEPTEMBER 2018]**

Classification Boosting in Imbalanced Data

Sinta Septi Pangastuti[1a], Kartika Fithriasari[1b]*, Nur Iriawan[1c], and Wahyuni Suryaningtyas[2d]

[1] Statistics Department, Institut Teknologi Sepuluh Nopember Jl. Arif Rahman Hakim, Surabaya 60111 Indonesia and Statistics Department, Faculty of Mathematics and Natural Sciences, Padjadjaran University Jl. Raya Bandung-Sumedang Km. 21, Jatinangor 45363, INDONESIA. E-mail: sintaseptip@gmail.com[a] ; sinta.septi@unpad.ac.id[a] ; kartika_f@statistika.its.ac.id[b] ; nur_i@statistika.its.ac.id[c] ,
[2] Doctoral Candidate at Statistics Department, Institut Teknologi Sepuluh Nopember Jl. Arif Rahman Hakim, Surabaya 60111 Indonesia and Mathematics Education Program Study, Faculty of Teacher Training and Education, Muhammadiyah University of Surabaya Jl. Sutorejo No. 59, Surabaya 60113, INDONESIA. E-mail: wahyuni.pendmat@fkip.um-surabaya.ac.id[d]
* Corresponding Author: kartika_f@statistika.its.ac.id
Received: 21st April 2019 Revised : 6th August 2019 Published: 30th September 2019
DOI : https://doi.org/10.22452/mjs.sp2019no2.4

ABSTRACT Most existing classification approaches assumed underlying training data set to be evenly distributed. However, in the imbalanced classification, the training data set of one majority class could far surpass those of the minority class. This becomes a problem because it's usually produces biased classifiers that have a higher predictive accuracy over the majority class, but poorer predictive accuracy over minority class. One popular method recently used to rectify this is the SMOTE (Synthetic Minority Over-Sampling Technique) which combines algorithms at data level. Therefore, this paper presents a novel approach for learning and imbalanced data sets, based on a combination of the SMOTE algorithm and the boosting procedure by focusing on a two-class problem. The Bidikmisi data set is imbalanced, because the distribution of majority class examples is 15 times the number of minority class examples. All models have been evaluated using stratified 5-fold cross-validation, and the performance criteria (such as Recall, F-Value and G-Mean) are examined. The results show that the SMOTE-Boosting algorithms have a better classification performance than the AdaBoost.M2 method, as the g-mean value increases 4-fold after the SMOTE method is used. We can say that SMOTE-Boosting algorithm is quite successful when taking advantage of boosting algorithms with SMOTE. When boosting affects the accuracy of the random forest by focusing on all data classes, the SMOTE algorithm alters the performance values of the random forest only in minority classes.

Keywords: Boosting, G-mean, Imbalanced classification, SMOTE

1. INTRODUCTION

Data mining is a method often used to determine the hidden relationship between variables (Han et al, 2006). There are mixtures of prevalent information mining undertakes inside the instructive information mining, e.g., grouping, bunching, anomaly location, affiliation standard, expectation, and so forth. In recent years, many applications of data mining are used to handle cases with large data or big data. In this study, the application of data mining was conducted by using the local data of Bidikmisi scholarships. Suryaningtyas et al. (2018) showed that the Bidikmisi grantee status Binary type (0 and 1) and then performed the classification analysis with Bayesian Bernoulli Mixture regression and Bayesian binary logistic regression. Cahyani et al. (2018) used Regression and Neural Network analysis to classify the acceptance of Bidikmisi scholarships and showed that the classifier is not good enough for imbalance data case. As the data collection and storage technology has made it possible to organize a huge amount of data, the class imbalance issue

has received worthy consideration in the classification problems. Imbalance class for a binary classification problem occurs when one class (majority class) highly exceeds the number of another class (minority class).

The classification technique aims to find a decision function that accurately predicts the class of testing data derived from the same distribution function as the data for training. The large class is called the majority class (negative class) while the smaller class is called the minority class (positive class). Under such conditions, most classifiers are biased towards the major class since the classification engine will be inclined to predict the major class and ignore the minor class (Japkowicz & Stephan, 2002). Imran et al. (2016) used three re-sampling techniques: SMOTE (Synthetic Minority Oversampling Technique), ROS (Random over Sampling), and RUS (Random under Sampling) with three different classifiers and trained them with the rebalanced data. There are several approaches to learning methods used to overcome the problem of imbalanced data; one of them is the ensemble method. The ensemble method, in principle, combines a set of classifiers that are trained in order to create a better classifier model that makes the ensemble classifier more accurate than the original classifier in performing a classification (Han et al, 2012). According to Schapire in Leaes et al. (2017), one approach that can be used to improve the performance of classification on imbalanced data is boosting. Boosting can improve performance by exploiting classification errors, which involves using the base classifier. We used the SMOTE-Boosting algorithm (Chawla et al, 2003) which provides good performance. SMOTE-Boosting modifies the Adaptive Boosting algorithm (Freud & Schapire, 1995) by employing the SMOTE algorithm in each iteration. The purpose of SMOTE is to increase the probability of selecting hard-to-class samples, derived from the minor class, into the training data in each iteration so as to make the base classifier to focus more on minor class observations. This will certainly improve the accuracy of classification of minority classes.

This paper is structured as follows. In section two, a brief explanation of ensemble methods and its algorithm are given. The performance evaluation for imbalanced data is discussed in section three, followed by results and discussion in section four. The conclusion is presented in section five.

2. ENSEMBLE METHODS

In this section, the ensemble method that is used for imbalanced data set for the Bidikmisi scholarship is presented. The ensemble classification method combines a collection of classifications to create a single composite model to provide better accuracy. Experimental studies such as Bühlamann and Hothorn (2007) showed that predictions from composite models provide better results compared to single-model predictions. This ensemble method has become popular in the last few decades, with some of the most popular combining techniques include Boosting (Freud & Schapire, 1996), Cost sensitive boosting (CSB) by Ting (2000) and Cost-Sensitive boosting algorithm (Sun et al, 2005). The most recent algorithm is SMOTEBoost, which successfully utilizes the benefits of both boosting and the SMOTE algorithm for an imbalanced dataset (Chawla et al, 2003), will be explained in the next section.

2.1 Adaptive Boosting M2 Algorithm

Boosting, which was introduced by Schapire in Leaes et al. (2017), is one of the ensemble methods used to improve the performance of a learning algorithm by combining a collection of weak classifiers to form a strong end classifier. Adaptive boosting is one of the boosting algorithm introduced by Freud & Schapire (1995). In this paper, we use well-known modifications that have been employed in imbalanced domains: AdaBoost.M2 (Schapire & Singer, 1999).

Note that this algorithm cannot deal with the imbalanced problem directly; it has to be combined with another technique as its base classifier. The goodness of a base classifier is measured based on the pseudo-loss, as seen in Algorithm 1.

Algorithm 1 AdaBoost.M2

Input: $(\mathbf{x}_i, y_i),...,(\mathbf{x}_m, y_m)$ where $\mathbf{x}_i \in \mathbf{X}$ and $y_i \in \mathbf{Y} = \{1,...,k\}$

Given: $B = \{(i, y) : i \in \{1,...,m\}, y \neq y_i\}$

1. Train weak learner or base classifier use D_t distribution
   $D_1(i, y) = 1/|B|$ for $(i, y) \in B$
2. for $t = 1, 2, ..., T$
3. Compute weak hypothesis $h_t : X \times \mathrm{Y} \to [0,1]$ with pseudo-loss, Equation (1).

$$\varepsilon_t = \frac{1}{2} \sum_{(i,y) \in B} D_t(i, y)\left(1 - h_t(x_i, y_i) + h_t(x_i, y_i)\right) \quad (1)$$

   If $\varepsilon_t > 0.5$, then the learning process stops.

4. Set $\beta_t = \dfrac{\varepsilon_t}{(1 - \varepsilon_t)}$ (2)
5. Update weight value

$$D_{t+1}(i, y) = \frac{D_t(i, y)}{Z_t} \times \beta_t^{\left(\frac{1}{2}\right)\left(1 + h_t(x_i, y_i) - h_t(x_i, y_i)\right)} \quad (3)$$

where Z_t is a normalization contact that makes $\sum_{i=1}^{m} D_{t+1}(i, y) = 1$

Output: Boosted classifier

$$H(x) = \arg\max_{y \in Y} \sum_{t=1}^{T} \left(\log \frac{1}{\beta_t} \right) h_t(x, y) \quad (4)$$

As mentioned earlier, this algorithm needs base classifier as its weal learner. In this paper, both AdaBoost.M2 and SMOTE-Boosting used the random forest as a base classifier.

2.2 SMOTE-Boosting Algorithm

The algorithm SMOTE-Boosting was proposed by Chawla et al. (2003). SMOTE-Boosting combines the SMOTE algorithm and standard boosting procedures, utilizing SMOTE to improve minority class predictions and utilizing boosting to avoid sacrificing accuracy over the entire data set. SMOTE is one method of dealing with imbalanced data proposed by Chawla et al. (2002). The basic idea of SMOTE is to increase the number of samples in the minor class to equal the major class, by generating synthetic data based on the nearest neighbor, k-nearest neighbor, where the nearest neighbor is selected based on the Euclidean distance between the two data. Suppose the given data with p variable is $\mathbf{x}^T = [\mathrm{x}_1, \mathrm{x}_2, ..., \mathrm{x}_p]$ and $\mathbf{z}^T = [\mathrm{z}_1, \mathrm{z}_2, ..., \mathrm{z}_p]$, then the Euclidean distance d(x,z) is defined as follows:

$$d(\mathbf{x}, \mathbf{z}) = \sqrt{(\mathrm{x}_1 - \mathrm{z}_1)^2 + (\mathrm{x}_2 - \mathrm{z}_2)^2 + ... + (\mathrm{x}_p - \mathrm{z}_p)^2}. \quad (5)$$

Synthetic data generation is done by using the following equation:

$$\mathbf{x}_{syn} = \mathbf{x}_i + (\mathbf{x}_{knn} - \mathbf{x}_i)\gamma. \quad (6)$$

Synthetic samples are generated in the following way: Take the difference between variable vector (sample) under consideration $(\mathbf{x}_i)$ and its nearest neighbor $(\mathbf{x}_{knn})$. Multiply this difference by a random number between 0 and 1 (γ), and add it to the variable vector under consideration $(\mathbf{x}_i)$. This cause the selection of a random point along the line segment between two specific variables, and so this approach effectively forces the decision region of the minority class to become more general.

SMOTE method is also used to handle continuous and nominal mixed data sets, and it is known as SMOTE-NC. Based on research by Chawla et al (2002) using data Adult from USI repository, the dataset has 6 continuous variables and 8 nominal variables. The SMOTE and SMOTE-NC algorithm are used to approve the dataset. In this study, 10 nominal variables and one continuous variable were used. Based on the study using data from the UCI repository, data showed that SMOTEBoost is able to achieve higher F-values than AdaCost, due to SMOTE's ability to improve coverage of the minority class.

The SMOTE-NC (*Synthetic Minority Oversampling Technique – Nominal Continous*) algorithm is described as follows:

1. Median Computation: Compute the median standard deviation of continuous variable for minority classes. If the nominal variables differ between a sample and its potential nearest neighbor, then this median is included in the Euclidean distance computation. The median is used to override differences in nominal variables by an amount that is related to the typical difference in continuous variable values.
2. Nearest Neighbor (*k*) Computation: Compute the Euclidean distance between the variable vector in which the nearest neighbor's *k* are being identified (minority class sample), and other variable vectors (minority class sample) using the continuous variable space. For each different nominal variables between the considered variable vector and its potential nearest-neighbor, including the median standard deviation calculated earlier, in the calculation of Euclidean distance.
3. Creating Synthetic Samples: The continuous variable of the new synthetic data for minority classes are created using the same SMOTE approach as described earlier. The nominal variables are given the value occurring in the majority of the *k*-nearest neighbors.

The purpose of merging the SMOTE and AdaBoost.M2 algorithm is to increase the True Positive (TP) rate. SMOTE-Boosting successfully combines AdaBoost.M2 and SMOTE, while AdaBoost.M2 tries to improve the accuracy of the classifier by focusing on the "difficult to classify" observations that come from both classes, SMOTE tries to improve the performance of the classifier only on observations in minority classes. Therefore, in several consecutive boosting iterations, SMOTE-Boosting was able to make wider decision areas for minority classes than the standard boosting method. SMOTE-Boosting initially used the iteration procedure from AdaBoost.M2, by Freud & Schapire (1995). In the AdaBoost.M2 iteration procedure, the classification result of the classifier component is first brought into the form of probability [0,1] for later use in calculating pseudo loss. SMOTE-Boosting introduces synthetic instances just before Step 3 of AdaBoost.M2 (Algorithm 1), as seen in Algorithm 2.

Algorithm 2 SMOTE-Boosting

Input: $(\mathbf{x}_i, y_i),...,(\mathbf{x}_m, y_m)$ where $\mathbf{x}_i \in \mathbf{X}$ and $y_i \in \mathbf{Y} = \{1,...,k\}$

Given: $B = \{(i, y) : i \in \{1,...,m\}, y \neq y_i\}$

1. Train weak learner or base classifier use D_t distribution
   $D_1(i, y) = 1/|B|$ for $(i, y) \in B$
2. for $t = 1, 2,..., T$
3. Modify distribution D_t by creating N synthetic examples from minority class using the SMOTE algorithm
4. Compute a weak hypothesis $h_t : X \times Y \to [0,1]$ with pseudo-loss, Equation (7).
   $$\varepsilon_t = \sum_{(i,y) \in B} D_t(i, y)\left(1 - h_t(x_i, y_i) + h_t(x_i, y)\right) \quad (7)$$
   If $\varepsilon_t > 0.5$, then the learning process stops.
5. Set
   $$\beta_t = \frac{\varepsilon_t}{(1 - \varepsilon_t)} \quad (8)$$
6. Update weight value:
   $$D_{t+1}(i, y) = \frac{D_t(i, y)}{Z_t} \times \beta_t^{\left(\frac{1}{2}\right)(1 + h_t(x_i, y_i) - h_t(x_i, y_i))} \quad (9)$$
   where Z_t is a normalization contact that makes $\sum_{i=1}^{m} D_{t+1}(i, y) = 1$

Output: Boosted classifier
$$H(x) = \arg\max_{y \in Y} \sum_{t=1}^{T} \left(\log \frac{1}{\beta_t}\right) h_t(x, y) \quad (10)$$

2.3 Tools and Techniques Used

In this paper, Data Mining techniques are used for the prediction of Bidikmisi data set. The techniques are classification using Random Forest algorithm, combined with SMOTE and boosting algorithm. For implementation of all these classification tasks we have used ebmc package in R, and Minitab.

3. PERFORMANCE EVALUATION FOR IMBALANCED DATASET

Actual data and predictive predicted data from the classification model is presented using a confusion matrix, which contains information about the actual data class represented in the matrix row and the prediction data class in the column (shown in Table 1). Traditionally, the accuracy rate has been the most commonly used empirical measure. However, in the case of the imbalanced class where the majority class is 90% of the total population, the classification results will achieve high accuracy because it only sees the majority class. It is clear that in the case of imbalanced, the accuracy of classification is not sufficient as a standard criterion measure. According to Joshi et al in Bayisa et al (2018), the value of metrics, such as recall, precision and F-value have been used to understand the performance of learning algorithms in minority classes. Based on Table 1, recall and F-value can be calculated as follows:

$$\text{Recall}=\frac{\text{TP}}{(\text{TP}+\text{FN})}. \tag{11}$$

$$\text{Precision}=\frac{\text{TP}}{(\text{TP}+\text{FN})}. \tag{12}$$

$$\text{F-Value}=\frac{2(\text{Recall}\times\text{Precision})}{\text{Recall}+\text{Precision}}. \tag{13}$$

Table 1: Confusion matrix.

	Predictive Positive Class	Predictive Negative Class
Real Positive Class	True Positive (TP)	False Negative (FN)
Real Negative Class	False Positive (FP)	True Negative (TN)

The recall value provides information on how minority classes are identified, but maybe at the expense of precision through misclassification of the majority class. The commonly used sensitivity and specificity are taken to measure the performance of each algorithm on the imbalanced data sets. They are defined as:

$$\text{Specificity}=\frac{\text{TN}}{(\text{TN}+\text{FP})}. \tag{14}$$

$$\text{Sensitivity}=\frac{\text{TP}}{(\text{TP}+\text{FN})}. \tag{15}$$

$$\text{G-Mean}=\sqrt{\text{Sensitivity}\times\text{Specificity}}. \tag{16}$$

To perform an overall performance evaluation, geometric mean (G-mean) can be used. G-mean is the geometric average of Recall (Sensitivity) and Specificity. According to Li et al. (2008), several studies use G-mean measurements to evaluate the performance of algorithms on imbalanced data problems because this measure combines sensitivity and specificity by taking the geometric mean. If all positive classes are unpredictable, then the G-mean will be zero, so expect a classification algorithm to reach a high G-mean value.

4. RESULTS AND DISCUSSION

In this paper, the experiments were performed on the Bidikmisi data set summarized in Table 2 and Table 3. The dataset contains 10829 records and 11 variables of 2017 East Java Bidikmisi Scholarship applicants. The data attributes can be classified as demographic attributes (such as occupation, education, housing ownership, area of residential land, area of residential building, *etc*).

Table 2: Summary of the Bidikmisi data set.

Data set	Number of majority class instances	Number of minority class instances	Number of classes
Bidikmisi	10143	686	2

*Source: Kemenristekdikti of the Republic of Indonesia

Table 3: Summary of the nominal and continuous variables

Number of nominal variables	10
Number of continuous variable	1

*Source: Kemenristekdikti of the Republic of Indonesia

The Bidikmisi data set has a mixture of both nominal and continuous variables (see Table 3), so SMOTE-NC is used to obtain synthetic data. Based on Figure 1, the distribution of class categories showed that there is an imbalanced data. The number of majority class examples is 15 times the number of minority class examples, so we increase the SMOTE parameter N value to 1500. Then we obtained the new train and test data sets by stratified 5-fold cross-validation.

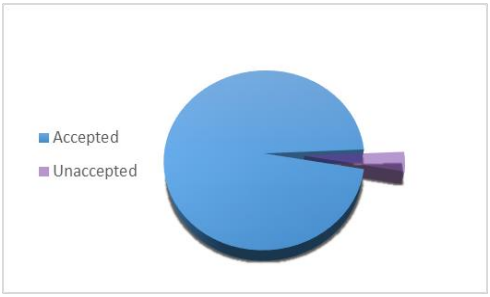
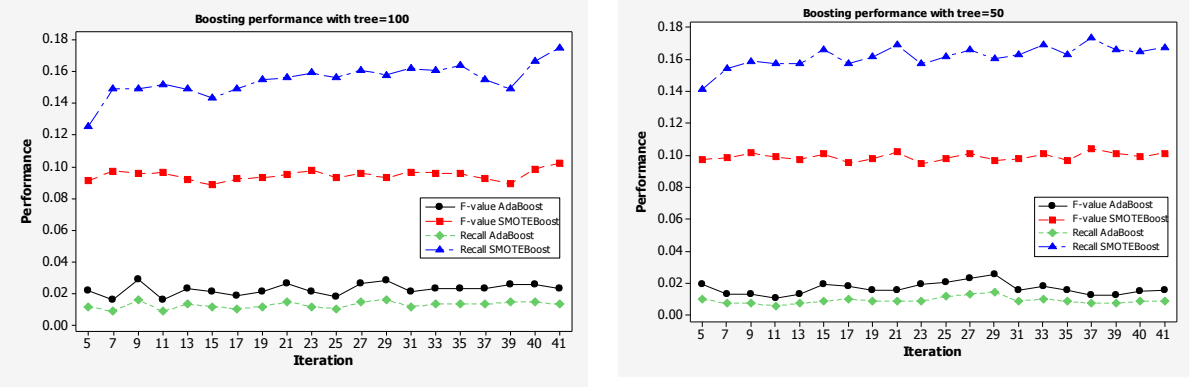


Figure 1: Data characteristics based on student status.

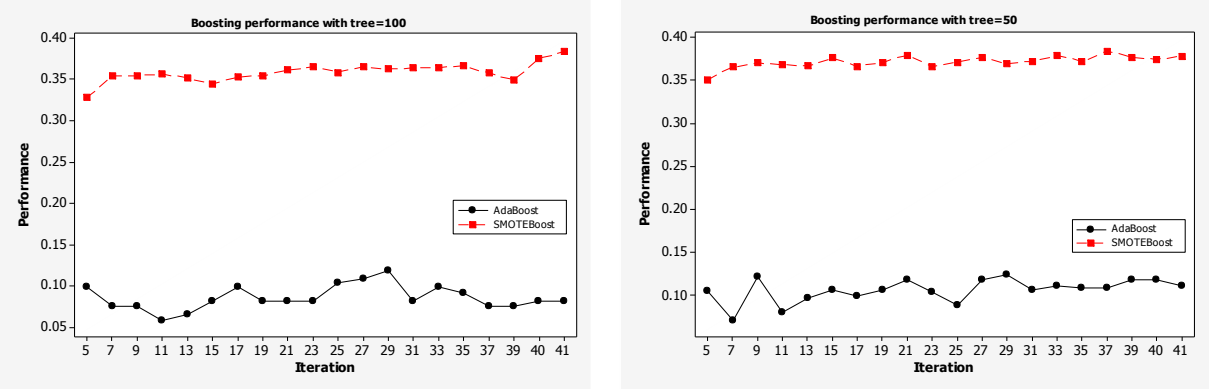
As shown in Figure 1, the majority class is accepted as student status with 97% and only 3% is minority class, i.e., student's status is not accepted. Such a condition would cause the classifier to be biased against the majority class, meaning that the classification engine would tend to predict the majority class and ignore the minor class. Therefore, the ensemble classification method is expected to be able to handle the problem created. The experimental result for the Bidikmisi data set is presented in Figures 2 to 3.

Performance evaluation for classification Bidikmisi data set in this study used several criteria to support decision making. Figures 2 and 3 show the results of methods and performance in experiments with different iteration presented as follows.



(a) (b)

Figure 2: Recall and F-value of the Bidikmisi data set when boosting random forest: (a) Tree=100 and (b) Tree=50 are applied.



(a) (b)

Figure 3: G-mean of the Bidikmisi data set when boosting random forest: (a) Tree=100 and (b) Tree=50 are applied.

The analysis of Figure 2 and 3 shows the behavior of the ensemble method with respect to the different number of iterations. It is apparent that the SMOTE-Boosting achieved higher *f*-value than AdaBoost.M2. We also compared the boosting algorithm with different random forest trees.

The recall value shown in Figure 2 shows the behavior of the ensemble method with respect to the number of iterations. It can be seen that the value of the recall tends to be stable or show an ascending pattern. The recall value corresponds to a true positive and a false negative (recall = TP / (TP + FN)). Then false negative will have a greater value than true positive, due to the increase of predicted minor class to the majority class. As for the value of *f*-value, it is the geometric average of the precision and recall value. SMOTE embedded within the boosting procedure additionally improved the recall achieved by the boosting procedure, thus increasing the F-value. SMOTE, as a part of SMOTE-Boosting, allows the learners to broaden the minority class scope, while the boosting, on the other hand, aims at reducing the number of false positives.

G-mean is the geometric mean value of the recall value for each class. Since the value of the recall tends to experience an ascending pattern during the initial phase of the iteration, the G-mean value tends to have the same pattern. It can be seen in Figure 3 that the G-mean of SMOTE-Boosting method is higher than AdaBoost.M2, where the highest G-mean value is 38.45%, earned when the number of iterations was 41 with tree=100.

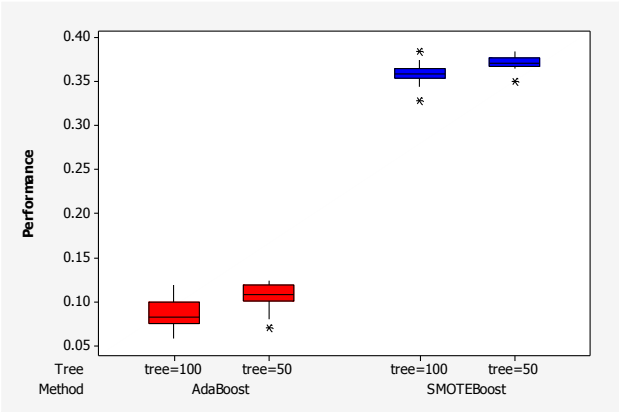


Figure 4: Clustered boxplot boosting performance of G-mean.

Figure 4 presents a boxplot of the G-mean values generated in each model. The G-mean value generated using the SMOTE-Boosting, shown in the figure by the blue box, was higher compared to AdaBoost.M2. The G-mean value is generated using the AdaBoost.M2, shown by the red box in the figure. The median for both methods seems similar, but the variations of G-mean generated by the SMOTE-Boosting algorithm tend to be smaller than the AdaBoost.M2, ranges from 35% to 37%. It is also shown for different trees where each method gives an almost the same results. This indicates that the G-mean performance generated by SMOTE-Boosting is more stable than AdaBoost.M2.

5. CONCLUSION

All models have been evaluated using stratified 5-fold cross-validation, and the performance criteria for each method are examined. The algorithm used is SMOTE-Boosting based on SMOTE algorithm integration in standard boosting procedures. The results of the imbalanced classes show that the SMOTE-Boosting ensemble algorithms show better classification performance than the AdaBoost.M2 method. It can be said that SMOTE-Boosting methods are quite successful when taking advantage of boosting algorithms with SMOTE. While boosting affects the accuracy of the random forest by focusing on all data classes, the SMOTE algorithm alters the performance values of the random forest only in minority classes.

6. ACKNOWLEDGEMENT

This research was supported by DPRM-DIKTI under scheme PUPT, project No. 1049/PKS/ITS/2018. The author thanks the Kementerian Riset, Teknologi, dan Pendidikan Tinggi for funding and to anonymous references for their useful suggestions.

7. REFERENCES

Bayisa, L.F., Liu, X., Garpebring, A. & Yu, J. (2018). Statistical learning in computed tomography image estimation. *The International Journal of Medical Physics Research and Practice,* 45(12): 5450-5460.

Bühlaman, P., & Hothorn, T. (2007). Boosting Algorithms: Regularization, Prediction and Model Fitting. *Statistical Science,* 22(4): 477-505.

Cahyani, N., Fithriasari, K., Irhamah & Iriawan, N. (2018). On the comparison of deep learning neural network and binary logistic regression for classifying the acceptance status of

bidikmisi scholarship applicants in east java. *MATEMATIKA: Malaysian Journal of Industrial and Applied Mathematics*, 34 (Special Issue): 83-90.

Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16: 321-357.

Chawla, N.V., Lazarevic, A., Hall, L.O. & Bowyer, K.W. (2003). SMOTEBoost: Improving the prediction of the minority class in boosting. *Proceedings of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases*, Cavtat-Dubrovnik, Croatia, 22-26 September, 107–119, Springer.

Freund, Y. & Schapire, R. E. (1995). A decision-theoretic generalization of on-line learning and an application to boosting. *Proceedings of the 2nd European Conference on Computational Learning Theory*, Barcelona, Spain, 13-15 March, 23-37, Springer.

Freund, Y. & Schapire, R. (1996). Experiments with a new boosting algorithm. *Proceedings of the 13th International Conference on Machine Learning*, 325-332.

Han, J., Kamber, M. & Pei, J. (2006). *Data Mining Concepts and Techniques 2nd Edition*. USA: Kaufman Publisher.

Han, J., Kamber, M. & Pei, J. (2012). *Data Mining Concepts and Techniques 3rd Edition*. USA: Kaufman Publisher.

Imran, M., Afroze, M., Sanampudi, SK., & Qyser, AAM. (2016). Data mining of imbalanced dataset in educational data using Weka tool. *International Journal of Engineering Science and Computing*, 6(6): 7666-7669.

Japkowicz, N. & Stephen, S. (2002). The Class Imbalance Problem: A Systematic Study. *Intelligent Data Analysis*, 6(5), 203-231.

Leaes, A., Fernandes, P., Lopes, L. & Assunção, J. (2017). Classifying with AdaBoost.M1: The training error threshold myth. *Proceedings of the Thirtieth International Florida Artificial Intelligence Research Society Conference*, Marco Island, Florida, 22-24 May.

Li, X., Wang, L. & Sung, E. (2008). AdaBoost with SVM-based component classifiers. Engineering Applications of Artificial Intelligence, 21(5) 785-795. From University of Wollongong Publications: http://ro.uow.edu.au/eispapers/602.

Schapire, R. & Singer, Y. (1999). Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37: 297–336.

Sun, Y., Wong, A.K.C. & Wang, Y. (2005). Parameter inference of cost-sensitive boosting algorithm. *Proceedings of the 4th International Conference Machine Learning and Data Mining in Pattern Recognition*, Leipzig, German, 9-11 July, pp. 21-30, Springer.

Suryaningtyas, W., Iriawan, N., Fithriasari, K., Ulama, BSS., Susanto, I., & Pravitasari, AA. (2018). On the Bernoulli Mixture Model for Bidikmisi Scholarship Classification with Bayesian MCMC. *Journal of Physics: Conference Series*, 1090: 1-8.

Ting, K. (2000). A Comparative Study of Cost-Sensitive Boosting Algorithms. *Proceedings of 17th International Conference on Machine Learning*, Stanford, CA, pp. 983-990.

Malaysian Journal of Science 38 (Special Issue 2): 46 - 54 (2019)

**THE INTERNATIONAL SEMINAR ON MATHEMATICS IN INDUSTRY (ISMI)
AND THE INTERNATIONAL CONFERENCE ON THEORETICAL AND APPLIED STATISTICS (ICTAS)
ISMI-ICTAS18 [4-6 SEPTEMBER 2018]**

An Outlier Detection Method for Circular Linear Functional Relationship Model Using *Covratio* Statistics

Nurkhairany Amyra Mokhtar[1a], Yong Zulina Zubairi[2b]*, Abdul Ghapor Hussin[1c] & Nor Hafizah Moslim[3d]

[1] Faculty of Defence Sciences and Technology, National Defence University of Malaysia, Kem Sungai Besi, 57000 Kuala Lumpur, MALAYSIA. E-mail: khairany.amyra@gmail.com[a] ; abdulghapor@gmail.com[c]
[2] Centre for Foundation Studies in Science,
University of Malaya, 50603 Kuala Lumpur, MALAYSIA. E-mail: yzulina@um.edu.my[b]
[3] Institute of Graduate Studies, University of Malaya, 50603 Kuala Lumpur, MALAYSIA. Email: moslimnorhafizah@gmail.com[d]
* Corresponding Author: yzulina@um.edu.my
Received: 21st April 2019 Revised : 6th August 2019 Published: 30th September 2019
DOI : https://doi.org/10.22452/mjs.sp2019no2.5

ABSTRACT The existence of outlier may affect data aberrantly. However, outlier detection problem has been frequently discussed for linear data but limited on circular data. Thus, this paper discusses an outlier detection method on circular data. We focus on circular data with equal error concentration parameters where the data is studied using linear functional relationship model. In this paper, the data and the error terms are distributed with von Mises distribution. We modify the *covratio* statistics in which the correction factor is applied to the estimation of concentration parameter. We develop the cut-off equation based on the 5% upper percentile of the *covratio* statistics and the power of performance of outlier detection is examined by a Monte Carlo simulation study. The simulation result shows that the power of performance increases when the concentration and the level of contamination increase. The applicability of the proposed method is illustrated by using the wind direction data collected from the Holderness Coastline at the Humberside Coast in North Sea, United Kingdom.

