Prediction Model of Air Pollutant Levels Using Linear Model with Component Analysis

Arie Dipareza Syafei, Akimasa Fujiwara, and Junyi Zhang

Abstract—The prediction of each of air pollutants as dependent variable was investigated using lag-1 (30 minutes before) values of air pollutants (nitrogen dioxide, NO$_2$, particulate matter 10µm, PM$_{10}$, and ozone, O$_3$) and meteorological factors and temporal variables as independent variables by taking into account serial error correlations in the predicted concentration. Alternative variables selection based on independent component analysis (ICA) and principal component analysis (PCA) were used to obtain subsets of the predictor variables to be imputed into the linear model. The data was taken from five monitoring stations in Surabaya City, Indonesia with data period between March-April 2002. The regression with variables extracted from ICA was the worst model for all pollutants NO$_2$, PM$_{10}$, and O$_3$ as their residual errors were highest compared with other models. The prediction of one-step ahead 30-mins interval of each pollutant NO$_2$, PM$_{10}$, and O$_3$ was best obtained by employing original variables combination of air pollutants and meteorological factors. Besides the importance of pollutants interaction and meteorological aspects into the prediction, the addition spatial source such as wind direction from each monitoring station has significant contribution to the prediction as the emission sources are different for each station.

Index Terms—Linear regression, principal component regression, independent component regression, air quality prediction, generalized least square.

I. INTRODUCTION

Air pollution prevention has been the leading concern of cities in most developing countries, in particular in Surabaya where the vehicle ownerships increase sharply every year. As a result, the emission of traffic-related pollutants e.g., NO$_2$, PM$_{10}$ increase. Furthermore, the reaction of NO$_2$ with NO will result in the ozone (O$_3$) formation. Therefore, it is mandatory to keep the concentration of these pollutants below permissible level in which the concentration will not affect humans health. High concentration of NO$_2$ and PM$_{10}$ is known to affect human health, whereas O$_3$ is responsible for photochemical smog [1]. High concentration of PM$_{10}$ increase the risk of cardiovascular and respiratory diseases [2]. The prediction model is required to ensure these limits are not surpassed, and if not, the information of the prediction will be crucial for future environmental policies.

In recent time, there have been many attempt to analyze the concentration of air pollutants and explore them to build short-term forecast of concentrations. Linear and non-linear models were developed, however, there was no significance difference noted between non-linear and linear models [3]. Reference [4] used a forecasting model called hierarchical technique to predict CO, NOx, and dust fall. Reference [3] compared five linear models to predict daily mean of PM$_{10}$ concentrations in one site in Oporto Metropolitan Area. However, spatial variability were not concerned on that study and the regression with variables obtained independent component analysis performed the worse. Reference [5] employed Artificial Neural Network (ANN) to predict CO, NO$_2$, PM$_{10}$ and O$_3$ concentrations and the performance was better compared with multiple linear regression. On this research, wind direction was considered as independent variables but they did not separate the effect of wind direction to each prediction of pollutant, moreover, serial error correlation due to time series model was not taken into account which might cause result bias. Reference [6] also used ANN to predict pollutants, but they noted less accuracy for O$_3$ prediction in Tehran, Iran.

The time series is an appropriate model which avoids the problems of geographical aspects. However, the trends observed in a pollution data presents serial error autocorrelation which generates problems in interpretation, analysis, and prediction [7]. Many researchers have performed the forecasting by regression technique but unfortunately many authors did not account for serial error autocorrelation.

Moreover, in a regression analysis, the correlation between independent variables (multicollinearity) may pose a serious difficulty in the interpretation of which predictors are the most influential to the response variables [8]. One way to remove such multicollinearity is using component analysis method, in this case widely used a Principal Component Analysis (PCA), and the newly emerged one Independent Component Analysis (ICA). Even though these two methods have their own approach, the goal is similar is to build components that are statistically independent with each other. In regression analysis, this is particularly very useful and become good input as predictors in a regression model since they optimize spatial patterns and remove complexity due to multicollinearity [8], [9]. ICR and PCR have been widely used in particular for plant study [9], dam deformation study [10], air pollutants in subway [11], air quality management [2], [3], [12], and O$_3$ prediction [1].

