PREDICTION OF OZONE (O3) VALUES USING SUPPORT VECTOR REGRESSION METHOD

Chasandra Puspitasari[1], Nur Rokhman[2], Wahyono[3]

[1] Computer Science, School of Computer Science, Bina Nusantara University
[2,3] Master Program in Computer Science and Electronics, Fakultas MIPA, Universitas Gadjah Mada
[1] chasandra.puspitasari@binus.ac.id, [2] nurrokhman@ugm.ac.id, [3] wahyo@ugm.ac.id

Abstract

A large number of motor vehicles that cause congestion is a major factor in the poor air quality in big cities. Ozone (O3) is one of the main indicators in measuring the level of air pollution in the city of Surabaya to find out how air quality. Prediction of Ozone (O3) value is important as a support for the community and government in efforts to improve the air quality. This study aims to predict the value of Ozone (O3) in the form of time series data using the Support Vector Regression (SVR) method with the Linear, Polynomial, RBF, and ANOVA kernels. The data used in this study are 549 primary data from the daily average of ozone (O3) value of Surabaya in the period 1 July 2017 - 31 December 2018. The data will be used in the training and testing process until prediction results are obtained. The results obtained from this study are the Linear kernel produces the best prediction model with a MAPE value of 21.78% with a parameter value $\lambda = 0.3$; $\varepsilon = 0.00001$; cLR = 0.005; and C = 0.5. The results of the Polynomial kernel are not much different from the Linear kernel which has a MAPE value of 21.83%. While the RBF and ANOVA kernels each produce a model with MAPE value of 24.49% and 22.0%. These results indicate that the SVR method with the kernels used can predict Ozone values quite well.

Kata kunci : Support Vector Regression, prediction, Ozon.

1. Introduction

Air quality is becoming increasingly feared by today's society. Low air quality can cause a complex effect for the community, shown by an increasing number of people breathing disease among the community. In the world, the ozone concentration on the surface is quite high especially in dense areas with varying levels of increase (Tim Penyusun, 2014). This is especially happening in industrial areas and in areas with high traffic density. Also in Indonesia, at major cities such as Surabaya City, the ozone the troposphere (O3) is one of the most dominant air pollutant indicators with maximum concentration compared to other indicators (Arifien et al., 2012).

In the past five years, researchers have been conducting in-depth studies on the approximate methods of air quality value and the prediction of air pollution in various regions. Some of the research was conducted Some of the research was conducted to predict ozone (O3) pollutants in Surabaya City using the artificial neural network method, comparing JST modelling multivariate with 7 predictor variables and time series with 2 predictor variables (Arifien et al., 2012). In the study, modelling JST multivariate had better results than univariate due to information from other parameter values on more multivariate modelling and change patterns can be easily mapped by Network.

However, often the availability and provision of variable data related to predicting ozone pollutant values is still an obstacle, so advanced research must take advantage of the univariate approach by detecting certain patterns of data in the past.

Predicted air pollution levels have indeed been popular and became an important topic in recent years. It is important to build an early warning system so that it can provide estimates and also warnings for local health factors both by medical practitioners and local governments (Gupita, Aisjah and Arifin, 2017). Therefore, it takes a method that can be used to make predictions or forecasting and result in good accuracy and low model error rates. Time series predictions are possible because the future depends on the past or analog because there is a relationship between the future and the past (Caraka, Yasin and Diponegoro, 2017). Another popular learning method that has been implemented in the predictive time series is the Support Vector Machine (SVM) method. SVM is a method that has advantages in optimizing pattern recognition system with the ability to classify the pattern or data is good and can also be used to predict a value (Bonita and Muflikhah, 2018). In the case of time series predictions are used in the development of the SVM method namely Support Vector Regression (SVR). Although some research suggests that SVR performs well in many time series prediction issues, it is difficult to select the most

suitable amount of training data to build the optimal SVR model for predictive data (Suzuki et al., 2014).

In this study was supported by Support Vector Regression (SVR) method by applying a sequential learning algorithm to predict the data time series ozone (O3). The sequential learning algorithm was first introduced by reference (Vijayakumar and Wu, 1999) to handle the search for the Lagrange multiplier values in the training process, providing an optimal solution and compute time faster. In the research (Bonita and Muflikhah, 2018), the sequential learning algorithm of SVR can predict the price of coal well as indicated by MAPE by 8.38% using ANOVA kernel. In the SVR algorithm, there are multiple kernels for non-linear regression cases and each kernel will provide different accuracy, along with the performance of different algorithms in each case. The research was conducted by implementing four SVR kernels namely Linear, Polynomial, RBF, and ANOVA separately. The results of the SVR performance are obtained based on the models produced both in determining the parameter values of each kernel used, as well as the window size in the learning process.

