Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 179 (2021) 747–753

Procedia Computer Science

www.elsevier.com/locate/procedia

5th International Conference on Computer Science and Computational Intelligence 2020

Support Vector Regression for Predicting the Number of Dengue Incidents in DKI Jakarta

Ivan Noverlianto Tanawi[a], Valentino Vito[a], Devvi Sarwinda[a], Hengki Tasman[a], Gatot Fatwanto Hertono[a,] *

[a]*Department of Mathematics, Faculty of Mathematics and Natural Sciences (FMIPA), Universitas Indonesia, Depok 16424, Indonesia*

Abstract

Dengue fever is a disease caused by the dengue virus, which is spread by *Aedes aegypti* and *Aedes albopictus* mosquitoes. According to the WHO, as a tropical country, Indonesia is a country at high risk for dengue. Dengue can spread to other people through mosquito bites. Weather factors, such as temperature, humidity, and rainfall, affect the number of dengue incidents. It is important to predict the number of dengue incidents so that the government and the people will be ready to prevent a dengue outbreak when the number of dengue incidents is predicted to be high. In this paper, we predict the number of dengue incidents in DKI Jakarta using support vector regression, with weather and the previous number of incidents as predictor variables. These predictor variables are determined by analyzing the time lag between each predictor variable and the number of incidents by using cross-correlation. Models for prediction are compared by Root Mean Squared Error and Mean Absolute Error. The result shows that support vector regression with linear kernel is quite good, and is in fact better than the radial kernel, for predicting the number of dengue incidents.

Keywords: Regression; support vector regression; machine learning; dengue; prediction

* Corresponding author. Tel.: +62-815-802-3055
E-mail address: gatot-f1@ui.ac.id

10.1016/j.procs.2021.01.063

1. Introduction

Dengue fever is an infectious disease spread by the dengue virus. This disease is commonly found in tropical and sub-tropical urban regions. Dengue is first discovered in Indonesia in the year 1968. In the years that followed, the number of dengue incidents increased from 58 cases in 1968 to 158,912 cases in 2009[7]. Hence, predicting the number of incidents concerning dengue is vital for reducing its impact.

In a previous research, Hasanah and Susanna[5] showed that rainfall, humidity, and temperature affect the number of dengue incidents in DKI Jakarta. This motivates us to use weather variables as predictor variables for predicting the number of dengue incidents. In another study, a linear regression model was provided in [3] for the same problem. On the other hand, Guo et al.[4] found that in analyzing incidence data of dengue fever in China, a prediction model constructed using support vector regression with linear kernel is more accurate compared to other models. Their work motivates us to use support vector regression for modeling the problem in DKI Jakarta. We also utilize cross-correlation on the weather data in DKI Jakarta similar to the analysis previously done by Withanage et al. [8]. In addition, we try to use not just the linear kernel, but also the radial kernel which was not studied in [4]. This is so that we can compare the performance between the two kernels using Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE).

In the present paper, we apply support vector regression with a linear kernel to weather and incidence data to create a model for predicting the number of dengue incidents in DKI Jakarta. We divide the province to five regions, namely North, South, East, West, and Central Jakarta, and then we analyze each region separately. Firstly, we try to find an acceptable time lag between weather variables and the dependent variable (that is, the number of dengue incidents) using cross-correlation. After dividing the data into training and testing data, we use support vector regression to predict future incidents based on the training data and analyze its accuracy using the testing data. In addition, we compare the performance between the linear kernel and radial kernel on their prediction accuracy.

2. Research approach

In this section, we discuss various methods of analyzing the available data on the five regions in DKI Jakarta. The incidence data on dengue fever are provided by the Jakarta Health Department. The weather data, including data on rainfall, humidity, and temperature, are provided by the Meteorology, Climatology, and Geophysical Agency (BMKG). The missing values in the data are imputed by the overall mean values, and the data are then converted into weekly data with 455 weeks from January 6, 2009 to September 25, 2017. The methodology flow chart of our research is shown in Fig. 1.

