Statistical Learning for Predicting Dengue Fever Rate in Surabaya

by Siana Halim

Submission date: 14-Jul-2020 11:20AM (UTC+0700)
Submission ID: 1357283644
File name: JTI_FL_Jun20_Revised_NEW_Turn.docx (3.57M)
Word count: 4499
Character count: 25033

Statistical Learning for Predicting Dengue Fever Rate in Surabaya

Siana Halim*[1], Felecia[1], Tanti Octavia[1]

Abstract: Dengue fever happening most in tropical countries and considered as the fastest spreading mosquito-borne disease which is endemic and estimated to have 96 million cases annually. It is transmitted by Aedes mosquito which infected with a dengue virus. Therefore, predicting the dengue fever rate as become the subject of researches in many tropical countries. Some of them use statistical and machine learning approach to predict the rate of the disease so that the government can prevent that incident. In this study, we explore many models in the statistical learning approaches for predicting the dengue fever rate. We applied several methods in the predictive statistics such as regression, spatial regression, geographically weighted regression and robust geographically weighted regression to predict the dengue fever rate in Surabaya. We then analyse the results, compare them based on the mean square error. Those four models are chosen, to show the global estimator's approaches, e.g. regression, and the local ones, e.g. geographically weighted regression. The model with the minimum mean square error is regarded as the most suitable model in the statistical learning area for solving the problem. Here, we look at the estimates of the dengue fever rate in the year 2012, to 2017, area, poverty percentage, precipitation, number of rainy days for predicting the dengue fever outbreak in the year 2018. In this study, the pattern of the predicted model can follow the pattern of the true dataset.

Keywords: Global Moran I statistics; Local Moran I statistics; Regression, Spatial Regression, Geographically Weighted Regression.

Introduction

Dengue fever happening most in tropical countries and considered as the fastest spreading mosquito-borne disease, which is endemic and estimated to have 96 million cases annually. It is transmitted by Aedes mosquito which infected with a dengue virus. There are several factors that cause dengue fever: the failure to control the Aedes mosquito populations, uncontrolled urbanization, and high population growth [1]. Dengue Fever most commonly happens in the urban environment. Indonesia, with its tropical climate and high humidity, has a high possibility for Dengue transmission. Indonesia reported as the second largest with dengue fever cases among 30 endemic countries [2]. Dengue fever increased rapidly over the past 45 years in Indonesia, with victims shifting from young children to older age groups [3]. The increasing number of dengue fever cases is more likely followed by the spread of the cities infected in all 34 provinces in Indonesia. From a total of 497 cities in Indonesia, about 80% reported the dengue fever cases in 2017 [2].

Therefore, predicting the dengue fever outbreak has become the subject of researches in many tropical countries. Some of those researchers predicted dengue fever outbreak in Srilanka [4], in Thailand [5], in the Northwest Coast of Yucatan, Mexico and San Juan, Puerto Rico [6]. These groups of researchers modelled the dengue fever outbreak using neural network approaches. In Indonesia, Mahdiana *et al.* [7] predicted dengue hemorrhagic fever (DHF) using vector autoregressive spatial autocorrelation (varsa). Mahdiana et al. [7] used five years dataset from Sleman, a district in Central Java, Indonesia, to predict the DHF outbreaks. In the model, they include min, max and average temperature, average humidity, and rainfall and irradiation time.

In Surabaya's case, additional to the weather condition, we also explore the population density, the precipitation and the poverty percentage as the factors that may affect the DHF. In the previous study, we [8] used the geographically weighted regression to predict the dengue fever outbreak in Surabaya, to continue the exploration, in this paper we model the outbreak prediction using statistical learning approaches.

Methods

We describe the data by clustering the location based on the Dengue fever rate, poverty rate and the number of rainy days. We use partition around medoids clustering. Partitioning around medoids (PAM) is a clustering algorithm which is based on k-medoids instead of k-means. In the k-medoids clustering, each cluster is represented by one of the data points in the cluster. These points are named as cluster medoids. A medoid is an object within a cluster in which average dissimilarity between the point and all other members of the cluster is minimal [17]. The k-medoids is a robust alternative to k-means clustering. The PAM algorithm [18], is the most common algorithm for performing the k-medoids clustering. It consist of five steps: (1) Select k object to become a medoid, (2) calculate the dissimilarity matrix, (3) assign each object to its closest medoid, (4) for each cluster search if any object of the cluster can decrease the average dissimilarity coefficient if it does, select the object, (5) if there is one medoid changed then go to (3) else end the algorithm.