2. Methods

In this section, the proposed method is explained in detail. This Includes data descriptions, preprocessing of data, and methods in predicting ozone values with Sequential Learning SVR.

2.1 Ozon (O3) Pollutant

Ozone (O3) is a compound in the air other than oxygen that has properties as a very strong oxidizing after Fluor, oxygen, and oxygen fluoride (OF2). The chemical elements contained in the ozone particle is three oxygen (O3). O3 is known as the ambient air O3 that is located in the troposphere layer, which is an atmospheric layer that is 10 km — 16 km from the Earth's surface. O3 Ambient air is a harmful air pollutant for health. At a high concentration O3 results in respiratory tract disorders, eye irritation, and impaired nerve function.

In addition to causing adverse impacts on human health, ozone polluters can cause economic harm due to material or material (textile, rubber, wood, metal, paint, etc., agricultural decline and ecosystem damage such as diminished biodiversity. Oxidants in the air include ozone (more than 90%), nitrogen dioxide, and peroxyacetylnitrate (PAN). Since most of the oxidants are ozone, the ambient air monitoring is expressed as ozone levels. In a large city environment such as in the city of Surabaya, pollutant O3 has become a problem of environment generated by various conditions and activities of its inhabitants. Increasing population increase has also impacted the increasing human activity in fulfilling the necessities of life such as industrial activity and

the mobilization of motor vehicles. Both activities are inevitable because the vehicle has become the main tool for the movement of Community activities today. These activities are the main suber producing harmful substances causing pollution one of which is ozone pollutant (O3).

$$R(x,y) = \begin{cases} 1 & jika \ f(x,y) >= T \\ 0 & sebaliknya \end{cases} \quad (1)$$

2.2 Data Description

In this research used daily data ozone value from SUF 7 located in Kebonsari area is in Kebonsari Village, Surabaya City. The area is close to the Surabaya-Porong toll road, so SUF 7 has the highest concentration of pollutants because it has the highest volume of vehicles. The ozone data used in this study in the period of 1 July 2017 – 31 December 2018 was 549 data. Figure 1 shows the fluctuations in the recorded ozone value every day. At the beginning of July 2017 recorded high ozone value, this is due to the return of the vehicle after the holiday with Eid al-Fitr day on 27 – 30 June 2017. So that vehicle traffic passing in Kebonsari area especially near the Surabaya-Porong toll is quite high.
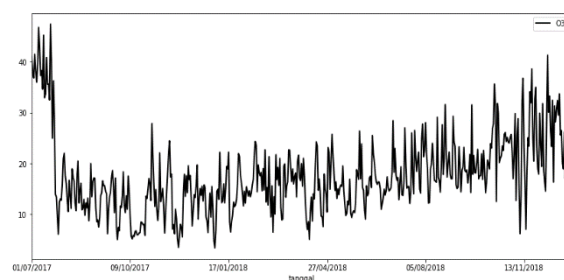

Figure 1. Ozone data plot period July 1, 2017 – Dec 31, 2018

As can be seen in Figure 1, it can be observed that the value of ozone is increasing by the end of 2018. The increasing value of ozone at that time was influenced by the increase in vehicle traffic on the Surabaya-Gempol section which has been seen since the end of 2018, especially during Christmas and New Year holidays. Another thing that also led to the increase in vehicle traffic is to start operating toll relocation Porong-Gempol (Porong-Kejapanan) since the end of 2018, so that vehicle traffic increased up to 36 percent.

The air quality value data in the Kebonsari region is not fully recorded by the tools at the ambient air quality monitoring station (SUF 7). This is so there is a missing value on the ozone data that the author received from the City Environment Office of Surabaya. Missing value is information that is unavailable for an object or case. Missing value on the recording of pollutant values in Surabaya occurs when the recording device has

calibration. Maintenance and calibration are done so that the air quality monitoring equipment is functioning normally. In the ozone data period used in this study, three missing values were on 14 to 16 November 2017.