Keywords: circular data, linear functional relationship model, outlier detection, *covratio* statistics

1. INTRODUCTION

Circular observation is specified by the angle from the initial direction to the point on the circle which corresponds to the observation, after an initial direction and an orientation of the circle have been chosen. It is measured in degrees or radians (Mardia & Jupp, 2000). Some examples of circular data include the orientation of fracture planes, the wind direction and the direction of the ocean current (Fisher, 1993).

The most useful distribution on the circle is said to be the Von Mises distribution (Mardia & Jupp, 2000). The probability density function of the distribution is

$$g(\theta;\mu,\kappa)=\frac{1}{2\pi I_0(\kappa)}e^{\kappa\cos(\theta-\mu)} \quad (1)$$

where $I_0(\kappa)$ is the modified Bessel function of the first kind and order zero, which can be defined by:

$$I_0(\kappa)=\frac{1}{2\pi}\int_0^{2\pi}e^{\kappa\cos\theta}d\theta \quad (2)$$

for $0\le\theta<2\pi$ and $\kappa>0$ where μ is the mean direction and κ is the concentration parameter.

Functional relationship model for

circular variables may be used to compare circular data where both of the variables are subjected to errors (Hassan et al., 2010). In this model, both X and Y variables are subject to random errors δ_i and ε_i, respectively. The errors are distributed with Von Mises distribution of $\delta_i \sim VM(0,\kappa)$ and $\varepsilon_i \sim VM(0,\kappa)$. Caires and Wyatt (2003) developed a model with the desired symmetry of the functional relationship model. The model is named the linear functional relationship model and is written in the form

$$Y = \alpha + X \left(\mathrm{mod}\, 2\pi\right) \tag{3}$$

where α is the rotation parameter.

Outlier detection has important applications such as fraud detection and robust analysis, among others (Mokhtar et al., 2018). *Covratio* statistics have been developed for linear data (Ghapor et al., 2014). However, for circular data, the method is somewhat limited, especially for linear functional relationship model. Therefore, this paper discusses the *covratio* statistics used for outlier detection in a linear functional relationship model for circular variables.

Section 2 describes the method used to obtain the cut-off equation to detect the outlier. Section 3 shows the results of the power of performance of the cut-off equation obtained in detecting the outlier. Section 4 illustrates the applicability of the proposed method and the conclusion is deduced in Section 5.

2. METHODS

2.1 Maximum Likelihood Estimation of Parameters of Von Mises Distribution

The log-likelihood function of the Von Mises distribution, based on observations x and y is given by

$$\log L(\alpha,\kappa,\nu,X;x,y) = -2n\log 2\pi - n\log I_0(\kappa) - n\log I_0(\nu) + \kappa\sum_{i=1}^{n}\cos\left(x_i - X_i\right) + \nu\sum_{1}^{n}\cos\left(y_i - \alpha - X_i\right) \tag{4}$$

where $I_0(\kappa)$ is the modified Bessel function of the first kind and order zero and κ is the concentration parameter.

In this case, the ratio of concentration parameter $\lambda = \frac{\nu}{\kappa}$ is fixed as 1 thus $\kappa = \nu$. Therefore, the log-likelihood function of the Von Mises distribution then becomes

$$\log L(\alpha,\kappa,X;x,y) = -2n\log 2\pi - 2n\log I_0(\kappa) + \kappa\sum_{i=1}^{n}\cos\left(x_i - X_i\right) + \kappa\sum_{i=1}^{n}\cos\left(y_i - \alpha - X_i\right) \tag{5}$$

The estimate of the rotation parameter for this LFRM is then

$$\hat{\alpha} = \begin{cases} \tan^{-1}\left\{\dfrac{S}{C}\right\} & \text{when } S>0, C>0 \\ \tan^{-1}\left\{\dfrac{S}{C}\right\}+\pi & \text{when } C<0 \\ \tan^{-1}\left\{\dfrac{S}{C}\right\}+2\pi & \text{when } S<0, C>0 \end{cases} \tag{6}$$

where $S=\sum_{i=1}^{n}\sin\left(y_i-\hat{X}_i\right)$ and $C=\sum_{i=1}^{n}\cos\left(y_i-\hat{X}_i\right)$. Meanwhile,

$$\hat{\kappa}=A_1^{-1}\left(\frac{1}{n}\left\{\sum_{i=1}^{n}\cos\left(x_i-\hat{X}_i\right)+\sum_{i=1}^{n}\cos\left(y_i-\alpha-\hat{X}_i\right)\right\}\right) \tag{7}$$

$A_1(x)$ is a function that behaves rather like $\left(\frac{2}{\pi}\right)\tan^{-1}x$ and so $A_1^{-1}(x)$ is like $\tan\left(\frac{\pi x}{2}\right)$ (Dobson, 1978). In this model, we use the approximation for the estimation of the concentration parameter κ that has been given by Fisher (1993) for the case of equal error concentration as a piecewise function of :

$$A_1^{-1}(w)=\begin{cases} 2w+w^3+\dfrac{5}{6}w^3 & \text{when } w<0.53 \\ -0.4+1.39w+\dfrac{0.43}{(1-w)} & \text{when } 0.53\le w<0.85 \\ \dfrac{1}{w^3-4w^2+3w} & \text{when } w\ge 0.85 \end{cases} \tag{8}$$

where

$$w=\frac{1}{n}\left\{\sum_{i=1}^{n}\cos\left(x_i-\hat{X}_i\right)+\sum_{i=1}^{n}\cos\left(y_i-\hat{\alpha}-\hat{X}_i\right)\right\}.$$

Caires and Wyatt (2003) noted that, in the circular case, the estimation of a concentration parameter (whose inverse is equivalent to the variance for linear data) needs to be corrected by dividing it by 2 as it suggests a consistent estimator of κ (Caires & Wyatt, 2003). It has been proposed that $\tilde{\kappa}=\frac{\hat{\kappa}}{2}$ gives a better approximation to the value of κ.

X_i is solved iteratively by some initial guess. Suppose $\hat{X}_{i0}$ is an initial estimate of $\hat{X}_i$. Then $x_i-\hat{X}_i=x_i-\hat{X}_{i0}+\hat{X}_{i0}-\hat{X}_i=\left(x_i-\hat{X}_{i0}\right)+\Delta_i$ where $\Delta_i=\hat{X}_{i0}-\hat{X}_i$. Thus, $y_i-\hat{\alpha}-\hat{X}_i=(y_i-\hat{\alpha}-\hat{X}_{i0})+\Delta_i$. Hence, the partial derivative equation above becomes: $\sin\left(x_i-\hat{X}_{i0}+\Delta_i\right)+\sin(y_i-\hat{\alpha}-\hat{X}_{i0}+\Delta_i)=0$

when Δ_i is small, then $\cos\Delta_i \approx 1$ and $\sin\Delta_i \approx \Delta_i$. Therefore, the variable X may be estimated by:

$$\hat{X}_{i1} \approx \hat{X}_{10} + \frac{\sin\left(x_i - \hat{X}_{i0}\right) + \sin\left(y_i - \hat{\alpha} - \hat{X}_{i0}\right)}{\cos\left(x_i - \hat{X}_{i0}\right) + \cos\left(y_i - \hat{\alpha} - \hat{X}_{i0}\right)} \tag{9}$$

2.2 Covratio Statistics to Detect Outlier in the LFRM for Circular Variables Assuming Equal Error Concentration Parameters

The outlier is an observation that lies outside the pattern (Belsley et al., 1980). Various outlier detections have been employed. The *covratio* statistics have long been used to identify outlier in linear regression models via a row deletion approach. Some studies on detecting outliers have been discussed by Abuzaid et al. (2011), Ibrahim et al. (2013), Ghapor et al. (2014) and Rambli et al. (2016). The *covratio* statistic is used to measure the effect of removing the observation based on the determinantal ratio given by:

$$COVRATIO_{(-i)} = \frac{|COV|}{|COV_{(-i)}|} \tag{10}$$

where $|COV|$ is the determinant of covariance matrix for the full set and $|COV_{(-i)}|$ is the determinant of covariance matrix for the reduced data set by excluding the *i-th* row.

In 2015, Mokhtar et al. (2015) studied the LFRM with the assumption of equal error concentration parameter and the covariance matrix of the parameter in the model is as given.

$$\operatorname{cov}\begin{bmatrix}\hat{\alpha}\\ \tilde{\kappa}\end{bmatrix} = \begin{bmatrix}\dfrac{1}{2nA_1'(\tilde{\kappa})} & 0\\ 0 & \dfrac{2}{n\tilde{\kappa}A_1'(\tilde{\kappa})}\end{bmatrix} \tag{11}$$

where $\operatorname{var}(\hat{\alpha}) = \dfrac{1}{2nA_1'(\tilde{\kappa})}$ and $\operatorname{var}(\tilde{\kappa}) = \dfrac{2}{n\tilde{\kappa}A_1'(\tilde{\kappa})}$, and $A_1'(\tilde{\kappa}) = 1 - \dfrac{A_1(\tilde{\kappa})}{\tilde{\kappa}} - [A_1(\tilde{\kappa})]^2$ is the first derivative $\dfrac{dA_1(\tilde{\kappa})}{d\tilde{\kappa}}$ for the function $A_1(\tilde{\kappa})$ which is the ratio of first and zeroth order Bessel functions.

Here, we propose that the *covratio* statistics method in which the correction factor is applied to estimate the concentration parameter. Therefore, the determinant of the covariance matrix for this model becomes

$$|COV| = \frac{1}{n^2\tilde{\kappa}[A_1'(\tilde{\kappa})]^2} \tag{12}$$

The observation with $\left|COVRATIO_{(-i)}-1\right|$ that exceeds the cut-off points will be identified as an outlier. The steps in determining the cut-off point are discussed in the next section.

2.3 Simulation Study to Determine the Cut-Off Equation using Covratio Statistics

In detecting the outlier, a cut-off point is needed as an indicator to examine the power of performance of the proposed *covratio* statistics. Therefore, a Monte Carlo simulation study is performed with different values of sample size and error concentration parameter. In this part, the number of simulation $s=500$.

Without the loss of generality, the variable X is generated from the Von Mises distribution of $VM(2,3)$ and the value of $\alpha=\frac{\pi}{4}=0.7854$. The values of the concentration parameters of the error term used in this study are $\kappa=3,5,10$ and 15. For each value of κ, the sample size $n=20,\ 30,\ 50,\ 70,\ 100,130$ and 150 are considered for the simulation with the assumption of $\kappa=\nu$.

The process was repeated for 500 simulations and the 5% upper percentiles of the maximum $\left|COVRATIO_{(-i)}-1\right|$ were obtained. The values of the upper 5% percentiles were then used to construct a cut-off equation in identifying the outlier for the LFRM for equal error concentration parameters. Table 1 shows the values of 5% upper percentile to consider the 95% confidence level.

Table 1: The values of the upper 5% percentile of $\left|COVRATIO_{(-i)}-1\right|$

n	$\kappa=3$	$\kappa=5$	$\kappa=10$	$\kappa=15$
20	0.482631	0.456601	0.427134	0.440119
30	0.345093	0.341747	0.327810	0.325791
50	0.247297	0.245086	0.213416	0.217016
70	0.194456	0.204701	0.162582	0.163800
100	0.134861	0.201583	0.126035	0.119783
130	0.113857	0.161380	0.104580	0.097920
150	0.101447	0.147530	0.091868	0.083569

The arithmetic mean of the values for the respected n are calculated and by finding the best fit of using least square method and thus a power function is obtained. Thus, the fit of the cut-off equation is obtained with $y=3.7586n^{-0.71}$. Figure 1 shows the plot of the best fit and the power fit has a good fit of R^2 almost equal to 1.

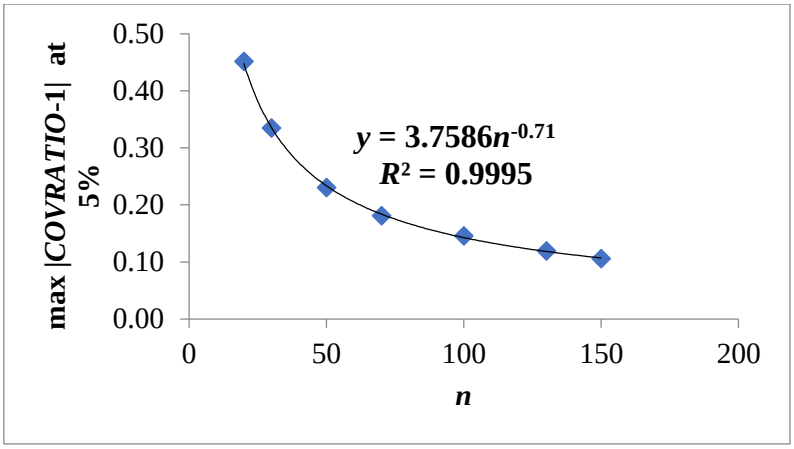


Figure 1: The power series graph to determine the cut-off equation at 5% significance level for equal error concentration

2.4 *Simulation Study to Assess the Power of Performance of Covratio Statistics in Outlier Detection*

To assess the feasibility of the proposed method, the power of performance is examined through a Monte Carlo simulation study with a number of simulations, $s = 500$. The steps below were carried out to identify the power of performance of the proposed *covratio* statistics in detecting the outlier.

Step 1: The values of X variable were generated from the distribution $VM(2,3)$ with the size of $n = 30, 70, 100$ and 130 and $\kappa = 10, 15$ and 20. An observation X_d^* is then contaminated with some levels of contamination ω where the level of the contamination was $0 \le \omega \le 1$ using the formula $X_d^* = X_i + \omega\pi \pmod{2\pi}$.

Step 2: Values of Y were found according to the generated X values. The variables X and Y were considered with the generated random error terms of $\delta_i \sim VM(0,\kappa)$ and $\varepsilon_i \sim VM(0,v)$, respectively where $\kappa = v$.

Step 3: The variables were fitted to the LFRM and the concentration parameter was estimated with the correction factor mentioned in Section 2.1.

Step 4: Calculate the value of $COVRATIO_{(-i)}$ and find $\left|COVRATIO_{(-i)} - 1\right|$ for all i. If the value of $\left|COVRATIO_{(-i)} - 1\right|$ exceeded $y = 3.7586n^{-0.71}$, then the ith observation is marked as a contaminated observation.

Step 5: The percentage of correct detection of the outlier was calculated as the power of performance.

3. RESULTS

Table 2 shows the simulation results in assessing the power of performance of $\left|COVRATIO_{(-i)} - 1\right|$ in LFRM for circular variables assuming equal error concentration parameters. The percentage of correct detection of the outlier is calculated when the maximum value of $\left|COVRATIO_{(-i)} - 1\right|$ exceeds $y = 3.7586n^{-0.71}$.

Table 2: The power of performance of $\left|COVRATIO_{(-i)} - 1\right|$

n	ω	κ =10	κ =15	κ =20
30	**0.2**	4.60	6.40	9.20
	0.4	21.60	41.80	60.60
	0.6	66.20	85.00	99.00
	0.8	94.20	99.80	100.00
	1.0	99.20	100.00	100.00
70	**0.2**	1.80	3.40	6.00
	0.4	14.20	31.20	54.60
	0.6	64.00	86.40	97.40
	0.8	93.60	99.60	100.00
	1	99.40	100.00	100.00
100	**0.2**	2.80	1.60	4.20
	0.4	10.80	33.00	45.80
	0.6	64.00	86.40	97.40
	0.8	92.20	99.20	100.00
	1	98.80	100.00	100.00
130	**0.2**	1.00	1.20	2.60
	0.4	10.20	25.00	46.40
	0.6	49.20	84.20	96.80
	0.8	88.00	99.40	100.00
	1	98.60	100.00	100.00

As an example, Figure 2 shows the plot of the power of performance for $n = 130$.

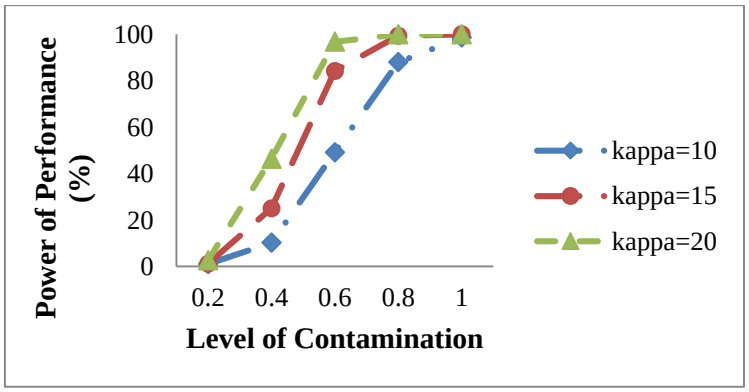


Figure 2: Power of performance for *covratio* statistics in detecting outlier for $n = 130$

The simulation study shows that the power of performance increases as the concentration parameter and the level of contamination increase. The highest concentration parameter with the highest level of contamination gives the highest power of performance where the value is 100%. Therefore, the *covratio* statistic method used in this study, in which the covariance matrix is derived with some correction factor for the maximum likelihood estimation, is sufficient to detect outlier in circular data with error terms.

4. APPLICATION ON REAL WIND DIRECTION DATA

The proposed method is applied to a real wind direction data set with sample size of $n = 129$ obtained from Holderness Coastline, Humberside Coast, United Kingdom (Hussin et al., 2004). The data is collected over the period of 22.7 days. Variable *x* is the data of the wind direction measured by high frequency radar system and developed by UK Rutherford and Appleton Laboratories, using the pulse radar operating at frequency of 24.2-27 MHz. Meanwhile for the variable *y*, the data was measured using an anchored wave buoy. Previous researchers of circular statistics such as Mokhtar et al. (2018) and Hussin et al. (2013) have used this data to illustrate the presence of outliers. It is worthwhile to note that the values of error concentration parameters of the variables *x* and *y* are assumed as equal. They have established that observations 38 and 111 as outliers of the data set.

Figure 3 shows the scatterplot of the values of $\left|COVRATIO_{(-i)} - 1\right|$ for all 129 observations in the data.

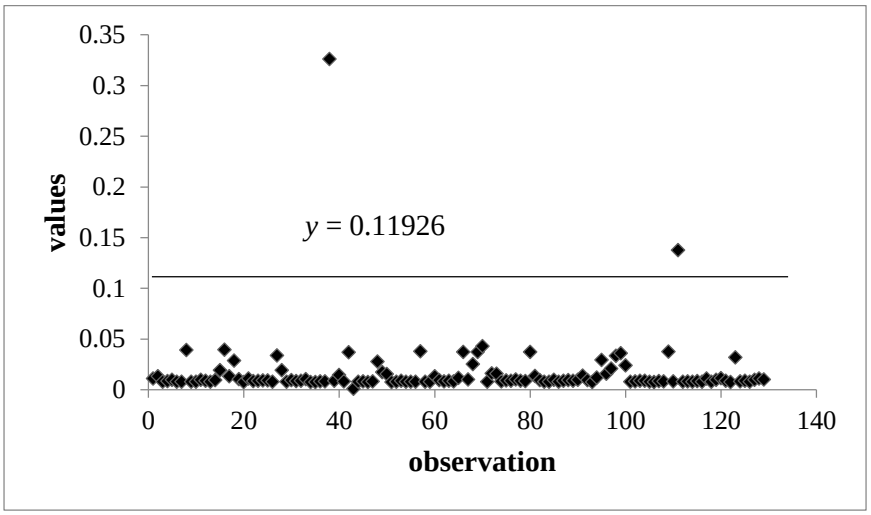


Figure 3: Values of $\left|COVRATIO_{(-i)} - 1\right|$ for all 129 observations of the real wind direction data.

Based on the scatterplot, it can be seen that the values of $\left|COVRATIO_{(-i)} - 1\right|$ for observations 38 and 111 exceeded the cut-off equation $y = 3.7586n^{-0.71} = 0.11926$. Therefore, observations 38 and 111 were detected as the outliers of the data set. The value of $\left|COVRATIO_{(-i)} - 1\right|$ for these two observations exceed the cut-off equation $y = 3.7586n^{-0.71}$.

5. CONCLUSION

This paper proposed an outlier detection method for the linear functional relationship model of circular variables with equal error concentration parameters. The covariance matrix is derived with some correction factor applied to the maximum likelihood estimation to obtain the *covratio* statistics of the model. This study considers a 95% confidence level, and the cut-off equation developed is to be at 5% significant level, read as $y = 3.7586n^{-0.71}$. The pattern in the power of performance in the simulation study shows that this method is adequate to detect the outlier that exists in a circular data.

6. ACKNOWLEDGEMENT

We would like to thank National Defence

University of Malaysia, University of Malaya (Grant GPF006H-2018), the Ministry of Education Malaysia and GE STEM grant (vot no. 07397) for supporting this work.

7. REFERENCES

Abuzaid, A., Mohamed I, Hussin, A. G. and Rambli, A. (2011). Covratio statistics for simple circular regression model. *Chiang Mai Journal of Science*, 2011. 38 (3): 321-330.

Belsley, D. A., Kuh, E and Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley.

Caires, S. and Wyatt, L. R. (2003). A linear functional relationship model for circular data with an application to the assessment of ocean wave measurement. *Journal of Agricultural, Biological, and Environmental Statistics, Biological and Environmental Statistics*, 8 (2): 153-169.

Dobson, A. (1978). Simple approximations for the Von Mises concentration statistic. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 27(3): 345-347.

Duan, L., Xu, L., Liu, Y. and Lee, J. (2009). Cluster-based outlier detection. *Annals of Operations Research*, 168(1): 151-168.

Fisher, N. I. (1993). *Statistical Analysis of Circular Data*. United Kingdom: Cambridge University Press.

Ghapor, A. A, Zubairi, Y. Z, Mamun, A. S. M. A. and Imon, A. H. M. R. (2014). On detecting outlier in simple linear functional relationship model using *covratio* statistic. *Pakistan Journal of Statistics*, 30(1): 129-142.

Hassan, S. F, Hussin, A. G. and Zubairi, Y. Z. (2010). Estimation of functional relationship model for circular variables and its application in measurement problem. *Chiang Mai Journal of Science*, 37(2): 195-205.

Hussin, A. G., Fieller, N. R. J. and Stillman, E. C. (2004). Linear regression for circular variables with application to directional data. *Journal of Applied Science & Technology*, 8(1 & 2): 1-6.

Hussin, A.G., Abuzaid, A.H., Ibrahim, A.I.N. & Rambli, A. (2013). Detection of outliers in the complex linear regression model. *Sains Malaysiana*, 42(6): 869–874

Ibrahim S, Rambli A, Hussin A G and Mohamed I. (2013). Outlier detection in a circular regression model using covratio statistic. *Communication in Statistics-Simulation and Computation*, 42(10): 2272-2280.

Mardia, K. V. and Jupp, P. E. (2000). *Directional Statistics*. New Jersey: John Wiley & Sons.

Mokhtar, N. A., Zubairi, Y. Z. and Hussin, A. G. (2015). A simple linear functional relationship model for circular variables and its application. *Proceedings of the 9th International Conference on Renewable Energy Sources (RES '15)*, Kuala Lumpur, Malaysia, pp. 57-63.

Mokhtar, N. A., Zubairi, Y. Z., & Hussin, A. G. (2018). A clustering approach to detect multiple outliers in linear functional relationship model for circular data. *Journal of Applied Statistics*, 45(6): 1041-1051.

Rambli, A., Abuzaid, A. H. M. , Mohamed, I. B. and Hussin, A. G. (2016). Procedure for detecting outliers in a circular regression model. *PLOS ONE*, 11(4): e0153074.

Spatial Bayesian Model Averaging to Calibrate Short-Range Weather Forecast in Jakarta, Indonesia

Niswatul Qona'ah[1a*], Sutikno[1b], Purhadi[1c]

[1] Department of Statistics, Faculty of Mathematics Computing and Data Science, Institut Teknologi Sepuluh Nopember, Kampus Sukolilo, Surabaya 60111, INDONESIA. E-mail: niswatulqonaah@gmail.com[a] ; sutikno@statistika.its.ac.id[b] ; purhadi@statistika.its.ac.id[c]
* Corresponding Author: niswatulqonaah@gmail.com
Received: 21st April 2019 Revised : 6th August 2019 Published: 30th September 2019
DOI : https://doi.org/10.22452/mjs.sp2019no2.6

ABSTRACT Bayesian Model Averaging (BMA) is a statistical post-processing method to calibrate the ensemble forecasts and create more reliable predictive interval. However, BMA does not consider spatial correlation. Geostatistical Output Perturbation (GOP) considers spatial correlation among several locations altogether. It has spatial parameters that modifies the forecast output to capture spatial information. Spatial Bayesian Model Averaging (Spatial BMA) is a method which combines BMA and GOP. This method is applied to calibrate the temperature forecast at 8 stations in Indonesia that is previously predicted by Numerical Weather Prediction (NWP). Temperature forecasts of BMA are used to obtain simulated spatially correlated error that modify temperature forecasts. Spatial BMA is able to calibrate the temperature forecast better than raw ensemble whose coverage comes closer to the standard 50%. Based on Root Mean Square Error (RMSE) criteria, Spatial BMA is able to correct forecast bias NWP with RMSE value of 1.399° lower than NWP of 2.180°.

Keywords: BMA; Ensemble; GOP; NWP; Spatial BMA.

1. INTRODUCTION

In the last few years, BMKG (Meteorological, Climatological and Geophysical Agency) in Indonesia has already developed the numerical weather forecasting process using Numerical Weather Prediction (NWP) to support forecasters. But the NWP forecasting has high bias because it is measured on a global scale (homogeneous) and is unable to capture the dynamics fluctuating atmosphere (BMKG, 2011; Wilks, 2006). Statistical post-processing of NWP outcome using ensemble has been used for increasing the forecasting accuracy i.e. the NWP composite of some modeling methods, such as ARIMA, Neural Network, Principal Component Regression (PCR), etc.

However, the ensemble forecasting is often still underdispersive (the centered weather forecasting on a value with low variance). As a result, the forecast interval becomes narrow and the observations cannot be contained in the forecast interval, so the ensemble calibration is required (Schmeits and Kok, 2010). Several methods for forecast calibrating of ensembles are Bayesian Model Averaging (BMA), Geostatistical Output Perturbation (GOP) and Spatial BMA, which are a combination of BMA and GOP.

BMA is a method that combines the forecasts of all members of the ensemble based on weighted average, so it does not just consider the contribution of one model from any models as in most of statistical models (Raftery and Zheng, 2003). In the Meteorology

field, BMA is the most widely used method because its performance is quite satisfactory. However, this method has the disadvantage, i.e. only considering one location and ignoring spatial correlation.

One of the spatial-based weather forecasting methods is GOP. This method is able to generate large ensembles based on the identified spatial relationship of the model error. Then, the error is added to the forecasting result from a simple regression, to obtain a calibrated forecast which is capable in capturing the spatial phenomenon (Gel et al., 2004). However, GOP generally uses only one predictor from the NWP output, such as maximum temperature or minimum temperature.

Spatial BMA is a combination of BMA and GOP methods. This method is expected to overcome the weaknesses of BMA and GOP. Like in BMA, probability density function (pdf) predictive Spatial BMA is the weighted average of conditional pdf's centered on bias correction of ensemble member models, with weights that depend on the contribution of each member. In Spatial BMA model, conditional pdf is multivariate densities with covariance structures in order to consider the spatial structure from weather observation.

In addition, the Spatial BMA model parameters have to conform to the GOP model parameters. Spatial BMA method can be used to produce statistical ensembles from an entire area of simultaneous weather observation, of any size, and at minimum computing cost. In individual location, Spatial BMA can be reduced to BMA (Berrocal et al., 2007).

In this study, the forecast of air temperature in 8 meteorological stations in Jakarta, Indonesia was calibrated with Spatial BMA. Previously, each NWP parameter in the nine measurement grids was processed first with Principal Component Analysis (PCA) to reduce the dimension. Members of the Spatial BMA ensemble are obtained from Partial Least Square (PLS), Principal Component Regression (PCR) and Ridge regression. This study aims to calibrate air temperature forecasts, in order to obtain a better method of optimizing the NWP output and is expected to be used for short-term forecasting.

2. LITERATURE REVIEW

2.1 Ensemble Forecasting using Model Output Statistics (MOS)

Prediction ensemble system is a system that consists of several combinations of models that process a single deterministic outcome (deterministic forecast) (Park, 2006). NWP is one of the single deterministic forecast models commonly used by many countries. However, in processing, NWP forecasts still have a high bias, so it needs to be optimized by combining multiple NWPs to obtain accurate, precise and calibrated forecasts.