There are many approaches to modelling the dengue fever outbreak. They can be classified into two approaches, i.e., machine learning and statistical learning. In this study, we are exploring the statistical learning approach to find the most suitable model for predicting the DHF outbreaks. We applied several methods in the predictive statistics such as regression, spatial regression, geographically weighted regression and robust geographically weighted regression to predict the dengue fever outbreak in Surabaya. We then analyses the results, compare them based on the mean square error. Those four models are chosen, to show the global estimator's approaches, e.g. regression, and the local ones, e.g. geographically weighted regression. The model with the minimum mean square error is regarded as the most suitable model in the statistical learning area for solving the problem.

We also test the spatial correlation of the dengue fever outbreak rate in each Puskesmas, we used the spatial local Moran I statistics. Anselin [19] suggested the local Moran's I statistics for identifying local clusters and local spatial outliers. The Local Moran's I statistics can be formulated as:

$$I_i = \frac{n(y_i-\bar{y})\sum_{j=1}^{n} w_{ij}(y_j-\bar{y})}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \qquad (1)$$

where y is the variable of interest. w_{ij} is the weight for the i, j observations.

We then modeled the data based on the statistical learning. The first model is the well-known regression model that can be formulated as

$$Y = X\beta + \varepsilon \qquad (2)$$

where X is the independent variables, Y is the number of dengue fever infected in each location, β is the global parameter and ε is the random error. The global parameter means for the whole locations they will have the same β. In the regression model, we all know that the error term should be independent. Moreover, in the regression, the spatial dependencies of the dataset do not appear in the model. Therefore, to accommodate those spatial dependencies, the spatial models should be considered.

First, we consider the first-order Spatial Autoregressive (SAR) model

$$Y = \rho \boldsymbol{W} Y + \boldsymbol{X}\beta + \boldsymbol{\varepsilon} \qquad (3)$$

where W is the spatial weigh matrix, (see Anselin [19] for the detail)

Another spatial class model is Spatial Error Model (SEM) model

$$Y = \boldsymbol{X}\beta + (\boldsymbol{I}_n - \lambda \boldsymbol{W})^{-1}\boldsymbol{\varepsilon} \qquad (4)$$

where $\boldsymbol{I}_n$ is identity matrix, λ is a scalar parameter, $\boldsymbol{W}$ is the spatial weight matrix, $\varepsilon \sim N(0, \boldsymbol{I}_n\sigma^2)$ is a vector of disturbance (see Anselin [19] for the detail)

The last spatial class model used in this paper is Spatial Durbin Model (SDM). The SDM concerns about the spatial heterogeneity (see Anselin [19] for the detail)

$$Y = \boldsymbol{X}\beta + (\boldsymbol{I}_n - \lambda \boldsymbol{W})^{-1}\boldsymbol{u} \qquad (5)$$

But $\boldsymbol{X}$ and $u = \boldsymbol{X}\gamma + \boldsymbol{\varepsilon}$ are correlated.

The spatial autoregressive models [20] have assumption that the structure of the models remains constant, i.e., there is no local variations in the parameter estimates. The GWR [21] allows the estimated parameters vary locally.

The geographically weighted regression models [21] can be formulated as

$y_i = X\beta_i + \varepsilon$ (6)
where i is the location in which the local parameters will be estimated.

The β_i is the parameters at the location i and can be estimated as
$\beta_i = (X'W_iX)^{-1}X'W_iy$ (7)
where w_{ij} is the weight for the observation at location i and location j and formulated as the Gaussian function
$w_{ij} = e^{\left(\frac{-d_{ij}}{h}\right)^2}$ (8)
The d_{ij} is the Euclidean distance between the location i and location j, while h is the bandwidth. The bandwidth h can be selected such that the root mean square prediction error is minimum.