2.3 Preprocessing Data

At this stage, the process is done to prepares the data to be processed by the system. Preprocessing data aims to form training data in order to simplify the training process and can help to produce a good model (Jacobus and Winarko, 2014). Data that have been obtained from the source is then processed as needed from system input.

2.3.1 Linear Interpolation

The process that was done first after the data entered the system was to interpolate the missing value of ozone data using linear interpolation techniques. Interpolation is a method that is used to determine the value that is between the knowns data points (Wahab, 2017). In the data used in this research, there are three missing value that is from 14 to 16 November 2017. Here is the equation formula to look for the linear interpolation result value.

$$f(x) = y = \frac{y_1 - y_0}{x_1 - x_0}(x - x_0) + y_0$$

2.3.2 Data Normalization

Normalization is a transformation done on data to evenly distribute data, and scale data into specific ranges. Data normalization can also accelerate the training process time on the same scale. There are several kinds of data normalization methods, one of which is the normalization of Min-Max. Min-Max normalization can be done by transforming the data into a certain scale, such as [0, 1] or [-1, 1]. Min-Max Normalization formula is:

$$x_i' = \frac{x_i - x_{min}}{x_{max} - x_{min}}(new\_x_{max} - new\_x_{min}) + new\_x_{min}$$

2.3.3 Sliding Window

In this study, the attribute used is considerable daily data ozone value (O3). The predicted ozone value (O3) can be seen as a time series problem. Suppose T symbolizes the current time, then the equation of time series:

$$y(t) = (x_1, x_2, x_3, x_4, ..., x_t)$$

is a series of O3 values measured at the same time interval up to t that is in the hourly time interval.

Time series predictions in research can be defined as an issue to predict every 1 hour of the next value.
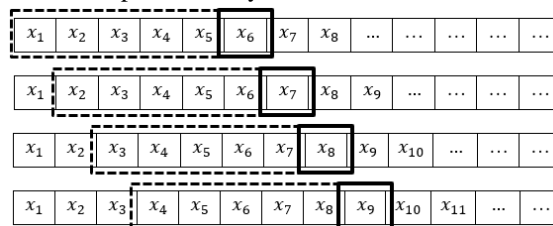

Figure 2. Illustration of dataset formation with sliding window

The sliding window process shown in Figure 2 is a dataset with a window size of 5. The sliding window process is done by selecting the first segment, then the next segment is selected from the end of the first segment. This process is repeated until all-time series data is segmented. In this process, each segment on the sliding window accumulates time series data historically to predict O3 value in the next 1 hour.

2.3.4 Support Vector Regression

Support Vector Regression (SVR) is the expansion of Support Vector Machine (SVM) that applied to regression problems and delivers output in the form of continuous data. The optimal hyperplane in the SVM can be found by laying it into Quadratic Programming (QP) complex problems (Sanusi et al., 2014). Form Primal form that has been very difficult to solve, will be transformed into a dual form that will only contain the value of alpha. Various algorithms have been developed to look for these alpha values. However, these algorithms take a long time, especially if used for large data because the algorithm uses numerical quadratic programming as an inner loop (Platt, 1998).

To address these issues, Vijayakumar and Wu provide a simple solution that is the sequential learning algorithm to handle the search for the Lagrange multiplier (α and α*) values in the training process, so that it can Provide optimal solution and faster computing time compared to conventional SVR. Here are the steps of the sequential learning algorithm for SVR (Vijayakumar and Wu, 1999):

1. Initialize the SVR and kernel parameters, maximum iteration, and set the value of α_i and α_i^* to 0.
2. Generate the hessian matrix (R_{ij}).
   $R_{ij} = K(\mathbf{x}_i, \mathbf{x}_j) + \lambda^2$ for $i, j = 1,...,n$
   where:
   $K(\mathbf{x}_i, \mathbf{x}_j)$ = kernel function
   λ = scalar variable of SVR
   $\mathbf{x}$ = the data
3. For each data training process, calculate the error (E_i) value and update the value of lagrange multiplier (α_i and α_i^*).
   a) $E_i = y_i - \sum_{j=1}^{n}(\alpha_i^* - \alpha_i) R_{ij}$

where:

E_i = Error value to i

y_i = Actual value to i

α_i, α_i^* = *Lagrange multiplier* value

b) $\delta\alpha_i^* = \min\{\max[\gamma(E_i - \varepsilon), -\alpha_i^*], C - \alpha_i^*\}$

