RESEARCH PAPER

Data-driven analysis using fuzzy time series for air quality management in Surabaya

Didiet Darmawan¹, Mohammad Isa Irawan², Arie Dipareza Syafei³
¹Dept. of Technology Management, Faculty of Business and Technology Management, Sepuluh
Nopember Institute of Technology, Indonesia
²Dept. of Mathematics, Faculty of Mathematics and Natural Sciences, Sepuluh Nopember Institute of
Technology, Indonesia
³Dept. of Environmental Engineering, Faculty of Civil Engineering, Environment,
and Earth, Sepuluh Nopember Institute of Technology, Indonesia

Abstract. One of the environmental issues that affect human health is air pollution. As the second largest city in Indonesia, economic development and infrastructure construction in Surabaya lead to the increasing role of industry and motor vehicle use which is proportional to increasing fuel oil consumption. This condition ultimately leads to the deterioration of air quality. Pollutants that contribute to air pollution such as CO, SO₂, O₃, NOₓ, and particulate matter PM₁₀ can directly affect human health. This study aimed to analyze, monitor, and predict air pollutant concentrations which have been recorded by the Environment Agency of Surabaya City using Fuzzy Time Series. The result determined the MAPE of the pollutant parameters with scores of 23.6% (NO₂), 19.5% (CO), 22.75% (O₃), 9.96% (PM₁₀), and 3.6% (SO₂).

Keywords: fuzzy time series; forecasting; air pollution

1. Introduction

Surabaya as one of the major cities in Indonesia with a population of 2,943,528 has many kinds of environmental issues (BPS Kota Surabaya, 2015). One of the environmental issues, which becomes a major concern for big cities in Indonesia and affects human health, is air pollution.

Economic development in Surabaya city has been increasing the role of industry and vehicle usage which leads to increasing fuel oil consumption (Rismaharini, 2016). These activities contribute to air pollution, for example one liter of fuel which is used for the combustion process will generate 30 grams of nitrogen oxides (NOₓ) and 100 grams of carbon monoxide (CO) (Hickman et al., 1999). These pollutants influence the health of the human respiratory tract whenever the level exceeds the safe limits. Epidemiological studies have concluded the close connection between the level of urban air pollution and

* Corresponding author. E-mail: didiet.d@gmail.com
DOI 10.22515/sustinere.jes.v1i2.13
the occurrence of respiratory diseases (Priyana & Abadi, 2000), for example the case of NO₂ pollution. When the level of NO₂ at 250 μg/m³ increases to 500 μg/m³, it will cause a respiratory malfunction as gaseous NO₂ will make it difficult for humans to take a breath (Ramadhan et al., 2015).

To mitigate the impact of air pollution, it is important to manage environmental and air quality (Rismaharini, 2016). Thus, air quality management efforts have been carried out by supervising and monitoring all activities that have the potential to contaminate the air (Pemerintah Kota Surabaya, 2008). Pollutant monitoring activities have been carried out by the Environment Agency (BLH) and the findings are reported to the public in the form of pollutant Index to measure air pollutants, such as nitrogen dioxide (NO₂), carbon monoxide (CO), ozone (O₃), particulates (PM₁₀), and sulfur dioxide (SO₂).

In this study, fuzzy time series based on the average interval was used to establish prediction based on the time series data record from Quality Monitoring Stations of the Environment Agency in Surabaya. The researcher used five parameters: nitrogen dioxide (NO₂), carbon monoxide (CO), ozone (O₃), particulates (PM₁₀), and sulfur dioxide (SO₂) as variables to be forecasted using fuzzy time series.