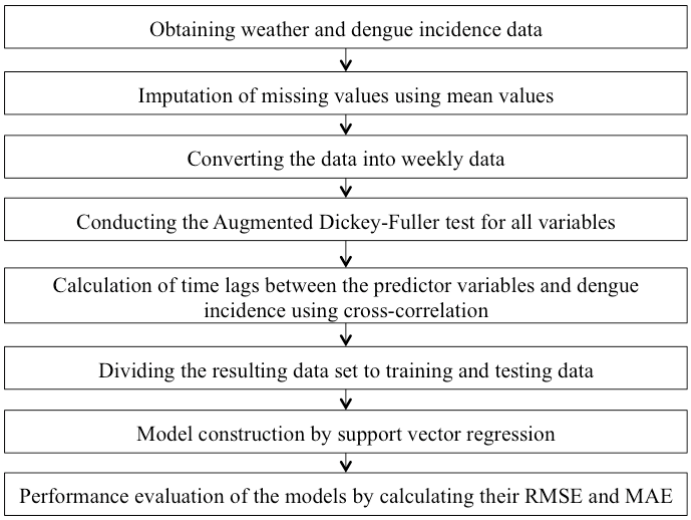

Fig. 1. The methodology flow chart of this research

All the variables used in this research, including weather data and incidence data, are in the form of a weekly time-series. We shall need to calculate the appropriate time lag between each predictor variable Y_t with the dependent variable X_t of dengue incidence. The predictor variables used are previous incidence, average temperature, cumulative rainfall, and average relative humidity. Afterwards, we employ support vector regression to construct the prediction model.

2.1. Cross-correlation and the Augmented Dickey-Fuller test

To calculate the time lag between two stationary time series $X = \{X_t\}$ and $Y = \{Y_t\}$ of length N, we define the sample cross-correlation at lag k as

$$r_k(X,Y) = \mathrm{Corr}(X_t, Y_{t-k}) = \frac{\sum(X_t - \bar{X})(Y_{t-k} - \bar{Y})}{\sqrt{\sum(X_t - \bar{X})^2}\sqrt{\sum(Y_t - \bar{Y})^2}}.$$

The appropriate predictor variable for X_t would then be Y_{t-k} such that the absolute value $|r_k(X,Y)|$ is maximized. If $r_k(X,Y)$ is positive (resp. negative), then it would suggest that there is a positive (resp. negative) correlation between the variables X_t and Y_{t-k}. We restrict the value of k to $1 \le k \le 8$, which means that the time lag must be between one to eight weeks.

We use the Augmented Dickey-Fuller test to check whether a time series is stationary. In effect, we calculate the Dickey-Fuller test statistic DF from the time series data and consult the table of critical values for the Dickey-Fuller t-distribution. If DF is less than the critical value, then we can assume that the time series is stationary. For more detail on cross-correlation and the Augmented Dickey-Fuller test, see [2].

2.2. Support vector regression

The model for support vector regression can be stated by the linear model

$$y(\boldsymbol{x}) = \boldsymbol{w}^T\phi(\boldsymbol{x}) + b \tag{1}$$

where $\boldsymbol{w}$ is a weight parameter, ϕ is a feature transformation and b is a scalar[1]. Suppose that the training data set consists of n input vectors $\boldsymbol{x}_i$ and n target values t_i. The primal optimization problem for support vector regression is

$$\min_{\boldsymbol{w},b,\xi,\hat{\xi}} \left(C\sum_{i=1}^{n}(\xi_i + \hat{\xi}_i) + \frac{1}{2}\|\boldsymbol{w}\|^2 \right) \tag{2}$$

subject to constraints

$$\begin{gathered} \xi_i \ge 0, \\ \hat{\xi}_i \ge 0, \\ t_i \le y(\boldsymbol{x}_i) + \varepsilon + \xi_i, \\ t_i \ge y(\boldsymbol{x}_i) - \varepsilon - \hat{\xi}_i, \end{gathered}$$

where $\xi_i, \hat{\xi}_i$ are slack variables and $C > 0$ is the penalty parameter. This optimization problem (2) can be stated in its dual form as follows:

$$\max_{\boldsymbol{a},\hat{\boldsymbol{a}},\boldsymbol{u},\hat{\boldsymbol{u}}} \inf_{\boldsymbol{w},b,\xi,\hat{\xi}} \left(C\sum_{i=1}^{n}(\xi_i + \hat{\xi}_i) + \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{n}(\mu_i\xi_i + \hat{\mu}_i\hat{\xi}_i) - \sum_{i=1}^{n} a_i(\varepsilon + \xi_i + y_i - t_i) - \sum_{i=1}^{n}\hat{a}_i(\varepsilon + \hat{\xi}_i - y_i + t_i) \right),$$

where $a_i, \hat{a}_i, \mu_i, \hat{\mu}_i \geq 0$ are Lagrange multipliers. By solving this Lagrangian problem (see [1] for details), we have a support vector regression model expressed in form of (1) as:

$$y(\boldsymbol{x}) = \sum_{i=1}^{n} (a_i - \hat{a}_i)\, K(\boldsymbol{x}_i, \boldsymbol{x}) + b,$$

where $K(\boldsymbol{x}, \boldsymbol{y}) = \phi(\boldsymbol{x}) \cdot \phi(\boldsymbol{y})$ is the kernel function.