$\delta\alpha_i = \min\{\max[\gamma(-E_i - \varepsilon), -\alpha_i], C - \alpha_i\}$

where:

$\delta\alpha_i^*, \delta\alpha_i =$ The changes in α_i dan α_i^* value

γ = Learning rate

ε = Epsilon parameter (insensitivity bound)

C = The complexity parameter

c) $\alpha_i^* = \alpha_i^* + \delta\alpha_i^*$

$\alpha_i = \alpha_i + \delta\alpha_i$

4. Repeat the sequence process in step 3 to:
   - Reach the maximum number of iterations, or
   - Meet the stop condition of $max\ (|\delta\alpha_i^*|) < \varepsilon$ and $max\ (|\delta\alpha_i|) < \varepsilon$.

5. Calculate the regression function $f(\mathrm{x})$ as the result of the prediction can be formulated as follows:

$$f(\mathbf{x}) = \sum_{i=1}^{n} (\alpha_i^* - \alpha_i) K(\mathbf{x}_i, \mathbf{x}_j)$$

2.3.5 Kernel Function

Kernel is a function to map data into a higher-dimensional vector space. The use of kernel can help to calculate predictions by only analyzing existing problems. The basic idea of the kernel method is to map the x data to a higher-dimensional feature space with the $\Phi(\mathbf{x})$ function, so that in the new vector space hyperplane can be constructed. Furthermore, the calculations to find their vector support points depend on the dot product of the data that has been transformed into the new dimension space, which is $\Phi(\mathbf{x}_i).\Phi(\mathbf{x}_j)$. Because of the difficulty of finding the transformation function of Φ, then proposed the dot product calculation replaced with kernel function $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i).\Phi(\mathbf{x}_j)$, in which the function defines Φ transformation implicitly. This is what is then called the "kernel trick".

These kernels including Linear, Polynomial, RBF, and ANOVA were used in this study to build SVR models comparison.

1. *Linear*

$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \cdot \mathbf{x}_j$

where:

i, j = 1, 2, 3, . . ., n

$\mathbf{x}_i^T$ = Transpose matrix of the input value to-i

$\mathbf{x}_j$ = Matrix of the input value to-j

2. *Kernel Polynomial*

$$K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T \cdot \mathbf{x}_j + 1)^d \ ; \ i, j = 1,2,3,\ldots, n$$

where:

d = *degree* parameter of the kernel

3. *Kernel RBF (Radial Basis Function)*

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right)$$

where:

σ = *sigma* parameter of the kernel

4. *Kernel ANOVA (Analysis of Variance)*

$$K(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{i,j=1}^{n} \exp\left(-\gamma(\mathbf{x}_i - \mathbf{x}_j)^2\right)\right)^d$$

where:

d = *degree* parameter of kernel

γ = gamma parameter of the kernel

2.3.6 Prediction Process

In this study will be constructed model of ozone prediction as the main indicator of air pollution in the city of Surabaya. The method used to build predictive models uses the Support Vector Regression (SVR) method with sequential learning by obtaining a Lagrange multiplier value (αi and αi values). Prediction of air pollution levels is done using daily data of ozone value. This prediction is used to predict the air quality value in subsequent days according to the ratio of the training data and test data used. The expected output is to get the best-predicted result value with the smallest MAPE value. The best-predicted value results can be obtained by testing the number of windows, amount of data, ratio of training data and test data used, as well as testing on the SVR parameters and kernel parameters.

(10)
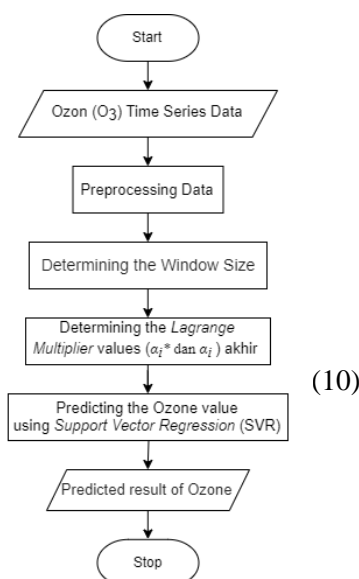
Figure 3. Ozone Prediction Model

In Figure 3 the overall stages of the design of the ozone Prediction model using the Support Vector Regression (SVR) method are divided into 6 stages. (1) Daily ozone data Input time series, (2) Data preprocessing to remove missing value, (3) Specify the window size by performing the data mapping using the sliding window technique, (4) Specifying the value of Lagrange multiplier (αi^* and αi) to form a regression function, (5) Predict the value of ozone using the generated SVR model, and (6) the output is ozone prediction data.