To identify and to reduce the effects of the outliers in GW regression, then various robust GW regression has been proposed. Two of them are described in [21]. The first robust model re-fits a GW regression with a filtered data set that has been found by removing observations that correspond to large externally studentised residuals of an initial GW regression fit. An externally studentised residual for each regression location i is defined as [22]:
$r_i = \frac{e_i}{\hat{\sigma}_{-i}\sqrt{q_{ii}}}$ (9)

where e_i is the residual at location i; $\hat{\sigma}_{-i}$ is a leave-one-out estimate of $\hat{\sigma}$; and q_{ii} is the i th element of $(I-S)(I-S)^T$. Observations are deemed outlying and filtered from the data if they have $|r_i| > 3$. The second robust model, iteratively down-weights observations that correspond to large residuals. This (non-geographical) weighting function w_r on the residual e_i is typically taken as:

$$w_r(e_i) = \begin{cases} 1, & \text{if } |e_i| \le 2\hat{\sigma} \\ [1-(|e_i|)^2]^2, & \text{if } 2\hat{\sigma} \le |e_i| \le 3\hat{\sigma} \\ 0, & \text{otherwise} \end{cases} \quad (10)$$

Observe that both approaches have an element of subjectivity, where the filtered data approach depends on the chosen residual cut-off and the iterative (automatic) approach depends on the chosen down-weighting function, with its associated cut-offs.

Results and Discussions

The data were collected data from 63 community health centers (pusat kesehatan masyarakat) in Surabaya. Community health centers are government- mandated community health clinics located across Indonesia [XX], and provide healthcare for population on sub-district level (Kelurahan). In Surabaya, each community health center provides healthcare for one up to two sub-districts level. The community health center recorded diseases the often affects the community. One of the diseases is the dengue fever. Surabaya government focus more on preventive action to reduce dengue fever outbreak. Therefore, the health center will promote dengue prevention through environmental cleaning programs, especially during wet season [23]. This study will help the Surabaya government to predict outbreaks in the neighborhoods of the center of DHF outbreak.

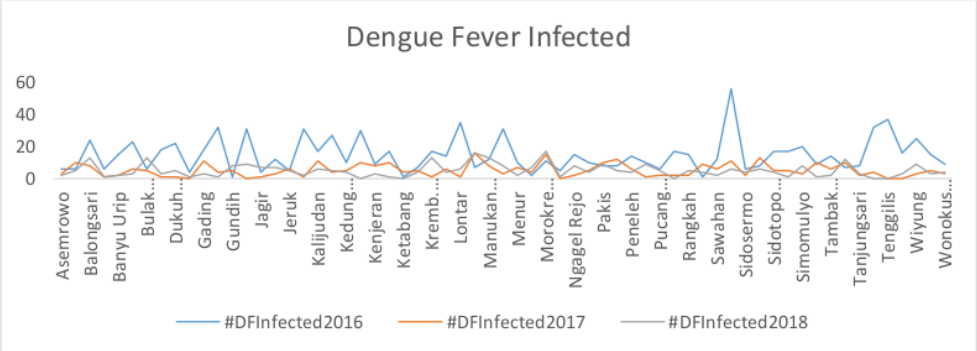

Figure. 1 The numbers of Dengue Fever Infected from 2016-2017 in each community health centers.

We collected data of the number of rainy days in a year, precipitation, maximum and minimum temperature, maximum and minimum humidity, population density, and poverty percentage in each Surabaya's district [24]. The weather is collected from the Perak II Meteorology Station Surabaya. The Data in reported in Surabaya in numbers (*Surabaya dalam Angka* [24]) monthly. The poverty percentage is calculated per family. It is the percentage of total family in a sub-district to the number of considered as poor families by the government.

The summary statistics for the Surabaya in 2018 can be seen in Table 1.

