Modeling of daily PM$_{2.5}$ concentration based on the principal components regression in South and Central Jakarta

S Alfiandy$^{1*}$, R H Virgianto$^{2*}$ and A S Putri$^2$

1 Department of Climatology, School of Meteorology Climatology and Geophysics (STMKG), Jakarta, 15221, Indonesia
2 Department of Climatology, School of Meteorology Climatology and Geophysics (STMKG), Jakarta, 15221, Indonesia

*E-mail: solihalfiandy@gmail.com; rista.virgianto@stmkg.ac.id.

Abstract. Jakarta as a megacity, has many roles in the distribution of PM$_{2.5}$ pollutant concentrations, especially from the activities of vehicle users. This research estimates daily PM$_{2.5}$ concentration from January to June 2018 at two points, each in the South Jakarta and Central Jakarta based on the principal components regression from daily averaged air temperature, rainfall, relative humidity, and surface wind speed. We retained the first three in principal components to account for over 90% of the variance in the climatological variables. The results show the daily concentration of PM$_{2.5}$ in South Jakarta has the highest correlation coefficient of 0.32 in March 2018 with a $p$-value of 0.079 and MAE value of 10.28. Meanwhile, Central Jakarta has the highest correlation coefficient of 0.46 calculated from daily PM$_{2.5}$ estimation in January 2018 with a $p$-value of 0.013 and MAE value of 8.27. Based on verification results at the two observation points from January to June 2018, it can be concluded that the estimation of the daily PM$_{2.5}$ concentration based on the principal component regression method in the South Jakarta showed slightly better than in South Jakarta.

1. Introduction

According to Mukono[1], air pollution is the increase of physical or chemical substances or substrates into the normal air environment that reaches a certain amount, so that it can be quantified or measured and may bring various impact on humans, animals, vegetation, and material. In addition, air pollution can also define as the change in the atmospheric composition due to the contamination of natural or artificial substances into the atmosphere. Pollutants such as dust, smoke, gas, fog, odor, smoke, or vapor occur in varying amounts, having characteristics and time that have the potential to change the state of the atmosphere[2]. Particulate Matter (PM) in the atmosphere comes from anthropogenic and natural sources[3].

According to Chow[4], the problem of air pollution is caused by solid particles TSP (Total Suspended Particulate) or total floating particles with a maximum diameter of around 45 $\mu$m, PM$_{10}$ particles with diameters of less than 10 $\mu$m and PM$_{2.5}$ with diameters of less than 2.5 $\mu$m. PM is one of the 12 parameters of air pollutants contained in the Republic of Indonesia Government Regulation No. 41 of 1999, this PM has the most dangerous impact on human health because of its ability to enter the deepest respiratory system. Particles measuring 2.5 $\mu$m to 10 $\mu$m can penetrate the lungs without filtering by hair in the nose. Particles under 2.5 $\mu$m or also called PM$_{2.5}$ if inhaled cannot be filtered in the upper respiratory system and will penetrate the deepest part of the lungs[5]. In addition to these human health...
impacts, elements associated with PM$_{2.5}$ can be stored on land and in coastal waters with potentially adverse consequences for ecosystem health[6]. According to Shendrikar and Steinmetz[7], high levels of PM$_{2.5}$ have been shown to result in reduced visibility which can affect transportation security.

Rushayati et al. [8] explain that related to high concentrations of air pollutants will cause a long wave backlash from the Earth's surface to be trapped by air pollutants, resulting in an increase in air temperature. According to Purnami[9], the results of maximum temperature modeling using principal component regression with residual modeling proved to be able to overcome the violation of homogeneous assumptions of variance sequences. Correlation between observational data with alleged data for model verification and model validation results in values that are not significantly different from the principal component regression models. That is, the assumptions in multiple regression are fulfilled, so it can be said that this is an advantage in this model. One assumption for multiple regression is the absence of multicollinearity. Multicollinearity is the existence of a perfect linear relationship between free variables in the model[10]. Multicollinearity is indicated by the value of Variance Inflation Factors (VIF) of more than 10[11]. One way to deal with the problem of multicollinearity is to perform a principal component regression.

In general, besides vehicle factors and human activities, meteorological factors also greatly influence the PM$_{2.5}$ concentrations, for example when air temperatures rise and wind speeds are relatively strong, then pollutants will also be carried out to spread until the wind conditions reach the minimum speed or calm[12]. The next meteorological factors are air humidity and rainfall, when the humidity is high and the air is wet, the pollutants that float in the air will fuse with the wet air, resulting in the deposition of air pollutants to the surface[13]. The principal objective of this study is to estimate the concentration of PM$_{2.5}$ by using the principal component regression with air temperature, rainfall, air humidity and surface wind as input variables at two observation points in South and Central Jakarta.