The ensemble forecast is performed by an integrated approach of several statistical modeling methods that process the same NWP output and is known as MOS. The study used 3 statistical modeling methods as ensemble members, i.e. Partial Least Square (PLS), Principal Component Regression (PCR) and Ridge Regression.

i. Partial Least Square (PLS)

PLS models the relationship between the response **Y** and the predictor **X** based on the latent variables simultaneously (Wold et al., 2001). PLS corresponds to PCR, i.e. forming a matrix of latent component **T** in (1) of size $n \times c$ as a linear transformation of the predictor matrix **X**:

$$\mathbf{T} = \mathbf{XW} \tag{1}$$

with **W** is being the weighted matrix $p \times c$. Index n denotes the number of observations, p denotes the number of predictors, and c denotes the number of latent components. The latent component **T** is as a random variable **X** and is used to predict **Y**. The response **Y** of size $q \times c$ with q is the number of responses. When **T** is formed, then **Q'** is obtained by the least squares method based on (2):

$$\mathbf{Q'} = (\mathbf{T'T})^{-1}\mathbf{T'Y}. \tag{2}$$

Based on Wold et al. (2001), one way to model **Y** is:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{F}. \tag{3}$$

By applying substitution involving (3), we obtain a weighted **B** in (4):

$$\begin{aligned} \mathbf{XB} + \mathbf{F} &= \mathbf{TQ'} + \mathbf{F} \\ \mathbf{XB} &= \mathbf{XWQ'} \\ \mathbf{B} &= \mathbf{WQ'} = \mathbf{W}(\mathbf{T'T})^{-1}\mathbf{T'Y}. \end{aligned} \tag{4}$$

Thus, it is obtained (5) to guess the response:

$$\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\mathbf{B}} = \left[\mathbf{TW}^{-1}\mathbf{W}(\mathbf{T'T})^{-1}\mathbf{T'Y}\right] = \left[\mathbf{T}(\mathbf{T'T})^{-1}\mathbf{T'Y}\right]. \tag{5}$$

ii. Principal Component Regression (PCR)

Suppose that $\mathrm{x}^{'} = \left[x_1\ x_2 \ldots x_p\right]$ is a predictor vector of a matrix **X** of size $n \times p$. If the matrix **A** is an orthogonal matrix $p \times p$ with the m-th column containing the m-th eigenvector from $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ and assumes $m \leq p$, then the PC score for each observation is as in (6):

$$\mathbf{Z} = \mathbf{XA} \tag{6}$$

where element (i) from **Z** represents the score of the m-th PC for the i-th observation where $i = 1,2, \ldots, n$. Based on the orthogonal properties of **A** where $\mathbf{AA'} = \mathbf{I}$, multiple linear regression with correlated predictors can be converted into PCR as in (7):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \tag{7}$$

with $\mathbf{XA} = \mathbf{Z}$ and $\mathbf{A}'\boldsymbol{\beta} = \boldsymbol{\gamma}$.

iii. Ridge Regression

Based on Draper and Smith (1992), ridge regression uses a non-negative constant λ to calculate the more efficient regression coefficients, while reducing the singularity due to multicollinearity. (8) is used to calculate the Ridge regression coefficient:

$$\hat{\boldsymbol{\beta}}_{ridge} = \left(\mathbf{X'X} + \boldsymbol{\lambda}\mathbf{I}\right)^{-1}\mathbf{X'y}\,. \tag{8}$$

So that obtained the estimated model based on (9):

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}_{ridge}\,. \tag{9}$$

The λ constant can be selected intuitively (by default) or by the cross-validation technique based on the lowest RMSE. The higher λ causes the $\hat{\boldsymbol{\beta}}_{ridge}$ coefficient closer to 0 or causes the parameter to have less effect on the response variable.

2.2 Spatial Bayesian Model Averaging (Spatial BMA)

Spatial Bayesian Model Averaging (Spatial BMA) is a method for post-processing ensembles statistically which is a combination of BMA (Raftery and Zheng, 2003) and GOP (Gel et al., 2004). Like BMA, pdf predictive Spatial BMA is the weighted average of conditional pdf that centered on bias correction of ensemble member models, with the weights which are associated with the contribution of each member.

The Spatial BMA method considers the weather location $\mathbf{Y} = \{Y(s) : s \in S\}$, where S is a quite large set of locations but is finite and conditional on an ensemble,

$$\mathbf{F_1} = \{f_1(s) : s \in S\}, ..., \mathbf{F}_M = \{f_M(s) : s \in S\}$$

of M weather forecasts simultaneously, rather than just a single deterministic weather forecast. The Spatial BMA predictive pdf for weather forecasts is

$$g(\mathbf{Y}|\mathbf{F}_1, ..., \mathbf{F}_M) = \sum_{m=1}^{M} w_m g_m(\mathbf{Y}|\mathbf{F}_m) \tag{10}$$

where w_m is the BMA weight, equal to the probability that member of m is the best among the members of the forecast ensemble, and $g_m(\mathrm{Y}|\mathrm{F}_m)$ is a conditional pdf of $\mathbf{Y}$ if member of m is known to be the best. In practice, conditional pdf is multivariate densities centered on the forecast member's bias correction, $\beta_{0,m}\mathbf{1} + \beta_{1,m}\mathbf{F}_m$, and has covariance structures spatially $\sum_m$. This condition can be denoted by:

$$\left(\mathbf{Y} \mid \mathbf{F}_m\right) \sim MVN\left(\beta_{0,m}\mathbf{1} + \beta_{1,m}\mathbf{F}_m, \Sigma_m\right) \tag{11}$$

In Equation (11),

$$\Sigma_m = \frac{\sigma^2}{\rho_m^2 + \tau_m^2}\Sigma \tag{12}$$

where σ^2 is BMA variance, Σ is spatial structure of GOP covariance matrix, each of ρ_m^2 and τ_m^2 is a GOP covariance parameter.

Like GOP, Spatial BMA has a multivariate predictive pdf for weather forecasts. The requirement of ensemble member *m* can be the best as described in equation (13):

$$\mathbf{Y} \mid \mathbf{F}_m = \beta_{0,m}\mathbf{1} + \beta_{1,m}\mathbf{F}_m + \mathbf{E}_{1m} + \mathbf{E}_{2m} \tag{13}$$

where each of $\mathbf{E}_{1m}$ and $\mathbf{E}_{2m}$ denotes a part of continuous and discrete of the conditional error.

Here is the algorithm for getting members of Spatial BMA ensemble:

1) Take as many $m \in \{1,\ldots,M\}$ samples, with probabilities given by BMA weights $w_1, w_2, \ldots, w_M$. This is to get members of the dynamic ensemble.
2) Simulate the realization of continuous and discrete parts, $\mathbf{E}_{1m}$ and $\mathbf{E}_{2m}$, on the conditional error of each conditional pdf.
3) Use the right-hand side in equation (13) to get a bias-corrected of weather forecast $\beta_{0,m}\mathbf{1} + \beta_{1,m}\mathbf{F}_m$, with conditional error simulation, $\mathbf{E}_{1m}$ and $\mathbf{E}_{2m}$.

Furthermore, we get weather forecast ensemble of Spatial BMA, with various ensemble sizes which we want, and with minimum computing costs (Berrocal et al., 2007).

2.3 Goodness of Fit Model Evaluation

Goodness of fit models that aim to calibrate weather forecasts are not adequately measured if by using only RMSE. The other measures are also needed to check the level of bias correction and sharpness forecasts ensemble, i.e. CRPS and coverage (Feldmann, 2012).

2.3.1 Root Mean Square Error (RMSE)

The RMSE as an indicator of accuracy in (14) is obtained from the square root of MSE, which is the sum of the squares of the difference between the forecast and observation values.

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{14}$$

where *n* is the number of observation.

2.3.2 Continuous Rank Probability Score (CRPS)

CRPS, calculated based on (15), is used to check how precise the predictive intervals produced by the calibration methods, such as Spatial BMA. The lower the CRPS value, the more reliable the predictive interval (Feldmann, 2012).

$$\text{CRPS} = \frac{1}{n}\sum_{i=1}^{n}\text{crps}\left(F_i, y_i\right) = \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty}\left[F_i^{forecast}(y) - F_i^{obs}(y)\right]^2 dy \tag{15}$$

where *n* is number of observation, *i* is the time period (e.g. daily), $F_i^{forecast}(y)$ is the predictive cumulative distribution function (cdf) at time i^{th}, and $F_i^{obs}(y)$ is the empirical cdf at time i^{th} (Anggraeni, 2013). If the threshold forecast < observation, then $F_i^{obs}(y) = 0$, and 1 if the threshold forecast ≥ observation.

2.3.3 Coverage

The sharpness of the ensemble forecasts can be identified through coverage in (16). If the observation is in the ensemble range, then the observation is identified to be in the coverage, where the coverage standard is derived from

$$\frac{M-1}{M+1} \times 100\% \tag{16}$$

The ensemble forecast is called to be calibrated if the value of coverage closes to the standard of calculation in (16).

2.4 Temperature

The atmospheric temperature is the measure of the temperature at various levels in the Earth's atmosphere that can be affected by solar radiation, humidity and altitude, thus the effect is a complex relationship among the biosphere, the lithosphere and the atmosphere (Tanudidjaja, 1993). Energy is constantly moving from the surface to the air above it which causes heat transfer.

3. METHODOLOGY

3.1 Data Source

The data used in this study are the secondary data from Meteorology, Climatology and Geophysics Agency (BMKG), i.e. the data of CCAM (conformal cubic atmospheric model) NWP data from 1st of January 2009 to 31st of December 2010 or 708 days. The location of research focus is meteorological station, i.e. Kemayoran, Priok, Cengkareng, Pondok Betung, Curug, Dermaga, Tangerang and Citeko.

3.2 Research Variables

Response variable in this study is air temperature of the observation data, i.e. maximum temperature (Celsius). Predictors are air temperature forecasts which are modeled by PLS, PCR, and Ridge regression. Predictors are obtained from the outcome of the CCAM NWP parameter, shown in Table 1, which were initially reduced by PCA.

Table 1: NWP parameters (BMKG, 2011).

NWP Parameter (code)	Level	Unit
Surface Pressure Tendency (dpsdt)	surface	hPa
Water Mixing Ratio (mixr)	1, 2, 4	g/kg
Vertical Velocity (omega)	1, 2, 4	knot
PBL *depth* (pblh)	surface	meter
Surface Pressure (ps)	surface	hPa
Mean Sea Level Pressure (psl)	surface	hPa
Screen Mixing Ratio (qgscm)	surface	g/kg
Relative Humidity (rh)	1, 2, 4	%
Precipitation (rnd)	surface	mm
Temperature	1, 2, 4	Celcius
Maximum Screen Temperature (tmaxcr)	surface	Celcius
Minimum Screen Temperature (tmincr)	surface	Celcius
Pan Temperature (tpan)	surface	Celcius
Screen Temperature (tscrn)	surface	Celcius
Zonal Wind (u)	1, 2, 4	knot
Friction Velocity (ustar)	surface	m/sec
Meridional Wind (v)	1, 2, 4	knot
Geopotential Height (zg)	1, 2, 4	meter

Besides the 7 parameters that are measured in the different pressure levels, 11 other parameters were measured only in the surface level with a height of ± 2 meters above sea level. Thus, the number of NWP parameters is 32 parameters. Then, each of the 32 parameters is measured on the nine grid (3 x 3) measurements, so there are 288 parameters in total.

3.3 Steps of Analysis

The steps to apply Spatial BMA to obtain a calibrated temperature forecast are as follows:

a. Standardize data, both for observation data and NWP parameter data.
b. Reduce the dimensions of each NWP parameter using PCA based on the covariance matrix, then we get PC scores.
c. Predict the air temperature with temperature observation as response and PC score as predictor.

Forecasts are generated from PLS, PCR, and Ridge regressions.

d. Calibrate the ensemble forecasts for air temperature using BMA based on 30-day training window, starting from estimation of regression coefficient of $\beta_{0,m}$ and $\beta_{1,m}$ and calculating BMA calibrated forecast on certain day.
e. Model the air temperature forecast using spatial modeling as in GOP, with BMA forecast results as a predictor based on 30-day training, starting from analyzing empirical semivariogram, obtaining forecasting temperature that has been added with spatial errors effect until goodness examination of Spatial BMA model.

4. RESULTS AND DISCUSSION

4.1 Standardizing Data and Dimension Reduction using PCA

The NWP parameters to be processed as predictors have various measurement units. Therefore, it is necessary to standardize using scaled and center method. It aims to minimize the difference of measurement scale between NWP parameters so that the model formed will be balanced.

Furthermore, NWP parameters need to be reduced using PCA as there is an indication of spatial relationship between grids in a NWP parameter. PCA also aims to simplify modeling and is expected to shorten the computation process without reducing precision and accuracy. For the Dermaga meteorological station, each NWP parameter produces 1 to 2 components. Resulting in a total of 41 components of 32 NWP parameters. Similarly with other meteorological stations, it also produces 1 to 2 components on each NWP parameter. The variability of NWP parameters explained by the PC varies from 80% to 100%.

Thus, the dependency of weather conditions between grids in a NWP parameter is relatively high. The ensemble member models will be discussed in detail for Dermaga station. Meanwhile, the other stations have the same steps. Furthermore, 41 components of NWP parameter will be involved in MOS modeling in which PC scores from each of the 41 selected components are used as predictors to obtain forecasts of ensemble members.

4.2 PLS Regression for Maximum Temperature

Predicted residual error sum of square (PRESS) value are obtained from validation data that are generally randomly selected. This is to see if a regression model has accurate forecasting capabilities. For Dermaga station, the lowest PRESS value for maximum temperature is in the PLS model with 16 components, which is 0.641. After the optimal number of components is obtained and used as a latent component contributing to the modeling, the next step is to apply the PLS regression for maximum temperature. Table 2 shows the regression coefficient of PLS (in the standardized form).

Table 2: The regression coefficient of PLS for maximum temperature.

Predictors	**T_{MAX}**	**Predictors**	**T_{MAX}**
PC.dpsdt	-0.112	PC.tmaxscr	-0.538
PC.mixr1	0.130	PC.tminscr	-0.128
PC.mixr2	0.139	PC.tpan	0.092
⋮	.	:	:
PC.temp1	-0.115	PC.zg2	0.020
PC.temp2	0.318	PC1.zg4	-0.0002
PC.temp4	0.017	PC2.zg4	0.067

The next step is to form the PLS model based on the regression coefficient in Table 2. However, the model is still in the form of the PC that can be returned to the variable form in nine grids. To return to the nine grids variable, multiply the PC with each eigenvector. Here is a summary of the eigenvector for each 41 components in Dermaga station with the first column showing the eigenvector of dpsdt (surface pressure tendency) and the last column is the eigenvector of zg4 (geopotential height).

$$\mathbf{E} = \begin{array}{c} \begin{array}{cccccccccc} \text{dpsdt} & & \text{mixr1} & \text{mixr2} & \text{mixr4} & \text{omega1} & \text{omega1} & \ldots & \text{zg4} & \text{zg4} \end{array} \\ \begin{bmatrix} -0,3333 & -0,3411 & -0,3353 & -0,3337 & -0,2908 & -0,4402 & \cdots & -0,2007 & 0,4516 \\ -0,3333 & -0,3220 & -0,3339 & -0,3346 & -0,2546 & -0,5264 & \cdots & -0,3958 & -0,2278 \\ -0,3333 & -0,3131 & -0,3239 & -0,3301 & -0,2129 & -0,5401 & \cdots & -0,4396 & 0,0187 \\ -0,3334 & -0,3494 & -0,3396 & -0,3370 & -0,3826 & 0,0026 & \cdots & -0,0704 & 0,5045 \\ -0,3334 & -0,3565 & -0,3432 & -0,3395 & -0,3931 & 0,1631 & \cdots & -0,4394 & -0,0633 \\ -0,3334 & -0,3478 & -0,3393 & -0,3355 & -0,3614 & 0,1186 & \cdots & -0,4080 & 0,1937 \\ -0,3333 & -0,3270 & -0,3260 & -0,3233 & -0,3439 & 0,2151 & \cdots & 0,1027 & 0,4975 \\ -0,3333 & -0,3226 & -0,3281 & -0,3337 & -0,3556 & 0,3005 & \cdots & -0,2504 & -0,4224 \\ -0,3333 & -0,3174 & -0,3301 & -0,3323 & -0,3593 & 0,2454 & \cdots & -0,4149 & 0,1481 \end{bmatrix} \end{array}$$

4.3 PCR for Maximum Temperature

Similar to PLS regression, the step before modeling the weather with PCR regression is to determine the optimal number of components from 41 components of the Dermaga station. Although there are similarities in the initial procedure, but PCR does not select the optimal component based on the lowest RMSE such as PLS regression. For PCR, the selected component is the cumulative number of components which is capable to represent a predictor variance of at least 80%, it means that the PC should be able to explain data variability of at least 80% (Johnson and Wichern, 2007). It was found that 8 of the 41 components were able to represent more than 80% predictor variance, so it was decided that the optimum components number for the maximum temperature PCR model were 8.

After knowing the optimal number of components to be used in the modeling, the next step is obtaining the PCR regression for maximum temperature. The PCR regression coefficients (in the standardized form) are shown in Table 3

Table 3: The regression coefficient of PCR for maximum temperature.

Predictor	**T_{MAKS}**	**Predictor**	**T_{MAX}**
PC.dpsdt	0.011	PC.tmaxscr	-0.072
PC.mixr1	-0.028	PC.tminscr	-0.054
PC.mixr2	0.015	PC.tpan	-0.073
⋮	⋮	⋮	⋮
PC.temp1	-0.058	PC.zg2	-0.011
PC.temp2	-0.049	PC1.zg4	-0.038
PC.temp4	-0.028	PC2.zg4	0.016

Furthermore, to obtain the maximum temperature forecasts is to form the PCR model based on the regression coefficient in Table 3, then do the same steps as PLS to obtain the model in the form of variables in nine grids.

4.4 Ridge Regression for Maximum Temperature

After obtaining ensemble members from PLS and PCR models, then weather modeling with Ridge regression is able to reduce the multicollinearity effect. Unlike the previous two methods, this method does not require the selection of many optimal components. This method uses a constant λ to minimize the impact of the singularity of $\mathbf{X}^T\mathbf{X}$. Figure 1 assists the visual determination of the constant λ.

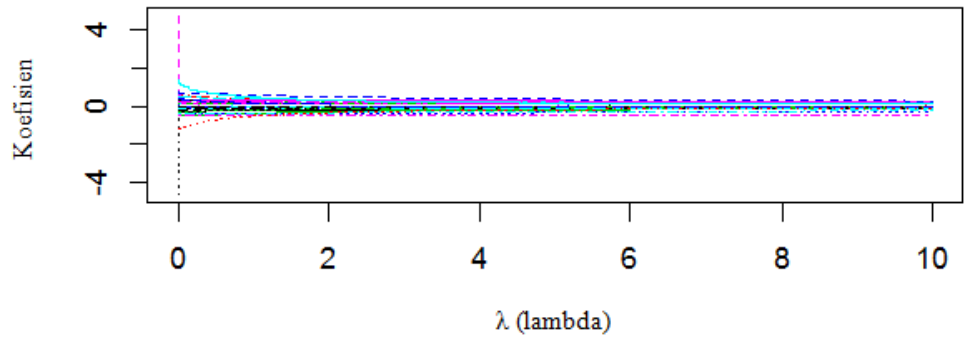


Figure 1: Convergence of ridge regression coefficients.

Determination of convergence based on Figure 1 should be avoided because it is subjective. However, in this case, the constant λ remains non convergent even when λ approaches 100. So it is decided to use visual assistance only in determining λ. Based on Draper and Smith (1992), the determination of the lower and upper limits for λ is not strictly determined. However, the upper limit should not be greater than 10 or 20 to rid of the regression coefficients that are insignificant because they are close to 0. Based on Figure 1, it is indicated that the regression coefficients for maximum temperature to converge is when λ is 9 or greater. Table 4 is the Ridge regression coefficient for Dermaga station with λ = 9.

Table 4: The ridge regression coefficient for maximum temperature.

Predictor	T_{MAKS}	Predictor	T_{MAX}
PC.dpsdt	-0.075	PC.tmaxscr	-0.513
PC.mixr1	0.123	PC.tminscr	-0.108
PC.mixr2	0.102	PC.tpan	0.053
⋮	⋮	⋮	⋮
PC.temp1	-0.108	PC.zg2	0.022
PC.temp2	0.256	PC1.zg4	0.001
PC.temp4	0.008	PC2.zg4	0.079

After the regression coefficient is obtained in the previous stage, then the next step is to form a model based on the regression coefficient in Table 4.

4.5 Description of Ensemble Member Weather Forecast

After obtaining the model to be used to predict the air temperature, the next step is to compare the forecast results of each ensemble member, i.e. PLS, PCR, and Ridge, and temperature observation values. Descriptive analysis is performed to see how well the ensemble member predictions before being calibrated by Spatial BMA. Figure 2 below shows the first 100-day trend of 2009 from the ensemble member's predictions and maximum temperature observations for the Dermaga station.

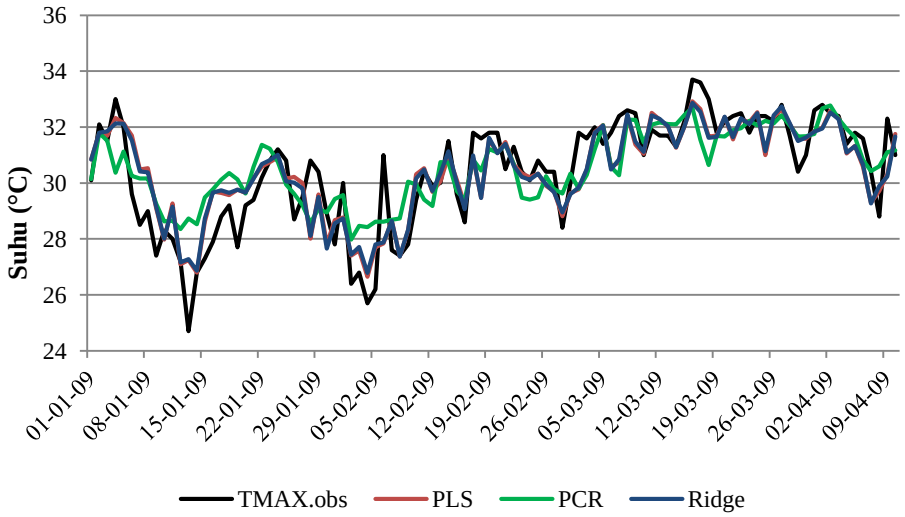


Figure 2 : Forecasts trend of ensemble members and observation of maximum temperatures.

Figure 2 indicates that each ensemble member, either PLS, PCR or Ridge, has been able to follow the general pattern of maximum temperature which is, if the maximum temperature trend increases then forecasts follow the rise, and the same if the temperature trend decreases, the forecasts will also decrease. However, the problems occur are under-fitting (predictions below the value of observation) or over-fitting (predictions are above the value of observation) that consistently occur on the same day. Although

forecasts ensemble members can capture the temperature patterns that occur, but the forecast generated is far from the observation. Therefore, it is necessary to calibrate the model to produce more accurate and precise weather forecasts.

4.6 Calibrating Weather Forecast using Spatial BMA

Based on the previous discussion, it is indicated that the results of ensemble forecasting still have a fairly low accuracy. Therefore, statistical processing methods are needed to calibrate the forecasting results to make the forecast bias lower. Calibration is done to make adjustments to the variance, to obtain a more reliable forecast with a proportional variance and has a narrower predictor interval.

Calibration of weather forecast on 8 sites individually using BMA was done by Luthfi (2017). Based on Raftery et al. (2005), it is said that calibration with BMA will result in better forecasts if the ensemble range has a significant correlation with the degree of forecasting error. Ensemble range is the difference between the maximum and minimum of an ensemble. However, the significant correlation does not necessarily guarantee the calibrated ensemble forecasts which are no longer under-dispersive or over-dispersive and can be identified from the Verification Range Histogram (VRH).

Figure 3 is used to identify whether the raw ensemble forecasts are under-dispersive or over-dispersive. A U-shaped histogram indicates an under-dispersive ensemble, while a histogram resembling a normal distribution curve indicates an over-dispersive ensemble. In this case, the raw ensemble consists of PLS, PCR, and Ridge which are the result of simultaneous ensemble forecasts.

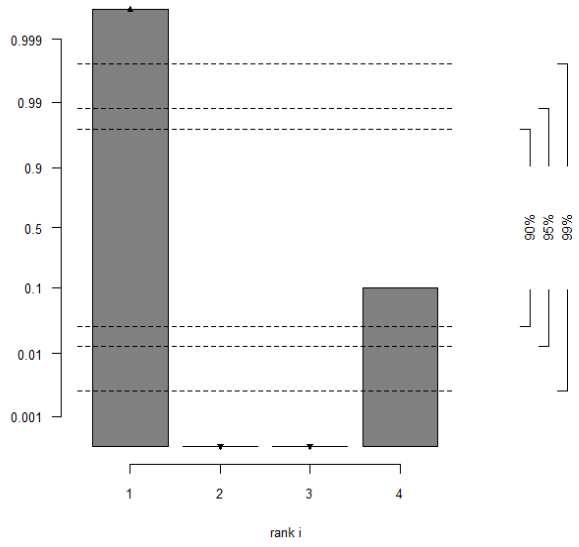


Figure 3: VRH raw ensemble of Dermaga site for maximum temperature 1st Jan ’09 - 30th Jan’09.

Figure 3 shows that the raw ensemble forecasts for maximum temperatures are still under-dispersive because the shape of the histograms. This indicates that there are still many maximum temperature observations that are outside the ensemble range, i.e. the difference between the maximum and minimum of an ensemble forecast. Based on Figure 3, the coverage of ensemble is 19.23%. This value is still far below the standard, which is 50%. It means that MOS ensemble forecasts are still under-dispersive. This can make the

predictive intervals less precise, so it is needed to calibrate using Spatial BMA.

Based on Luthfi (2017), it is indicated that BMA as a non-spatial approach that can calibrate the ensemble forecasts quite well. By applying the Spatial BMA method, it is expected to capture the spatial phenomena that occur and also utilize information from ensemble forecasts, so it can result in accurate and reliable weather forecasts. The first step in Spatial BMA modeling is to simultaneously regress all stations of each ensemble member to the observation so as to obtain regression bias coefficient, that is, $\beta_{0,m}$ and $\beta_{1,m}$ for each member *m*. The regression coefficient and weight for each member *m* and time *t* are the same for 8 sites. For example, Table 5 presents the regression coefficients, weights, forecasts that will be predictors in the Spatial BMA model.

Table 5: Parameter estimates for ensemble member model of BMA, maximum temperature, Kemayoran, 8th of February 2009.

Model	β_0	β_1	*w*	Ensemble Member Forecasts (°C)	Obs. (°C)	Simultaneous BMA (°C)
PLS	0.27	0.97	0.771	29.26		
PCR	-0.77	0.99	0.095	28.20	29.30	28.60
Ridge	0.31	0.97	0.134	2.23		

Table 5 shows that PLS has the highest contribution to forecasts for maximum temperatures due to *w* weights of 0.771, higher than PCR and Ridge whose respective weights are 0.095 and 0.134. Based on Table 5 it can be said that the simultaneous BMA forecast accuracy on February 8, 2009 did not differ significantly with the ensemble member forecasts. Furthermore, the forecast result of the simultaneous BMA used as a predictor for the formation of Spatial BMA model with response is the observation of maximum temperature. Before that, we calculated the value of Moran's I and *p*-value to see the significance of spatial dependencies. Table 6 presents the values of Moran's I and *p*-value.

Table 6: Parameter estimates for ensemble member model of BMA, Maximum temperature, Kemayoran, 8th of February 2009.

Weather Element	Moran's I	st.dev	*p*-value
T_{MAX}	0.135	0.141	0.048

The Moran's I and p-value in Table 6 indicate that there are spatial dependencies for maximum temperatures at the significance level of $\alpha = 0.05$. Positive Moran's I values indicate that air temperatures at adjacent sites tend to have a higher relationship than distant locations. Having proven spatial dependencies between 8 sites, the first step in Spatial BMA modeling is to obtain the estimates of parameter β_0 and β_1, and also residual estimation to form empirical semivariogram. Estimation of parameters and presented in Table 7.

Table 7: Parameter estimation of spatial BMA model for maximum temperature.

Weather Element	$\hat{\beta}_0$	$\hat{\beta}_1$	$\mathbf{SE}\left(\hat{\beta}_0\right)$	$\mathbf{SE}\left(\hat{\beta}_1\right)$
T_{MAX}	0.542	0.990	0.741	0.025

Based on Table 7, it can be seen that the β_0 parameters in the T_{MAKS} weather element are not significant, while the β_1 parameters are significant. Although there are insignificant parameters, this does not affect the coverage and predictive interval width of the Spatial BMA model. Furthermore, semivariogram formed empirical exponential model of spatial parameters ρ^2, τ^2 and r estimated based on the iterative method of L-BFGS. The semivariogram is formed from residual model Spatial BMA based on the estimation of β_0 and β_1 from Table 7. Figure 4 is a semivariogram formed.

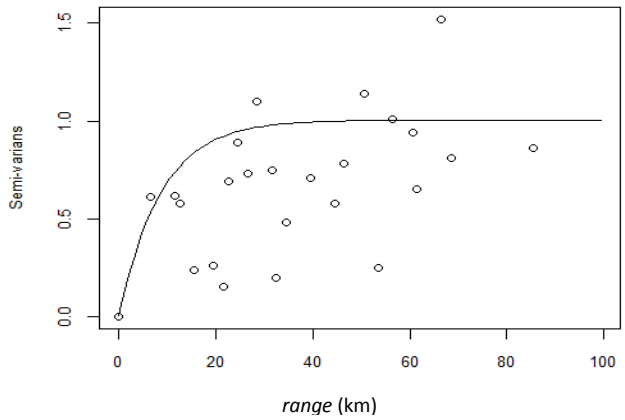


Figure 4: Empirical semivariogram for maximum temperature.

In Figure 4, the semivariogram value is constant after the range of 8.69 km or more, with a sill value (nugget + partial sill) of 1.005. A larger sill value may cause the estimation variance to become larger. This means that there is a possibility that the precision of the maximum temperature forecast is high so that it impacts the interval of Spatial BMA model that can be evaluated with CRPS. The Spatial BMA model runs a simulation process by modifying residuals to get calibrated weather forecasts. The process is run to get 99 member realization of the ensemble. The reason for the use of 99 realizations according to Gel et al. (2004) is to accommodate the observed value when Spatial BMA is said to be calibrated when the t^{th} temperature observation falls between the two percentiles. Table 8 presents the RMSE and predictive interval of the Spatial BMA model on 2nd of March, 2009.

Table 8 : RMSE of maximum temperature forecasts using Spatial BMA and NWP, 2nd of March, 2009.

Stamet	Obs. (°C)	NWP (°C)	Spatial BMA (°C)	RMSE Spatial BMA	RMSE NWP
Kemayoran	33	29.87	31.38	1.399°	2.180°
Priok	32.4	29.58	31.18		
Cengkareng	31.9	29.77	31.04		
Pd. Betung	34	29.59	32.57		
Curug	33.2	29.23	31.41		
Tangerang	33	29.67	32.13		
Citeko	26	29.65	24.84		
Dermaga	31.8	29.69	29.76		

Based on Table 8, it can be said that Spatial BMA is able to correct forecast bias NWP. This is indicated by the lower RMSE forecast of Spatial BMA than the NWP forecast at the maximum temperature parameter. In addition, the coverage value is used as the calibrated indicator of forecasts. Forecast results are said to be calibrated if the value of coverage approaches the standard coverage value. The coverage standard is the percentage comparison between the number of ensembles minus 1 by the number of ensembles plus 1. In this case, according to the number of simulated ensemble realizations, the number of ensembles is 99. So the standard coverage used is 98%. This coverage indicator can be visualized with Verification Rank Histogram (VRH). Figure 5 represents the VRH for the forecast of maximum temperature using Spatial BM.