Table 1. Surabaya Statistics 2018

	Min	Mean	Average
Population (thousand)	13617	49427.44	94440
Number of family	4127	14659.48	29055
Density (thousand/Km2)	1618.65	21743.46	141407.7
Poverty percentage (%)	3.32	16.15746	45.99
Area (Km2)	0.915	2.001	14.4
#Rainy day (days/month)	9.83	13.99	16
Precipitation (mm/month)	174.42	482.9797	530.308
Max Humidity per month	70	73.3873	80
Min Humidity per month	46.08	53.14	57.83
Max Temperature	28.21	33.3	34.43
Min Temperature	23.11	26.29	28.73
Six years in education percentages (Elementary school)	1.09	17.1254	35.96
Nine years in education percentages (Middle high school)	5.17	12.83492	38.54
Twelve years in education percentages (High school)	10.25	26.41603	37.82
The total percentage for less than or equal to twelve years in education	31.85007	56.37647	73.80726
More than twelve years in education percentage (University)	2.98	11.64683	23.8

*Summarized from Surabaya in numbers [24]

Clustering

First, we assumed there is any relationship between poverty and the number of dengue fever incidence (DFI) or the dengue fever rate. To prove our assumption, we then cluster the location based on the dengue fever rate (DFR), and the poverty percentage (based on [24]). The DFR is calculated as the number of dengue fever infected in the location at year divided by populations in the same location at the same year for 10,000 people, i.e.:

$$DFR_i(t) = \frac{\#DFI_i(t)}{\#Population_i(t)} 10{,}000$$

While the number of optimal clusters is determined based on the optimum average silhouette width. Rousseeuw [25] defined the silhouette value as a measure of similarity of an object to its own cluster compare to the other clusters. Based on the optimum average silhouette, we get the optimal number of clusters is two (see Figure 3). We then used the partitioning around medoids with the number of clusters equal to two to cluster the regions with respect to the poverty percentage, DFI dan DFR. As a result, we get a Surabaya map which is clustered based on the poverty percentage, DFI dan DFR (Figure 4)

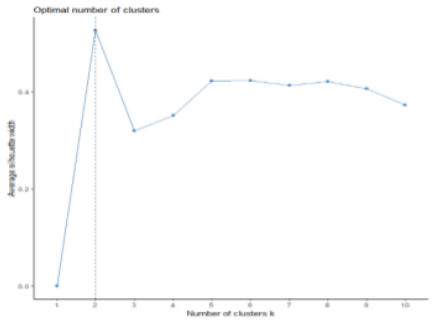

Figure 3. Number of optimum cluster

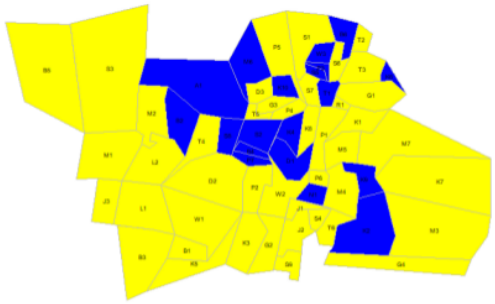

Figure 4. Map of Surabaya which is clustered based on the poverty percentage, DFI dan DFR

Table 2. Cluster summary for poverty percentage to DFR and DFI

Cluster	Poverty%	DFR2018	DFI2018
1	32.54	1.18	5.37
2	11.75	1.15	4.98
T_Test (P-value)		0.457	0.356

We then test the hypothesis to prove the DFR and DFI between those two clusters are significantly different. The P-value of t-Test: Two-Sample shows that nor mean value of DFR or DFI 2018 is significantly different (see Table 2). We can conclude that poverty percentage does not influence the number of dengue fever incidence nor dengue fever rate.

In 2018, the three highest DFR was Tambak Wedi, Putat Jaya and Medokan Ayu. In Putat Jaya there were 9 DF infected, but since the population only 15155 people then per 10000 citizens, the DFR is 5.93. While in Putat Jaya, the DFI is 17 (the highest one in number), since the population is 44913 people, then per 10000 citizens the DFR is only 3.78. It is lower compared to Tambak Wedi. For Medokan Ayu, the DFI is 16, the population is 57647, and the DFR is 2.78 per 10000 citizens.