2. Literature review
2.1. Fuzzy set and fuzzy time series

The basic principle of fuzzy sets is to represent a value that does not just mean false (0) or true (1) (Zadeh, 1978; Bocklisch et al., 2000; Kusumadewi dan Purnomo, 2004; Tauryawati dan Irawan, 2014). Fuzzy set membership values lie in the range from 0 to 1, which means that the interpretation of fuzzy sets can be used to represent each value based on opinions or decisions (Alpasan et al., 2012; Hansun, 2012; Anan et al., 2015; Hakim, 2015; Akbar et al., 2016). However, there are values that lie between right and wrong. In other words, the truth value of an item is not only the right and the wrong (Zhang and Zhu, 2013; Casillas et al., 2013; Beratu & Sediyono, 2013). In the terms of fuzzy sets, there is a universe of discourse and the whole value is allowed to be operated in fuzzy variables. A universe of discourse is a set of real numbers that always rise or increase monotonically from left to right. The value universe of discourse can be shown either as a negative or positive number (Babuska, 1999; Chen & Hwang, 2000; Kieso, 2002; Anggriani, 2012). For example, the universe of discourse for the sample variable temperature [-7°C, 50°C].

Fuzzy time series prediction data are a method that uses the concept of fuzzy sets as the basis for calculation (Bocklisch et al., 2000; Kusumadewi & Purnomo, 2004; Poulsen, 2009). The method is used to capture the pattern of past data and predict the future data (Casillas et al., 2013; Rachmawansah, 2014). The definitions of fuzzy time series according to Song and Chissom (1993) are:

Definition 1: \( Y(t) \) (\( t = 0, 1, 2, \ldots \)) is a subset of \( R \). Let \( Y(t) \) is the set of rules described by the fuzzy set \( \mu(t) \). If \( F(t) \) consists of \( \mu(i)(i = 1, 2, \ldots) \), \( F(t) \) is called fuzzy time series \( Y(t) \).

Definition 2: if \( F(t + 1) = A_i \) and \( F(t) = A_j \), a fuzzy logical relationship can be described as \( A_i \rightarrow A_j \), where \( A_i \) and \( A_j \) are the left side and the right side of the fuzzy logical relationship.
2.2. Air quality monitoring

Air quality monitoring is an activity to calculate the pollutant contents in the air (Chen & Hwang, 2000; Dwiputra & Syafei, 2015). Calculation of the air pollutant contents is carried out in order to get a general idea of how the pollution occurs. In general, the purpose of the air quality monitoring is to provide a strong scientific basis for any development activities that require a cost-effective policy and rules to conquer air pollution, determine the air quality standards and thresholds, evaluate the potential impacts of air pollution on ecosystems and the environment, and meet the air quality requirements (Ahmet & Van Dijk, 1994; Priyana & Abadi, 2011).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Time of exposure (hours)</th>
<th>Standard quality (µg/Nm³)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SO₂</td>
<td>24</td>
<td>262</td>
</tr>
<tr>
<td>CO</td>
<td>8</td>
<td>22,600</td>
</tr>
<tr>
<td>NO₂</td>
<td>24</td>
<td>92,5</td>
</tr>
<tr>
<td>O₃</td>
<td>1</td>
<td>200</td>
</tr>
<tr>
<td>PM₁₀</td>
<td>24</td>
<td>260</td>
</tr>
</tbody>
</table>

Source: Regulation of the Governor of East Java No. 10 of 2009

3. Research Methods

This study used fuzzy time series as the method to design and analyze historical data series of air pollution parameters which were collected from fixed monitoring stations. Data on nitrogen dioxide (NO₂) and carbon monoxide (CO) were collected from SUF 6 located in Wonorejo. Data on ozon (O₃) were collected from SUF 1 at Prestasi Park and data on PM₁₀ were collected from SUF 3 in Sukomanunggal. Those data were recorded from 8–21 July 2013. Meanwhile, data on sulfur dioxide (SO₂) from SUF 4 in Gayungan were collected from 20 July–2 August 2013. All data were used to represent the actual values of the parameters every 30 minutes. Additionally, in this paper all data were multiplied by 100 to simplify the calculation and divided back by the original value to determine the Mean Square Error (MSE) and the Mean Absolute Percentage (MAPE).

In this study, fuzzy time series used is as follows:

1) Determine the data interval. Calculate the absolute value of the data difference value where \( D_i (i = 1, \ldots n-1) \), so we can get:

\[
\sum_{i=1}^{n-1} |D_{i+1} - D_i|
\]  

Then, divide the summary with the total amount data

2) To determine the interval basis, divide the result from point (2) by 2.