2. Data and Methods

2.1. Site locations and data collections
We used the hourly PM$_{2.5}$ concentration datasets at two observation points in South Jakarta and Central Jakarta as shown in figure 1. The PM$_{2.5}$ datasets were downloaded from AirNow Website (www.airnow.gov) with the data period from 1 January 2016 to 31 December 2017.
The daily air temperature, relative humidity, 10-meters wind speed, and rainfall datasets were obtained from the Kemayoran Meteorological Station to represent the Central Jakarta area and from the Pondok Betung Climatological Station to represent the South Jakarta area. Same as PM$_{2.5}$ concentration datasets, all meteorological datasets were taken from 1 January 2016 to 31 December 2017.

2.2. Principal component analysis

The Principal Component Analysis (PCA) is a procedure for reducing data dimensions through the transformation of correlated origin variables into a new set of uncorrelated variables[14]. The new variables are called the Principal Components (PCs). Suppose the random vector $\mathbf{x}' = [x_1, x_2, \ldots, x_p]$, which consists of observations as many as $p$ variables, then PC is a linear combination of these variables which is a new coordinate system resulting from the rotation of the original system $x_1, x_2, \ldots, x_p$ as coordinate axis. The new axis $(Z_1, Z_2, \ldots, Z_p)$ is the direction with maximum variability that gives a simpler variation structure and $Z_1, Z_2, \ldots, Z_p$ is the uncorrelated PC[15]. The number of principal components formed is the same as the number of original variables. Reduction or simplification of dimensions is carried out with the percentage criteria for the diversity of data explained by the first few principal components. To determine the number of components to be used, we can use a criterion of the minimum cumulative proportion of variance at 80%[15].

If the principal components are derived from normal multivariate populations with random vectors $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ and average vectors $\mathbf{\mu} = (\mu_1, \mu_2, \ldots, \mu_p)$ and covariance matrix $\Sigma$ with characteristic roots (eigenvalue) i.e. $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ we get the linear combination of the principal components as follows:

$$Y_1 = e'_1 \mathbf{X} = e'_1 X_1 + e'_2 X_2 + \cdots + e'_p X_p$$

$$Y_2 = e'_2 \mathbf{X} = e'_2 X_1 + e'_2 X_2 + \cdots + e'_p X_p$$

$$Y_p = e'_p \mathbf{X} = e'_p X_1 + e'_p X_2 + \cdots + e'_p X_p$$

Then,

$$\text{Var} \ (Y_i) = e'_i \Sigma e_i$$

and

$$\text{Cov} \ (Y_i, Y_k) = e'_i \Sigma e_k$$

where: $i, k = 1, 2, \ldots, p$.

The requirement to form the principal component which is a linear combination of the $\mathbf{X}$ variable in order to have a maximum variance is to choose an eigen vector that is $\mathbf{e} = (e_1, e_2, \ldots, e_p)$ such that $\text{Var} \ (Y_i) = e'_i \Sigma e_i$ maximum and $e'_i e_i = 1$. The $i$-th principal component is a linear combination of $e'_i \mathbf{X}$ which maximizes $\text{Var} \ (e'_i \mathbf{X})$ with the terms $e'_i e_i = 1$ and $\text{Cov} \ (e'_i e_k)$ equal to 0 for $k < 1$. The principal components do not correlate and have the same variation with the eigen of $\Sigma$. The eigen of the variance matrix $\Sigma$ is a variance of the principal component $\mathbf{Y}$, so the variance - covariance matrix of $\mathbf{Y}$ is:

$$\Sigma = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{bmatrix}$$

The total variance of original variables will be the same as the total variance explained by the principal components, therefore:

$$\sum_{j=1}^{p} \text{Var}(X_j) = (\Sigma)^T = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sum_{j=1}^{p} \text{Var}(Y_i)$$

Dimension reduction of the original variables is done by taking a small number of components that are able to explain the largest part of the variance of data[16]. If the principal component is taken as
many as \( q \) component, where \( q < p \), then the proportion of total variance that can be explained by the \( i \)-th principal component is:

\[
\frac{\lambda_i}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} \quad i = 1, 2, \ldots, p
\]  

(8)