Secondly, we also assumed that precipitation and number of rainy days influence the DFI and DFR. Using a similar approach, we have three clusters, and the summary of the clusters is given in Table 3. Using one-way ANOVA, we test the DFR and DFI with respect to cluster 1, 2 and 3. The cluster 1 has average precipitation 156.09(mm/year) and the average number of rainy days 13.42 (days/month) the DFR is 1.09 and DFI is 4.94. The DFR for each cluster is significantly different (p-value 0.018) while the DFI is not significantly different (p-value 0.706). We can conclude that precipitation and number of rainy days influence the dengue fever rate. The regions which are registered in cluster 3 are Sidotopo Wetan, Tambak Rejo, and Tanah Kali Kedinding. Those three regions have a precipitation rate of 129.86 mm/year and an average number of rainy days per month 13.92. Among those three regions, the highest DFR is in Tambak Rejo (5.94, DFI = 9), then followed by Tanah Kali Kedinding (1.31, DFI = 7) and Sidotopo Wetan (0.88, with DFI = 5)

Table 3 Cluster summary for precipitation, number of rainy days to DFR and DFI

Cluster	Precipitation	#RainyDays	DFR2018	DFI2018
1	156.09	13.42	1.09	4.94
2	177.30	14.61	1.08	5.07
3	129.86	13.92	2.71	7.00
One-way anova(Pvalue) (one-way			0.018	0.706

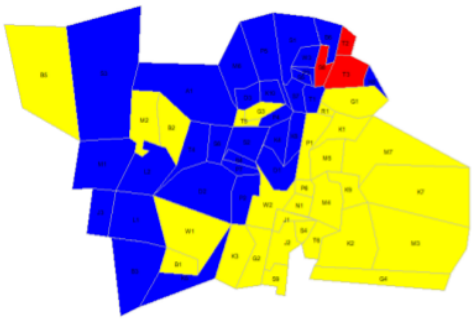

Figure 5. Map of Surabaya which is clustered based on the Precipitation, number of rainy days to DFI and DFR

Global Linear and Spatial Model

The summary statistics and the clustering above give us a clear description of the dengue fever rate and the number of infected persons that occurred in Surabaya. In this section, we want to explore the global and spatial models to predict the DFR.

At the first step, we use the global Moran's I statistics to test the data are under randomization (Ho) or have spatial dependencies. The test shows that the data significantly have spatial dependencies (p-value = 0.0154. To see which districts are spatially correlated strongly, we then use the local Moran's I statistics. Figure 6 shows that there are two districts which have very strong spatial correlation, they are Tenggilis, and both districts have 13 dengue fever infected, and the rate is 2.34 per 10000 citizens. This correlation means that the DFR in Tenggilis highly influenced its neighbourhoods. Tenggilis has the highest DFR in the neighbourhood. The Tenggilis neighborhoods are Sidosermo (DFR = 0.74, DFI = 3), Menur (DFR = 0, DFI = 0), Kalirungkut (DFR = 1.18, DFI = 6). Since the data has dependencies spatially globally and locally, we then use the spatial models for predicting the DFR.

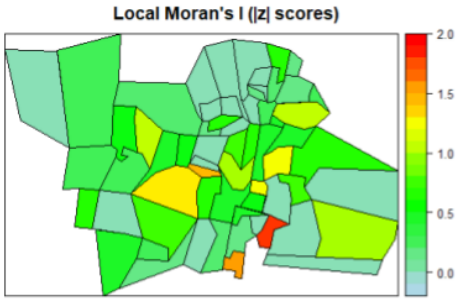

Figure 6. Local Moran's I statistics

To predict the dengue fever rate in 2018 globally, we used simple linear regression. We regressed the DBR2018 to the area, poverty percentage, precipitation, number of rainy days and the dengue fever rate from 2012-2017. As a result, the global model shows only estimated parameters of DFR in 2013, 2016 and 2017 are significant for predicting the DFR2018. The models also not good, since the R2 only 37 percent. So, we cannot use this model for predicting the DFR 2018. The linear regression and the spatial models give similar results (Table 4) shows the summary statistics of those models. Among external parameters used to models the dengue fever rate in 2018, only three variables are significant, they are, DFR 2013, DFR 2016 dan DFR 2017. The external variables such as the area, poverty percentage, precipitation, number of rainfall days rate are not significant.