3) The next step is to form the Universe of Discourse (U) using the following equation:

\[
(D_{\text{max}} - D_{\text{min}}) / \text{interval length}
\]
4) Where \( D_{\text{max}} \) is the biggest data and \( D_{\text{min}} \) is the minimum data. Then \( U \):

\[
U = \{U_1, U_2, U_3, \ldots, U_n\}
\]

(3)

5) Which will form as follows \( U_i = \{D_{\text{min}}, x_1\}, U_2 = \{x_i, x_2\}, \ldots, U_n = \{x_{n-1}, D_{\text{max}}\} \), where \( x_1 < x_2 < \ldots < x_{n-1} \)

6) To define each member of fuzzy set \( A \), which has been previously divided is described below:

\[
A_i = fA(U_1)/U_1 + fA(U_2)/U_2 + \ldots + fA(U_n)/U_n
\]

(4)

Where \( fA \) is the membership function of \( A_i \), \( fA_i : U \rightarrow [0,1] \). \( fA_i(U) \), then constructed where fuzzy set \( A_i \) a \( k \) where \( 1 \leq i \leq k \). For \( U_n \) equal to \( A_p \), where \( 1 \leq j \leq n \), in order to obtain the fuzzy sets as follows:

\[
\begin{align*}
A_1 &= a_{11}/U_1 + a_{12}/U_2 + \ldots + a_{1n}/U_n \\
A_2 &= a_{21}/U_1 + a_{22}/U_2 + \ldots + a_{2n}/U_n \\
A_k &= a_{ki}/U_1 + a_{ki}/U_2 + \ldots + a_{kn}/U_n
\end{align*}
\]

7) Define the fuzzy logical relationship, if \( Y(t) (t = 0, 1, 2, 3, \ldots) \) is the set of universe which is described as fuzzy set \( U(t) \). If \( F(t) \) consists of \( U(t) (i = 1, 2, 3, \ldots) \), \( F(t) \) is called fuzzy time series on \( Y(t) \).

\[
\text{If } F(t) \text{ then } Y(t)
\]

(5)

8) Let \( F(i) = A_i \) and \( F(i+1) = A_j \). The relationship between the two consecutive samples, \( F(i) \) and \( F(i+1) \), referred to as a fuzzy logical relationship, can be denoted by \( A_i \rightarrow A_j \), where \( A_i \) refers to as the left or current state and \( A_j \) is called the right side or the next state.

9) To combine fuzzy logical relationships into a Fuzzy Logical Relationship Group (FLRG), start from the left side which has the same group. E.g. \( (A) : A_i \rightarrow A_{j1}, A_{j1} \rightarrow A_{j2} \) and \( A_i \rightarrow A_{j2} \). All of the three relationships of fuzzy logic can be grouped, and assumed as the same group, so only one is taken.

10) The defuzzification process, e.g. from the FLRG result, generates: \( A_i \) (current state), \( A_{j1}, A_{j1}, A_{j2}, A_{j2}, \ldots, A_{jp} \) (next state), then the calculation process for the Fuzzy Logical Relationship Group produces \( A_{j1}, A_{j1}, A_{j1}, A_{j2}, A_{j2}, \ldots, A_{jp} \) where the maximum membership value of \( A_{j1}, A_{j1}, A_{j1}, A_{j2}, A_{j2}, \ldots, A_{jp} \) is the midpoint of \( U_1, U_2, U_3, \ldots, U_p \), thus the calculation is \( (m_1, m_2, \ldots, m_p)/p \).

Then, calculate the forecast error rate using MAPE and MSE.

\[
\text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} |A_i - F_i|A_i
\]

(6)

\[
\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (A_i - F_i)^2
\]

(7)
Where $A_i$ is the actual value of the data to $-i$, $F_i$ is the resulting value prediction of the data to $-i$, and $n$ is the number of data.