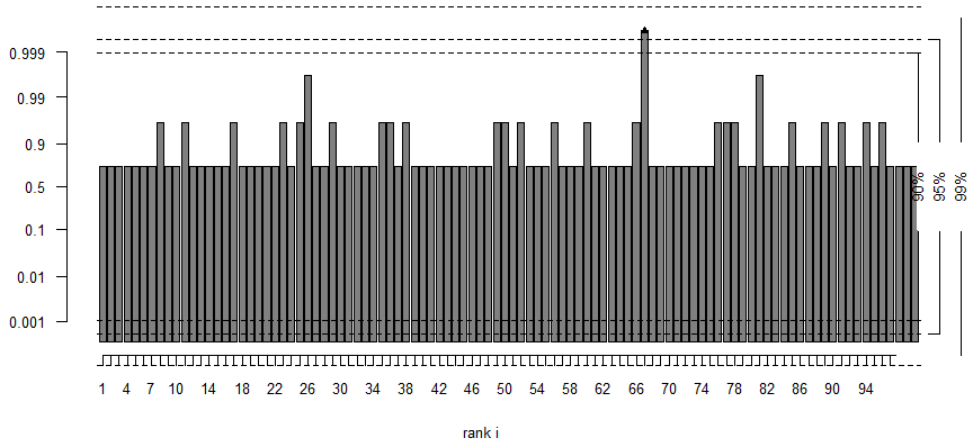


Figure 5: VRH Spatial BMA forecasts, number of ensembles = 99.

Based on Figure 5, it is known that the coverage forecasts for Spatial BMA T_{MAKS} is 87.94% which is the percentage of the number of observations that are in 2nd rank to 98th rank. The coverage value of T_{MAKS} parameters is close to 98%. This indicates that Spatial BMA is sufficiently able to calibrate the maximum temperature forecast parameters. Furthermore, in order to be able to compare directly with the forecast results before calibration, the number of simulations of the ensemble realization is 3, in accordance with the number of ensemble members before calibration. Thus, the default coverage value used is 50%. Figure 6 is a

verification Rank Histogram (VRH) of the T_{MAKS} forecasts result using Spatial BMA with the number of ensembles of 3.

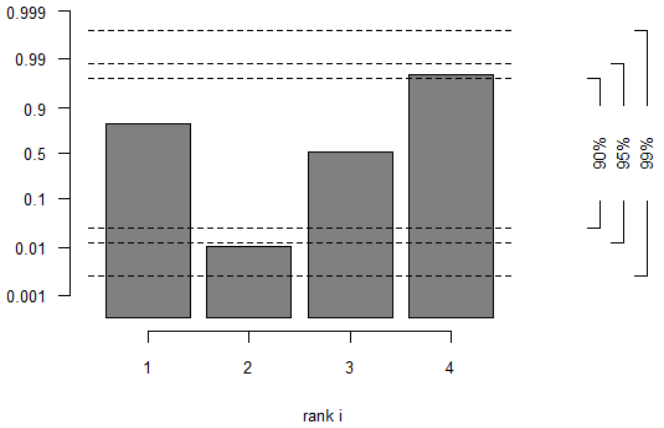


Figure 6: VRH Spatial BMA forecasts, number of ensembles = 3.

Figure 6 shows that the Spatial BMA forecast coverage for the T_{MAKS} parameter is 48.63%, which is the percentage of the number of observations entered in 2nd rank and 3rd rank. The value is close to the 50% coverage standard. This indicates that the Spatial BMA with the number of ensembles of 3 has been sufficiently able to calibrate the results of the maximum temperature forecast. This indication is supported by the value of forecast coverage before being calibrated which is still far from the standard that is 19.23%.

Based on RMSE and coverage standards, it can be said that Spatial BMA method is able to improve the sharpness of the ensemble forecasts by making more observation values in the ensemble range. But in terms of accuracy, Spatial BMA has not been able to correct the forecast bias. This is supported by the CRPS values of Spatial BMA that is higher than the Simultaneous BMA. Table 9 shows the comparison of CRPS values between the Spatial BMA model and the Simultaneous BMA model (before spatially calibrated).

Table 9: CRPS value of Spatial BMA and Simultaneous BMA.

Response	CRPS BMA Simultaneous	CRPS Spatial BMA	
		99 *Ensemble*	**3 *Ensemble***
T_{MAX}	0.562	0.580	0.763

Table 9 shows that the CRPS value of Simultaneous BMA for maximum temperature is smaller than the CRPS of Spatial BMA either with 3 ensembles or 99 ensembles. This indicates that the Spatial BMA model has not been able to correct forecast bias and improve the predictive pdf of the Simultaneous BMA model. Table 9 also shows that the CRPS model value of Spatial BMA with 99 ensembles is smaller than the Spatial BMA model CRPS with only 3 ensembles. So that, for the Spatial BMA model it is better to use a large number of simulated realization

ensembles in order to calibrate the forecast optimally.

5. CONCLUSION

Based on coverage standards, it can be said that Spatial BMA method is able to improve the sharpness of the ensemble forecasts by making more observation values in the ensemble range. Coverage of Spatial BMA forecasts for maximum temperature increased to closer coverage standard (50%), i.e. 48.63% from previous 19.23%. In addition, based on RMSE value, Spatial BMA is able to correct forecast bias of NWP prediction with RMSE value of 1.399° lower than NWP of 2.180°.

6. ACKNOWLEDGEMENT

The entire data used in this study were supported by the Meteorology, Climatology and Geophysics Agency (BMKG) of Indonesia. The fund for this study is supported by the Ministry of Research, Technology and Higher Education of Indonesia for the grant of the National Strategic Research 2018.

7. REFERENCES

Anggraeni, D. (2013). Kalibrasi Peramalan *Ensemble* Data Curah Hujan Dengan Metode *Ensemble Model Output Statistics* (EMOS) dan *Bayesian Model Averaging* (BMA). *Tesis*. Insitut Teknologi Sepuluh Nopember, Surabaya.

Berrocal, V.J, Raftery, A.E., and Gneiting, T. (2007). Combining Spatial Statistical and Ensemble Information in Probabilistic Weather Forecast. *Monthly Weather Review AMS,* 135: 1386-1402.

BMKG. (2011). Kajian dan Aplikasi Model CCAM (Conformal Cubic Atmospheric Model) *untuk Prakiraan Cuaca Jangka Pendek Menggunakan MOS (Model Output Statistics).* Jakarta: Pusat Penelitian dan Pengembangan BMKG.

Draper, N.R. and Smith, H. (1992). *Applied Regression Analysis Second Edition*. New York: John Wiley and Sons, Inc.

Feldmann, K. (2012). Statistical Postprocessing of Ensemble Forecasts for Temperature: The Importance of Spatial Modeling. *Diplomarbeit*. Ruperto-Carola University of Heidelberg, Germany.

Gel, Y., Raftery, A.E., and Gneiting, T. (2004). Calibrated probabilistic mesoscale weather field forecasting: The Geostatistical Output Perturbation (GOP) method (with discussion). *Journal of the American Statistical Association*, 99 (467): 575–583.

Johnson, R.A. dan Wichern, D.W. (2007). *Applied Multivariate Statistical Analysis 5th Edition*. New Jersey: Prentice Hall.

Luthfi, M. (2017). Bayesian Model Averaging dan Geostatistical Output Perturbation untuk Prakiraan Cuaca Jangka Pendek Terkalibrasi. *Thesis*, Insitut Teknologi Sepuluh Nopember, Surabaya.

Park, Y.Y. (2006). Recent development of ensemble forecast system. *ASEAN-ROK Cooperation Training Workshop for the Use of Numerical Weather Prediction Products*, KMA, Seoul, South Korea, 93-177.

Raftery, A.E. and Zheng, Y. (2003). Discussion: Performance of Bayesian Model Averaging. *Journal of the*

American Statistical Association, 98: 931-938.

Raftery, A.E., Gneiting, T., Balabdoui, F. and Polakowski, M. (2005). Using Bayesian Model Averaging to Calibrate Forecast Ensembles. *Monthly Weather Review AMS*, 133: 1155-1174.

Schmeits, M.J. and Kok, K.J. (2010). A Comparison between Raw Ensemble Output, (Modified) Bayesian Model Averaging and Extended Logistic Regression Using ECMWF Ensemble Precipitation Forecast. *Monthly Weather Review AMS*, 138: 4199-4211.

Tanudidjaja. (1993). *Ilmu Pengetahuan Bumi dan Antariksa*. Jakarta: Penerbit Departemen Pendidikan dan Kebudayaan.

Wilks, D.S. (2006). *Statistical Methods in the Atmospheric Sciences 2nd Edition.* Boston: Elsevier.

Wold, S., Sjöström, M., and Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58: 109-130

Malaysian Journal of Science 38 (Special Issue 2): 73 - 83 (2019)

**THE INTERNATIONAL SEMINAR ON MATHEMATICS IN INDUSTRY (ISMI)
AND THE INTERNATIONAL CONFERENCE ON THEORETICAL AND APPLIED STATISTICS (ICTAS)
ISMI-ICTAS18 [4-6 SEPTEMBER 2018]**

Multivariate CUSUM Control Chart Based on the Residuals of Multioutput Least Squares SVR for Monitoring Water Quality

Hidayatul Khusna[1a], Muhammad Mashuri[1b], Suhartono[1c], Dedy Dwi Prastyo[1d*], Muhammad Ahsan[1e]

[1]Department of Statistics, Institut Teknologi Sepuluh Nopember, Surabaya 60111, INDONESIA.
E-mail: khusna16@mhs.statistika.its.ac.id[a]; m_mashuri@statistika.its.ac.id[b]; suhartono@statistika.its.ac.id[c]; dedy-dp@statistika.its.ac.id[d]; ahsan16@mhs.statistika.its.ac.id[e]
*Corresponding Author: dedy-dp@statistika.its.ac.id[d]
Received: 21st April 2019 Revised: 6th August 2019 Published: 30th September 2019
DOI : https://doi.org/10.22452/mjs.sp2019no2.7

ABSTRACT Monitoring serially dependent processes using conventional control charts yields a high false alarm rate. Multioutput Least Squares Support Vector Regression (MLS-SVR) has the capability to encompass the cross-relatedness between output variables by learning multivariate output variables simultaneously. This research aims to develop a Multivariate Cumulative Sum (MCUSUM) control chart based on the residual obtained from the MLS-SVR model for monitoring autocorrelated data. The inputs of the MLS-SVR are selected using the significant lag of a partial autocorrelation function. The proposed control chart is applied to monitor water quality data and it can detect the assignable causes in those data caused by a broken pipeline.

Keywords: autocorrelated, control chart, multioutput least squares SVR, multivariate CUSUM, water quality.

1. INTRODUCTION

One of the most useable tools in statistical process control is the control chart (Woodall & Montgomery, 1999). Many assumptions need to be fulfilled in order to propose a control chart. Most of the traditional charts assume that the observations are not dependent and satisfy a multivariate normal distribution. The violation of the independence assumption affecting the control chart performance has been investigated by many researchers. Harris & Ross (1991) proved that autocorrelation affects the Average Run Length (ARL) of an Exponentially Weighted Moving Average (EWMA) and Cumulative Sum (CUSUM) control chart. Moreover, a CUSUM control chart drew an incorrect conclusion when applied to autoregressive AR(1) and moving average MA(1) data (Johnson & Bagshaw, 1974). Serially dependent data might lead to an incorrect out-of-control signal and break the effectiveness of a control chart (Noorossana & Vaghefi, 2006). Applying the conventional control chart for monitoring the autocorrelated processes will produce a high false alarm rate and reduce its ability to detect a shift of the process (Psarakis & Papaleonida, 2007).

There are two different procedures that deal with monitoring autocorrelated data. First is monitoring serially dependent data using a modified control limit of the conventional control chart (Vanbrackle & Reynolds, 1997). Another approach is developing a residual-based control chart. This second procedure applies a time series model to the serially dependent data then uses a residuals component, i.e., the difference between the actual value and the forecast value, to monitor the process. The obtained residuals are independent, so it is possible to

monitor the process using the residuals component. Yashchin (1993) suggested applying the modified control chart for a low level of autocorrelation, but the residual-based control chart for a high level of autocorrelation.

Many researchers have proposed different procedures for monitoring multivariate time series data. Chan & Li (1994) and Charnes (1995) improved the performance of Hotelling's T^2 control chart for a serially dependent process. Kalgonda & Kulkarni (2004) presented a multivariate control chart based on the residual of a Vector Autoregressive (VAR) process. Theodossiou (1993) introduced a Vector Autoregressive and Moving Average (VARMA)-based MCUSUM control chart. Kramer & Schmid (1997) proposed a Multivariate EWMA (MEWMA) control chart (Lowry et al., 1992) for a multivariate autocorrelated process. Śliwa & Schmid (2005) developed a residual-based MEWMA control chart to monitor a cross-covariance matrix of a multivariate time series process. Furthermore, Wororomi et al. (2014) and Khusna et al. (2018) developed a residual based MEWMA control chart for monitoring the mean vector shift of multivariate autocorrelated data.

When modeling a real application using the traditional time series method, it is usually hard to satisfy the particular assumption. The application of the traditional time series method also needs great expertise due to the complex structure of the autocorrelated data. In order to overcome these limitations, some researchers recommend the utilization of Support Vector Regression (SVR) as an alternative method (Sato et al., 2008; Thissen et al., 2003). The SVR algorithm has two main advantages. First, it is able to tackle both linear and nonlinear patterns data. Second, it is more reproducible since it can yield a globally optimal solution. Khediri et al. (2010) proved that a multivariate control chart based on the residuals of SVR is more effective than a multivariate control chart based on the residuals of an Artificial Neural Network (ANN). Issam & Mohamed (2008) proved that, in comparison to ANN and VAR, an SVR-based control chart performs better.

Least Squares SVR (LS-SVR) is developed by replacing the quadratic programming problem in the SVR algorithm with a linear programming problem (Vapnik, 1998; Vapnik, 2000). Furthermore, LS-SVR changes the inequality constraints in the SVR formula for equality ones (Suykens & Vandewalle, 1999; Suykens et al., 2002). The linear equation in LS-SVR is simple to solve and useful in computational time-saving. Xu et al. (2013) proposed multioutput LS-SVR (MLS-SVR) by combining the idea of Multioutput SVR (M-SVR) (Tuia et al., 2011) and multiresponse regression (Liu et al., 2009). In order to cover the Hierarchical Bayes intuition, Xu et al. (2013) proposed an MLS-SVR algorithm which has the ability to overcome the differences in slope function for each output variable.

Several researchers have developed a MCUSUM control chart for monitoring multivariate autocorrelated data. Bodnar & Schmid (2007) proposed both a modified MCUSUM control chart (Crosier, 1988; Ngai & Zhang, 2001; Pignatiello & Runger, 1990) and a VARMA-based MCUSUM control chart. Hwang (2016) presented an MLS-SVR-based MCUSUM (Healy, 1987) control chart by considering the covariance of the error term. Hwang (2016) developed a new MLS-SVR which is different from the one proposed by Xu et al. (2013). The residuals of the MLS-SVR model resulting from in-control processes are assumed to satisfy a multivariate normal distribution as well as a white noise assumption. Therefore, the control limit of the MLS-SVR-based MCUSUM chart is equivalent to the control limit of the conventional MCUSUM chart (Healy, 1987). Hwang (2016) pointed out that the ARL of a MCUSUM chart based on the residuals of the MLS-SVR model outperforms

the ARL of a MCUSUM chart based on the residuals of the VARMA and LS-SVR model.

The objective of this research is to develop an MLS-SVR (Xu et al., 2013) based MCUSUM control chart proposed by Crosier (1988) instead of the one proposed by Healy (1987). The proposed control chart is then applied to monitor water quality data. Water turbidity and the chlorine residual are two critical quality characteristics in water manufacturing processes. These quality characteristics are recorded hourly, in which observations show serial dependency. The rest of this paper is organized as follows. Section 2 describes the MLS-SVR algorithm while section 3 presents the proposed control chart. The application of the MLS-SVR-based MCUSUM control chart for monitoring water quality data is shown in section 4. Finally, section 5 summarizes the results found in this work and presents future research.

2. MULTIOUTPUT LEAST SQUARES SUPPORT VECTOR REGRESSION

The basic algorithm of LS-SVR only learns the mapping from the input to a single output. If a problem solved using the LS-SVR algorithm consists of a multioutput case, then the cross-relatedness among output is disregarded. This problem inspired (Xu et al., 2013) to develop the MLS-SVR algorithm, a multitask learning method which is useful to capture the relation between outputs. An observable output variable is defined as $\mathbf{Y}=[y_{ij}]\in R^{n\times m}$, for $i=1,2,\ldots,n$ observations and $j=1,2,\ldots,m$ output variables. Let $\{(\mathbf{x}_1,\mathbf{y}^1),(\mathbf{x}_2,\mathbf{y}^2),\ldots,(\mathbf{x}_n,\mathbf{y}^n)\}$ be a specific independent and identically distributed sample, where $\mathbf{x}_i\in R^d$, $\mathbf{y}^i\in R^m$, and d defines the dimension of the input variables. Let $\varphi: R^d\to R^h$ be a mapping function to some higher dimensional Hilbert space with h dimension. All the MLS-SVR parameters are assumed to be associated with $\varphi(\mathbf{x})$, so that vector $\mathbf{w}_j\in R^h$, for $j\in N_m$ can be rewritten as $\mathbf{w}_j=\mathbf{w}_0+\mathbf{v}_j$, where the mean vector $\mathbf{w}_0\in R^h$. The small value of vector $\mathbf{v}_j\in R^h$ for $j\in N_m$ indicates that the output variables are similar to each other. The mean vector $\mathbf{w}_0$ carries commonality information, while vector $\mathbf{v}_j$ carries specialty information.

Estimating the vector $\mathbf{w}_0\in R^h$, matrix $\mathbf{V}=(\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_m)\in R^{h\times m}$, and parameter $\mathbf{b}=(b_1,b_2,\ldots,b_m)\in R^m$ can be simultaneously obtained by minimizing the objective function with constraints as follows (Xu et al., 2013):

$$\begin{aligned}&\min J(\mathbf{w}_0,\mathbf{V},\Xi)=\frac{1}{2}\left(\mathbf{w}_0^T\mathbf{w}_0\right)+\frac{\gamma''}{2m}trace\left(\mathbf{V}^T\mathbf{V}\right)+\frac{\gamma'}{2}trace\left(\Xi^T\Xi\right),\\&\text{s.t }\mathbf{Y}=\mathbf{Z}^T\mathbf{W}+repmat(\mathbf{b}^T,n,\mathbf{1})+\Xi,\end{aligned}\tag{1}$$

where $\mathbf{Z}=\left(\varphi(\mathbf{x}_1),\varphi(\mathbf{x}_2),\ldots,\varphi(\mathbf{x}_n)\right)\in R^{h\times n}$. The matrix $\Xi=(\xi_1,\xi_2,\ldots,\xi_m)\in R_+^{n\times m}$ contains the slack variables whereas the matrix $\mathbf{W}=\left(\mathbf{w}_0+\mathbf{v}_1,\mathbf{w}_0+\mathbf{v}_2,\ldots,\mathbf{w}_0+\mathbf{v}_j,\ldots\mathbf{w}_0+\mathbf{v}_m\right)\in R^{h\times m}$ illustrates the MLS-SVR parameter. The constants $\gamma',\gamma''\in R_+$ are regularized parameters. The optimization problem in Equation (1) has the following Lagrange function:

$$L(\mathbf{w}_0,\mathbf{V},\mathbf{b},\Xi,\mathbf{A})=J(\mathbf{w}_0,\mathbf{V},\Xi)-trace\left(\mathbf{A}^T\left(\mathbf{Z}^T\mathbf{W}+repmat\left(\mathbf{b}^T,n,\mathbf{1}\right)+\Xi-\mathbf{Y}\right)\right),\tag{2}$$

where matrix $\mathbf{A}=\left(\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\ldots,\boldsymbol{\alpha}_m\right)^T\in R^{n\times m}$ consists of the Lagrange multiplier.

Similar to LS-SVR, the MLS-SVR algorithm has a computational advantage by solving linear programming problems. The solutions of MLS-SVR are obtained by solving two sets of linear equation systems with the same positive definite matrix **M**, as stated in step 9 of Algorithm 1 (Xu et al., 2013).

Algorithm 1. Calculation of MLS-SVR solutions

1. Save the outputs of MLS-SVR as $\mathbf{y}=\left(\mathbf{y}_1^T,\mathbf{y}_2^T,\ldots,\mathbf{y}_m^T\right)^T\in R^{mn}$.
2. Specify the kernel function parameter σ, regularized parameters γ'' and γ'.
3. Calculate the matrix containing kernel function $\mathbf{K}=\mathbf{Z}^T\mathbf{Z}\in R^{n\times n}$ from the inputs of MLS-SVR.
4. Specify matrix $\mathbf{N}=blockdiag\left(\overbrace{\mathbf{1}_n,\mathbf{1}_n,\ldots,\mathbf{1}_n}^{m}\right)\in R^{mn\times m}$.
5. Specify matrix $\boldsymbol{\Omega}=repmat\left(\mathbf{K},m,m\right)\in R^{mn\times mn}$.
6. Specify matrix $\mathbf{Q}=blockdiag\left(\overbrace{\mathbf{K},\mathbf{K},\ldots,\mathbf{K}}^{m}\right)\in R^{mn\times mn}$.
7. Calculate the matrix $\mathbf{M}=\boldsymbol{\Omega}+(\gamma')^{-1}\mathbf{I}_{mn}+\left(m/\gamma''\right)\mathbf{Q}\in R^{mn\times mn}$.
8. Calculate the matrix $\mathbf{G}=\mathbf{N}^T\mathbf{M}^{-1}\mathbf{N}\in R^{m\times m}$.
9. Calculate ϑ and ν from $\mathbf{M}\vartheta=\mathbf{N}$ and $\mathbf{M}\nu=\mathbf{y}$.
10. Calculate $\mathbf{G}=\mathbf{N}^T\vartheta$.
11. Find a solution from $\mathbf{b}=\mathbf{G}^{-1}\vartheta^T\mathbf{y}$ and $\boldsymbol{\alpha}=\nu-\vartheta\mathbf{b}$.

Supposing that $\tilde{\boldsymbol{\alpha}}=\left(\left(\tilde{\alpha}_1^{'}\right)^T,\left(\tilde{\alpha}_2^{'}\right)^T,\ldots,\left(\tilde{\alpha}_\ell^{'}\right)^T\right)^T$ and $\tilde{\mathbf{b}}$ are the solutions for the MLS-SVR model, the decision function of MLS-SVR can be formulated as

$$\begin{aligned}\hat{f}(\mathbf{x})&=\varphi(\mathbf{x})^T\tilde{\mathbf{W}}+\left(\tilde{\mathbf{b}}\right)^T=\varphi(\mathbf{x})^T repmat\left(\tilde{\mathbf{w}}_0,1,m\right)+\varphi(\mathbf{x})^T\tilde{\mathbf{V}}+\left(\tilde{\mathbf{b}}\right)^T\\&=\varphi(\mathbf{x})^T repmat\left(\sum_{j=1}^{m}\mathbf{Z}\tilde{\boldsymbol{\alpha}}_j,1,m\right)+\frac{m}{\gamma''}\varphi(\mathbf{x})^T\mathbf{Z}\left(\tilde{\mathbf{A}}\right)+\tilde{\mathbf{b}}^T\\&=repmat\left(\sum_{j=1}^{m}\sum_{i=1}^{n}\tilde{\alpha}_{ij}K\left(\mathbf{x},\mathbf{x}_i\right),1,m\right)+\frac{m}{\gamma''}\sum_{i=1}^{n}\tilde{\alpha}_i K\left(\mathbf{x},\mathbf{x}_i\right)+\tilde{\mathbf{b}}^T,\end{aligned}\tag{3}$$

where $K\left(\mathbf{x},\mathbf{x}_i\right)$ is a kernel function. This research employs the Radial Basis Function (RBF) kernel function. A grid search method (Hsu et al., 2016) is utilized to identify the proper hyper-parameter of the MLS-SVR model. The optimal pair of hyper-parameters σ, γ', and γ'' is selected based on the criterion of the Minimum Mean Squared Error (MSE) value. The grid search method is carried out using all possible combinations of the kernel function parameter, $\sigma\in\left\{2^{-15},2^{-13},\ldots,2^{3}\right\}$, as well as regularized parameters, $\gamma''\in\left\{2^{-10},2^{-8},\ldots,2^{10}\right\}$ and $\gamma'\in\left\{2^{-5},2^{-3},\ldots,2^{15}\right\}$.

3. MCUSUM CONTROL CHART BASED ON THE RESIDUALS OF MLS-SVR

The observable output variables that

satisfy multivariate autocorrelated data are defined as $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_j, \ldots, \mathbf{y}_m$, where $\mathbf{y}_j = (y_{1j}, y_{2j}, \ldots, y_{nj})^T$ with $j = 1, 2, \ldots, m$ the number of output variables. Each output variable is assumed to have a significant partial autocorrelation function (PACF) until lag $p_1, p_2, \ldots, p_m$ so that the input variables of the MLS-SVR model are selected as

$$\mathbf{x} = \left(\mathbf{y}_{1,(i-1)}, \ldots, \mathbf{y}_{1,(i-p_1)}, \ldots, \mathbf{y}_{j,(i-1)}, \ldots, \mathbf{y}_{j,(i-p_j)}, \ldots, \mathbf{y}_{m,(i-1)}, \ldots, \mathbf{y}_{m,(i-p_m)} \right). \quad (4)$$

Let $\hat{f}(\mathbf{x}_j)$ be the decision function of the MLS-SVR model utilizing the optimal parameters as in Equation (3). The vector of the residual can be computed as $\mathbf{e}_j = \mathbf{y}_j - \hat{f}(\mathbf{x}_j)$. Hence, the $n \times m$ residual matrix $\mathbf{e}$ consists of e_{ij}, where $i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, m$.

Crosier (1988) presented an MCUSUM control chart in the form of the Hotelling T statistic, which is usually known as the Cumulative Sum of T (COT). If the residuals of the MLS-SVR model follow multivariate normal distribution $N_m(\boldsymbol{\mu}_e, \mathbf{V})$ then it can be transformed into Hotelling T statistics as follows:

$$T_i = [(\mathbf{e}_i - \boldsymbol{\mu}_e)^{\mathrm{T}} \mathbf{V}^{-1} (\mathbf{e}_i - \boldsymbol{\mu}_e)]^{1/2}. \quad (5)$$

Therefore, MLS-SVR residuals-based MCUSUM statistics can be calculated with the following equation:

$$C_i = \max[0, T_i - k + C_{i-1}], \quad (6)$$

where the initial value $C_0 = 0$ and the reference value $k > 0$. The MLS-SVR residuals-based MCUSUM control chart detects an out-of-control signal if the C_i statistic is greater than the upper control limit H.

The inputs of the MLS-SVR model are selected based on in-control processes (observations from Phase 1) using Equation (4); then the hyper-parameters are optimized using the grid search method so that the residuals satisfy the white noise condition. Once this assumption is satisfied, the value of H is estimated as in Crosier (1988). The application of MCUSUM on the MLS-SVR residual will ensure the stability and adaptability of the monitoring process. The optimal parameters and hyper-parameters obtained from Phase I can be utilized directly in the Phase II monitoring process. Furthermore, the researcher can adjust the level of tightness by setting reference value k in the MCUSUM statistics.

4. RESULTS AND DISCUSSION

This paper presents the application of the proposed MLS-SVR-based MCUSUM chart to monitor water quality data. Two prominent quality characteristics in the drinking water manufacturing process are the water turbidity and the chlorine residual. The concentration of dissolution and the existence of particles in a liquid are usually referred to as turbidity. Chlorination is affixing chlorine into contaminated water and is principally intended for killing the microbes. Turbidity is measured using Nephelometric Turbidity Units (NTU) whereas the chlorine residual is measured in ppm unit. Drinking water is safe from bacteria if it has a minimum of 0.2 ppm chlorine residual. However, the smell and taste of water are affected by exaggerated chlorine affixation.

This research monitors both water turbidity and chlorine residual data of the water manufacturing process in Surabaya, Indonesia. The water quality data in Phase I were monitored hourly from 19 August to 29 August 2016. Figure 1 exhibits the plots of the PACF for water quality data in Phase I with a 5% significance limit. The PACF of water turbidity ($\mathbf{y}_1$) is significant at lag 1, lag 2, lag 7, and lag 8, whereas the PACF of the chlorine residual ($\mathbf{y}_2$) is significant at lag 1, lag 3, and lag 9. Therefore, $\mathbf{y}_{1(i-1)}, \mathbf{y}_{1(i-2)}, \mathbf{y}_{1(i-7)}, \mathbf{y}_{1(i-8)}, \mathbf{y}_{2(i-1)}, \mathbf{y}_{2(i-3)}, \mathbf{y}_{2(i-9)}$ are chosen as the inputs of the MLS-SVR model for Phase I. These selected inputs along with the optimal combination of hyper-parameters $\gamma' = 2^{-5}$, $\gamma'' = 2^{-8}$, and $\sigma = 2^{3}$ produce the minimum MSE$= 6.05 \times 10^{-4}$. It is important to know that this research utilizes the Radial Basis Function (RBF) kernel function with a parameter σ.

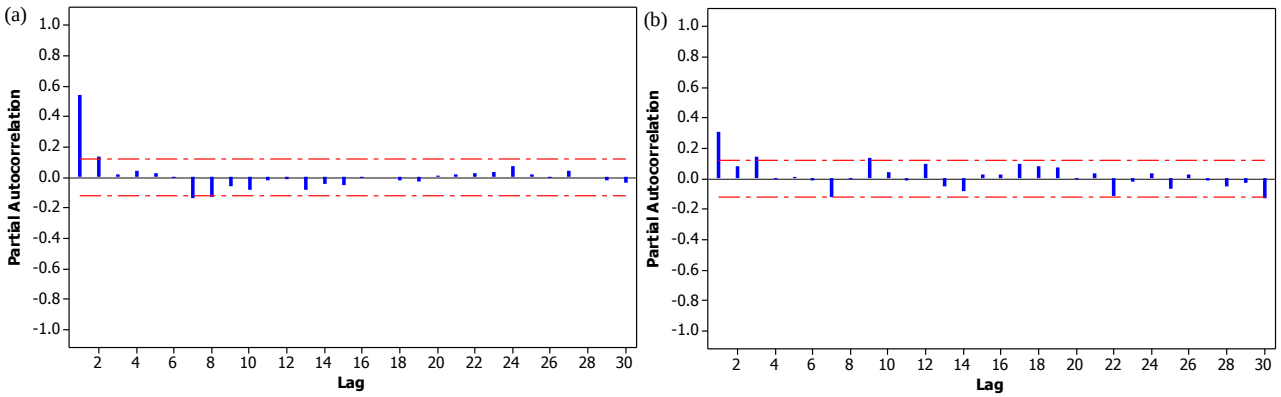


Figure 1: Partial autocorrelation function plots of (a) water turbidity and (b) chlorine residual in Phase I.