Tabel 4. The statistics summary for global model using linear regression and spatial models

Model	LM	SAR	SEM	SDM
Intercept	2.1509	2.2232	2.1908	2.2232
Area	0.0405	0.0417	0.0432	0.0417
PovertyPercentage	-0.0115	-0.0115	-0.0116	-0.0115
Precipitation	-0.0048	-0.0049	-0.0046	-0.0049
RainfallDays	-0.0054	-0.0554	-0.0590	-0.0554
DFR2013***	0.0285	0.0285	0.0286	0.0285
DFR2014	-0.0062	-0.0059	-0.0059	-0.0059
DFR2015	-0.0093	-0.0096	-0.0098	-0.0096
DFR2016***	-0.0386	-0.0388	-0.0389	-0.0388
DFR2017***	0.0483	0.4851	0.4825	0.4851
R2	0.3748			
AIC		170.3100	170.3200	170.3100
MSE	0.5974	0.5971	0.5972	0.5971
Moran Residual Test (p-value)	0.4527	0.4108	0.4205	0.4108

LM – Linear Model, SAR – Spatial Auto Regressive, SEM – Spatial Error Model, SDM – Spatial Durbin Model

The p-value of Moran residual test shows that for all models, the residual is randomly distributed, i.e., the spatial process promoting the residual pattern of values is a random chance.

To improve the model performance, we then use the Geographically Weighted Regression (GWR) to model the Dengue Fever Outbreak Rate.

Geographically Weighted Regression

The geographically weighted regression (GWR) models permit the parameters to estimate locally in each district in which the community health centers locate. Table 5 presents the summary of GWR coefficient estimates at data points. The number of rainfall days, precipitation, max and min humidity, and temperature are not varied too much since those community health centers are located in the same climate. Therefore, we only look at the local coefficient estimates at the dengue fever rate (DFR) 2016 and 2017 and poverty percentage. The global parameter weight resulting from the GWR model is the same as the parameter weight resulting from the linear model (compare Table 4 and Table 5).

Tabel 5. Summary of GWR coefficient estimates at data points

	Global	Min.	1st Qu.	Median	3rd Qu.	Max.
Intercept	2.1509	0.0076	1.3770	2.1073	2.9258	3.7791
Area	0.0405	-0.0364	0.0001	0.0127	0.0313	0.1018
PovertyPercentage	-0.0115	-0.0149	-0.0129	-0.0103	-0.0075	-0.0016
Precipitation	-0.0048	-0.0148	-0.0088	-0.0032	0.0002	0.0110
RainfallDays	-0.0538	-0.1156	-0.0812	-0.0663	-0.0534	-0.0211
DFR2013	0.0285	0.0062	0.0175	0.0229	0.0307	0.0391
DFR2014	-0.0062	-0.0130	-0.0109	-0.0086	-0.0054	0.0022
DFR2015	-0.0093	-0.0161	-0.0090	-0.0027	0.0036	0.0129
DFR2016	-0.0386	-0.0500	-0.0405	-0.0331	-0.0242	-0.0113
DFR2017	0.4835	0.3592	0.4396	0.5079	0.5860	0.6941
MSE	0.4368					
Moran Residual Test (p-value)	0.4894					

The local coefficients estimate for each explanatory variable are depicted in Figure 5.
In this research, we decided to divide the local coefficients into three intervals. We want to analyze the low, middle and high local coefficient intervals with respect to the regions. It is well known that the absolute value of the coefficient parameter indicates the strongness of the relationship between the response to the respective explanatory variables. In contrast, the sign of the coefficient parameter indicates the direction of that relationship.
Therefore, in this research, we used two different ways in defining the interval

(1) If the whole local coefficients are positive or negative, we then divide the local coefficient into three intervals, from the lowest coefficient (in absolute value) to the highest one (in absolute value).

(2) If some of the coefficients are negative, and some of them are positive, we divide the interval from the most negative to zero, and then zero to the highest positive. Between those two intervals, we divide longer one into two again so that in overall, we will have three intervals.