Meanwhile, the principal component of the correlation matrix is done if the data has been transformed into the standard \((Z)\) form. This transformation is carried out on the data that the units of observation are not the same. If the observed variables are measured on a scale with large differences or the units of measurement are not the same, then these variables need to be standardized first. The standard variable \((Z)\) is obtained from the transformation of the original variable in the following matrix[17]:

\[
Z = (V^{1/2})^{-1} (X - \mu)
\]  

(9)

\( V^{1/2} \) is the standard deviation matrix with the principal diagonal element is \((\alpha_{ii})^{1/2}\) while the other elements are zero. Expected value of \( E(Z) = 0 \) and its variance is:

\[
\text{Cov}(Z) = (V^{1/2})^{-1} \cdot \Sigma(V^{1/2})^{-1} = p
\]  

(10)

Thus, the principal component of \( Z \) can be determined from the eigen vectors obtained through the correlation of the original variable \( p \). To find the eigen and determine the weighting vector the same as in the matrix \( \Sigma \). While the trace matrix correlation \( p \) will be equal to the number of \( p \) variables used. The selection of the principal components used is based on a cumulative proportion of total variance of at least 90% [18].

2.3. Principal component regression

We use a simple linear regression method using the first principal components that have a cumulative proportion of total variance of at least 90% to simulate daily PM\(_{2.5}\) concentration as a dependent variable for each observation point. The equation used is as follows:

\[
Y = a + b_1 Z_1 + b_2 Z_2 \ldots + b_n Z_n
\]  

(11)

with:

- \( Y = \text{PM}_{2.5} \) estimation
- \( Z = \text{Principal components} \)
- \( a = \text{constant} \)
- \( b = \text{regression coefficient} Z \) with respect to \( y \)

2.4. Model verification

We run model verification to assess the performance of the PM\(_{2.5}\) estimation model based on principal components regression on meteorological variables in South and Central Jakarta. We used a Pearson correlation coefficient as a measure of reliability between the estimated PM\(_{2.5}\) and its observations[19]. We further calculated the Mean Absolute Error (MAE) to measure the accuracy of a forecasting model. MAE values represent the average absolute error between the forecast results \((y_i)\) with the actual value \((o_i)\). Here is the equation for MAE: [20]

\[
\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - o_i|
\]  

(12)
3. Results and Discussion

3.1. The Relation between PM$_{2.5}$ concentration to the meteorological variables

To find out the relation between PM$_{2.5}$ concentration and air temperature, rainfall, air humidity, and surface wind speed, we calculated the Pearson correlation coefficient between the PM$_{2.5}$ concentration with the mentioned meteorological parameters as follows:

| Location          | Air temperature | Rainfall  | Relative humidity | Surface wind speed | Location |
|-------------------|-----------------|-----------|-------------------|--------------------|----------|
| South Jakarta     | 0.39            | -0.17     | -0.16             | -0.46              |          |
| Central Jakarta   | 0.39            | -0.19     | -0.16             | -0.34              |          |

As displayed on table 1 above, at both PM$_{2.5}$ observation points, show the highest positive correlation coefficient between daily PM$_{2.5}$ concentration with air temperature with a value of 0.39. The highest negative correlation coefficient obtained from daily PM$_{2.5}$ concentration and surface wind speed in which correlation coefficient is -0.46 in South Jakarta and -0.34 in Central Jakarta. Furthermore, the correlation coefficients between daily PM$_{2.5}$ concentration and rainfall and also relative humidity show low negative values for both regions.

Positive correlation coefficients of daily PM$_{2.5}$ concentration with air temperature and otherwise negative values with relative humidity indicate that warmer days with lower humidity have an impact in PM$_{2.5}$ concentration increase at both observation points. Then for the negative correlation coefficient of daily PM$_{2.5}$ concentration with daily rainfall denotes that rain effects to decrease of PM$_{2.5}$ concentration by wet deposition. In addition, negative correlation coefficients of daily PM$_{2.5}$ concentration with surface wind speed signifies that stronger wind speed will disperse the PM$_{2.5}$ at a location to its surroundings so the concentration will be lower.

3.2. Results of estimated daily PM$_{2.5}$ concentration using the principal components regression model.

The following table exhibits the principal components based on the principal components regression model of meteorological variables at two site locations.

From the results indicated in table 2, we chose the first three principal components to use in the regression model based on the cumulative proportion of total variance with a minimum value of 90%. The following figure 2 compare the estimated and observed values of the daily PM$_{2.5}$ concentration form 1 January to 30 June 2018.