Figure 4 is the steps of learning process which used to obtain predictive results using the SVR sequential learning algorithm are based on the process that has been made. In this study, the first step to predict is to preprocess the data, which means the data is interpolated. It is a process of replacing or giving value to the missing data using a linear interpolation method. The next step is to normalize the data using the min-max normalization (0-1) method. The train data and test data then normalized to speed up the calculation. After the preprocessing data is completed, the next step is to determine the size of the window that will be used in the learning process. This stage is important because the size of the window can be used as an algorithm reference in recognizing patterns of the data in the past.

The next step is the process of dividing data (split data) into 2 parts, namely train data and test data. The ratio of data sharing in this study was 2:1 or 67% for train data while the rest was for test data. Next, in the Hessian matrix mapping phase, the Linear, Polynomial, RBF, and ANOVA kernels are implemented separately by initializing the SVR parameters including cLR, C, ε, λ, on each kernel parameter. Furthermore, the learning process uses sequential learning towards the train data. Sequential learning is done until the maximum iteration is met. After sequential learning is completed and obtained αi^* and αi (Lagrange multiplier), the regression function is calculated against the test data to obtain the predicted ozone value. The predicted ozone value of the test data is obtained after denormalization to return the value and compare the result with the actual value. Then, the predicted result of each kernel usage is evaluated and compared with error calculations using the Mean Absolute Percentage Error (MAPE). The best final SVR Model obtained by evaluating the MAPE value is with the lowest MAPE value through the grid search.
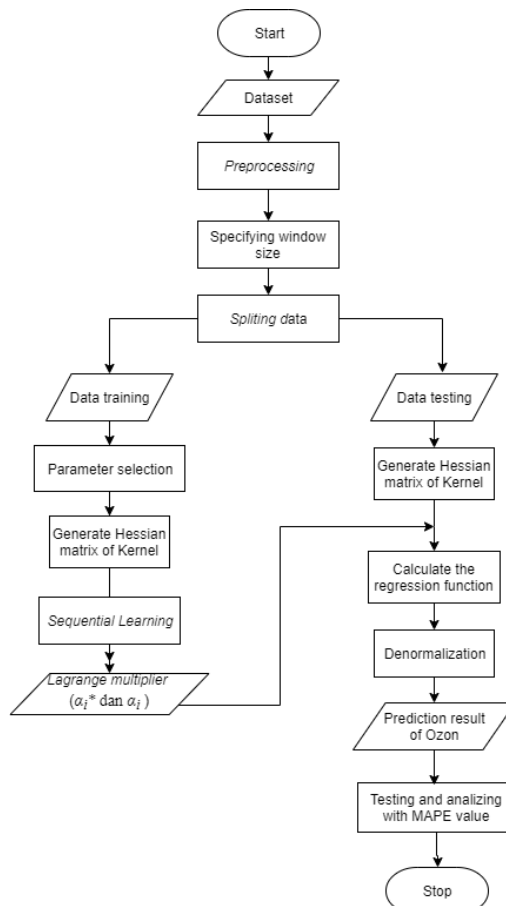

Figure 4. Flowchart diagram for Ozon Prediction with Support Vector Regression

3. Results and Discussion

The results are derived from experiments that have been performed on each kernel. The test parameter is attempted based on the constraints of the parameter value that has been defined. Based on this test, there will be a pair of the best parameter values of each kernel based on the smallest MAPE value. This principle of testing is similar to the concept of grid search which is the method of finding the best parameter of pairs by calculating the MAPE value iteratively. This test will by default use a window size of five with the training data ratio composition test data of 2:1. The best result of this test will then be used on the next test comparing to the other window sizes.
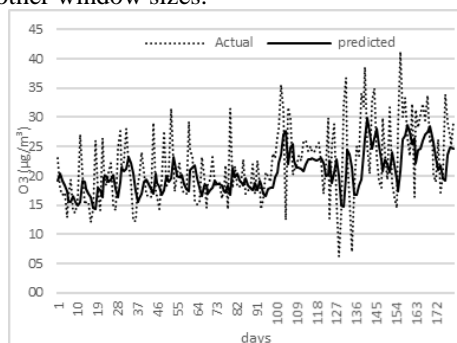