4. Results and Discussions

4.1 The Forecasting of NO$_2$

The time series data from the actual data of NO$_2$ were ranged and added up as the following Fuzzy time series method to determine the sum of absolute difference values and the interval length of the data. The researcher afterwards created fuzzification to show the Fuzzy Logical Relationship for the parameter NO$_2$. The next step was combining the data into a group relationship or defuzzification in order to determine the forecast data, MAPE, and MSE.

The actual data of NO$_2$ were plotted into a graph shown in Figure 1. It shows the data series of NO$_2$ ($\mu$g/m$^3$) from 8–21 July 2013. The resulting values of MAPE (23%) and MSE (13.8) were determined using Equations 6 and 7 from the Fuzzy Logical Group Relationship which is shown in Table 1. MSE was determined to examine the average false possibility of the forecasting. Due to the result, 13.8 point of MSE means that the forecasting of NO$_2$ is reliable to do because the error number is low. It is supported by the resulting MAPE which indicates that the error percentage of the comparison between the actual data and the forecasting data is only 23% (77% accuracy).

![Figure 1. The time series plot for the actual data of NO$_2$](image)

### Table 2. The absolute values of data differences for the pollutant parameter NO$_2$, MAPE, and MSE of NO$_2$ prediction

<table>
<thead>
<tr>
<th></th>
<th>MAPE</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prediction</td>
<td>15847</td>
<td>9292</td>
</tr>
<tr>
<td></td>
<td>23%</td>
<td>13.8</td>
</tr>
</tbody>
</table>

Meanwhile, the researcher predicted NO$_2$ from the actual data and compared it to the air quality index (standard) which has been used in Surabaya. The comparison graph is...
shown in Figure 2. Based on the air pollutant index, the forecasting data as the result of the fuzzy time series method show that it could not be defined and used as the data of NO$_2$ monitoring prediction (the data could not show any information) because the index parameter of the actual data is below the limit that can be recorded by the monitoring device.

![Figure 2](image2.png)

**Figure 2.** The comparison graph for the pollutant NO$_2$ (8–21 July 2013)

### 4.2 The Forecasting of CO

Figure 3 shows the time series plot of the actual data of CO ($\mu$g/m$^3$) from 8–21 July 2013. Meanwhile, the resulting values of MAPE (19.5 %) and MSE (0.11) indicate that the forecasting of CO is valid or more reliable to do because the error number is very small (after comparing the actual data and the forecasting data using defuzzification).

![Figure 3](image3.png)

**Figure 3.** The time series plot for the data record of the pollutant parameter CO
Table 3. The absolute values of the data differences for the pollutant parameter CO, MAPE, and MSE from CO prediction

<table>
<thead>
<tr>
<th></th>
<th>MAPE</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>CO</td>
<td>13082.3</td>
<td>74.9</td>
</tr>
<tr>
<td></td>
<td>19.5%</td>
<td>0.11</td>
</tr>
</tbody>
</table>

The prediction of CO from the actual data and its comparison with the air quality index (standard) which has been used in Surabaya are shown in the graph as illustrated in Figure 4. Based on the comparison with the air pollutant index, the parameter CO from the forecasting data series can be lower than the actual data and it indicates the possibility of a good air quality condition.

4.3 The Forecasting of O₃

The actual data of O₃ were plotted into a graph as shown in Figure 5. It shows the data series of O₃ (µg/m³) from 8–21 2013. The resulting values of MAPE (22.75 %) and MSE (53.95) are shown in Table 4. MSE was determined to examine the average false possibility from the forecasting. Due to the result, 53.95 point of MSE means that the forecasting of O₃ is not reliable to do because the error number is high.