In this paper, we only consider the linear kernel $K(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x} \cdot \boldsymbol{y}$ and the gaussian radial basis function kernel $K(\boldsymbol{x}, \boldsymbol{y}) = \exp(-\gamma \|\boldsymbol{x} - \boldsymbol{y}\|^2)$.

2.3. Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE)

The RMSE and MAE, which are used to measure the performance of our models, are formulated as follows:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} \left(X_i - \hat{X}_i\right)^2}{n}}, \qquad \text{MAE} = \frac{\sum_{i=1}^{n} \left|X_i - \hat{X}_i\right|}{n},$$

where n is the number of observations, X_i is the true i-th observation, and $\hat{X}_i$ is the predicted i-th observation by the model[6].

3. Results and discussions

The Augmented Dickey-Fuller test is performed on four time-series variables in each region in DKI Jakarta, with $\boldsymbol{\alpha = 0.1}$. The results are that we can assume that every variable in every region is stationary, except for cumulative rainfall in South Jakarta. This variable is thus taken out of the consideration when constructing the regression model. Throughout most of this section, we focus on our results in Central Jakarta.

Table 1. The values of the cross-correlation between variables in Central Jakarta.

Lag (in weeks)	Cross-correlation between incidence and			
	previous incidence	average temperature	cumulative rainfall	average relative humidity
1	0.736	-0.036	0.074	0.268
2	0.686	-0.082	0.138	0.322
3	0.635	-0.162	0.172	0.386
4	0.632	-0.180	0.212	0.404
5	0.568	-0.208	0.255	0.416
6	0.513	-0.232	0.266	0.432
7	0.424	-0.220	0.279	0.433
8	0.377	-0.239	0.287	0.429

Values of the cross-correlation between each predictor variable and the number of dengue incidents in Central Jakarta is shown in Table 1. The time lag used for constructing the regression model is obtained from the lag maximizing the absolute value of the cross-correlation:

1. The lag between incidence and previous incidence is 1 week.
2. The lag between incidence and previous average temperature is 8 weeks.
3. The lag between incidence and previous cumulative rainfall is 8 weeks.
4. The lag between incidence and previous average humidity is 7 weeks.

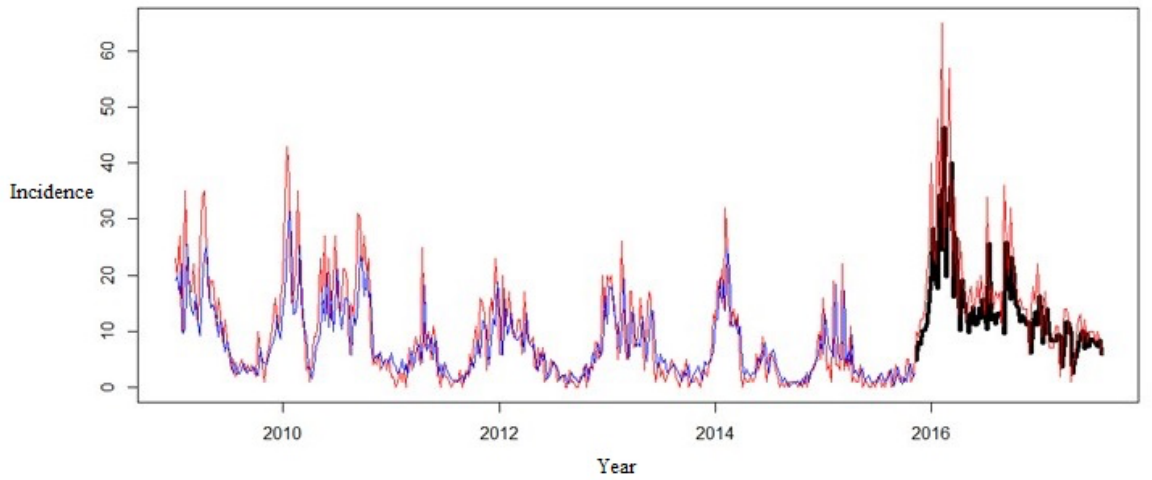

Fig. 2. The graph of incidence prediction with linear kernel ($C = 10$) in Central Jakarta using 80% training data. The red graph denotes the number of actual incidents, the blue graph denotes the prediction of training data, and the black graph denotes the prediction of test data

We can see that the lag chosen for the weather variables are greater than the lag chosen for previous incidence. This type of pattern also arises when considering the lag for other regions. Furthermore, we have a positive correlation between dengue incidence and previous incidence, cumulative rainfall, and average relative humidity. On the other hand, we have a negative correlation between dengue incidence and average temperature. This is also true for other regions in DKI Jakarta.