In this paper, we will predict one-step (next 30mins) ahead three pollutants, namely, NO$_2$, PM$_{10}$, and O$_3$ concentrations by including some spatial and temporal factors. Important variables which are also included are six air pollutants (NO$_2$, NO, O$_3$, SO$_2$, CO, PM$_{10}$), and meteorological factors (wind speed, wind direction for each station, solar gradation,
humidity, and temperatures). We employ a Generalized Least Square (GLS) model with taking into concern the series of error autocorrelation. As far as author concerns this study is the first one applied in current city and country and its ultimate benefit that we may be able to show the key factors of a model which can be applied further into other areas and regions, in particular within Indonesia.

II. MATERIALS AND DATA

We make use of 30-mins interval concentrations of NO, NO\textsubscript{2}, O\textsubscript{3}, SO\textsubscript{2}, CO, and PM\textsubscript{10}, as well as meteorological factors that consist of wind direction, wind speed (m/s), solar gradation (W/m\textsuperscript{2}), humidity (%), and temperatures (°C). These data were obtained from Air Quality Laboratory of Environmental Agency in Surabaya City, recorded from five monitoring stations installed. These five stations represent: city center (Ketagang Kali – Station 1), trading zone (Perak – Station 2), suburban (west side of Surabaya, Sukomanunggal – Station 3), near highway zone (Gayungsari – Station 4), and suburban (east side of Surabaya, Sukolilo – Station 5).

In the present study we attempt to predict NO\textsubscript{2}, PM\textsubscript{10}, and O\textsubscript{3} using GLS model using the data taken from March 2002 to April 2002, with total 14635 observations (five stations), as training set, whereas the test set, which is not used for parameter estimation, was taken from May 2002. Missing values were treated using Expectation Maximization based algorithm by package Amelia run through R open source program [13]. Table I shows each mean value, standard deviation, and median of each pollutant concentrations (in ug/m\textsuperscript{3}) from five stations in Surabaya. It can be seen that average emission that is related to traffic (NO\textsubscript{2}) was high in city center and low in suburban2 on the east side of Surabaya. With regards of PM\textsubscript{10}, high average concentration was found on suburban1 on west side of Surabaya. Interestingly, mean value of O\textsubscript{3} concentration was high on highway zone suggesting high reaction rate between NO and NO\textsubscript{2} due to traffic flow.

<table>
<thead>
<tr>
<th>Description</th>
<th>City Center</th>
<th>Trading</th>
<th>Suburban1</th>
<th>Near Highway</th>
<th>Suburban2</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO\textsubscript{2}</td>
<td>Min</td>
<td>1.603</td>
<td>0.335</td>
<td>1.175</td>
<td>0.48</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>12.309</td>
<td>11.577</td>
<td>10.013</td>
<td>16.024</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>5.487</td>
<td>4.745</td>
<td>4.85</td>
<td>5.266</td>
</tr>
<tr>
<td></td>
<td>Standard Deviation</td>
<td>1.518</td>
<td>1.703</td>
<td>1.503</td>
<td>1.758</td>
</tr>
<tr>
<td>PM\textsubscript{10}</td>
<td>Min</td>
<td>0.656</td>
<td>0.1</td>
<td>0.317</td>
<td>0.541</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>17.38</td>
<td>48.799</td>
<td>17.689</td>
<td>48.99</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>7.137</td>
<td>7.673</td>
<td>7.857</td>
<td>7.488</td>
</tr>
<tr>
<td></td>
<td>Standard Deviation</td>
<td>2.437</td>
<td>2.768</td>
<td>2.527</td>
<td>3.122</td>
</tr>
<tr>
<td>O\textsubscript{3}</td>
<td>Min</td>
<td>1.558</td>
<td>0.042</td>
<td>0.01</td>
<td>0.042</td>
</tr>
<tr>
<td></td>
<td>Standard Deviation</td>
<td>1.984</td>
<td>1.579</td>
<td>1.941</td>
<td>2.21</td>
</tr>
</tbody>
</table>