As demonstrated in figure 2, the estimated daily PM$_{2.5}$ concentrations give a higher value than the observation in January 2018 at both sites. Then for the following months such as February to June, the estimated daily PM$_{2.5}$ concentrations tend to underestimate the observed concentration in South Jakarta while they tend to overestimate in Central Jakarta. Overall, the estimated daily PM$_{2.5}$ concentrations show the same pattern as observation ones although the variability of estimated concentrations is lower than the observation.
Table 2. Principal components calculated from meteorological variables at observation points in South and Central Jakarta from 1 January 2016 to 31 December 2017.

| Location     | Month | Cumulative percentage of variance | Number of retained Z |
|--------------|-------|----------------------------------|----------------------|
|              |       | Z₁      | Z₂      | Z₃      | Z₄      |                     |
| South Jakarta| January| 0.545   | 0.806   | 0.966   | 1.000   | 3                   |
|              | February| 0.479   | 0.756   | 0.963   | 1.000   | 3                   |
|              | March  | 0.465   | 0.726   | 0.904   | 1.000   | 3                   |
|              | April  | 0.482   | 0.775   | 0.962   | 1.000   | 3                   |
|              | May    | 0.525   | 0.754   | 0.944   | 1.000   | 3                   |
|              | June   | 0.501   | 0.732   | 0.941   | 1.000   | 3                   |
| Central Jakarta| January| 0.483   | 0.761   | 0.938   | 1.000   | 3                   |
|              | February| 0.556   | 0.795   | 0.944   | 1.000   | 3                   |
|              | March  | 0.456   | 0.726   | 0.904   | 1.000   | 3                   |
|              | April  | 0.556   | 0.778   | 0.948   | 1.000   | 3                   |
|              | May    | 0.525   | 0.754   | 0.944   | 1.000   | 3                   |
|              | June   | 0.547   | 0.752   | 0.924   | 1.000   | 3                   |

Figure 2. Time series plot of estimated and observed daily PM₂.₅ concentration at observation points in South and Central Jakarta.

The following is a table of correlation results and monthly error per month from the relationship between the PM2.5 concentration values of observations and the predicted PM2.5 concentration values.
Table 3. Verification of daily PM2.5 concentration model in South and Central Jakarta from 1 January to 30 June 2018 which calculated for each month.

| Location       | Location       | Correlation coefficients | p-value | MAE   |
|----------------|----------------|--------------------------|---------|-------|
| South Jakarta  | January        | 0.32                     | 0.248   | 12.93 |
|                | February       | -0.04                    | 0.841   | 10.31 |
|                | March          | 0.32                     | 0.079   | 10.28 |
|                | April          | 0.31                     | 0.100   | 15.94 |
|                | May            | 0.32                     | 0.075   | 17.94 |
|                | June           | 0.34                     | 0.063   | 17.22 |
| Central Jakarta| January        | 0.46                     | 0.013   | 8.27  |
|                | February       | -0.01                    | 0.946   | 8.18  |
|                | March          | 0.22                     | 0.225   | 12.25 |
|                | April          | -0.09                    | 0.645   | 11.96 |
|                | May            | 0.17                     | 0.353   | 14.07 |
|                | June           | 0.11                     | 0.569   | 9.78  |

As shown in table 3, the verification results for each month indicate the highest significant correlation coefficient at 90% significance level occurred in March and May with the coefficient of 0.32 and the lowest MAE shown in March at South Jakarta. In general, the MAE between estimated and observed daily PM2.5 concentrations at South Jakarta ranged from 10.28 to 17.94. Meanwhile, there is no significant correlation coefficient at a 90% significance level between the estimated and observed daily PM2.5 concentration if we calculated for each month at Central Jakarta, though in January shows a high enough correlation coefficient of 0.46 with the MAE of 8.27. In general, the MAE between estimated and observed daily PM2.5 concentrations at Central Jakarta is lower than in South Jakarta ranged from 8.18 to 14.07.

Table 4. Verification of daily PM2.5 concentration model in South and Central Jakarta from 1 January to 30 June 2018.

| Location      | Correlation coefficients | p-value | MAE   |
|---------------|--------------------------|---------|-------|
| South Jakarta | 0.60                     | 0.000   | 14.28 |
| Central Jakarta | 0.46                   | 0.013   | 10.83 |

If we calculated along estimation period from 1 January to 30 June 2018 as shown in table 4, the estimated daily PM2.5 concentration gave a significant correlation coefficient at a 99% significance level only in South Jakarta with a positive correlation coefficient of 0.60 which is higher than in Central Jakarta which only 0.46. However, the model in South Jakarta gives higher MAE with a value of 14.28 than in Central Jakarta which value is only 10.83.