The graphs of the time series of dengue incidence and the prediction model obtained from support vector regression with linear and radial kernels are shown in Fig. 2 and Fig. 3, respectively. From both figures, the performance of the linear kernel is markedly better than the radial kernel. In particular, the radial kernel fails to predict the spike of incidents in early 2016, while the linear kernel manages to predict it. This makes the radial kernel less reliable to give information of potential outbreaks in the future. A more comprehensive look at the performance of both kernels with varying parameters is presented in Table 2.

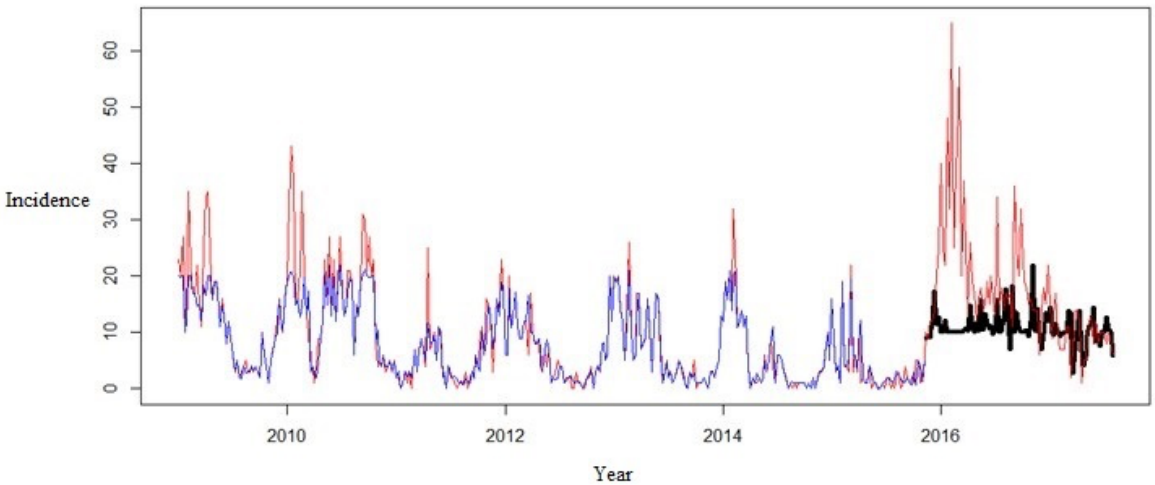

Fig. 3. The graph of incidence prediction with radial kernel ($C = 10, \gamma = 0.1$) in Central Jakarta using 80% training data. The red graph denotes the number of actual incidents, the blue graph denotes the prediction of training data, and the black graph denotes the prediction of test data

Table 2. Evaluation of incidence prediction in Central Jakarta with 95% training data.

Kernel	C	γ	RMSE Training	RMSE Test	MAE Training	MAE Test
Linear	0.01	-	6.6202	3.6678	4.1767	2.8115
Linear	0.1	-	6.6141	3.6872	4.1756	2.8198
Linear	1	-	6.6139	3.6860	4.1758	2.8186
Linear	10	-	6.6139	3.6774	4.1776	2.8195
Radial	0.01	0.1	10.2749	3.9081	7.3470	3.2972
Radial	0.01	0.5	10.2994	3.9236	7.3819	3.3159
Radial	0.01	1	10.3019	3.9240	7.3868	3.3166
Radial	0.1	0.1	9.9982	3.7760	6.9111	3.1262
Radial	0.1	0.5	10.1911	3.9071	7.2194	3.2955
Radial	0.1	1	10.2141	3.9112	7.2682	3.3028
Radial	1	0.1	8.5113	3.4070	5.2938	2.4883
Radial	1	0.5	9.1617	3.5732	5.9653	2.8776
Radial	1	1	9.4040	3.7691	6.1583	3.1380
Radial	10	0.1	4.6332	4.0628	1.3272	3.1230
Radial	10	0.5	4.7198	4.4944	1.6213	3.3221
Radial	10	1	4.7557	3.6609	1.3528	2.8527

We can see from Table 2 that the error for models obtained by the linear kernel tends to be less than the error for models obtained by the radial kernel. In addition, the radial kernel's performance is not consistent since its accuracy is highly dependent on its parameters. The linear kernel does not have this problem since its accuracy is roughly the same even though a different parameter is used. Based on this observation, we exclusively use the linear kernel to predict the number of dengue incidents for other regions. The accuracy of support vector regression with linear kernel for predicting incidents in other regions of DKI Jakarta is given in Table 3.