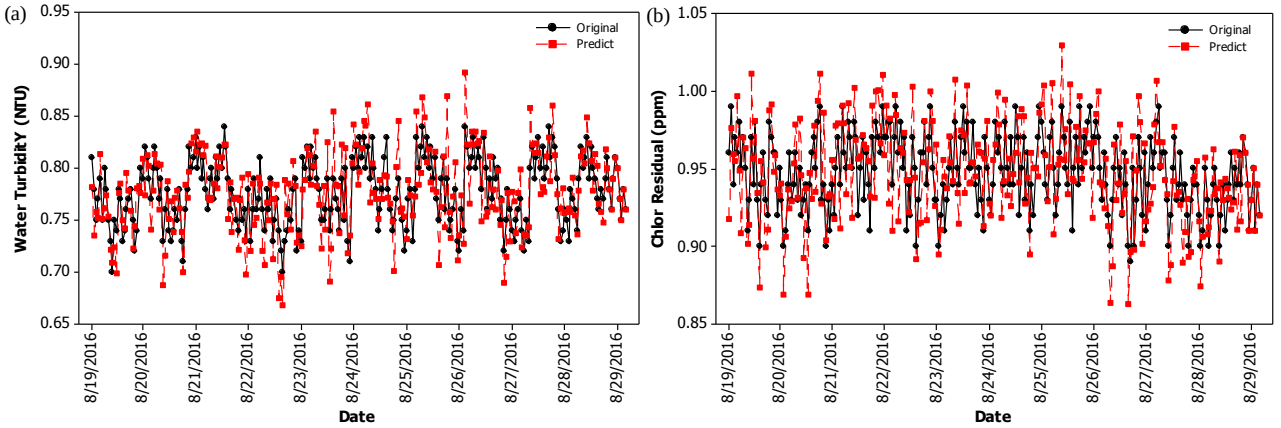


Figure 2: Time series plots of (a) water turbidity and (b) chlorine residual in Phase I.

The time series plots of water quality data in Phase I are displayed in Figure 2. The predicted value from the MLS-SVR model shows a similar pattern to the actual value. In addition, the residuals resulting from the MLS-SVR model for water quality data in Phase I satisfy the white noise assumption (confirmed by the ACF plots in Figure 3). The MLS-SVR residuals also fulfill a multivariate normal distribution. These residuals are then monitored using the MCUSUM control as displayed in Figure 4. It does not detect any out-of-control signal, so the upper control limit *H* and the parameters of the MLS-SVR-based MCUSUM control chart can be utilized in Phase II of the water manufacturing process.

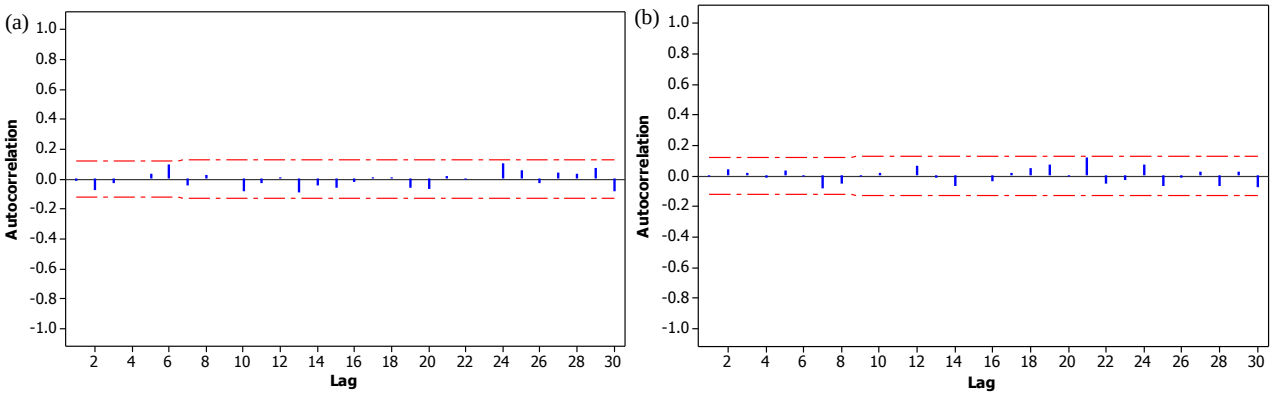


Figure 3: Autocorrelation function (ACF) plots of MLS-SVR residuals with 5% significance limit for (a) water turbidity and (b) chlorine residual in Phase I.

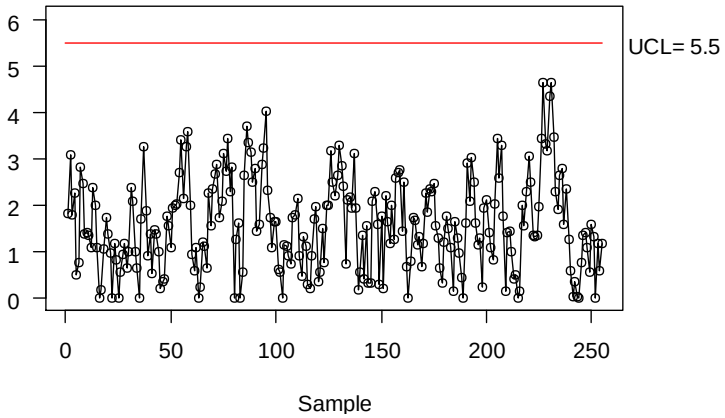


Figure 4: Monitoring water quality data in Phase I using MLS-SVR-based MCUSUM control chart with reference value k=0.5

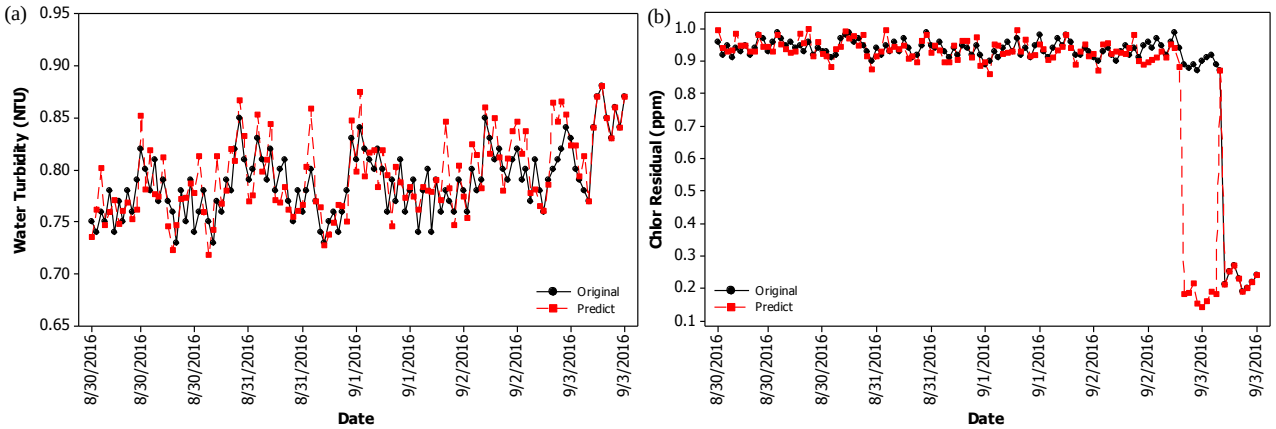


Figure 5: Time series plots of (a) water turbidity and (b) chlorine residual in Phase II.

The water quality data in Phase II were monitored hourly for five days starting from 30 August 2016. The time series plots of the actual water quality data and the predicted value from the MLS-SVR model in Phase II are displayed in Figure 5. The time series plots of actual water quality data starting from 10.00 AM on 3 September 2016 show an

unusual pattern. The increasing pattern in the water turbidity plot indicates more turbid water. On the contrary, the steeply decreasing shift in the chlorine residual plot points to the higher number of microbes in the water. Both MLS-SVR modeling needs some adaptation before it can follow the extreme changes in indicators reflect the worsened quality of the water. When the actual values are shifted significantly, the chlorine residual predictions are much less than the actual values (see Figure 5.b). This evidence indicates that actual data.

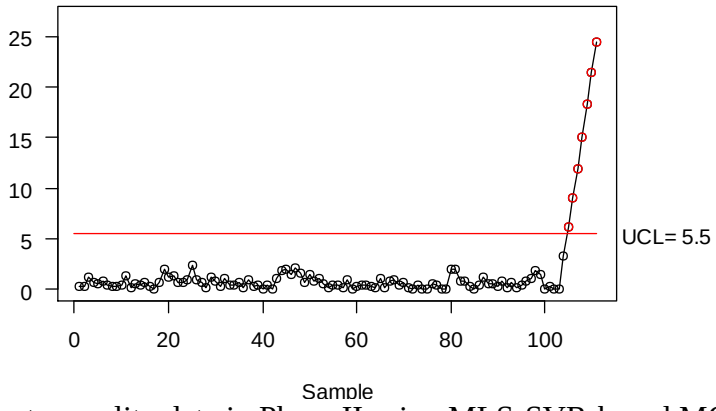


Figure 6: Monitoring water quality data in Phase II using MLS-SVR-based MCUSUM control chart with reference value k=0.5

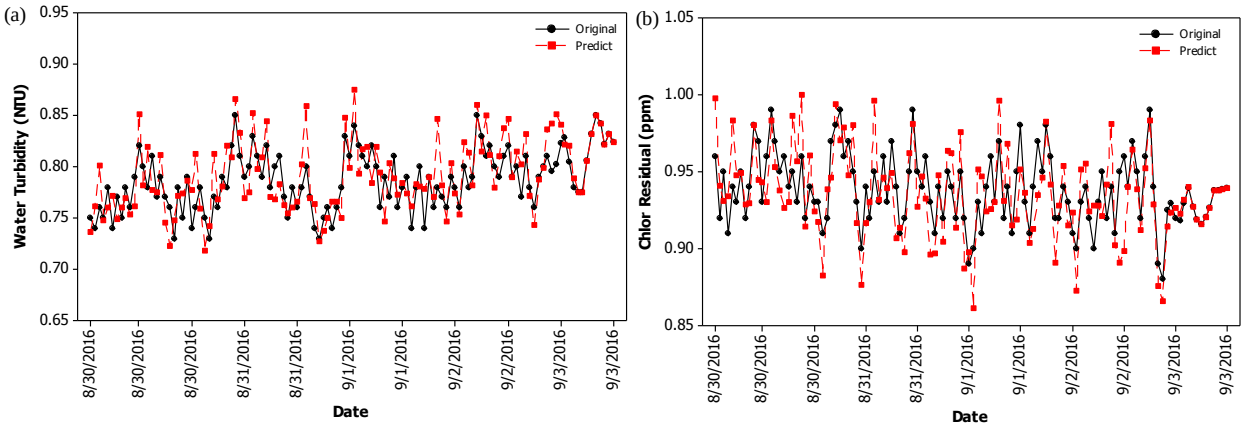


Figure 7: Time series plots of (a) water turbidity and (b) chlorine residual for updated Phase II.

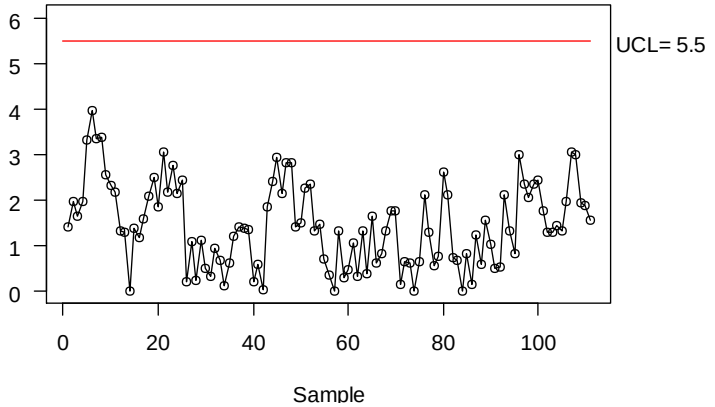


Figure 8: Monitoring updated water quality data in Phase II using MLS-SVR-based MCUSUM control chart with reference value k=0.5

The water quality data in Phase II are trained using the MLS-SVR algorithm by utilizing the optimal parameters and hyper-parameters produced in Phase I. As depicted in Figure 6, the last eight statistics of the MLS-SVR-based MCUSUM control chart in Phase II show significant out-of-control signals. This result indicates a serious problem in both the water turbidity and chlorine residuals data. It is necessary to look for the assignable cause in order to improve the water manufacturing process. At that time point, a broken pipeline was found in one of the water distribution areas. Pipeline maintenance and flow meter installation were then conducted in order to handle the water manufacturing problem.

The MLS-SVR-based MCUSUM control chart needs to be updated in order to fit the new observations in the further monitoring process. That is why the actual unusual pattern displayed in Figure 5 needs to be replaced with the predicted value from the MLS-SVR model. The updated water quality data in Phase II (see Figure 7) consist of actual water quality data in Phase II, except for the data starting from 10.00 AM on 3 September 2016. The actual data in those periods are replaced by the predicted values from the MLS-SVR model. The MLS-SVR-based MCUSUM control chart for the updated Phase II is exhibited in Figure 8. All of the MLS-SVR-based MCUSUM statistics are assigned as being in the in-control condition such that these updated data can be used in the next Phase I monitoring process.

5. CONCLUSION

This paper develops a MCUSUM control chart based on the residual of the MLS-SVR model for monitoring the mean vector of time series data. The inputs of the MLS-SVR model are determined based on the significant lag of the PACF. The appropriate inputs and the optimal combination of hyper-parameters yield the residuals of the MLS-SVR model that fulfill the white noise assumption. The MCUSUM control chart based on the residual of MLS-SVR that is used to monitor the water quality data indicates that corrective actions should be carried out in order to improve the water manufacturing process. Evaluating the performance of the proposed control chart using the Average Run Length (ARL) criterion may be useful in future research. Similarly, optimizing the SVR parameters using an evolutionary algorithm (Härdle et al., 2014) may also be useful in future work.

6. ACKNOWLEDGEMENT

This research was supported by Research, Technology, and Higher Education Ministry, the Republic of Indonesia through PMDSU Scheme [Grant number 128/SP2H/PTNBH/DRPM/2018].

7. REFERENCES

Bodnar O., and Schmid W. (2007). Surveillance of the mean behavior of multivariate time series, *Statistica Neerlandica* 61 (4):383–406.

Chan L K., and Li G-Y. (1994). A multivariate control chart for detecting linear trends, *Communications in Statistics-Simulation and Computation* 23 (4):997–1012.

Charnes J M. (1995). Tests for special causes with multivariate autocorrelated data, *Computers and Operations Research* 22 (4):443–453.

Crosier R B. (1988). Multivariate generalizations of cumulative sum quality-control schemes, *Technometrics* 30 (3):291–303.

Härdle W K., Prastyo D D., and Hafner C M. (2014). Support vector machines with

evolutionary model selection for default prediction. *The Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics*. Oxford University Press.

Harris T J., and Ross W H. (1991). Statistical process control procedures for correlated observations, *Canadian Journal of Chemical Engineering* 69 (1):48–57.

Healy J D. (1987). A note on multivariate CUSUM procedures, *Technometrics* 29 (4):409–412.

Hsu C W., Chang C C., and Lin C J. (2016). A practical guide to support vector classification. *National Taiwan University*.

Hwang C. (2016). Multioutput LS-SVR based residual MCUSUM control chart for autocorrelated process, *Journal of the Korean Data and Information Science Society* 27 (2):523–530.

Issam B K., and Mohamed L. (2008). Support vector regression based residual MCUSUM control chart for autocorrelated process, *Applied Mathematics and Computation* 201 (1–2):565–574.

Johnson R A., and Bagshaw M. (1974). The effect of serial correlation on the performance of CUSUM tests, *Technometrics* 16 (1):103–112.

Kalgonda A A., and Kulkarni S R. (2004). Multivariate quality control chart for autocorrelated processes, *Journal of Applied Statistics* 31 (3):317–327.

Khediri I B., Weihs C., and Limam M. (2010). Support vector regression control charts for multivariate nonlinear autocorrelated processes, *Chemometrics and Intelligent Laboratory Systems* 103 (1):76–81.

Khusna H., Mashuri M., Prastyo D D., and Ahsan M. (2018). Multioutput least square SVR based multivariate EWMA control chart, *Journal of Physics: Conference Series* 1028:12221. IOP Publishing.

Kramer H G., and Schmid L V. (1997). EWMA charts for multivariate time series, *Sequential Analysis* 16 (2):131–154.

Liu G., Lin Z., and Yu Y. (2009). Multi-output regression on the output manifold, *Pattern Recognition* 42 (11):2737–2743.

Lowry C A., Woodall W H., Champ C W., and Rigdon S E. (1992). A multivariate exponentially weighted moving average control chart, *Technometrics* 34 (1):46–53.

Ngai H-M., and Zhang J. (2001). Multivariate cumulative sum control charts based on projection pursuit, *Statistica Sinica* 747–766.

Noorossana R., and Vaghefi S J M. (2006). Effect of autocorrelation on performance of the MCUSUM control chart, *Quality and Reliability Engineering International* 22 (2), 191–197.

Pignatiello J J., and Runger G C. (1990). Comparisons of multivariate CUSUM charts, *Journal of Quality Technology* 22 (3):173–186.

Psarakis S., and Papaleonida G. (2007). SPC procedures for monitoring autocorrelated processes, *Quality Technology and Quantitative Management* 4 (4):501–540.

Sato J R., Costafreda S., Morettin P A., and Brammer M J. (2008). Measuring time series predictability using support vector regression, *Communications in Statistics-Simulation and Computation* 37 (6):1183–1197.

Śliwa P., and Schmid W. (2005). Monitoring the cross-covariances of a multivariate time series, *Metrika* 61 (1):89–115.

Suykens J A K., Van-Gestel T., De-Brabanter J., De-Moor B., and Vandewalle J. (2002). Least squares support vector machines. *World Scientific*.

Suykens J A K., and Vandewalle J. (1999). Multiclass least squares support vector machines, *IEEE* 2:900-903.

Theodossiou P. (1993). Predicting shifts in the mean of a multivariate time series process: an application in predicting business failures, *Journal of the American Statistical Association* 88:441–449.

Thissen U., Van-Brakel R., De-Weijer A P., Melssen W J., and Buydens L M C. (2003). Using support vector machines for time series prediction, *Chemometrics and Intelligent Laboratory Systems* 69 (1–2):35–49.

Tuia D., Verrelst J., Alonso L., Perez-Cruz F., and Camps-Valls G. (2011). Multioutput support vector regression for remote sensing biophysical parameter estimation, *IEEE Geoscience and Remote Sensing Letters* 8 (4):804–808.

Vanbrackle L N., and Reynolds M R. (1997). EWMA and CUSUM control charts in the presence of correlation, *Communications in Statistics-Simulation and Computation* 26 (3):979–1008.

Vapnik V N. (1998). *Statistical Learning Theory*. *John Wiley & Sons*.

Vapnik V N. (2000). *The Nature of Statistical Learning Theory* 8.

Woodall W H., and Montgomery D C. (1999). Research issues and ideas in statistical process control, *Journal of Quality Technology* 31 (4):11.

Wororomi J K., Mashuri M., Irhamah, and Arifin A Z. (2014). On monitoring shift in the mean processes with vector autoregressive residual control charts of individual observation, *Applied Mathematical Sciences* 8:3491–3499.

Xu S., An X., Qiao X., Zhu L., and Li, L. (2013). Multi-output least-squares support vector regression machines, *Pattern Recognition Letters* 34 (9):1078–1084.

Yashchin E. (1993). Performance of CUSUM control schemes for serially correlated observations. *Technometrics* 35 (1):37–52.

Clustering of Rainfall Distribution Patterns in Peninsular Malaysia using Time Series Clustering Method

Noratiqah Mohd Ariff[1a], Mohd Aftar Abu Bakar[1b]*, Sharifah Faridah Syed Mahbar[2c], Mohd Shahrul Mohd Nadzir[3,4d]

[1] School of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, MALAYSIA. E-mail: tqah@ukm.edu.my[a]; aftar@ukm.edu.my[b]
[2] Pusat Operasi Cuaca & Geofizik Nasional, Jabatan Meteorologi Malaysia, Kementerian Tenaga, Sains, Teknologi, Alam Sekitar & Perubahan Iklim, Jalan Sultan, 46667 Petaling Jaya, Selangor, MALAYSIA. E-mail: aridah_mahbar@met.gov.my[c]
[3] School of Environmental and Natural Resource Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600, UKM Bangi, Selangor, MALAYSIA.
[4] Centre for Tropical Climate Change System, Institute of Climate Change, Universiti Kebangsaan Malaysia, 43600, UKM Bangi, Selangor, MALAYSIA. E-mail: shahrulnadzir@ukm.edu.my[d]
* Corresponding Author: aftar@ukm.edu.my[b]
Received: 21st April 2019 Revised: 6th August 2019 Published: 30th September 2019
DOI : https://doi.org/10.22452/mjs.sp2019no2.8

ABSTRACT Time series clustering technique was used in this study to categorize the locations in Peninsular Malaysia according to the similarity of rainfall distribution patterns. Daily rainfall time series data from 12 meteorological observation stations across Peninsular Malaysia have been considered for this study. Four dissimilarity measure methods were examined and compared in terms of accuracy and suitability, namely Euclidean distance (ED), complexity-invariant distance (CID), correlation-based distance (COR) and integrated periodogram-based distance (IP). The average silhouette width (ASW) was used to determine the optimal group number for the rainfall time series data. Using Ward's hierarchical clustering method, this study found that the rainfall time series in Peninsular Malaysia can be divided into four regions of homogeneous climate zones. Based on the results, the IP was the most suitable dissimilarity measures for clustering rainfall time series data in Peninsular Malaysia, except during the Southwest Monsoon where the COR performed better.

Keywords: time series clustering, dissimilarity measures, rainfall patterns, Peninsular Malaysia.

1. INTRODUCTION

Accuracy in weather forecasting helps to contribute to the nation socioeconomic activities and development. The weather reports are used in planning and decision making for matters related to disaster management, water management, agriculture, industry and tourism. Clustering technique is one of the effective data mining techniques to extract useful information. It is important to identify the set of objects whose class is unknown in data mining. This has been applied in the study of taxonomy, agriculture, remote sensing and process control (Kavitha & Punithavalli, 2010), as well as meteorology study to determine and classify rainfall patterns (Munoz-Diaz & Rodrigo, 2004; Soltani & Modarres, 2006).

Time series clustering is a technique which can partition time series data into groups based on its similarity or distance. Time series clustering has been used for recognizing dynamic changes in time series, discovering patterns, prediction and

recommendation in many field of studies such as in climate, energy, environment, finance and medicine (Aghabozorgi et al., 2015; Rani & Sikka, 2012). Ahmad et al. (2013) used the hierarchical clustering approach to regionalise the daily rainfall data in Peninsular Malaysia. However, they do not consider the seasonal factor, which is crucial for the Malaysian climate. This was conducted by clustering the time series data only during the Northeast Monsoon (or Southwest Monsoon), instead of clustering the whole time series.

In this study, the rainfall time series data from 12 meteorological stations in Peninsular Malaysia from 1970 to 2014 (45 years) were analysed using clustering technique. This study examined and compared four dissimilarity measure methods used to cluster the rainfall time series in Malaysia according to homogenous climate zone.

2. RAINFALL DATA

Malaysia is a country located near to the equator, divided into two regions which are the Peninsular Malaysia and East Malaysia separated by the South China Sea. The climate is hot and humid throughout the year with heavy rainfalls. There are two monsoon seasons, the Southwest Monsoon (May to August), where the east coast of Peninsular Malaysia, west of Sarawak and east coast of Sabah have more rainfalls, and the Northeast Monsoon (November to February), where the rainfall occurrence is lesser at the east coast of Peninsular Malaysia. The total precipitation is between 2000 and 4000 mm annually.

Twelve Malaysian Meteorological Department (MMD) observation stations that cover three zones in Peninsular Malaysia were selected in this study. Details for each station and its location is depicted in Table 1 and Figure 1. The daily time series rainfall data from 1970 until 2014 were used in this analysis.

Table 1: Stations details, location and percentage of missing data for each station.

Station	Station Code	Latitude (°N)	Longitude (°E)	Mean Sea Level (MSL) (m)	Missing Data (%)
Alor Setar	48603	6.2	100.4	3.9	-
Bayan Lepas	48601	5.3	100.2667	2.5	-
Kota Bharu	48615	6.1667	102.3	4.4	-
Hospital Dungun	49476	4.7667	103.4167	3	2.46
Kuantan	48657	3.7667	103.2167	15.2	-
Mersing	48674	2.45	103.8333	43.6	-
Subang	48647	3.1333	101.55	16.6	-
Malacca	48665	2.2667	102.25	8.5	-
Hospital Baling	41545	5.6833	100.9167	52	0.42
Ipoh	48625	4.5667	101.1	40.1	-
Hospital Tapah	43421	4.2	101.2667	35	0.27
Sitiawan	48620	4.2167	100.7	6.8	-

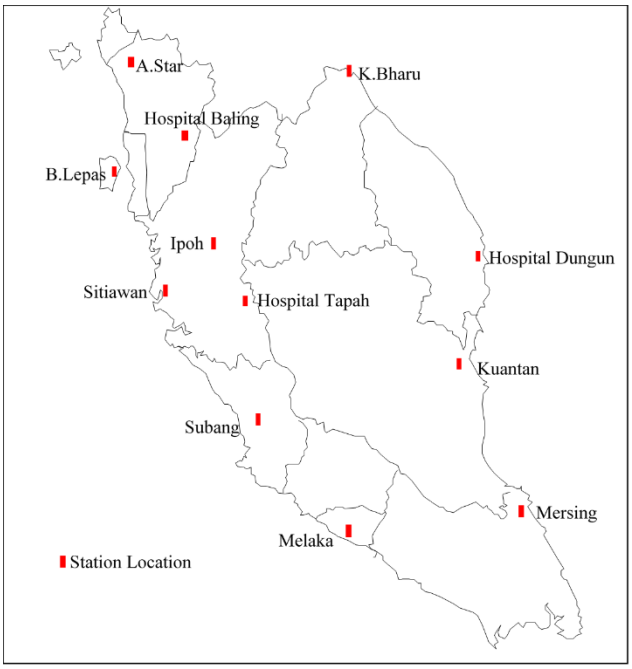


Figure 1: Locations of rain gauge stations.

3. TIME SERIES CLUSTER ANALYSIS

Cluster analysis is a technique that groups certain observations with similar characteristics or traits when the true group is unknown. Cluster analysis is applied in various data types, for example numerical data (Michinaka et al., 2011), image data (Arifin & Asano, 2006) and text data (Ariff et al., 2018). Time series clustering have been used in many areas of hydrology, such as to determine and group stations according to its homogeneous climate areas (DeGaetano,2001) or time frame according to a cluster that represents weather events or patterns (Ramos, 2001).

Generally, there are three types of time series, which are whole time series clustering, sub-sequence time-series clustering and time-point clustering (Aghabozorgi et al., 2015). For this study, only whole time series clustering will be considered since the purpose is to compare several meteorological observation stations rainfall time series data with respect to their similarity. Han et al. (2012) have classified clustering methods into five categories:

- partitioning method
- hierarchical method
- probabilistic model-based method
- density-based method
- grid-based method

The first three methods were used directly or modified for time series clustering. Partition clustering aims to separate set of objects into consistent group. At first, the objects will be placed randomly and later transferred into another cluster until being positioned in an almost similar group while for hierarchical clustering, each object is defined as a single group. Then, each object (group) will be merged to form a new one. The merging process continues until only one group is left.

In this study, Ward's hierarchical clustering was used to cluster the rainfall time series data in Peninsular Malaysia. Several studies have shown that Ward's approach is suitable for clustering the rainfall data since the clusters do not have to be equiprobable

which imply that the number of stations in each cluster does not have to be equal. (Ramos, 2001; Tennant & Hewitson, 2002; Crétat et al., 2012).

The use of Ward's method in hierarchical clustering is to minimise the loss of information resulted from the combination of clusters. At each stage, the combination of each pair of possible clusters is considered and the combination of two clusters will increase the sum of squared errors (SSE). Eventually, all clusters will be combined into one large cluster with larger SSE value.

3.1 Dissimilarity Measures

The most important step prior to algorithm clustering is to generate numerical similarity and dissimilarity measures to characterise relationships between data (Munoz-Diaz & Rodrigo, 2004; Prasanna, 2012). According to Lin & Li (2009), the similarity or dissimilarity between time series can be based on shape or structure concepts. The dissimilarity shape concept measures the similarity or dissimilarity based on the geometric of the series; this concept was commonly known as model free approachwhile the structure concept, also known as model based approach measure the dissimilarity based on the global underlying structure of the series.

Three model free dissimilarity measures have been selected in this study which are Euclidean distance (ED), correlation-based distance (COR) and integrated periodogram-based distance (IP). The complexity-invariant distance (CID) which is a model-based dissimilarity measure was also considered in this study.

Euclidean distance is the most common and easiest shape based dissimilarity measure for time series data. ED is calculated by

$$\mathrm{ED}(\mathbf{X_T}, \mathbf{Y_T}) = \sqrt{\sum_{t=1}^{T} \left(X_t - Y_t \right)^2}, \tag{1}$$

where $\boldsymbol{X}_T$ and $\boldsymbol{Y}_T$ are two different time series.