The interval is calculated from the min local coefficient – max local coefficient divided by three (See Table 4). For the Poverty percentage, the local coefficients are varied from -0.0149 to -0.0016. All local coefficients are negative. This sign shows that the highest the poverty percentage the lowest the estimate DFR2018 will be and vice versa. The color map of local coefficients with respect to the poverty percentage (PP) is presented in Figure 7a. As an example, we highlighted three regions in the red area with high poverty percentage but low DFR2018, and low poverty percentage but high DFR2018. Those regions are Bulak Banteng (PP = 30.24, DFR2018 = 0.95); Tambak Wedi (PP = 9.45, DFR2018 = 5.94); Wonokusumo (PP = 37.69, DFR2018=0.72). While for the regions in the blue area will not have strong differences as in the red area. For the precipitation, the local coefficients are varied from -0.0148 to 0.011. Some coefficients have negative signs, and others have positive signs. The color map for the negative signs is red and yellow, while the positive sign is blue. Three regions that we already highlighted in clustering based on precipitation section, i.e., Tambak Rejo, Tanah Kali Kedinding and Sidotopo Wetan have the negative signs (in the red area). They have a low precipitation rate but high DFR in 2018. The positive signs indicate that area with higher precipitation tends to have high DFR and vice versa.

By studying the local parameters estimate with respect to the external factors such as poverty percentage and precipitation we hope the Surabaya public health officials can use this information to prevent the spread of dengue fever in the coming year. Figure 8 presented the predicted and the true value of dengue fever rate in 2018. There are four regions in Surabaya that had zeros DFR in 2018. They are Jemursari, Jeruk, Menur, Mulyorejo and Siwalankerto. In the prediction those regions are not predicted as zero but consecutively are 0.53; 0.78; 0.51; 0.81; 0.11. This prediction could happen since the neighbourhood of these regions does not have zero DFR in previous years. For example, since the neighborhoods of Jemursari are Jagir (DFR2017 = 0.78); Sidosermo (DFR2017=0.74), Siwalankerto (DFR2017=0), Gayungan(DFR2017 =0.90), then the DFR of Jemursari in 2018 is predicted as 0.53. In overall, the mean square error of the forecast is 0.4368.

Tabel 4. Local coefficient estimate interval

	Red	Yellow	Blue
Poverty%	[-0.0149, -0.0105)	[-0.0105, -0.006]	[-0.006, -0.0016)
Precipitation	[-0.0148, -0.0074)	[-0.0074,0)	[0,0.011]