Table 3. Evaluation of incidence prediction in DKI Jakarta with 95% training data.

Region	C	RMSE Training	RMSE Test	MAE Training	MAE Test
East	0.01	6.6304	9.0966	4.3321	7.9919
East	0.1	6.5483	8.8393	4.3301	7.7620
East	1	6.5374	8.8029	4.3300	7.7274
East	10	6.5236	8.7618	4.3297	7.6891
West	0.01	12.7403	5.8418	8.4034	4.5970
West	0.1	12.7441	5.8629	8.4012	4.6335
West	1	12.7474	5.8602	8.4013	4.6334
West	10	12.7402	5.8650	8.4015	4.6284
North	0.01	4.9745	6.6133	3.4426	5.1323
North	0.1	4.9662	6.6077	3.4421	5.1355
North	1	4.9665	6.6079	3.4422	5.1354
North	10	4.9693	6.6084	3.4424	5.1321
South	0.01	9.7191	4.5668	6.1510	3.4727
South	0.1	9.6875	4.5629	6.1319	3.5056
South	1	9.6835	4.5550	6.1281	3.5073
South	10	9.6786	4.5507	6.1281	3.5092

We can see from Table 3 that support vector regression with linear kernel gives relatively small errors in predicting dengue incidents in other regions as well. As in the case of Central Jakarta, the choice of the penalty parameter C has little effect on the success of the model.

4. Concluding remarks

Support vector regression with linear kernel is quite good at predicting the number of dengue incidents in DKI Jakarta based on the RMSE and MAE. All our trials involving different penalty parameters C for the linear kernel showed relatively accurate results. The linear kernel tends to have a better accuracy for prediction than the radial kernel when considering the cross-correlation between variables. Therefore, we suggest that future research involving support vector regression and cross-correlation uses the linear kernel instead of radial. Additionally, from the cross-correlation analysis, we find a positive correlation between dengue incidence and previous incidence, cumulative rainfall, and average relative humidity. On the other hand, there is a negative correlation between dengue incidence and average temperature.

Further research can be done to improve the accuracy of prediction by employing other supervised learning methods or by using support vector regression with another type of kernel to compare its performance. Some kernels not studied in this paper include the polynomial kernel $K(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x} \cdot \boldsymbol{y} + 1)^d$ and the sigmoid kernel $K(\boldsymbol{x}, \boldsymbol{y}) = \tanh(\alpha \boldsymbol{x} \cdot \boldsymbol{y} + b)$.

Acknowledgements

This research is funded by Hibah PUTI Q3 UI No. NKB-1997/UN2.RST/HKP.05.00/2020. We would like to thank the Meteorology, Climatology, and Geophysical Agency (BMKG) and the Jakarta Health Department for their data sets, without which this research would not have been possible.

References

1. Bishop CM. *Pattern recognition and machine learning*. New York: Springer; 2006a.
2. Cryer JD, Chan K. *Time series analysis: with applications in R*. 2nd ed. New York: Springer; 2008.
3. Fakhruddin M, Putra PS, Wijaya KP, Sopaheluwakan A, Satyaningsih R, Komalasari KE, et al. Assessing the interplay between dengue incidence and weather in Jakarta via a clustering integrated multiple regression model. *Ecol Complex* 2019;**39**:100768.
4. Guo P, Liu T, Zhang Qin, Wang L, Xiao J, Zhang Qingying, et al. Developing a dengue forecast model using machine learning: A case study in China. *PLoS Neglect Trop D* 2017;**11**:e0005973.
5. Hasanah, Susanna D. Weather implication for dengue fever in Jakarta, Indonesia 2008-2016. *KLS* 2019;**4**:184.
6. Hyndman RJ, Koehler AB. Another look at measures of forecast accuracy. *Int J Forecasting* 2006;**22**:679–88.
7. Pangribowo S, Tryadi A, Indah IS (Eds.). Buletin jendela epidemiologi. 2nd vol. Jakarta: Kementerian Kesehatan Republik Indonesia; 2010.
8. Withanage GP, Viswakula SD, Nilmini Silva Gunawardena YI, Hapugoda MD. A forecasting model for dengue incidence in the District of Gampaha, Sri Lanka. *Parasite Vectors* 2018;**11**:262.