Pearson correlation coefficient is selected in this study as the correlation-based dissimilarity measure. Highly correlated values mean that the distance is close and the formulae Pearson correlation is given as follows;

$$\mathrm{COR}(\mathbf{X_T}, \mathbf{Y_T}) = \frac{\sum_{t=1}^{T} \left(X_t - \bar{X}_T \right)\left(Y_t - \bar{Y}_T \right)}{\sqrt{\sum_{t=1}^{T-1} \left(X_T - \bar{X}_T \right)^2} \sqrt{\sum_{t=1}^{T-1} \left(Y_T - \bar{Y}_T \right)^2}} \tag{2}$$

Periodogram method is used to determine the dominant time period and frequency for a time series. This technique is also used to analyse periodic data by transforming the data into frequency waves. De Lucas (2010) discussed the distance measure on cumulative periodogram known as integrated periodogram (IP). IP calculates the distance difference between two time series in terms of cumulative periodogram. The advantage of this method over the basic periodogram is it can determine the entire stochastic processes that occur in the time series sequence. The steps to calculate IP is given as

$$\mathrm{IP}(\mathbf{X_T}, \mathbf{Y_T}) = \int_{-\pi}^{\pi} \left| F_{X_T}(\lambda) - F_{Y_T}(\lambda) \right| d\lambda, \quad \lambda \in [-\pi, \pi] \tag{3}$$

where

$$F_{X_T}(\lambda_j) = C_{X_T}^{-1} \sum_{i=1}^{j} I_{X_T}(\lambda_i), \qquad C_{X_T} = \sum_i I_X(\lambda_i)$$

$$F_{Y_T}(\lambda_j) = C_{Y_T}^{-1} \sum_{i=1}^{j} I_{Y_T}(\lambda_i), \qquad C_{Y_T} = \sum_i I_Y(\lambda_i)$$

$$I_{X_T}(\lambda_k) = T^{-1} \left| \sum_{t=1}^{T} X_T e^{-i\lambda_k t} \right|^2$$

$$I_{Y_T}(\lambda_k) = T^{-1} \left| \sum_{t=1}^{T} Y_T e^{-i\lambda_k t} \right|^2$$

$$\lambda_k = \frac{2\pi k}{T}, \qquad k = 1, \ldots, n, \qquad n = \left[\frac{T-1}{2} \right]$$

with

T = vector length, $T \geq 1$
$I_{X_T}(\lambda_k)$ = periodogram for $\mathbf{X_T}$
$I_{Y_T}(\lambda_k)$ = periodogram for $\mathbf{Y_T}$.

Batista et al. (2014) introduced the CID time series measure, which improves the classification and clustering accuracy without compromising the efficiency. CID measures the complexity difference between two time series. It is a ratio of complexity of one time series to another (the less complex one). Complexity correction factor (CF) will be closer to one if both series have similar complexity level or greater than one if the complexity level of both series is different. CID is calculated by

$$\mathrm{CID}(\mathbf{X_T}, \mathbf{Y_T}) = \mathrm{ED}(\mathbf{X_T}, \mathbf{Y_T}) \times \mathrm{CF}(\mathbf{X_T}, \mathbf{Y_T}) \tag{4}$$

where

$$\mathrm{CF}(\mathbf{X_T}, \mathbf{Y_T}) = \frac{\max\left(\mathrm{CE}(X_T), \mathrm{CE}(Y_T)\right)}{\min\left(\mathrm{CE}(X_T), \mathrm{CE}(Y_T)\right)}$$

and the complexity estimate is

$$\mathrm{CE}(\mathbf{X_T}) = \sqrt{\sum_{t=1}^{T-1} \left(X_t - X_{t+1}\right)^2}\,. \tag{5}$$

3.2 Average Silhouette Width

The optimal number of clusters, *k*, for a dataset is determined in clusterisation process. Out of several ways to determine the *k* value in this study, the average silhouette width (ASW) was selected.

At first, the average distance for each subject in similar cluster is calculated. Cluster member with the lowest distance shows that the difference between subjects is minimal and can be clustered together. Then, the average distance for each subject will be compared to the average distance of neighbouring cluster members. The difference in ratio obtained from the member's dissimilarity point in the same cluster to the nearest neighbouring cluster is known as the silhouette value. The overall silhouette value is calculated by looking for the average silhouette of each member. This measure the similarity level of cluster members. The ASW value obtained is used to determine the optimal cluster number, *k*, of a dataset.

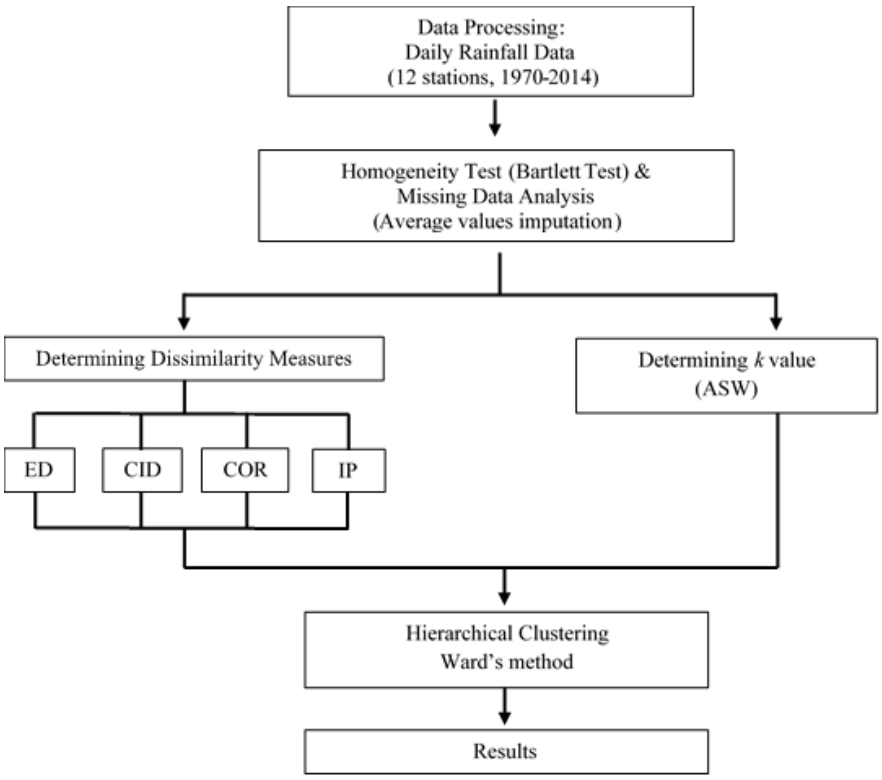


Figure 2: Methodology flowchart.

Table 2: Bartlett's test results of the yearly rainfall time series (1970-2014) according to each zone.

Station	Bartlett's K-squared test statistics	d.o.f	*p*-value
East Zone:			
1. Kota Bharu			
2. Hospital Dungun	1.2171	3	0.7489
3. Kuantan			
4. Mersing			
Northwest Zone:			
1. Alor Setar			
2. Hospital Baling	2.5603	2	0.278
3. Bayan Lepas			
West I Zone:			
1. Ipoh			
2. Hospital Tapah	0.6254	2	0.7315
3. Subang			
West II Zone:			
1. Sitiawan	0.0883	1	0.7663
2. Melaka			

4. ANALYSIS AND RESULTS

Figure 2 summarise the flow of analysis process in this study. After rainfall data was processed, Bartlett's test was used to check the homogeneity of variance of the time series data. This is to ensure that the data is of high quality to make sure the results are highly reliable. According to Table 2, the *p*-value for all the tests is not significant. Thus, this no evidence of unequal homogeneity variance within the stations in each cluster. This may imply that the time series data in each clusters have no inhomogeneity issue.

The results for each dissimilarity measure used in clusterisation using Ward's method were compared where the value closer to one is regarded as the most suitable dissimilarity measure for the time series data. At each station, there are three sets of time series data, representing the overall time series data and time series for both monsoon seasons. Using the ASW value, the optimal number of clusters is $k = 4$ (Figure 3).

The dissimilarity measure results are summarised in Table 3, in which IP is the best dissimilarity measure for cluster analysis of the entire and NEM time series data. The values obtained are closer to one, showing that the real data cluster partitioning is reflected from the model. CID is also suitable for NEM time series while for SWM, COR is the best dissimilarity measure. The simplest dissimilarity measure, ED, does not provide better results for any time series.

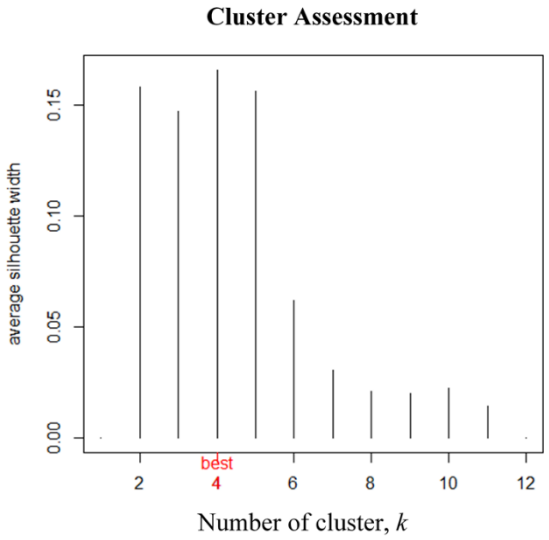


Figure 3: Optimal number of clusters from the ASW method.

Figure 4 shows the hierarchical clustering in the form of dendrograms of the overall time series using all dissimilarity measures. The *y*-axis refers to the difference or dissimilarity between each cluster where the longer the vertical line, the larger the difference between clusters. From this, it is shown that cluster analysis results using IP and COR dissimilarity measures are almost similar. The percentage of stations of each cluster for all types of dissimilarity measures are tabulated in Table 4. Geographical factor and the station locations play a role in determining the clusters, as shown by the clusterisation map in Figure 5, which is based on the IP clusterisation

.

Table 3: Dissimilarity measure results of overall, Northeast Monsoon (NEM) and Southwest Monsoon (SWM) rainfall time series data.

Dissimilarity Measure	**Time Series**		
Distance	**Overall**	**NEM**	**SWM**
ED	0.4977	0.4977	0.5583
CID	0.5857	0.6167	0.425
IP	0.7292	0.6167	0.6167
COR	0.6792	0.5778	0.6786

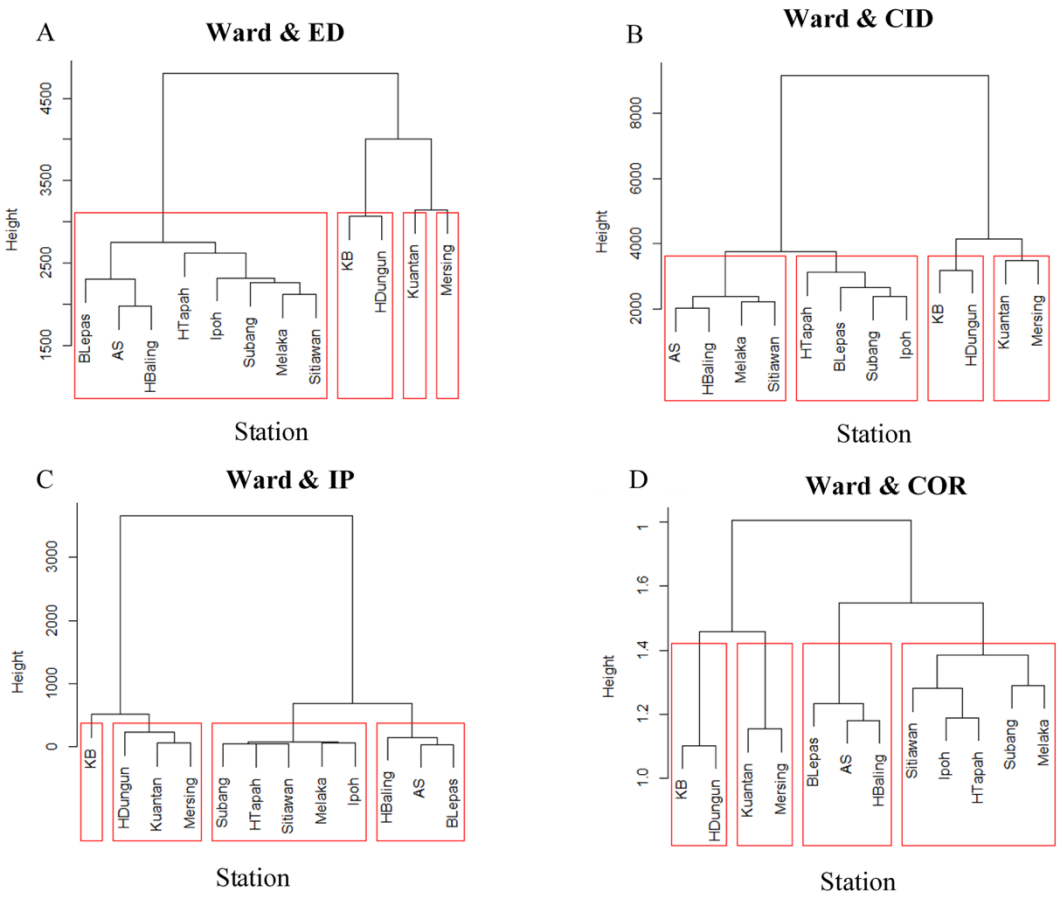


Figure 4: Dendrograms of time series data in 12 stations with dissimilarity measures: ED (A), CID (B), IP (C) and COR (D).

Table 4: The percentage number of stations for each cluster with different dissimilarity measures.

Cluster	Dissimilarity Measures			
	ED	**CID**	**IP**	**COR**
#1	66.70%	33.30%	41.70%	41.70%
#2	16.70%	33.30%	25.00%	25.00%
#3	8.30%	16.70%	25.00%	16.70%
#4	8.30%	16.70%	8.30%	16.70%

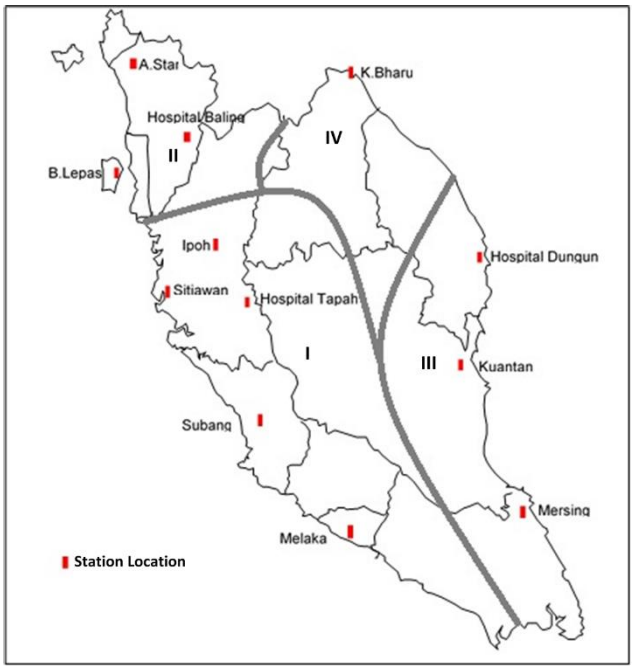


Figure 5: Clusterisation map of the Peninsular Malaysia rainfall data using IP dissimilarity measure based on the overall data.

Figure 6 shows the cluster analysis dendrograms of NEM time series data where CID and IP distance measures produce similar results. During NEM, the east coast area of Peninsular Malaysia receives a lot of rain, thus influencing the cluster analysis results. The percentage number of stations of each cluster for NEM time series data is illustrated in Table 5 and the clusterisation map is depicted in Figure 7.

The cluster analysis dendrograms of SWM time series data is shown in Figure 8, which is different than the NEM time series data. For SWM, the occurrence of rain is lower than the NEM, this significantly influences the determination of clusters than the NEM time series data clusters. For the SWM time series data, the percentage number of stations of each cluster is illustrated in Table 6 and the clusterisation map is depicted in Figure 9.

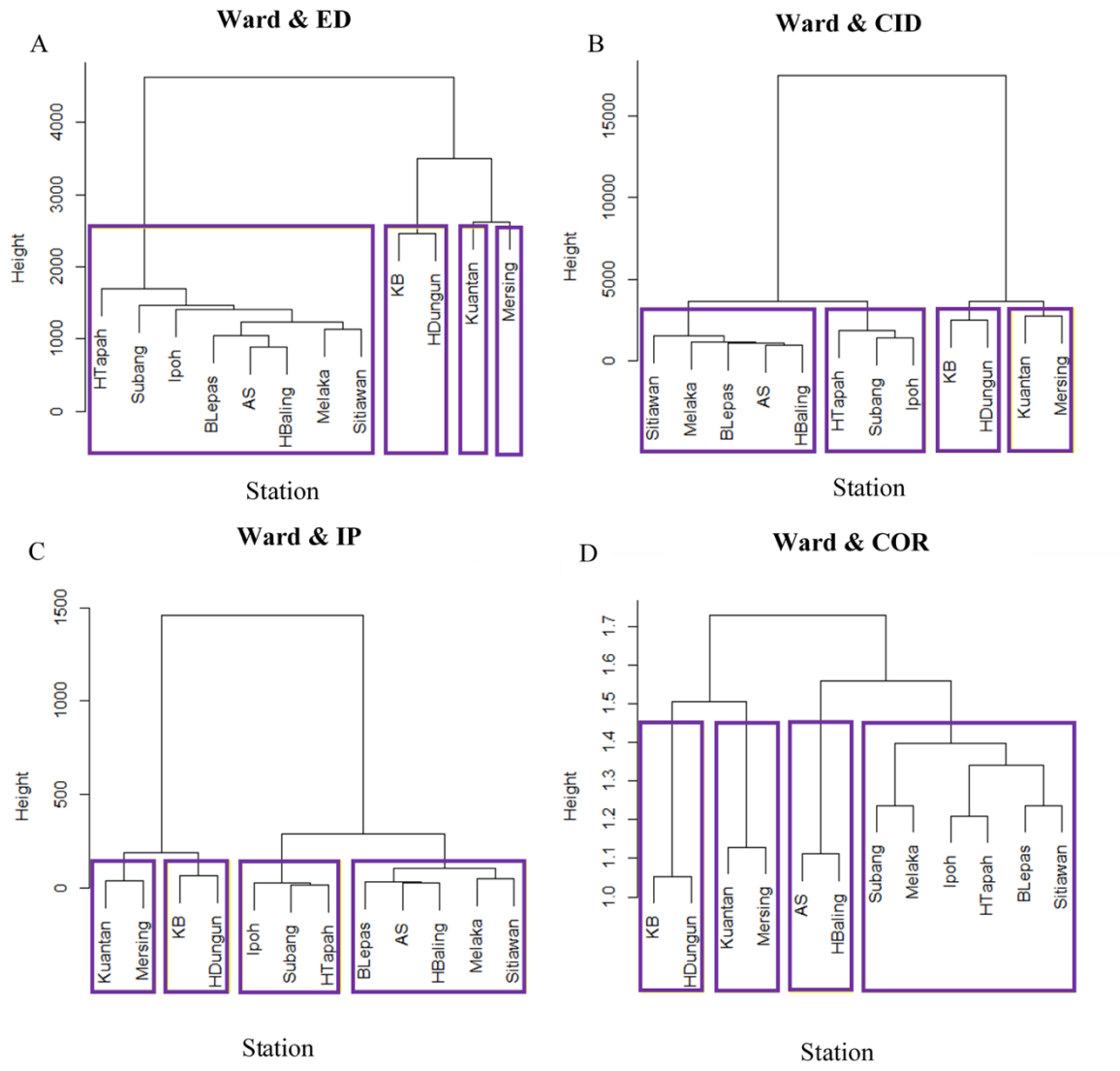


Figure 6: Dendrograms of the Northeast Monsoon (NEM) time series data in 12 stations with dissimilarity measures; ED (A), CID (B), IP (C) and COR (D).

Table 5: The percentage number of stations for each cluster with different dissimilarity measures for the NEM data.

Cluster	Dissimilarity Measure Distance			
	ED	CID	IP	COR
#1	66.70%	41.70%	41.70%	50.00%
#2	16.70%	25.00%	25.00%	16.70%
#3	8.30%	16.70%	16.70%	16.70%
#4	8.30%	16.70%	16.70%	16.70%

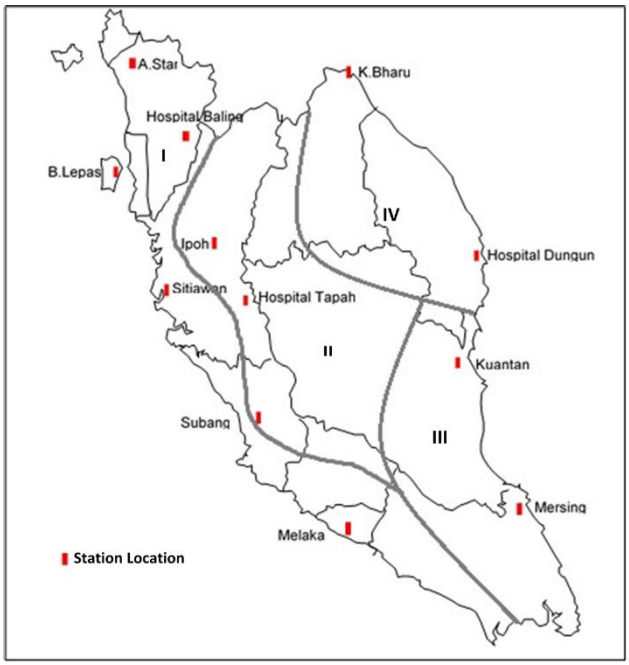


Figure 7: Clusterisation map of the Peninsular Malaysia rainfall data using CID or IP dissimilarity measure based on the NEM data.

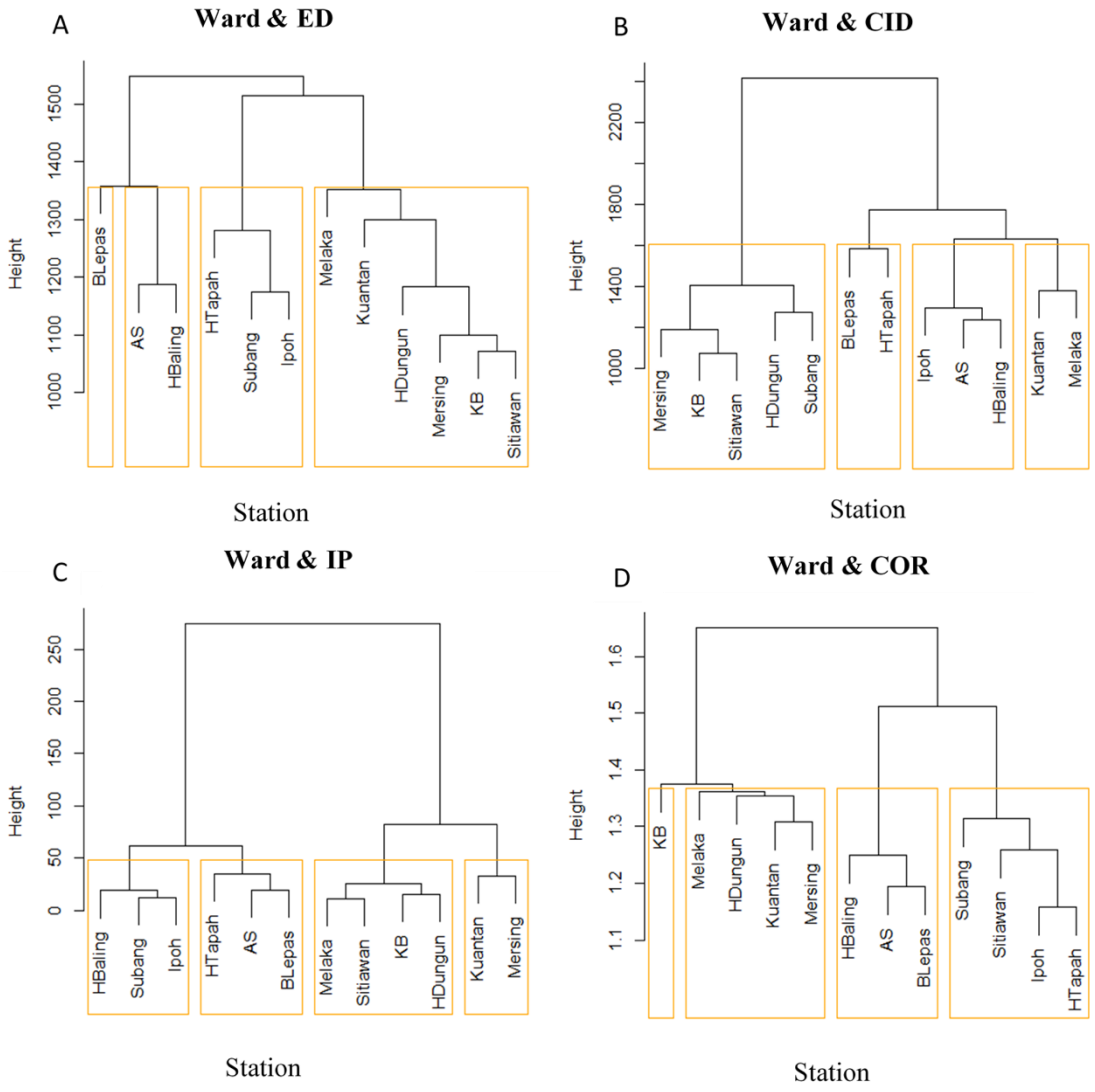


Figure 8: Dendrograms of the Southwest Monsoon (SWM) time series data in 12 stations with dissimilarity measures; ED (A), CID (B), IP (C) and COR (D).

Table 6: The percentage number of stations of each cluster with different dissimilarity measures for the SWM data.

Cluster	Dissimilarity Measure Distances			
	ED	CID	IP	COR
#1	50.00%	41.70%	33.30%	33.30%
#2	25.00%	25.00%	25.00%	33.30%
#3	16.70%	16.70%	25.00%	25.00%
#4	8.30%	16.70%	16.70%	8.30%

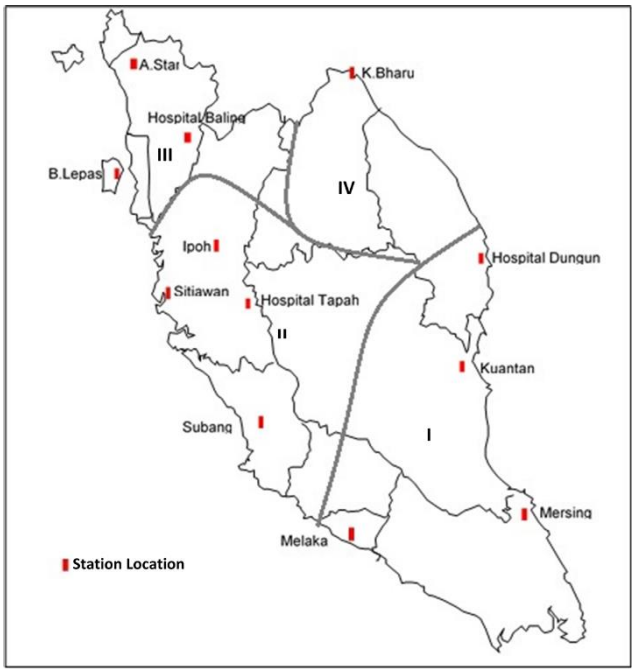


Figure 9: Clusterisation map of the Peninsular Malaysia rainfall data using COR dissimilarity measure based on the SWM data.

5. CONCLUSION

This study shows that the time series clusterisation can be used to study the rainfall pattern in Malaysia. Given that the Malaysian climate has two monsoon seasons, the cluster analysis should be done separately. For Peninsular Malaysia, the optimal cluster number is four, in which Peninsular Malaysia is divided into four homogeneous climate zones, especially the northwest and east coast regions. Factors such as geographical region, locations and precipitation rate play a role in determining the clusters. IP dissimilarity measure is the most suitable measure for the analysis on the overall time series data, while IP and CID are for the NEM data and COR is for the SWM. From the results, it is concluded that Ward's method is useful to cluster the Malaysian rainfall time series data. This approach can be extended by using other clustering techniques such as wavelet clustering (Singhal & Seborg, 2005) and can be used for storm event clustering (Ariff et al., 2016).

6. ACKNOWLEDGEMENT

The authors would like to thank Universiti Kebangsaan Malaysia for allocating the research grants (GGPM-2015-026 and GGPM-2017-124) and its facilities for this research.

7. REFERENCES

Aghabozorgi, S., Shirkhorshidi, A.S. & Wah, T.Y. (2015). Time-series clustering–A decade review. *Information Systems,* 53: 16-38.

Ahmad N.H., Othman I.R. & Deni S.M. (2013). Hierarchical cluster approach for regionalization of Peninsular Malaysia based on the precipitation amount. *Journal of Physics: Conference Series,* 423(1): 12-18.

Ariff N.M., Bakar M.A.A. & Rahmad M.I. (2018). Comparative study of

document clustering algorithms. *International Journal of Engineering and Technology (UAE),* 7(4): 246-251.

Ariff N.M., Jemain A.A. & Bakar M.A.A. (2016). Regionalization of IDF curves with L-moments for storm events. *International Journal of Mathematical and Computational Sciences*, 10: 217-223.

Arifin A.Z. & Asano A. (2006). Image segmentation by histogram thresholding using hierarchical cluster analysis. *Pattern Recognition Letters,* 27(13): 1515-1521.

Batista G.E., Keogh E.J., Tataw O.M. & De Souza V.M. (2014). CID: an efficient complexity-invariant distance for time series. *Data Mining and Knowledge Discovery*, 28(3): 634-669.

Crétat J., Richard Y., Pohl B., Rouault M., Reason C. & Fauchereau N. (2012). Recurrent daily rainfall patterns over South Africa and associated dynamics during the core of the austral summer. *International Journal of Climatology,* 32(2): 261-273.

De Lucas D.C. (2010). *Classification Techniques for Time Series and Functional Data*. Universidad Carlos III de Madrid. Doctoral dissertation.

DeGaetano A.T. (2001). Spatial grouping of United States climate stations using a hybrid clustering approach. *International Journal of Climatology,* 21(7): 791-807.

Han J., Pei J. & Kamber M. (2012). *Data Mining: Concepts and Techniques 3rd Edition*. Waltham, M.A.: Morgan Kaufmann Publishers.

Kavitha V. & Punithavalli M. (2010). Clustering time series data stream–a literature survey. *International Journal of Computer Science and Information Security*, 8(1):289-294.

Lin J. & Li Y. (2009). Finding Structural Similarity in Time Series Data Using Bag-of-Patterns Representation. In Proceedings of *the 21st International Conference on Scientific and Statistical Database Management*, 461-477.

Michinaka T., Tachibana S. & Turner J.A. (2011). Estimating price and income elasticities of demand for forest products: cluster analysis used as a tool in grouping. *Forest Policy and Economics,* 13(6): 435-445.

Munoz-Diaz D. & Rodrigo F.S. (2004). Spatio-temporal patterns of seasonal rainfall in Spain (1912-2000) using cluster and principal component analysis: comparison. *Annales Geophysicae,* 22(5): 1435-1448.

Prasanna K.A.V.L. (2012). Performance evaluation of multiviewpoint-based similarity measure for data clustering. *Journal of Global Research in Computer Science*, 3(11): 21-26.