GLSs were then fitted, with square-root transformed of each pollutant (NO\textsubscript{2}, O\textsubscript{3}, or PM\textsubscript{10}) as the dependent variable. The explanatory variables were functions of lag-1 30-mins interval of pollutant levels: NO\textsubscript{2}, NO, O\textsubscript{3}, SO\textsubscript{2}, CO, PM\textsubscript{10}, wind speed, solar gradation, humidity, temperatures, status of day (weekends, workdays as base reference), peak time of morning and afternoon session (non-peak time as base reference with peak time morning is between 630am to 9am and in the evening between 430pm to 7pm), holidays, spatial covariates of zones: trading, suburban1, highway, suburban2, with city center as base reference, and wind direction. For wind direction we created eight variables representing direction as dummy variables. They are north, northeast, east, southeast, east, south, southwest, west, and northwest, with north as base reference. These variable values are different for each station, thus creating more 35 wind direction variables for model input. All air pollutants and wind speed were all square-root transformed as standard procedures to stabilize the variance. Each dependent variable is predicted by the interaction of other pollutants one-step backward (last 30-mins concentration). The three pollutants were considered in three separate GLS models because of their substantial correlation. On a second and third model, the variables of six air pollutants and four meteorological factors were replaced by components extracted from an ICA and PCA. Total there were 10 ICs and 10 PCs were obtained and used as predictor variabels along with other independent variables as described above.

III. MODELS

A. Generalized Least Square

We employ a Generalized Least Squares (GLS) model to formula the mixed linear effect of predictor variables towards the concentration of pollutants, following the equation:

$$ y = X\beta + \varepsilon $$ (1)

where $y$ is $n \times 1$ response variable (pollutant) and $X$ is an $n \times p$ matrix, $\beta$ is a $p \times 1$ vector of estimated parameters, and $\varepsilon$ is $n \times 1$ vector of errors. With the assumption that $\varepsilon \sim N_p(0, \sigma^2 I_n)$, we can estimate ordinary least square estimator of $\beta$:

$$ \beta_{GLS} = (X'X)^{-1}X'y $$ (2)

With covariance matrix

$$ V(\beta_{GLS}) = \sigma^2 (X'X)^{-1} $$ (3)
When the error covariance $\Sigma$ positive-definite and symmetric and its diagonal entries $\Sigma$ correspond to non-constant error variances, and nonzero off-diagonal entries are associated with correlated errors, we can estimate the log-likelihood of the model, given that $\Sigma$ is known:

$$\log L(\beta) = -\frac{n}{2} \log(\det \Sigma) - \frac{1}{2} (y - X\beta)'(y - X\beta)$$

(4)

The function is maximized by the GLS estimator of $\beta$:

$$\hat{\beta}_{GLS} = \left( X' \Sigma^{-1} X \right)^{-1} X' \Sigma^{-1} y$$

(5)

With covariance matrix:

$$V(\beta_{GLS}) = \left( X' \Sigma^{-1} X \right)^{-1}$$

(6)

However, in the application, the matrix of $\Sigma$ is not known and therefore must be estimated from the data with the regression coefficients, $\beta$.

In time series data, though, there is a concern of error correlation. Assuming that all errors have same expectation and same variance, the covariance of two errors depends on their separation $s$ in time:

$$C(e_i, e_{i+s}) = C(e_i, e_{i-s}) = \sigma^2 \rho_s$$

(7)

where $\rho_s$ is the error autocorrelation at lag $s$. The error-covariance matrix will become:

$$\Sigma = \sigma^2 \begin{bmatrix}
1 & \rho_1 & \rho_2 & \cdots & \rho_{n-1} \\
\rho_1 & 1 & \rho_1 & \cdots & \rho_{n-2} \\
\rho_2 & \rho_1 & 1 & \cdots & \rho_{n-3} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \cdots & 1
\end{bmatrix} = \sigma^2 \mathbf{P}$$

(8)

For stationary time-series, we apply first-order auto-regressive process, AR(1) for autocorrelated regression errors:

$$e_i = \phi e_{i-1} + v_i$$

(9)

Under this model, the $v_i$ is assumed to be Gaussian white noise, $\rho_1 = \phi e_{i-1}$, $\rho_2 = \phi^2$, and $\sigma^2 = \sigma^2 / (1 - \phi^2)$, along with the time run, the error autocorrelations $\rho_s$ will decay exponentially as $s$ increases to 0. A GLS model is run through a gls command under nlme library package within R open source software.