4. Conclusions
The highest correlation between PM2.5 concentration and air temperature, rainfall, humidity, and wind speed is the correlation between PM2.5 concentration and air temperature with a positive correlation.
value of 0.39. In this study, estimating PM$_{2.5}$ concentrations using predictors in the form of air temperature, rainfall, air humidity, and surface wind speed based on the main component regression method, overall yields a good correlation value of 0.60 for the region for the South Jakarta region and 0.59 for the Central Jakarta area.

Estimation results for the South Jakarta region, the highest correlation value occurred in March, which was positive 0.32 with a significant value of 0.079, which is significant at 9% significance level and the lowest MAE value with a value of 10.28 occurred in the same month. Then for the Central Jakarta region, the highest predicted correlation value occurred in January with a positive correlation value of 0.46 with a significant value of 0.013, which not significance at a 90% significance level and the lowest MAE value occurred in February with a value of 8.18. From the estimation results in both regions, it can be concluded that the results of the PM$_{2.5}$ concentration estimation based on the best main component regression method are in the South Jakarta. We can conclude that daily PM$_{2.5}$ estimation model based on principal components regression of meteorological variables shows high reliability - but lower - accuracy in South Jakarta, while low reliability - but higher - accuracy in Central Jakarta. This study may has limitations in estimating daily PM$_{2.5}$ concentration using meteorological parameters are there the choice meteorological factors that may be too few to estimate the PM$_{2.5}$ concentration and other variables may have a major influence in the variability of PM$_{2.5}$ concentration such as vehicles density in traffic at the observation site.

Acknowledgments
We thank the Director-General of The Agency for Meteorology, Climatology and Geophysics (BMKG) for the meteorological datasets access and Board Directors of School of Meteorology Climatology and Geophysics (STMKG) for funding this research. Further thanks to the EPA AirNow program for providing the PM$_{2.5}$ concentration datasets.

References
[1] Mukono H 2011 Aspek Kesehatan Pencemaran Udara: Airlangga University Press
[2] Flagan R C and Seinfeld J H 2012 Fundamentals of air pollution engineering: Courier Corporation
[3] Miranda R and Tomaz E 2008 Atmospheric Research 87 147-57
[4] Chow J C 1995 Journal of the Air & Waste Management Association 45 320-82
[5] Dockery D W, Pope C A, Xu X, Spengler J D, Ware J H, Fay M E, Ferris Jr B G and Speizer F E 1993 New England journal of medicine 329 1753-9
[6] Gao Y, Nelson E, Field M, Ding Q, Li H, Sherrell R, Gigliotti C, Van Ry D, Glenn T and Eisenreich S 2002 Atmospheric Environment 36 1077-86
[7] Shendrikar A D and Steinmetz W K 2003 Atmospheric Environment 37 1383-92
[8] Rushayati S B, Dahlen E N and Hermawan R 2010 Ameliorasi Iklim melalui Zonasi Hutan Kota Berdasarkan Peta Sebaran Polutan Udara. In: Forum Geografi, pp 73-84
[9] Purnami N 2012 Regresi Komponen Utama dengan Pemodelan Sisaan ARCH/GARCH
[10] Juanda B 2009 Ekonometrika pemodelan dan pendugaan Bogor: IPB Press
[11] O’Brien R M 2007 Quality & quantity 41 673-90
[12] Yura E A, Kear T and Niemeier D 2007 Atmospheric Environment 41 8747-57
[13] Matsuda K, Fujimura Y, Hayashi K, Takahashi A and Nakaya K 2010 Atmospheric Environment 44 4582-7
[14] Su C-T and Tong L-I 1997 Total quality management 8 409-16
[15] Johnson R A and Wichern D W 2002 Applied multivariate statistical analysis Vol 5: Prentice hall Upper Saddle River, NJ
[16] Kantardzic M 2011 Data mining: concepts, models, methods, and algorithms John Wiley & Sons
[17] Larose D T and Larose D T 2006 Data mining methods and models vol 12: Wiley Online Library
[18] Jolliffe I T 2002 Principal components in regression analysis Principal component analysis 167-98
[19] Eisinga R, Te Grotenhuis M and Pelzer B 2013 *International Journal of Public Health* **58** 637-42

[20] Wilks D S 2011 *Statistical methods in the atmospheric sciences* Vol 100 (Oxford: Academic Press)