Ramos M.C. (2001). Divisive and hierarchical clustering techniques to analyse variability of rainfall distribution patterns in a Mediterranean region. *Atmospheric Research*, 57(2):123-138.

Maharaj E.A., D'Urso P. & Galagedera D.U. (2010). Wavelet-based fuzzy clustering of time series. *Journal of Classification,* 27(2): 231-275.

Rani, S. & Sikka, G. (2012). Recent techniques of clustering of time series data: a survey. *International Journal of Computer Applications,* 52(15): 1-9.

Soltani S. & Modarres R. (2006). Classification of spatio-temporal pattern of rainfall in Iran using a hierarchical and divisive cluster analysis. *Journal of Spatial Hydrology,* 6(2): 1-12.

Tennant W.J. & Hewitson B.C. (2002). Intra-seasonal rainfall characteristics and their importance to the seasonal prediction problem. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 22(9): 1033-1048.

Malaysian Journal of Science 38 (Special Issue 2): 100 - 112 (2019)

**THE INTERNATIONAL SEMINAR ON MATHEMATICS IN INDUSTRY (ISMI)
AND THE INTERNATIONAL CONFERENCE ON THEORETICAL AND APPLIED STATISTICS (ICTAS)
ISMI-ICTAS18 [4-6 SEPTEMBER 2018]**

Calibrating Weather Forecasting in Indonesia: The Geostatistical Output Perturbation Method

Sutikno[1a], Purhadi[1b], Imam Mukhlash[2c], Kartika Nur 'Anisa'[1d]*, Urip Haryoko[3e], Hastuadi Harsa[3f]

[1] Department of Statistics, Faculty of Mathematics Computing and Data Science, Institut Teknologi Sepuluh Nopember, Kampus Sukolilo, Surabaya 60111, INDONESIA. E-mail: sutikno@statistika.its.ac.id[a] ; purhadi@statistika.its.ac.id[b] ; kartika.nuranisa9@gmail.com[d]
[2] Department of Mathematics, Faculty of Mathematics Computing and Data Science, Institut Teknologi Sepuluh Nopember, Kampus Sukolilo, Surabaya 60111, INDONESIA. E-mail: imamm@matematika.its.ac.id[c]
[3] Centre for Research and Development Indonesian Agency for Meteorology, Climatology, and Geophysics, Jakarta 10720, INDONESIA. E-mail: urip.haryoko@bmkg.go.id[e] ; hastuadi@gmail.com[f]
* Corresponding Author: kartika.nuranisa9@gmail.com[d]
Received: 21st April 2019 Revised: 6th August 2019 Published : 30th September 2019
DOI : https://doi.org/10.22452/mjs.sp2019no2.9

ABSTRACT The Numerical Weather Prediction (NWP) was developed by the Meteorological, Climatological, and Geophysical Agency in Indonesia for the purpose of weather forecasting, however, it comes with a high level of bias. This purpose of this study therefore was to improve this model with the use of Geostatistical Output Perturbation (GOP), implemented in the conformal-cubic atmospheric model (CCAM) on NWP data from the eight meteorological stations in Indonesia, i.e. Kemayoran, Priok, Cengkareng, Pondok Betung, Curug, Dermaga, Tangerang and Citeko stations. The findings indicated exponential as the best distribution model for analyzing temperature in Indonesia using GOP. Also, locations which are considerably far away from other locations could have significant impact on the accuracy of the weather forecasts. In this case, Citeko station has quite different characteristics location considering the fact that it is located on higher elevation compared with other stations. Therefore, the exclusion of Citeko station produced better forecasting in terms of accuracy and precision, increasing to about twice the result when the station was included in the analysis.

Keywords: GOP, NWP, spatial, weather forecasting.

1. INTRODUCTION

Accurate information on weather forecasting is very vital considering the fact that weather conditions directly and indirectly impact on human's activities such as determining the harvesting period of crops (agriculture), time of fishing (fishery), and feasibility of flight or voyage (transportation). Also, accurate and up to date weather forecasting has the capacity to minimize the risk of disasters due to weather or hydrometeorology.

The weather condition of Indonesia is unique due to its location between the Pacific and Indian oceans, as well as the monsoon climates with dynamic weather and atmospheric conditions (Tjasyono & Harijono, 2008). Therefore, the process of analyzing information about weather forecasting in the short, medium or long-term should be done continuously in order to find the most appropriate methods to capture the weather characteristics in each region.

A few years back, the Meteorological, Climatological and Geophysical Agency, Indonesia, (BMKG) developed the weather forecasting process using Numerical Weather

Prediction (NWP) to support the method already in use. However, forecasting through NWP has a high level of bias because it is measured on a global scale (homogeneous) and unable to capture the dynamics fluctuations in the atmosphere (BMKG, 2011; Wilks, 2006). Therefore, to improve its accuracy, the output of NWP must be with the statistical post-processing.

Also, a study conducted by Safitri & Sutikno (2012) showed that forecasting using NWP directly could result to bias outcomes. The authors went ahead and also used the Model Output Statistics (MOS) method and concluded that it improved the NWP model by 86%, this model was observed in four stations. Another study by Narendra, Sutikno, & Purhadi (2017) used Ensemble Model Output Statistics (EMOS) as the post-processing technique to solve the bias forecasting and under-dispersion problems in Cengkareng station. However, previous studies did not consider the spatial correlation patterns which might impact on the forecasting results. Therefore, this study considered that aspect and observed that the stations have dependent relation with one another.

The Geostatistical Output Perturbation (GOP) is a method which reconstructs error from the outcome of the deterministic forecast, and not the input, to produce ensemble forecasts of any size (Gel, Raftery, & Gneiting, 2004). Estimating the parameters for GOP is a common geostatistics modeling, which involves using maximum likelihood estimation (MLE) to estimate regression parameter with empirical semivariogram for the purpose of identifying the spatial correlation. According to Cressie (1985), the GOP spatial parameters are estimated by weighted least square from the semivariogram. Then, the forecasting from linear regression is added by an error which is simulated based on the spatial estimator in order to give the calibrated forecasting (Berrocal, Raftery, & Gneiting, 2007). Also, Feldmann (2012) found that GOP has the capacity to calibrate better temperature forecasts compared with the non-spatial methods, although GOP only uses the modification and simulation of one deterministic forecast. Therefore, the countries which do not have sufficient funds to develop enough NWPs, for example Indonesia, are only able to produce weather forecasts based on the ensemble.

Therefore, this study uses the GOP as the implemented method on weather data. In analyzing weather forecasting, the GOP considers the spatial correlation of all the locations or stations simultaneously. Although GOP only is based on one deterministic forecasting model, which has the spatial parameter for improving the forecasting result that could not identify the spatial cases. In this study, temperature was used as the response variable because it is the element of weather that has enough relation to rain event, air pressure, as well as humidity.

The section 2 of this study describes the Geostatistical Output Perturbation (GOP) method, spatial dependencies, variables used in the study, and the model evaluation, including the Root Mean Square Error (RMSE), Continuous Rank Probability Score (CRPS) and Coverage. The method was applied in section 3 to forecast temperature and show the results of the analysis. Finally, section 4 presents the conclusion and outlines some plans for future studies.

2. METHOD AND MATERIALS

2.1 Geostatistical Output Perturbation

Statistically, S is a considerable set of observational and relatively large locations, where $t = 1, 2, ..., T$ and considering the multivariate aspect between the locations, comes $\mathbf{y}_t = [y_{1t}, y_{2t}, ..., y_{st}]'$ which is the weather element vector observed in all observational

locations with size $s \times 1$ and $\mathbf{x}_t = \left[x_{1t}, x_{2t}, ..., x_{st}\right]'$ which is the weather forecast vector. Therefore, the GOP model is given as (1),

$$\mathbf{y}_t = \beta_0 \mathbf{1} + \beta_1 \mathbf{x}_t + \boldsymbol{\varepsilon}_t \tag{1}$$

where **1** is $s \times 1$ vector of which all elements are 1 and $\boldsymbol{\varepsilon}_t = \left[\varepsilon_{1t}, \varepsilon_{2t}, ..., \varepsilon_{st}\right]'$. The GOP model error of (1) follows the normal distribution with mean 0 and covariance Σ which depends on the spatial covariance structure (Gel, Raftery, & Gneiting, 2004). If $C\left(\mathrm{s}_i, \mathrm{s}_j\right)$ in (2) are the stationary and isotropy functions, then the element value of (*i*,*j*) Σ is obtained from

$$\frac{1}{2}\mathrm{var}\left(\varepsilon\left(\mathrm{s}_i\right) - \varepsilon\left(\mathrm{s}_j\right)\right) = \rho^2 + \sigma^2\left(1 - C\left(\mathrm{s}_i, \mathrm{s}_j\right)\right) \tag{2}$$

with ρ^2 is nugget effect, i.e. the measurement of error variance and the size of the spatial diversity in the certain distance that still exerts an influence $\rho^2 + \sigma^2$.is sill, i.e. total observed variation of the variable and the range, *r* is the distance between two observations that could be considered independent (Cressie, Statistics for Spatial Data Revised Edition, 1993; Karl & Maurer, 2010). The semivariogram is shown in Figure 1. Also, (2) the error dependency can be identified through the exponential semivariogram as represented in (3).

$$\gamma\left(\mathbf{d}\right) = \rho^2 + \sigma^2\left(1 - \exp\left(-\mathbf{d}/r\right)\right) \tag{3}$$

where **d** is $\left\|\mathrm{s}_i - \mathrm{s}_j\right\|$ that represents the Euclidean distance between the set of pairs of locations s_i and s_j and *r* is the range (in km) indicating the range of distances where the spatial correlation error begins to decrease significantly (Tanudidjaja, 1993). The other models are presented as Gaussian (4) and Spherical (5) semivariogram as shown below.

$$\gamma\left(\mathbf{d}\right) = \rho^2 + \sigma^2\left(1 - \exp\left(-\mathbf{d}^2/r^2\right)\right) \tag{4}$$

$$\gamma\left(\mathbf{d}\right) = \rho^2 + \sigma^2\left(\frac{3}{2}.\frac{\mathbf{d}}{r} - \frac{1}{2}.\frac{\mathbf{d}^3}{r^3}\right) \tag{5}$$

2.2 Spatial Dependencies

One of the indicators normally used to identify spatial dependencies in data is the Moran's I which is based on a distance matrix **W**. This approach uses certain distance (in km, meter, mile, etc) which is estimated to have the spatial influence between two locations. According to Anselin (1998), the element matrix **W** is given 1 when the distance is under cut-off part, otherwise, it is given 0. Also, the proximity location is indicated as being linked when the value is positive and otherwise when it is negative. The equation (6) below represents the Moran's I based on a distance matrix $\mathbf{W}_{s \times s}$ standardized, while **x** is the observed vector $s \times 1$.

$$I = \frac{(\mathbf{x}-\bar{\mathbf{x}})'\mathbf{W}(\mathbf{x}-\bar{\mathbf{x}})}{(\mathbf{x}-\bar{\mathbf{x}})'(\mathbf{x}-\bar{\mathbf{x}})} \tag{6}$$

The significance of Moran's I is tested through the normal approach (Cliff & Ord, 1981). Also, in the case of spatial autocorrelation, the null hypothesis (H_0) always states that there is no spatial clustering of the values associated with the geographic features in the study area. The hypothesis testing is shown in equation (7)

$$Z(I) = \frac{I - \mathrm{E}(I)}{\sqrt{\mathrm{var}(I)}} \sim N(0,1) \tag{7}$$

with $\mathrm{E}(I) = -\frac{1}{n-1}$; $\mathrm{var}(I) = \frac{n^2 S_1 - nS_2 + 3S_0^2}{(n^2-1)S_0^2} - \left(\mathrm{E}(I)\right)^2$ and $\mathrm{var}(I)$, $S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}$;

$S_1 = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(w_{ij}+w_{ji}\right)^2$; $S_2 = \sum_{i=1}^{n}\left(\sum_{j=1}^{n} w_{ij} + \sum_{j=1}^{n} w_{ji}\right)^2$. H_0 is rejected if $|Z(I)| > z_{1-\alpha/2}$ where $z_{1-\alpha/2}$ is quantile of $\left(1-\alpha/2\right)\times 100\%$ a normal standard distribution.

2.3 Evaluating Calibrated Model

The goodness of fit models does not give accurate measurements when only the root mean square error (RMSE) is used to calibrate the weather forecasts. According to Feldmann (2012), other methods are also needed to check the level of bias correction and sharpness of forecasts ensembles, such as the continuous rank probability score (CRPS) and coverage.

1. *Root Mean Square Error* (RMSE)

   The RMSE is an indicator of accuracy obtained from the (8) square root of MSE, which is the sum of the squares of the difference between the forecast and observed values.

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{8}$$

2. *Continuous Rank Probability Score* (CRPS)

   The CRPS is used (9) to check how precise the predictive intervals are produced by calibration methods. According to Feldmann (2012), the lower the value, the more reliable the predictive interval.

$$\mathrm{CRPS} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{crps}\left(F_i, y_i\right) = \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty}\left[F_i^{forecast}(y) - F_i^{obs}(y)\right]^2 dy \tag{9}$$

where n is the number of observations, i is time period (e.g. daily), $F_i^{forecast}(y)$ is predictive

CDF at time i, and $F_i^{obs}(y)$ is empirical CDF at time i (Anggraeni, 2013). If the threshold forecast < observation, then $F_i^{obs}(y)=0$, however, it is 1 when the threshold forecast ≥ observation.

3. *Coverage*

The sharpness of the ensemble forecasts can be identified through coverage as shown in equation (10). According to Moller (2014), observations are said to be in the coverage when within an ensemble range.

$$\frac{M-1}{M+1}\times 100\% \tag{10}$$

M denotes the ensemble member and the ensemble forecast is calibrated if the empirical coverage is much closer to the standard coverage.

2.4 Data and Variables

Secondary data are used in this study, obtained from Meteorology, Climatology and Geophysics Agency (BMKG). Data of conformal-cubic atmospheric model (CCAM) NWP collected from 1st January 2009 to 31st December 2010 (708 days). The locations of the research are the meteorological stations in Kemayoran, Priok, Cengkareng, Pondok Betung, Curug, Dermaga, Tangerang and Citeko (Luthfi, Sutikno, & Purhadi, 2018) as shown in Figure 1. The response variable in this study is observed air temperature, i.e. maximum and minimum temperature (Celsius).

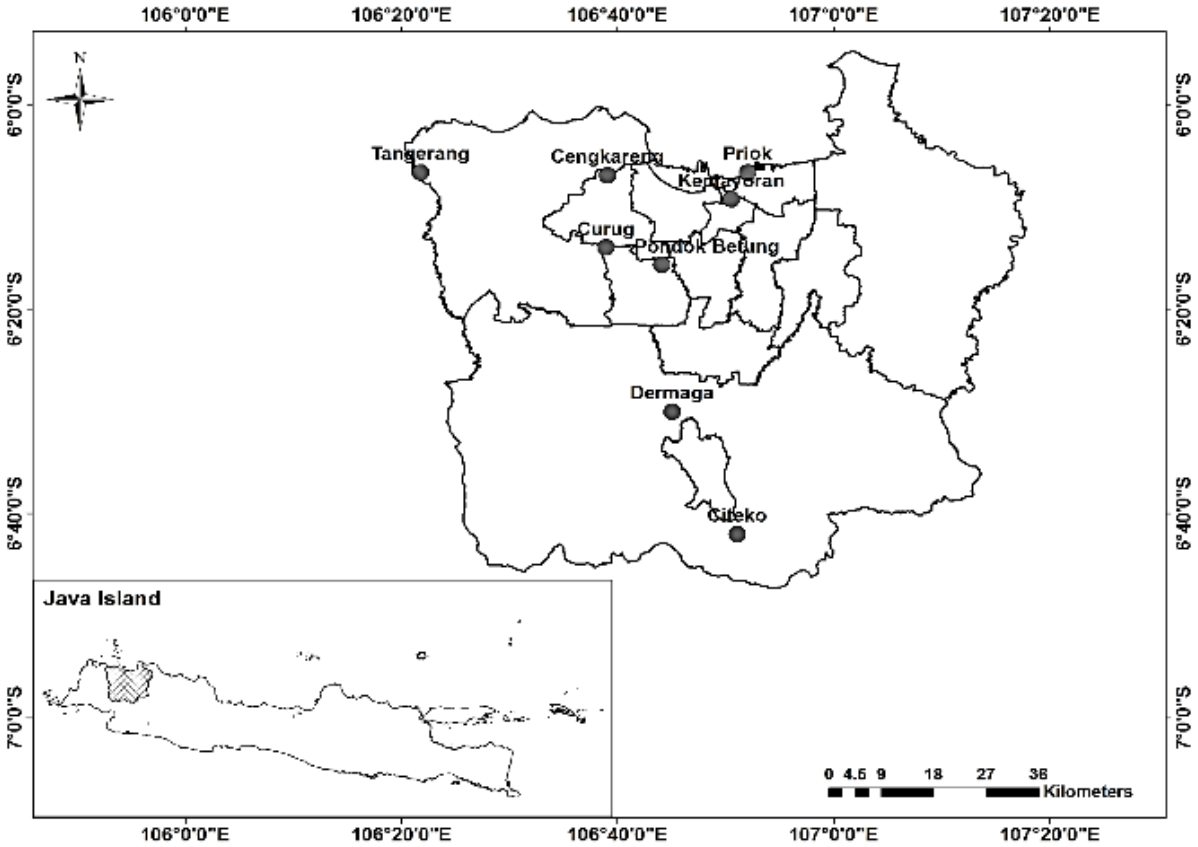


Figure 1: Meteorological stations for study region.

3. RESULTS AND DISCUSSION

3.1 Spatial Dependencies

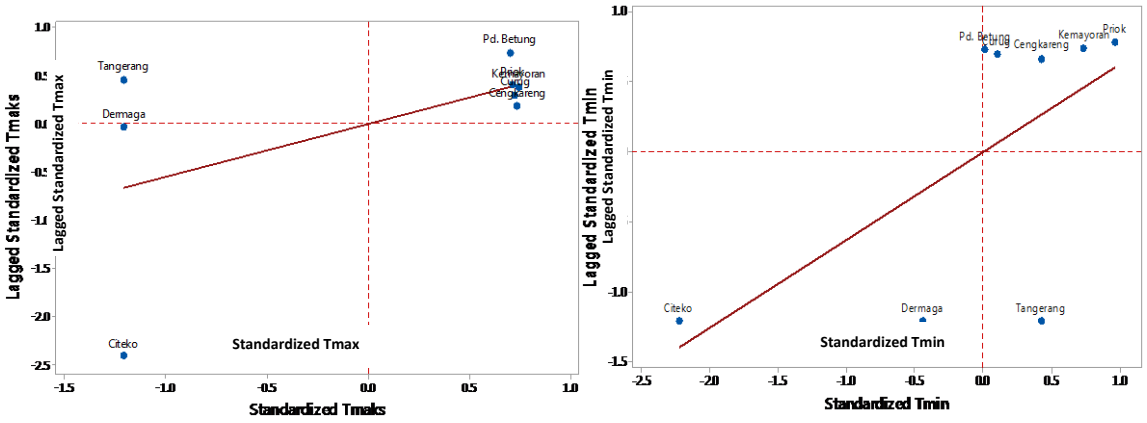
The Moran's I is used to test the significance of spatial dependencies of the maximum temperature between the eight meteorology stations. Table 1 shows that Citeko has the farthest distance from the other stations. Based on BMKG (2011) and the assumption of uniform elevation (height) of the station, the cut-off distance is set at 30km. The standardized weighted distance matrix was used in order to get the Morgan's I. This is to make its range to be between -1 and 1. The Moran's I of T_{MAX} is 0.135 (p-value = 0.048) and T_{MIN} is 0.379 (p-value=0.006). It means the spatial dependencies for T_{MAX} and T_{MIN} are statistically significant at $\alpha = 5\%$. It also means that the spatial strength of minimum temperature is higher than the maximum temperature (i.e. the Moran's I of T_{MIN} is higher than T_{MAX}). More so, the closer stations have stronger relationships than those further apart.

Figure 2 shows that the temperature in most stations is related to each other. Despite the fact that Citeko and Dermaga stations are on the same quadrant, there is a weak correlation between them because of the geographical distant position.

Table 1: Matrix of distance 8 meteorology stations (km).

Station	1	2	3	4	5	6	7	8
1	0	6,810	21,694	15,479	22,929	53,593	60,575	39,634
2	6,810	0	24,091	22,091	28,231	56,130	66,818	46,372
3	21,694	24,091	0	19,615	12,98	32,108	68,598	44,080
4	15,479	22,091	19,615	0	11,609	46,443	50,027	26,550
5	22,929	28,231	12,980	11,609	0	34,907	56,467	31,683
6	53,593	56,130	32,108	46,443	34,907	0	85,158	61,212
7	60,575	66,818	68,598	50,027	56,467	85,158	0	24,859
8	39,634	46,371	44,080	26,550	31,683	61,212	24,859	0

Note.:
1 : Kemayoran 2 : Priok 3 : Cengkareng 4 : Pondok Betung
5 : Curug 6 : Tangerang 7 : Citeko 8 : Dermaga



(a) (b)

Figure 2: Moran's I plot of spatial correlation among stations: (a) maximum temperature and (b) minimum temperature.

3.2 The Assessment of the Best Model

Efforts were made to find the best semivariogram model for applying the temperature in Indonesia using three different models, i.e. exponential, Gaussian and spherical. The models are compared based on empirical semivariogram as shown in Figure 3, then the model evaluation in Table 2, while the predictive checking by Verification Rank Histogram is presented in Figure 4.

Also, Figure 2 shows that spatial inconsistency exists on both the maximum and minimum temperature. On some distance pairs, there are few bins which have semivariance far greater than others. Since the patterns are seen in bins representing distance 50km or more, it could be because Citeko, Tangerang and Dermaga stations are geographically located far from others, as seen in Figure 2. This has been the biggest effect on the unexpected inconsistency.

Figure 3 shows the construction of the residual of the GOP model to obtain the exponential, Gaussian and spherical semivariogram. The semivariogram is roughly constant on reaching a distance of about 8.69km. This implies that the temperature between two locations has no dependency after 8.69km, with sill recording 3.65 for

maximum temperature and 2.53 for minimum temperature. The high amount of sill might result to a greater variance of estimation or influence the forecast. And despite the fact that the spherical semivariogram has the same range and sill values with other models for maximum and minimum temperature, it does not give the best semivariogram as shown in Figure 4c.

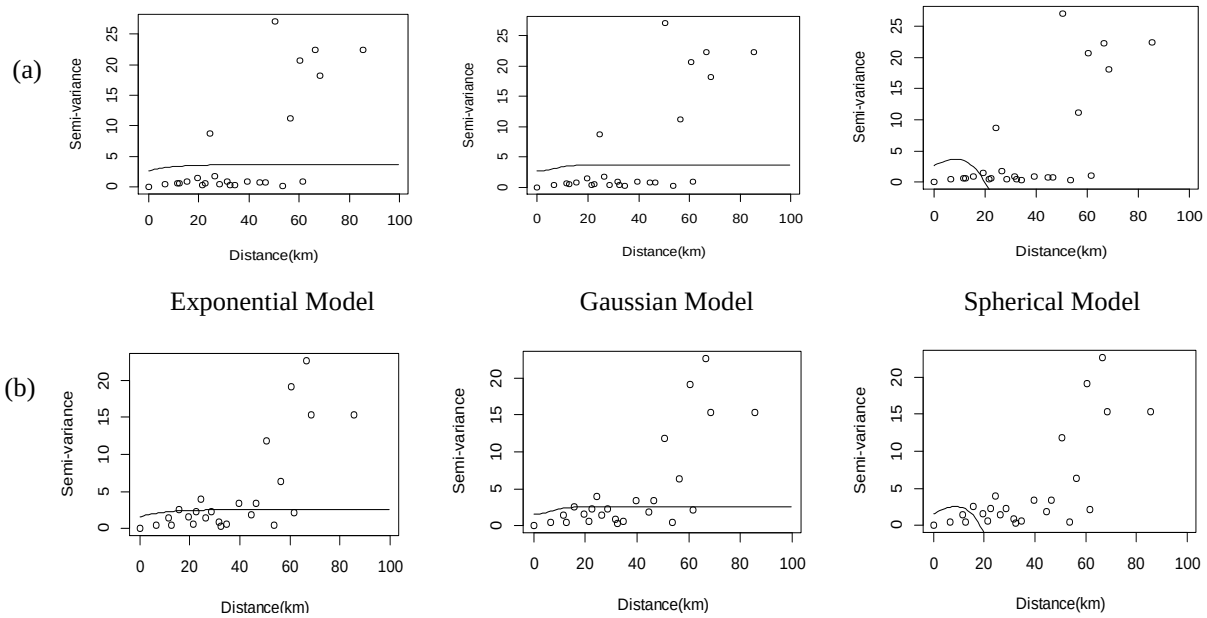


Figure 3: Empirical semivariogram over 30-day training period of temperature: (a) maximum temperature (b) minimum temperature.

Upon evaluation and comparing the GOP for the three models, the best model discovered was the exponential model as presented in Table 2.

Table 2: Comparison of evaluating GOP model, 31/1/2009 – 31/1/2010.

	Exponential			**Gaussian**			**Spherical**		
	RMSE (°C)	**CRPS**	**Coverage (%)**	**RMSE (°C)**	**CRPS**	**Coverage (%)**	**RMSE (°C)**	**CRPS**	**Coverage (%)**
T_{MAX}	3.044	1.540	75.526	3.066	1.544	75.589	3.056	1.540	75.461
T_{MIN}	2.665	1.423	68.732	2.673	1.429	68.510	2.677	1.426	68.676

As the other comparison, the predictive checking for GOP ensemble was applied for the predictions in all stations that used the three models from 1st January 2009 to 31st December 2010. The verification rank histograms shown in Figure 4 give relatively same results and the abscissa shows the number of GOP ensemble and probable ordinate.

Considering the three criteria as shown in figures 3 and 4, as well as in Table 2, the exponential distribution is the best model for applying the temperature in Indonesia.

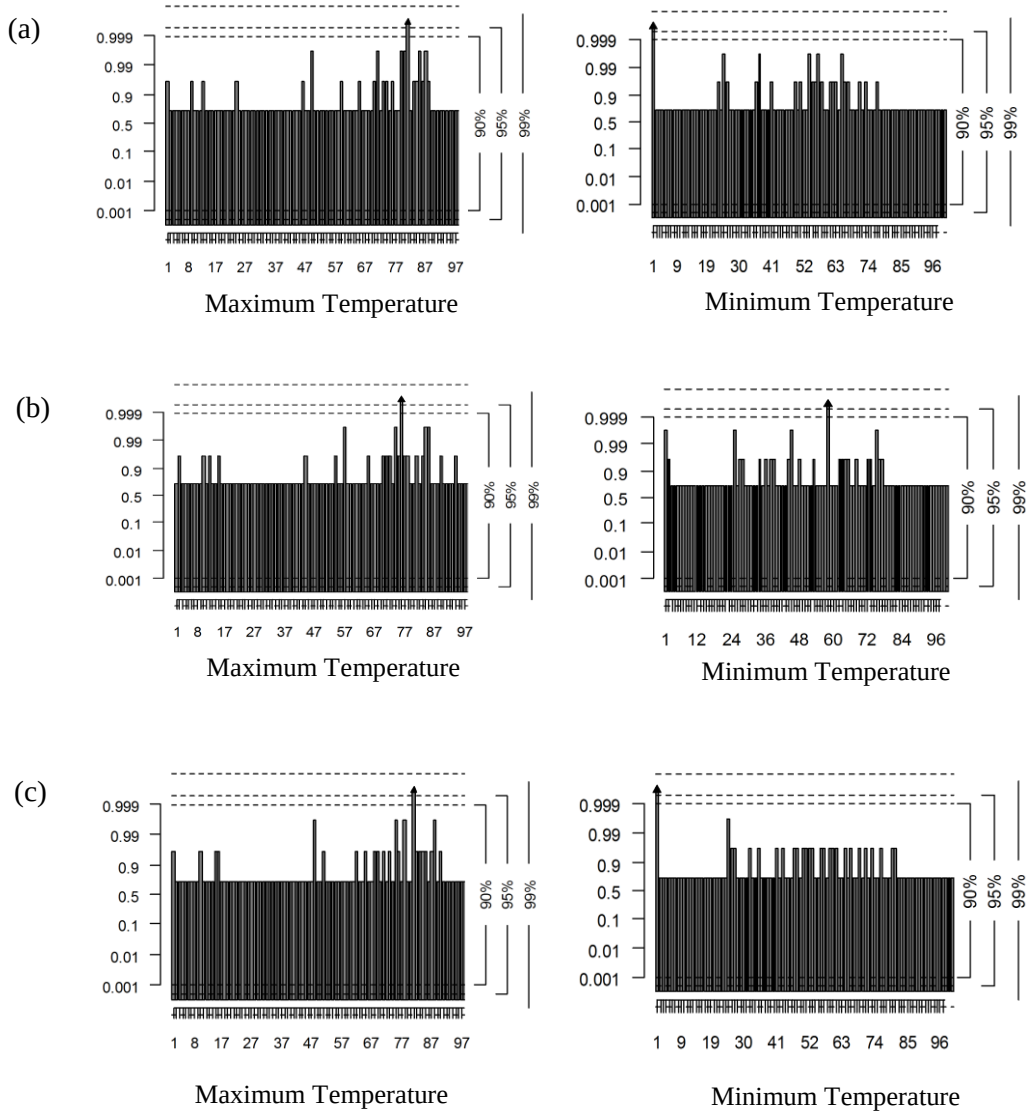


Figure 4: Verification rank histogram for GOP ensemble predictions, 31/1/2009 – 31/1/2010 (a) exponential, (b) gaussian, (c) spherical.

3.3 Model Temperature Forecast

The GOP models for both the maximum and minimum temperature with 30-day training period are presented as follow (11)

$$\begin{aligned} T_{MAXs,t} &= 1.785 + 0.959\text{tmaxscr}_{s,t} \\ T_{MINs,t} &= 25.963 - 0.131\text{tminscr}_{s,t} \end{aligned} \tag{11}$$

The GOP improves the bias correction rate as shown in Table 3, which reflected on the RMSE of GOP for maximum temperature (2.13°) which is lower (i.e. the model is better) than of NWP (2.18°), despite the fact that the RMSE of minimum temperature

forecasting by GOP is not better than that of NWP. A strong reason for this weakness with regards to the accuracy of GOP is its failure to forecast temperatures in Citeko and Dermaga stations, which are geographically far from the other stations. This problem is compounded by the observed value in minimum temperature which does not fit into the 90% predictive interval for Citeko station. The GOP is strongly vulnerable and risky in case of inadequate location of interest, inappropriate data properties, data mishandle, etc.