B. Independent Component Analysis

In ICA, the input variables are regarded as linear combinations of latent variables which are considered independent and non-Gaussian. ICA establishes independent components from original variables. The concept of ICA is regarded to be able in explaining more for variable relationship because independence is a high-order statistic that is in favor over orthogonality [9]. The GLS regression forms relationship between response variable ($y$) and the ICs from ICA along with other explanatory variables (e.g., days in week, season). A typical ICA model is expressed as:

$$X = \Sigma A$$

(10)

$X$ is observation matrix, derived through the mixing of an $n$-dimensional source matrix, $S = (s_1, \ldots, s_n)'$, with temporal dimension of 1, referred to ICs, with $n$ is independent components extracted. $A$ is the mixing matrix of dimension $n \times m$ or $m \times n$ where $m \leq n$. The objective of ICA is to estimate $A$ and $S$, knowing only the observations matrix $X$. The present study uses Fast ICA algorithm to estimate $A$ and $S$ from observations $X$. The $S$ components will be used as input variables in the model. A fast ICA function within R program was used to obtain ICs. For forecasting purpose, the following formula is used to obtain ICs for input to the prediction model:

$$XA^{-1} = S$$

(11)

$X$ is lag-1 independent variables whereas $A$ is the inverse of loading matrix obtained from training set data. There were 10 ICs obtained as input variables.

C. Principal Component Analysis

PCA creates principal components (PCs) that are orthogonal and uncorrelated and linear combinations of the original variables. The first PC is the one that has the largest portion of original data variability. A varimax rotation is commonly used to obtain rotated factor weight loadings that represent effect of each each variable in one particular PC. MPC regression examines a relationship between the output variable ($y$) and the PCs obtained from explanatory variables (air pollutants: NO, NO2, O3, SO2, CO, and PM10, and meteorological factors: wind speed, solar radiation, humidity, and temperatures). The estimation procedure is given in the following equations:

$$PC_{ij} = \sum_{k=1}^{m} w_{ik} x_{ij}$$

(12)

where $PC_{ij}$ is the PC score for $i$th component and $j$-th object. The loading weight is represented by $w_{ik}$ for $k$-th variable variable on the $i$-th component, and $x_{ij}$ is the standardized value of $k$-th variable for the $j$-th observation [14]. A PCA is run using prcomp function within R open source program. To obtain PCs for prediction purposes, the $x_{ij}$ were simply lag-1 independent variables.

IV. MODEL TERMS, FORECASTING, AND PERFORMANCE INDEXES

In the present study we define the term model as follow: Model 1 is a Generalized Least Square (GLS) model with original square-root transformed independent variables (air pollutants and meteorological factors). Model 2 refers to that predictor variables are replaced by components extracted from ICA. Model 3 refers to a model with the predictor
variables are replaced by components obtained from PCA. Forecasting of the model was done by comparing fitted values and observation values in Station 1 (city center) and Station 5 (suburban 2). The three models were compared and judged using these statistical performance index: mean error (ME), mean absolute error (MAE), root mean square error (RMSE), and $R^2$, that are commonly used in many literatures. ME is useful to obtain whether the fitted values overestimate or underestimate. MAE and RMSE measure the magnitude of difference between predicted values and observed values, the lower the better. The $R^2$ indicate percentage of which variance can be explained by variables.

V. RESULTS AND DISCUSSIONS

The correlation analysis of the raw concentration between pollutants and meteorological factors were displayed in Table II. The Pearson correlation reported that there was a linear correlation between the predicted pollutant concentrations and the independent variables. To evaluate correlation coefficients, the significance value was calculated with a significance level of 0.05, marked by values in bold.

We note that NO, CO, wind speed, solar radiation have strong correlation with concentrations of NO$_2$, NO and CO has positive correlation whereas wind speed and solar radiation have negative correlation. Higher wind speed and solar radiation allows for gas purification in the lower layer of atmosphere. It was noted that NO, wind speed, solar radiation, humidity, and temperatures have statistically significant correlation values with O$_3$. This is due to that the presence of O$_3$ are highly due to chemical reaction in the atmospheric layer. The negative correlation of NO and O$_3$ indicates that higher NO concentration along with NO$_2$ reaction cause O$_3$ depletion. On the other hand, wind speed and solar radiation have positive correlation with O$_3$ despite of opposite coefficient for humidity, lower humidity tend to increase O$_3$ presence. Since PM$_{10}$ is an inert gas particle, there was no significant variables which affect the concentration of PM$_{10}$ although different results were indicated by [3].