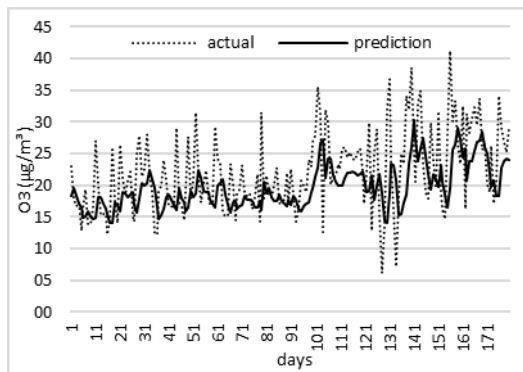Figure 5a. Prediction of Ozone with kernel Linear

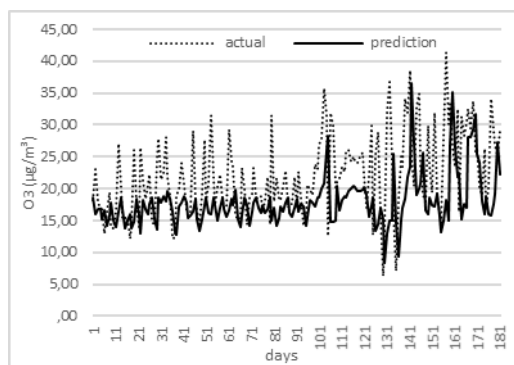Figure 5b. Prediction of Ozone with kernel Polynomial
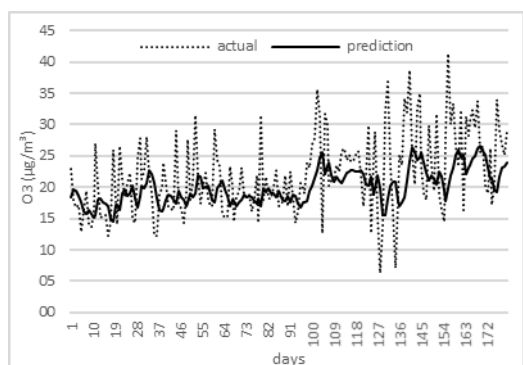

Figure 5c. Prediction of Ozone with kernel RBF

Figure 5d. Prediction of Ozone with kernel ANOVA

Figure 5a is a plot test result of a linear kernel with a window size of 5. In this test, the best value of parameter pairs with MAPE value is 22.4%, with the parameter value $\lambda = 0.3$; $\varepsilon = 0.00001$; cLR = 0.005; and C = 0.5. From these results, it can be seen that the Linear kernel can be quite well in predicting the ozone value by the SVR method. While the Polynomial kernel testing is obtained the optimal value of the parameter pairs with the MAPE value of 22.25%, i.e. with a parameter value d = 2; $\lambda = 0.01$; $\varepsilon = 0.00001$; cLR = 0.01; and C = 0.5. Figure 5b is a plot graph of the prediction value of the actual value. As seen in the test results shows that the magnitude of the d (degree) value greatly affects the performance of the SVR in predicting the ozone value, while the C value change does not affect the test results.

Subsequent experiments were conducted testing on the predicted ozone value by comparing the window size of the optimal parameter value generated on previous parameter testing. The experiments are performed on each kernel by comparing the results between the window size of 3, 5, 7 and 14. This test aims to determine the performance results of each kernel in predicting the ozone value using the Support Vector Regression method by calculating the MAPE value.

Table 1 Comparison of MAPE values based on window size of each kernel

Kernel	MAPE values (%) by window size			
	3	5	7	14
Linear	22,90	22,38	21,97	21,78
Polynomial	22,56	22,25	21,83	21,96
RBF	24,49	27,49	31,33	34,25
ANOVA	22,97	22,25	22,01	65,31

In the comparison table, it can be noted that Linear and Polynomial kernels have almost the same results because the difference in MAPE values in each test is very small. However, compared to predicted results with RBF and ANOVA Kernels, Linear and Polynomial kernels show better results. The best results are obtained from the SVR method with a Linear kernel with a MAPE value of 21.78% with a window size of 14 and a parameter value of $\lambda = 0.3$; $\varepsilon = 0.00001$; cLR = 0.005; and C = 0.5.