Then, a remodeling process was conducted without Citeko station in order to get better results. Based on this, the new GOP models for temperature with 30-day training period are given as (12)

$$\begin{aligned} T_{MAX\,s,t} &= 4.739 + 0.885\,\text{tmaxscr}_{s,t} \\ T_{MIN\,s,t} &= 21.894 + 0.074\,\text{tminscr}_{s,t} \end{aligned} \tag{12}$$

Table 3: RMSE of GOP forecast and NWP on January 31st, 2009.

Temperature	Station	Obs. (°C)	NWP (°C)	GOP (°C)	P_5 (°C)	P_{95} (°C)	RMSE of NWP	RMSE of GOP
T_{MAX}	Kemayoran	28.8	26.43	29.63	23.95	30.96	2.18°	2.13°
	Priok	28.7	26.55	31.03	24.34	29.77		
	Cengkareng	28.6	26.42	28.44	24.4	29.87		
	Pd. Betung	29.0	26.49	25.09	24.15	30.3		
	Curug	28.3	26.26	28.39	23.79	30.32		
	Tangerang	29.2	26.34	27.88	23.08	29.56		
	Citeko	25.0	26.77	26.76	23.95	30.58		
	Dermaga	27.8	26.73	24.66	24.33	30.32		
Temperature	**Station**	**Obs. (°C)**	**NWP (°C)**	**GOP (°C)**	**P_5 (°C)**	**P_{95} (°C)**	**RMSE of NWP**	**RMSE of GOP**
T_{MIN}	Kemayoran	23.8	22.6	25.86	20.84	25.75	1.57°	2.27°
	Priok	23.8	23.18	24.05	20.99	26.05		
	Cengkareng	23.4	22.55	22.46	20.75	25.46		
	Pd. Betung	23.4	22.38	24.12	20.16	25.53		
	Curug	23.5	22.19	23.02	20.49	25.69		
	Tangerang	23.3	22.39	23.89	19.73	25.39		
	Citeko	18.4	22.08	23.72	20.92	25.38		
	Dermaga	22.4	22.24	24.97	20.67	25.46		

The regression parameters β_0 and β_1 for the maximum temperature (12) without Citeko station, are both significant at $\alpha = 5\%$, however on involving the (11) Citeko station, only β_1 for maximum temperature and β_0 for minimum temperature are significant. This could affect the accuracy of the forecasting results

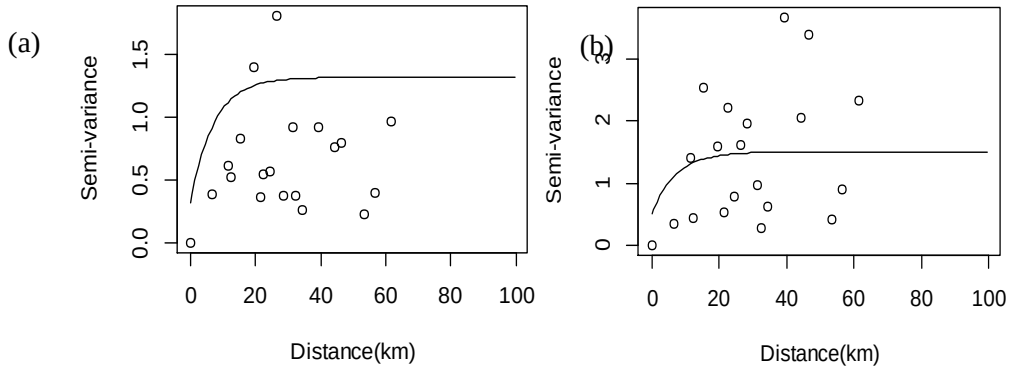


Figure 5: Empirical semivariogram of (a) maximum temperature. (b) minimum temperature over 30-day training period without Citeko station.

Figure 5 shows the adjustment of exponential semivariogram which is significant when compared with the semivariogram in Figure 3a which involves Citeko station. The semivariogram in Figure 5 could follow the spatial pattern from empirical semivariogram with range 7.14km and sill recording of 1.317km for maximum and 1.505km for minimum temperature. This is an indication that spatial simulation might be more capable of modifying residual data testing, thereby producing more accurate forecasting results

The evaluation of GOP model without Citeko station gives better results, i.e. the better accuracy (lower RMSE), more reliable (lower CRPS) and better calibrated (the coverage of minimum temperature is closer to 90%). In addition, the RMSE of NWP without Citeko station is better than that of NWP with Citeko, despite the fact that the RMSE of minimum temperature by GOP is not better than RMSE of NWP. What could be responsible for this is the inclusion of Dermaga station in this study, which is also geographically far from the other stations.

Considering Table 4, it is clear that the effects of far stations are significant on the accuracy and precision of both the GOP and NWP models for temperature forecasting. Aside that, Citeko station is located on a higher elevation compared with other stations, hence, possesses different geographic characteristics.

Table 4: Comparison of evaluating GOP and NWP, January 31st, 2009 – December 31st, 2010.

	With Citeko Station				**Without Citeko Station**			
	NWP model	**GOP model**			**NWP model**	**GOP model**		
	RMSE (°C)	**RMSE (°C)**	**CRPS**	***Coverage* (%)**	**RMSE (°C)**	**RMSE (°C)**	**CRPS**	***Coverage* (%)**
T_{MAX}	2.810	3.044	1.54	75.526	2.646	1.897	0.972	71.766
T_{MIN}	2.010	2.665	1.423	68.732	1.530	1.866	0.87	79.667

The predictive checking for GOP ensemble predictions in all stations (i.e. without Citeko station) in the period January 1, 2009 to December 31, 2010, were applied to assess its predictive performance.

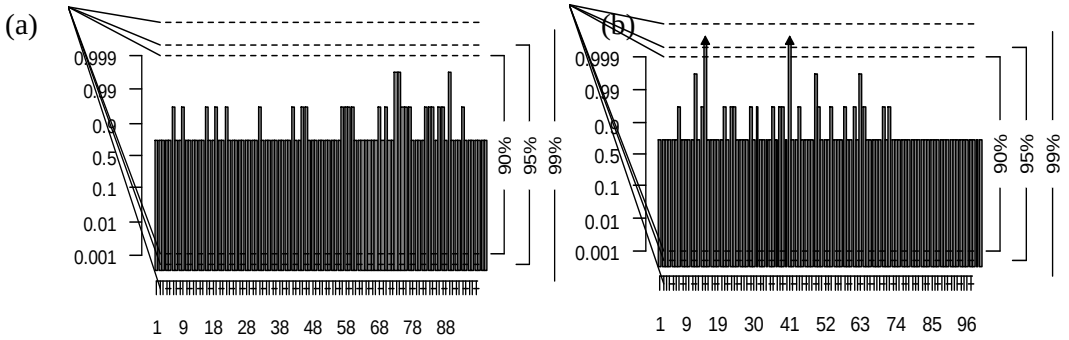


Figure 6: Verification rank histogram for GOP ensemble predictions, January 31st, 2009 – December 31st, 2010 without Citeko station: (a) maximum temperature and (b) minimum temperature

The abscissa in Figure 7 shows the number of GOP ensemble and probable ordinate. The verification rank histogram for maximum and minimum temperature without Citeko station, are relatively close to the ones in Figure 5 with Citeko, thereby indicating proper coverage of the prediction intervals at all levels. Although the verification rank histogram for minimum temperature seems not better with Citeko, the GOP result in Table 4 was used in the evaluation. Therefore, this study concludes that GOP ensemble without Citeko shows better predictions than with Citeko station.

4. CONCLUSION

This study proposes a GOP model for analyzing the spatial relation in eight stations for the purpose of obtaining better accuracy and precision of weather forecasting and to investigate the underlying analysis in those several stations. The findings showed that exponential is the best distribution model for analyzing both the maximum and minimum temperature in Indonesia using GOP. Also, the exclusion of Citeko station in the GOP modeling for temperature in Indonesia gives better results in terms of accuracy and precision, which increase by almost twice in all stations. Considering the fact that this study involves only eight stations, the GOP might not be totally accurate in estimating the optimum parameters at these stations, as better results could be obtained with more meteorological stations, at least a dozen, in quite close locations. This has the capacity to minimize inaccurate forecasting because locations geographically far from others have higher percentage of giving inaccurate forecasts.

5. ACKNOWLEDGMENT

The authors wish to thank the Meteorology, Climatology and Geophysics Agency (BMKG) of Indonesia for providing all the data used in this study. A big thank also to the Ministry of Research, Technology and Higher Education of Indonesia for providing the grant for the National Strategic Research 2018, of which the authors are beneficiaries.

6. REFERENCES

Anggraeni, D. (2013). Calibration forecasting rainfall ensemble data using output statistic model ensemble (EMOS) and Bayesian model averaging (BMA). Surabaya: *Thesis*, Institut Teknologi

Sepuluh Nopember.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Dordrecht: Kluwer Academic Publisher.

Berrocal, V., Raftery, A., and Gneiting, T. (2007). Combining spatial statistical and ensemble information in probabilistic weather forecast. *Monthly Weather Review AMS*, 135: 1386-1402.

BMKG. (2011). Study and model application of CCAM (conformal-cubic atmospheric model) model for short term weather forecast using MOS (model output statistics). Jakarta: Research and Development Center BMKG.

Cliff, A., and Ord, J. (1981). *Spatial Processes: Models & Applications*. London: Pion Limited.

Cressie, N. (1985). Fitting variogram models by weighted least squares. *Mathematical Geology*, 5(17): 563–585.

Cressie, N. (1993). *Statistics for Spatial Data Revised Edition*. New Jersey: John Wiley and Sons, Inc.

Feldmann, K. (2012). *Statistical Postprocessing of Ensemble Forecasts for Temperature: The Importance of Spatial Modeling*. Germany: Diplomarbeit, Ruperto-Carola University of Heidelberg.

Gel, Y., Raftery, A., and Gneiting, T. (2004). Calibrated probabilistic mesoscale weather field forecasting: The Geostatistical Output Perturbation (GOP) method (with discussion). *Journal of the American Statistical Association*, 99(467): 575–583.

Karl, J. W., and Maurer, B. A. (2010). Spatial dependence of predictions from image segmentation: A variogram-based method to determine appropriate scales for producing land-management information. *Ecological Informatics*, 5(3): 194–202.

Luthfi, M., Sutikno, and Purhadi. (2018). Calibrating weather forecast using Bayesian model averaging and geostatistical output perturbation. *Journal of Science*, 3(1): 1-9.

Moller, A. (2014). Multivariate and spatial ensemble postprocessing methods. Germany: *Dissertation*. Ruperto-Carola University of Heidelberg.

Narendra, R. D., Sutikno, and Purhadi. (2017). Ensemble model output statistics for short-term weather forecast. In *ICoMSE 2017: The 1st Annual International Conference on Mathematics, Science, and Education*. Malang, Indonesia.

Safitri, R., and Sutikno. (2012). Model output statistics with projection pursuit regression to predict minimum, maximum temperature and humidity. *Jurnal Sains dan Seni ITS*, 1(1): 2301-928X.

Tanudidjaja. (1993). *Earth and Space Science*. Jakarta: Publisher Department of Education and Culture.

Tjasyono, B., and Harijono, S. (2008). *Indonesian Meteorology Model 2 of Clouds and Monsoon Rain*. Meteorology and Geophysics Agency Jakarta.

Wilks, D. (2006). *Statistical Methods in the Atmospheric Sciences 2nd Edition*. Boston: Elsevier.

Malaysian Journal of Science 38 (Special Issue 2): 113- 119 (2019)

**THE INTERNATIONAL SEMINAR ON MATHEMATICS IN INDUSTRY (ISMI)
AND THE INTERNATIONAL CONFERENCE ON THEORETICAL AND APPLIED STATISTICS (ICTAS)
ISMI-ICTAS18 [4-6 SEPTEMBER 2018]**

Rainfall Forecast with Best and Full Members of the North American Multi-Model Ensemble

Defi Yusti Faidah[1,2a], Heri Kuswanto[2b*], Suhartono[2c] & Kiki Ferawati[2d]

[1] Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, Jalan Raya Bandung-Sumedang Km.21, Jatinangor, 45363, Sumedang, INDONESIA. E-mail: defi.yusti@unpad.ac.id[a]
[2] Department of Statistics, Faculty of Mathematics, Computation, and Data Science, Institut Teknologi Sepuluh Nopember, Sukolilo, Surabaya, INDONESIA. E-mail: heri_k@statistika.its.ac.id[b] ; suhartono@its.ac.id[c] ; kiki.ferawati@gmail.com[d]
* Corresponding Author: heri_k@statistika.its.ac.id[b]
Received: 21st April 2019 Revised : 6th August 2019 Published : 30th September 2019
DOI : https://doi.org/10.22452/mjs.sp2019no2.10

ABSTRACT The North American Multi-Model Ensemble (NMME) is a multi-model seasonal forecasting system consisting of models from combined US modelling centres. The NMME is expected to generate better rainfall prediction than a single model. However, the NMME forecasts are underdispersive or overdispersive, and calibration is needed to produce more accurate forecasting. This research examined the monthly rainfall data in Surabaya generated by nine NMME models and further calibrated them with bayesian model averaging (BMA). The purpose of this research was to assess the performance of the calibration results using the best four models and the full ensemble. The four models are CanCM3, CanCM4, CCSM3, and CCSM4, which were selected based on their skills. Both calibration results were evaluated using the continuous range probability score (CRPS) and the percentage of captured observations. The calibration with four models produced an average CRPS of 6.27 with 88.16% coverage, while with nine models an average CRPS of 5.23 with 92.11% coverage was obtained. This result suggests using the full ensemble to generate more accurate probabilistic forecasts.

Keywords: BMA, calibration, NMME

1. INTRODUCTION

The North American Multi-Model Ensemble (NMME) is a multi-model seasonal forecasting system consisting of coupled models from US modelling centres, including the NOAA National Centers for Environmental Prediction (NOAA/NCEP), the Center for Ocean-Land-Atmosphere Studies (COLA), the NOAA's Geophysical Fluid Dynamics Laboratory (NOAA/GFDL), the National Aeronautics and Space Administration/Global Modeling and Assimilation Office (NASA/GMAO), and Canadian modelling centres (Kirtman et al., 2014). Becker et al. (2014) examined the NMME's skill and verified it against observations globally. They found that, for the precipitation rate and sea surface temperature, the NMME's skill is higher than that of any single model, although there may be many regional and seasonal variations. The NMME usually makes better predictions than most, if not all, individual models. However, both the potential predictability and the real forecast skill vary depending on the geographical region and season.

The NMME involves two major processes. The first focuses on changing the seasonal and annual time scales into a monthly

scale. The second defines the most appropriate forecast parameters. Forecasting is performed every mid-month. Kirtman et al. (2014) explained that the multi-model approach using the NMME is more accurate than single-model forecasting. The NMME has been used extensively in previous research to verify forecasting results from the average monthly rainfall (Kuswanto, 2010; Wang et al., 2016), regional temperatures at 2 m above sea level (Becker et al., 2014), the sea surface temperature (Barnston et al., 2011; Kuswanto & Sari, 2013), seasonal rainfall (Ma et al., 2015), and seasonal droughts (Yuan & Wood, 2013).

A lot of researches showed that ensemble prediction systems have bias and hence, they have to be post-processed statistically to generate calibrated predictive distributions (Hamill & Colucci, 1997). Raftery et al. (2005) introduced Bayesian Model Averaging (BMA) with more recent extensions to quantitative precipitation (Sloughter et al., 2010), wind direction (Bao et al., 2013), and wind speed (Hamill & Colucci, 1997). The NMME's skill has never been investigated. This research has several goals. The first is to show that the NMME has bias. The second is to verify that BMA can improve the reliability and validity of the NMME. The last is to assess the performance of calibration results using the best four models and the full ensemble evaluated using the continuous range probability score (CRPS) and the percentage of captured observations in Surabaya.

2. LITERATURE REVIEW

2.1 North American Multi-Model Ensemble (NMME)

The NMME is a forecasting system consisting of coupled models from US and Canadian modelling centres. The NMME was launched in the United States (Kirtman et al., 2014) with real-time experimental operational forecasts from the NOAA or the NCEP. The multi-model ensemble approach has been shown to produce better prediction quality on average than any single model of the ensemble, motivating the NMME's undertaking (Doblas-Rayes et al., 2005; Gneiting et al., 2005; Hagedorn et al., 2004; Palmer, 2001; Smith et al., 2013). The models included in the NMME are CMC1-CanCM3 and CMC2-CanCM4 from CanSIPS, COLA-RSMAS-CCSM3 and COLA-RSMAS-CCSM4 from COLA, GFDL-CM2p1-aer04 from GFDL, ECHAM4p5-Anomaly and ECHAM4p5-DirectCoupled from IRI, and CFSv1 and CFSv2 from NCEP.

2.2 Bayesian Model Averaging (BMA)

Ensembles of numerical weather prediction models have been developed, in which multiple estimates of the current state of the atmosphere are used to generate probabilistic forecasts for future weather events. However, ensemble systems are uncalibrated and biased and thus need to be post-processed statistically, for which BMA is the preferred method. BMA was introduced by Raftery et al. (2005). The basic idea is that, for any given forecast ensemble, there is a best model or member, but we do not know which it is. In BMA, the overall forecast probability density function (pdf) is a weighted average of the forecast pdfs based on each of the individual forecasts. The weights are the estimated posterior model probabilities and reflect the models' forecast skill. The forecast f_k is then associated with a conditional pdf $g_k(y \mid f_k)$, which can be interpreted as the conditional pdf of y conditional on f_k, given that f_k is the best forecast in the ensemble. The BMA predictive model is:

$$p(y \mid f_1, f_1, \ldots, f_K) = \sum_{k=1}^{K} w_k g_k(y \mid f_k) \quad (1)$$

where f_k is an ensemble forecast from K models. w_k is the posterior probability of forecast k being the best one. The w_k's are probabilities, so they are non-negative and add up to 1. $g_k(y \mid f_k)$ is the gamma pdf with mean $\mu_k = \alpha_k \beta_k$ and standard deviation $\sigma_k = \sqrt{\alpha_k \beta_k}$, where α_k is the shape parameter and β_k is the scale parameter. Thus, $g_k(y \mid f_k)$ can be written as follows:

$$g_k(y \mid f_k) = \frac{1}{\beta_k^{\alpha k}\Gamma(a_k)} y^{a_k - 1} \exp\left(-\frac{y}{\beta_k}\right) \tag{2}$$

2.3 Continuous Range Probability Score (CRPS)

The calibrated ensemble generates estimated intervals in pdf form. The CRPS is a much-used measure of performance for probabilistic forecasts (Hersbach, 2000). It is derived from a quadratic measure of the difference between the forecast cumulative distribution function (cdf) and the empirical cdf of the observation. The formula of the CRPS can be written as follows:

$$CRPS = \frac{1}{n^f} \sum_{i=1}^{n} \int_{x=-\infty}^{\infty} \left(F_i^f(x) - F_i^0(x)\right)^2 dx \tag{3}$$

where $F_i^f(x)$ is the cdf from the forecast in the i-th period, $F_i^0(x)$ is the cdf from the observations in the i-th period, and n^f is the number of forecasts.

3. DATA AND METHODOLOGY

The data set used in this paper contains the monthly series of precipitation predictions from each individual model, which were downloaded from the official website of the NMME and the official website of the European Centre for Medium Range Weather Forecast (ECMWF). The data consist of monthly rainfall forecast results and the observed total rainfall in Juanda Surabaya. The two data sets have the same time periods, from 2003 to 2010. There are nine ensemble members, which are analysed as follows:

1. Evaluating the forecast model in the NMME data set against the real-time observations in the ECMWF data set using Root Mean Square Error (RMSE).
2. Calibrating the forecast models in the NMME data set with pre-process result data using the BMA approach. The calibration process using BMA will be examined for the window time (m) $m = 12$. The window time is the amount of data used to estimate the BMA parameters. The calibration is carried out in the following steps:
   - Starting a regression between forecasts as a predictor with the observation (dependent variable) using as many data as in the *m*-period before the calibrated period to obtain bias correction.
   - Based on equation (3), estimating the w_k for each ensemble member and variance with the expectation maximization

algorithm. w_k is the posterior probability of forecast k being the best one.

- After all the parameters have been obtained, then the calibrated forecast can be obtained.

3. Evaluating the model's reliability using the CRPS.

4. RESULT

4.1 Evaluation of the Rainfall Forecast Model

There are nine models to be calibrated in this research. However, they must be evaluated first to determine whether the individual models are reliable or not. In this research, the performance of the ensemble model is assessed using the R^2 to determine the accuracy of the forecast in relation to the observations. In addition, the RMSE is used to evaluate the goodness of the model. Table 1 presents the performance of the monthly precipitation in individual models based on the R^2 and RMSE.

Table 1: Performance of monthly precipitation individual models.

Ensemble Member	R^2	**RMSE**
CMC1-CanCM3	53.00%	6.95
CMC2-CanCM4	49.70%	7.72
COLA-RSMAS-CCSM3	37.20%	7.04
COLA-RSMAS-CCSM4	51.00%	6.83
GFDL-CM2p1-aer04	47.90%	8.27
ECHAM4p5-Anomaly	26.60%	4.66
ECHAM4p5-DirectCoupled	32.00%	5.28
CFSv1	5.90%	5.06
CFSv2	31.60%	2.78

Based on the values in Table 1, the best ensemble members are determined by comparing the R^2 value of each ensemble member with the observation data. The best are CMC1-CanCM3, CMC2-CanCM4, COLA-RSMAS-CCSM3, and COLA-RSMAS-CCSM3. CFSv2 has the smallest RMSE. The best model is selected using the R^2 due to the fact that the basic idea of BMA is to capture the uncertainty. The R^2 is used to explain how much variability in the observations that can be explained by the ensemble forecasting from each model.

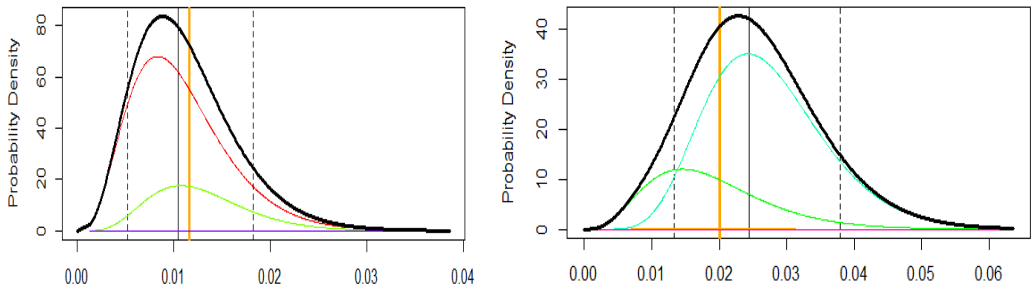
4.2 Calibration of Rainfall Forecasts

Based on the previous sub-section, the result of ensemble forecasting is still unreliable. Therefore, a post-processing method is needed to calibrate the ensemble model to produce better forecasts. The BMA reduces the mean bias value towards the observed value. In addition, it adjusts the variance to obtain a calibrated forecasting value. The first step is to determine the estimates of the parameter and to obtain the calibrated mean (μ) and variance (σ^2). Table 2 shows the BMA parameter along with the mean and standard deviation values for the period December 2010 for the lead time of one month.

Table 2: BMA Parameters at -7º South and 113º East.

	Model	μ	w	μ -Calibrated	σ^2 -Calibrated
Best Four Models	CMC1-CanCM3	0.0294884	7.89E-01	0.0312889	0.0050796
	CMC2-CanCM4	0.0314713	2.11E-01		
	COLA-RSMAS-CCSM4	0.0255515	1.90E-06		
	GFDL-CM2p1-aer04	0.0280928	1.11E-11		
Full Model	CMC1-CanCM3	0.0269645	1.72E-04	0.0277121	0.033588
	CMC2-CanCM4	0.0277123	1.00E+00		
	COLA-RSMAS-CCSM3	0.0254797	3.58E-10		
	COLA-RSMAS-CCSM4	0.0264381	1.46E-17		
	GFDL-CM2p1-aer04	0.027008	2.22E-14		
	ECHAM4p5-Anomaly	0.0256025	1.06E-12		
	ECHAM4p5-DirectCoupled	0.0253882	5.51E-14		
	CFSv1	0.0240376	8.78E-09		
	CFSv2	0.0254054	6.84E-12		

Based on Table 2, CMC1-CanCM3 has the largest weight of the best four models, 7.89E-01. CMC2-CanCM4, COLA-RSMAS-CCSM4, and GFDL-CM2p1-aer04 have weights of 0.211, 0.0000019, and 1.11E-11. This means that CMC1-CanCM3 makes a greater contribution to BMA, because its weight is larger than the others. On the contrary, COLA-RSMAS-CCSM4 and GFDL-CM2p1-aer04 do not contribute to BMA, because their weight is very small, while CMC2-CanCM4 has the largest weight in the full model and tends to be close to one. The larger weight indicates a greater contribution to BMA.



(a) (b)

Figure 1: BMA predictive pdf: (a) best four models; (b) full model.

Figure 1 shows the BMA forecasting results using the best four models and the full model. The orange vertical line indicates the observation data, and the black vertical line is the 95% confidence interval from the calibrated forecasting result. Based on Figure 1, BMA produces a reliable interval. This shows that the forecasting results of the best four models and the full ensemble are within the 95% confidence interval of the calibrated forecasting result. In addition, the forecasting interval is narrow, meaning that the forecasting

precision is better. The full model's pdf looks wider than that of the four models. This further shows the advantage of BMA, which can reduce underdispersiveness by attempting to adjust the variance, still covering the value of the observations.

4.3 CRPS Mean Value and Percentage of Captured Observations for Calibrated Forecasts using BMA

The purpose of the model evaluation is to determine which calibration method can provide better forecasting results, regarding both accuracy and density. The evaluation indicator uses the CRPS to compare the cdf between forecasting results and observation data. In addition, the evaluation of the calibrated forecast is assessed using the percentage of the captured observations. The CRPS and percentage of captured observations are shown in Table 3.

Table 3: CRPS Mean Value and Percentage of Captured Observations.

	CRPS	Percentage of Captured Observations
Best Four Models	6.27	88.16%
Full Model	5.23	92.11%

Table 3 shows that the full model has a smaller CRPS than the best four models. This indicates that the full model's forecasting results will tend to have better reliability and density and be closer to the observation values. In addition, the percentage of captured observations by the interval calibrated full model is higher than that of the best four models.

5. CONCLUSION

Based on the analysis, it can be concluded that the accuracy model of the best four models produces an average CRPS of 6.27 with 88.16% coverage, while with nine models an average CRPS of 5.23 with 92.11% coverage is obtained. This result suggests using all the ensemble members in order to generate more accurate probabilistic forecasts.

6. REFERENCES

Barnston, A.G., Tippett, M.K., L'Heureux, M.L., Li, S. and DeWitt, D.G. (2011). Skill of real-time seasonal ENSO model predictions during 2002–11: Is our capability increasing?. *Bulletin of the American Meteorological Society*, 93: 631-651.

Bao, L., Gneiting, T., Grimit, E.P., Guttorp, P. & Raftery, A.E. (2013). Bias correction and bayesian model averaging for ensemble forecasts of surface wind direction. *Monthly Weather Review*, 138: 1811-1821.

Becker, E., Van den Dool, D. & Zhang, Q. (2014). Predictability and forecast skill in NMME. *Journal of Climate*, 27(15): 5891-5906.

Doblas-Reyes, F.J., Hagedorn, R. & Palmer, T.N. (2005). The rationale behind the success of multi-model ensembles in seasonal forecasting-II calibration and combination. *Tellus A*, 57: 234-252.

Gneiting, T., Raftery, A.E., Westveld III, A.H. & Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 135(5): 1098-1118.

Hagedorn, R., Doblas-Reyes, F.J. & Palmer, T.N. (2004). The rationale behind the

success of multi-model ensembles in seasonal forecasting. *Tellus A*, 57: 219-233.

Hamill, T.M. & Colucci, S.J. (1997). Verification of Eta-RSM short-range ensemble forecast. *Monthly Weather Forecast*, 125: 1312-1327.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15: 59-570.

Kirtman, B.P., Min, D., Infanti, J.M., Kinter, J.L., Paolino, D.A., Zhang, Q. & Wood, E.F. (2014). The North American Multi Model Ensemble (NMME); Phase-1, Seasonal-to- Interannual Prediction; Phase-2, toward Developing Intraseasonal Prediction. *Bulletin of the American Meteorological Society*, 95(4): 585-601.

Kuswanto, H. (2010). New calibration method for ensemble forecast of non-normally distributed climate variables using meta-gaussian distribution. In: Chaerun, S.K. & Ihsanawati. (eds.): Science for Sustainable Development, *Proceeding of the Third International Conference on Mathematics and Natural Sciences, Bandung*, 23-25 November, pp. 932-939.

Kuswanto, H. & Sari, M.R. (2013). Bayesian model averaging with markov chain monte carlo for calibrating temperature forecast from combination of time series model. *Journal of Mathematics and Statistics*, 9(4): 349-356.

Ma, F., Ye, A., Deng, X., Zhou, Z., Liu, X., Duan, Q. & Gong, W. (2015). Evaluating the skill of NMME seasonal precipitation ensemble predictions for 17 hydroclimatic regions in continental China. *International Journal of Climatology*, 36(1): 132-144.

Palmer, T.N.A. (2001). Nonlinear dynamical perspective on model error: a proposal for nonlocal stochastic–dynamic parameterization in weather and climate prediction models. *Journal Meteorological Society*, 127: 685-708.

Raftery, A.E., Gneiting, T., Balabdoul, F. & Polakowski, M. (2005). Using bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133: 1155-1174.

Sloughter, J.M., Gneiting, T., Raftery, A.E. (2010). Probabilistic wind speed forecasting using ensembels and bayesian model averaging. *Journal of the American Statistical Association*, 105(489): 25-35.

Smith, D.M., Scaife, A.A., Boer, G.J., Caian, M., Doblas-Reyes, F.J., Guemas, V. & Wyser, K. (2013). Real-time multi-model decadal climate predictions. *Climate Dynamic*, 41**:** 2875-2888.

Wang, S., Zhang, N., Wu, L. & Wang, Y. (2016). Wind speed forecasting based on the hybrid ensemble empirical mode decomposition and GA-BP neural network method. *Renew Energy*, 94: 629-36.

Yuan, X. & Wood, E.F. (2013). Multimodel seasonal forecasting of global drought onset. *Geophysical Research*, 40(18): 4900-4905.