The performance of GLS model with original variables perform the best compared with ICR and PCR model for both internal validation (Table III) and forecasting (Table IV) for NO$_2$. Observing the performance for other pollutants, we found that the performance of ICR was the worst than PCR and GLS model. Specifically for O$_3$ prediction, we did not obtain comparable performance as those measured by [8] showing that the use of PCR alone yielded $R^2$ of 0.965, assuming we care less of interval concentration of O$_3$ they used. However, if we compare the performance of ICR and PCR in the training set (internal validation), ICR performed better than PCR for O$_3$ and PM$_{10}$ on two stations. This result may indicate better component extraction due to the nature of ICA process. Unfortunately, the performance of ICR in the forecasting was worse than PCR for all pollutants. These facts show the failure of the component analysis method in producing reliable surrogate variables for the model.

Observing the result of different stations, we may observe that the performance of ICR was worse than PCR in station 1 (city center) in terms of RMSE. On the other hand, in station 5, the difference of RMSE was not far. This may indicate the sensitivity of performance result highly depend how ICs perform component extraction from the training data set. This may impose a risk where if the particular station has concentrations which highly fluctuate with much noises, this may affect the components extracted, thus affect the prediction severely. The following factors may contribute to the errors of the models such as: a) errors in data itself (monitoring), b) high fluctuation of concentrations due to short interval time concentrations, c) a linear equation may not be enough to capture the true air quality because of complexity and uncertain system of urban air management.

Table V shows parameter estimates for the GLS model with NO$_2$, O$_3$, and PM$_{10}$. The independent variables are the original square-root transformed variables. For NO$_2$, among other independent variables, besides lag-1 NO$_2$ the concentrations of PM$_{10}$ and CO has been the most important predictor for NO$_2$, which on the opposite, CO also has high contributions to the concentration of PM$_{10}$. On the other hand, statistically significant as expected, the contribution between NO and NO$_2$ to the forecasting O$_3$ concentrations was noticed to be significant.

Furthermore, the increase of wind speed help mixing process of NO$_2$, PM$_{10}$, and O$_3$ whereas on the opposite, higher temperature will likely to increase the concentrations. Besides for O$_3$, solar radiation negatively affect the predicted concentrations of NO$_2$ and PM$_{10}$. The congestion level of traffic during peak time affect the emission of NO$_2$, and as a result, the concentrations during peak times are higher than non-peak time, in particular during morning session, also for PM$_{10}$ and O$_3$. In the evening time, the concentrations of O$_3$ tends to be lower than non-peak time. The concentrations are also lower during holidays and weekends for NO$_2$ and PM$_{10}$, but the opposite for O$_3$, indicating photochemical reaction of among contributor gases.

Observing the spatial location, it is interesting that the NO$_2$ concentrations were found to be higher in suburban 1 area
Sukomanunggal (west side of Surabaya) than city center. However, that particular location was highly-densed residentials that emit larger than other zones. Only in the east of Surabaya (suburban 2), the NO₂ concentrations were lower than other zones. In the trading zone, the predicted PM₁₀ concentration is higher in suburban 1 (Sukomanunggal) than other zones whereas we note higher estimate for O₃ in east side of Surabaya, Sukolilo region, as indicated by suburban 2 variable.

**TABLE III: INTERNAL VALIDATION OF EACH MODEL FOR NO₂**

<table>
<thead>
<tr>
<th>Station</th>
<th>Model</th>
<th>ME</th>
<th>MAE</th>
<th>RMSE</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO₂</td>
<td>1</td>
<td>0.0218</td>
<td>0.4742</td>
<td>0.6437</td>
<td>0.8004</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.3969</td>
<td>0.9462</td>
<td>1.2231</td>
<td>0.3496</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.7914</td>
<td>2.2566</td>
<td>2.7951</td>
<td>0.0057</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.7914</td>
<td>1.9074</td>
<td>2.3769</td>
<td>0.0134</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>-1.9166</td>
<td>2.1259</td>
<td>2.59</td>
<td>-04</td>
</tr>
</tbody>
</table>