From the test results that have been done on each kernel indicates that the window size has a significant role in the RBF and ANOVA kernels. This can be seen from the changes in the MAPE value in table 1. In the RBF kernel shows that the bigger the window size, the larger the MAPE value earned. So in this case, predictions with SVR methods with RBF kernels are better suited using small window sizes. Meanwhile, high enough MAPE spikes occur on ANOVA kernel testing at a window size of 14. But from the fourth testing window size of the kernel ANOVA the best results are obtained by a window size of 7 with a very small completion MAPE compared to the size of window 3 and 5.

In this research, it has been obtained that the four kernels used in the training and testing process of SVR methods can be used to predict the data time series ozone value using the sliding window technique. Although it can be categorized quite well, it still needs to be left with minimizing the MAPE error value in each kernel.

4. **Conclusion**

Based on the research that has been done, it can be concluded that of the four kernels implemented, namely Linear, Polynomial, RBF, and ANOVA Kernels, the best ozone value (O3) prediction model is produced by Linear kernel with

MAPE value of 21.79% at a window size of 14. The results showed that in this study the Support method of Vector Regression with Linear kernel could result in a predictive model of ozone value quite good with parameter value of $\lambda = 0.3$; $\varepsilon = 0.00001$; cLR = 0.005; and C = 0.5. Meanwhile, the results of the Polynomial kernel do not differ considerably from the Linear kernel, namely the MAPE of 21.83%. Meanwhile, RBF and ANOVA kernels produce the best MAPE values of 24.49% and 22% respectively.

However, the value of the parameters used in this study was limited. It needs to be added as a method to optimize parameter values to determine the most optimal parameter to improve SVR performance. And also, the training and testing data in this research still uses univariate time series data. In future work can be tested with multivariate time series data as an additional feature, especially concerning the indicators of other air pollution causes.

Daftar Pustaka:

Arifien, N. F. et al. (2012) 'Prediksi Kadar Polutan Dengan Menggunakan Jaringan Syaraf Tiruan (JST) Untuk Pemantauan Kualitas Udara di Kota Surabaya', Seminar Nasional Teknik Kimia Soebardjo Brotohardjono Ix, pp. 1–11.

Bonita, O. and Muflikhah, L. (2018) 'Comparison of Gaussian and ANOVA Kernel in Support Vector Regression for Predicting Coal Price', 2018 International Conference on Sustainable Information Engineering and Technology (SIET). IEEE, pp. 147–150.

Caraka, R. E., Yasin, H. and Diponegoro, A. W. (2017) 'Peramalan Crude Palm Oil (CPO) Menggunakan Support Vector Regression Kernel Radial Basis', Jurnal Matematika, 7(ISSN: 1693-1394). doi: 10.24843/JMAT.2017.v07.i01.p81.

Gupita, S. A. N., Aisjah, A. S. and Arifin, S. (2017) Prediksi Kadar Polutan Menggunakan Adaptive Neouro-Fuzzy Inference System (ANFIS) untuk Pemantauan Kualitas Udara di Kota Surabaya. Surabaya: Departemen Teknik Fisika, Fakultas Teknologi Industri, Institut Teknologi Sepuluh November.

Jacobus, A. and Winarko, E. (2014) 'Penerapan Metode Support Vector Machine pada Sistem Deteksi Intrusi secara Real-time', IJCCS (Indonesian Journal of Computing and Cybernetics Systems), 8(1), p. 13. doi: 10.22146/ijccs.3491.

Platt, J. (1998) 'Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines', (MSR-TR-98-14).

Sanusi, S. et al. (2014) 'Downscaling Modeling Using Support Vector Regression for Rainfall Prediction', TELKOMNIKA Indonesian Journal of Electrical Engineering, 12(8). doi: 10.11591/telkomnika.v12i8.6195.

Suzuki, Y. et al. (2014) 'Proposal to Sliding Sindow-based Support Sector Regression', Procedia Computer Science - Science Direct. Elsevier Masson SAS, 35, pp. 1615–1624. doi: 10.1016/j.procs.2014.08.245.

Tim Penyusun (2014) Laporan Status Lingkungan Hidup Daerah (SLHD) Kota Surabaya. Surabaya: Badan Lingkungan Hidup.

Vijayakumar, S. and Wu, S. (1999) 'Sequential Support Vector Classifiers and Regression', International Conference on Soft Computing, 86(February), pp. 610–619. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.60.8325&rep=rep1&type=pdf.

Wahab, M. A. (2017) 'Interpolation and Extrapolation', Topics in System Engineering. doi: 10.1201/9781315120195-4.