The prediction of O₃ from the actual data and its comparison with the air quality index (standard) which has been used in Surabaya are shown in the graph as illustrated in Figure 6. Based on the comparison with the air pollutant index, the parameter O₃ from the forecasting data series can be lower than the actual data and it indicates the possibility of a good O₃ condition. But, this result cannot be used due to a high MSE and the unreliable resulting value.
Figure 5. The time series plot for the data record of the pollutant parameter $O_3$

Table 4. The absolute values of data differences for the pollutant parameter $O_3$, MAPE, and MSE from $O_3$ prediction

<table>
<thead>
<tr>
<th></th>
<th>MAPE</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>15262.6</td>
<td>36202.6</td>
</tr>
<tr>
<td></td>
<td>22.75%</td>
<td>53.95</td>
</tr>
</tbody>
</table>

Figure 6. The comparison graph for the pollutant parameter $O_3$ (8–21 July 2013)

4.4 The Forecasting of PM$_{10}$

The actual data of PM$_{10}$ were plotted into a graph as shown in Figure 7. It shows the data series of PM$_{10}$ ($\mu$g/m$^3$) from 8–21 July 2013. The resulting values of MAPE (9.96 %) and MSE (5.43) are shown in Table 5, and they indicate that the forecasting of PM$_{10}$ is reliable to do because the error number is low. It is supported by the resulting MAPE which suggests that the error percentage of the comparison between the actual data and the forecasting data is only 9.96 %.
Figure 7. The time series plot for the data record of the pollutant parameter PM$_{10}$

Table 5. The absolute values of data differences for the pollutant parameter PM$_{10}$, MAPE, and MSE from PM$_{10}$ prediction

<table>
<thead>
<tr>
<th>MAPE</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>6347.77</td>
<td>3038.1</td>
</tr>
<tr>
<td>9.96%</td>
<td>5.43%</td>
</tr>
</tbody>
</table>

The prediction of PM$_{10}$ from the actual data and its comparison with the air quality index (standard) which has been used in Surabaya are shown in the graph illustrated in Figure 8. Based on the comparison with the air pollutant index, the parameter PM$_{10}$ from the forecasting data series can be lower than the actual data and it indicates the possibility of a good PM$_{10}$ condition.

Figure 8. The comparison graph for the pollutant parameter PM$_{10}$ (8–21 July 2013)
4.5 The Forecasting of SO$_2$

The actual data of SO$_2$ were plotted into a graph shown in Figure 7. It shows the data series of PM$_{10}$ (µg/m$^3$) from 20 July–2 August 2013. The resulting values of MAPE (3.6%) and MSE (875.524) are shown in Table 6. Due to the result, 875.524 point of MSE indicates that the forecasting of SO$_2$ is not reliable to do because the error number is high.

![Figure 7](image_url)

**Table 6.** The absolute values of data differences for the pollutant parameter SO$_2$, MAPE, and MSE from SO$_2$ prediction

<table>
<thead>
<tr>
<th></th>
<th>MAPE</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted</td>
<td>2430.7</td>
<td>587488.7</td>
</tr>
<tr>
<td></td>
<td>3.6%</td>
<td>875.524</td>
</tr>
</tbody>
</table>

![Figure 10](image_url)
The prediction of SO$_2$ from the actual data and its comparison with the air quality index (standard) which has been used in Surabaya are shown in the graph illustrated in Figure 10. Based on the comparison with the air pollutant index, the parameter SO$_2$ from the forecasting data series could not be defined and used as the data of SO$_2$ monitoring prediction due to a high MSE and the unreliable resulting value.

5. Conclusion

Fuzzy time series is a method that potentially becomes an effective tool for early warning system algorithms used to forecast air pollution parameters at least for the next three hours since the last record. The unreliable forecast data which are shown for the parameters O$_3$ and SO$_2$ occurred because of several factors neglected in the present study, such as wind direction, topology, and temperature.

Besides controlling the fuel factor, vehicle types, and vehicle volume, the government can establish environmental policies and implement some environmental control programs as follows: strengthening and tightening vehicle emission standard requirements; making investments in better mass transit to enable people to use public transport such as Bus, Railway, and Municipal Transport; and carrying out traffic efficiency management. Meanwhile, for short-term programs as a recommendation in anticipation of the increasing number of pollutants, the government should make any plans of alternative traffic routes or simply distribute masks.

References