**TABLE IV: PERFORMANCE OF FORECASTED VALUES IN STATION 1 AND 5 FOR EACH MODEL FOR NO₂**

<table>
<thead>
<tr>
<th>Station</th>
<th>Model</th>
<th>ME</th>
<th>MAE</th>
<th>RMSE</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO₂</td>
<td>1</td>
<td>0.0218</td>
<td>0.4742</td>
<td>0.6437</td>
<td>0.8004</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.3969</td>
<td>0.9462</td>
<td>1.2231</td>
<td>0.3496</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.7914</td>
<td>2.2566</td>
<td>2.7951</td>
<td>0.0057</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.7914</td>
<td>1.9074</td>
<td>2.3769</td>
<td>0.0134</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>-1.9166</td>
<td>2.1259</td>
<td>2.59</td>
<td>-04</td>
</tr>
</tbody>
</table>

**TABLE V: PARAMETER ESTIMATES OF NO₂, PM₁₀, AND O₃ WITH LAG-1 INDEPENDENT VARIABLES**

<table>
<thead>
<tr>
<th>Variables</th>
<th>NO₂</th>
<th>PM₁₀</th>
<th>O₃</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.656</td>
<td>-2.025</td>
<td>0.872 (1.763)</td>
</tr>
<tr>
<td>NO</td>
<td>-0.006</td>
<td>-0.0397</td>
<td>(-2.575)</td>
</tr>
<tr>
<td>NO₂</td>
<td>0.085</td>
<td>0.0414</td>
<td>(3.477)</td>
</tr>
<tr>
<td>O₃</td>
<td>0.005</td>
<td>0.744 (107.64)</td>
<td></td>
</tr>
<tr>
<td>SO₂</td>
<td>-0.0017</td>
<td>0.004 (0.852)</td>
<td></td>
</tr>
<tr>
<td>PM₁₀</td>
<td>-0.015</td>
<td>0.001 (0.198)</td>
<td></td>
</tr>
<tr>
<td>CO</td>
<td>0.298</td>
<td>0.0125 (2.532)</td>
<td></td>
</tr>
<tr>
<td>Wind Speed</td>
<td>-0.096</td>
<td>-0.366</td>
<td>-0.052</td>
</tr>
</tbody>
</table>

Solar Gradation: -0.001 (-11.375) -0.0005 (-5.440) 0.0002 (3.012)
Humidity: -0.006 (-5.502) 0.008 (3.041) -0.005 (-2.450)
Temperatures: 0.013 (12.21) 0.120 (6.674) 0.027 (2.092)

**Peak session time (non-peak time as base)**

Peak morning: 0.257 (9.404) 0.110 (2.506) 0.546 (18.086)
Peak afternoon: 0.257 (9.404) -0.015 (-0.325) -0.218 (-6.724)

**Holidays (non-holiday as base)**

Weekends: -0.137 (-5.912) -0.298 (-6.448) 0.010 (0.335)

**Wind direction (north as base for each station)**

Northeast1: -0.213 (-2.651) -0.235 (-1.731) -0.003 (-0.036)
Northeast2: 0.123 (1.644) 0.450 (3.518) 0.030 (0.339)
Northeast3: 0.328 (3.924) 0.198 (1.402) -0.061 (-0.650)
Northeast4: -0.493 (-6.690) -0.465 (-3.705) 0.079 (0.903)

**South1**: 0.452 (6.073) 1.015 (7.995) 0.025 (0.282)
**Southwest1**: 0.298 (2.933) 0.360 (2.106) -0.020 (-0.162)
**West1**: 0.123 (1.644) 0.450 (3.518) 0.030 (0.339)
**Northwest1**: 0.328 (3.924) 0.198 (1.402) -0.061 (-0.650)
**Northeast2**: -0.493 (-6.690) -0.465 (-3.705) 0.079 (0.903)
**East1**: -0.731 (-10.169) -0.662 (-5.372) 0.004 (0.044)
**South2**: -0.709 (-9.142) -0.589 (-4.335) 0.028 (0.305)
**Southeast2**: -0.303 (-5.392) -0.649 (-6.170) -0.002 (-0.026)
**South3**: -0.303 (-5.392) -0.649 (-6.170) -0.002 (-0.026)
**Southeast3**: -0.154 (-1.768) -0.454 (-3.054) -0.026 (-0.247)
**South4**: -0.023 (-0.315) 0.425 (3.399) -0.108 (-1.228)
**Southwest4**: -0.303 (-4.772) 0.291 (2.687) 0.110 (1.47)
**West4**: -0.672 (-9.153) -0.178 (-1.425) 0.100 (1.136)
**Northwest4**: -0.111 (-1.273) -0.320 (-2.157) 0.281 (2.674)
**Northeast4**: -0.135 (-1.443) -0.068 (-0.430) 0.064 (0.563)
**East4**: -0.347 (-4.785) -0.061 (-0.493) 0.016 (0.186)
**Southeast4**: -0.473 (-6.034) -0.272 (-2.042) -0.063 (-0.670)
**South4**: -0.123 (-1.437) 0.656 (4.531) 0.038 (0.366)
**Southwest4**: 0.031 0.841 0.037 (0.388)
The inclusion of wind direction has been proven to be helpful not only for predicting one-step ahead concentrations, but also to identify and estimate possible sources of emission by judging from where the wind dominantly flew. In city center, the direction of southeast is dominant contributor of NO₂ emission because on this direction, there is a big crowded road which includes a governor office. For trading center, the direction of southeast is dominant contributor of pollutants concentrations, and therefore it is necessary to separate these variables according to their monitoring station location. Holidays and weekends have important role for prediction because the traffic-related emissions tend to be lower during holidays and weekends, due to lower traffic flow. A similar approach in the present study could be extended by incorporating days within week to the data from other stations on other cities to establish a prediction. The information will be very helpful for decision makers in polishing environmental policy related to air quality and human’s health.

VI. CONCLUSIONS

There is a concern of adverse effect to humans health due to high concentration of pollutants which exceed the standard value. These events occur often and people should get alerted when this happens, thus making the short-term prediction of pollutant become crucial. Linear models with original variables, ICS, and PCs extracted from six pollutants (NO, NO₂, O₃, SO₂, CO, PM₁₀) and meteorological factors (wind speed, solar gradient, humidity and temperatures) were employed to predict 30-mins ahead of NO₂, PM₁₀, and O₃. In addition, we include serial error correlation computation in the model for model accuracy. As expected, the presence of NO has positive correlation with NO₂, aside with CO, wind speed and solar gradient. Furthermore, it was shown that meteorologica factors have high role in the formation of O₃. The faster wind speed will reduce the concentration of NO₂ while on the opposite will increase the concentration of O₃. This pattern is also found for humidity. Since PM₁₀ is relatively inert particle gas with less than 10μm, using the 30-mins data we obtained, no significant correlation was found with other variables.

During the test step, indicated by internal validation on two zones used for validation, the residual errors from GLS regression were generally lower than ICR and PCR for all pollutants observed (NO₂, PM₁₀, and O₃). Removing the correlation among variables by using predictor variables from ICA and PCA did not improve the performance in the forecasting step. This means the components obtained from these two methods may also remove the necessary pattern of the concentration profiles that affect the prediction. use of wind speeds and directions and air temperatures.

Using 30-mins interval concentrations of NO₂, PM₁₀, and O₃, we have indicated the effect of other pollutants effect and meteorological factors. The ultimate benefit on using such short interval data is we can exploit the benefit of wind direction instead of taking average if the data were hourly or even daily. Among meteorological factors, we find that the direction of wind highly affect pollutants concentrations, and therefore it is necessary to separate these variables according to their monitoring station location. Holidays and weekends have important role for prediction because the traffic-related emissions tend to be lower during holidays and weekends, due to lower traffic flow. A similar approach in the present study could be extended by incorporating days within week to the data from other stations on other cities to establish a prediction. The information will be very helpful for decision makers in polishing environmental policy related to air quality and human’s health.

REFERENCES

Arie Dipareza Syafei was born in Surabaya on January 19, 1982. He obtained his bachelor degree at the Department of Environmental Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia. His master degree was conferred by Graduate School of Environmental Engineering, National Taiwan University, Taiwan in 2007, specializing in environmental pollution, in particular water treatment using membrane technology. Now he is working and focusing on atmospheric purification and/or air pollution area.

His current position is a faculty staff in Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia at the Department of Environmental Engineering, since 2005. However, since 2011 until current, he has been a PhD candidate in the Graduate School of International Development and Cooperation (IDEC), Hiroshima University. His current and future research is air pollution and air quality monitoring assessment. Furthermore, he would incorporate the use of remote sensing to overcome spatial data limitation that is often faced by policy makers in developing countries.