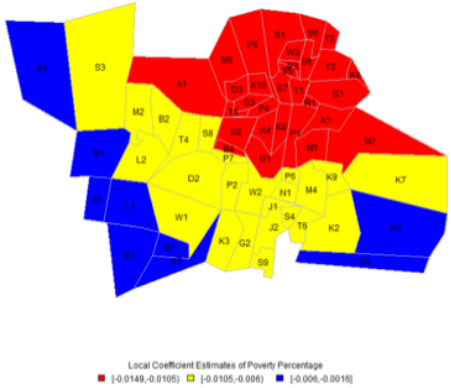

Figure 7a. Local Coefficient estimate for Poverty Percentage

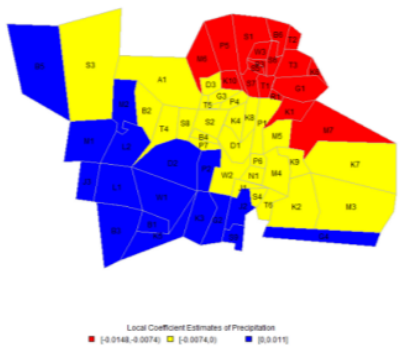

Figure 7b. Local Coefficient estimate for Precipitation

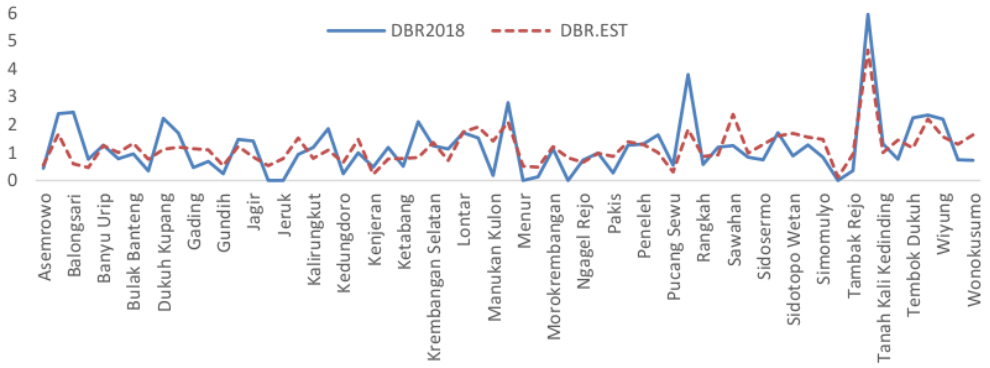

Figure 8. Dengue Fever Rate 2018 vs Predicted Dengue Fever Rate 2018

Conclusion

In this paper, we described the influences of poverty percentage, precipitation and number of rainy days to the 2018 dengue fever rate in Surabaya using partitioning around medoids approach. We then explored statistical learning to predict the DFR2018 using external factors and the DFR2012-DFR2017. The geographically weighted regression in the best model for solving this problem compared to the linear regression model and the spatial models. We also studying the characteristics of local parameters estimate with respect to poverty percentage and precipitation. There is no positive correlation between poverty to the DFR as we presumed in the beginning of the study, in fact the area with high poverty percentage yet has lower DFR compare to the area with the lowest poverty percentage. However, the precipitation has negative correlation even though it is not significant. It's meant the areas with high precipitation tend to have low DFR. This situation fit to the nature of the mosquitos. Mosquitoes breed in pools of water. As a result, we hope that the Surabaya public health officials can use this information to prevent the spread of dengue fever in the coming year. In future research, we will use the machine learning approach for solving the problems and present the result interactively as an app.

References

[1] Bhatt, S., Gething, P.W., Brady, O.J., Messina, J.P., Farlow, A.W., Moyes, C.L., Drake, J.M., Brownstein, J.S., Hoen, A.G., Sankoh, O., Myers, M.F., George D.B., Jaenisch, T., Wint, G.R., Simmons, C.P., Scott, T.W., Farrar, J.J., and Hay, S.I., The Global Distribution and Burden of Dengue, *Nature*, 496(7446), 2013, pp. 504–507.

[2] Haryanto, B., *Indonesia Dengue Fever: Status, Vulnerability, and Challenges*. IntechOpen: Current Topics in Tropical Emerging Diseases and Travel Medicine, 2018.

[3] Karyanti, M.R., Uiterwaal, C.S., Kusriastuti, R., Hadinegoro, S.R., Rovers, M.M., Heesterbeek, H., Hoes, A.W., Buijning-Verhagen, P., The Changing Incidence of Dengue Haemorrhagic Fever in Indonesia: A 45-year Registry-Based Analysis, *BMC Infectious Diseases*, 14, 2014, p.412

[4] Herath, P.H.M.N., Perera, A.A.I., and Wijekoon, H.P., Prediction of Dengue Outbreaks in Srilanka using Artificial Neural Network, *International Journal of Computer Applications*, 101,2014, pp. 1-5.

[5] Jongmuenwai, B., Lowanichchai, S., and Jabjone, S., Prediction Model of Dengue Hemorrhagic Fever Outbreak using Artificial Neural Networks in Northeast of Thailand, *International Journal of Pure and Applied Mathematics*, 118(8), 2018, pp3407-3417.

[6] Laureano-Rosario, A.E., Duncan, A.P., Mendez-Lazaro, P.A., Garcia-Rejon, J.E., Gomez-Carro, S., Farfan-Ale, J., Savic, D.A., and Muller-Karger, F.E., Application of Artificial Neural Networks for Dengue Fever Outbreak Predictions in the Northwest Coast of Yucatan, Mexico and San Juan, Puerto Rico, *Tropical Medicine Infectious Disease*, 3(5), 2018, pp. 1-16.

[7] Mahdiana, D., Winarko, E., Ashari, A., and Kusnanto, H., A Model for Forecasting the Number of Cases and Distribution Pattern of Dengue Hemorrhagic Fever in Indonesia, *International Journal of Advanced Computer Science and Applications*, 8(11), 2017, pp. 143-150.

[8] Halim, S., Octavia, T., Felecia, and Handojo, A. Dengue Fever Outbreak Prediction in Surabaya using a Geographically Weighted Regression, *Times-Icon Proceeding*, Thailand Dec, 2019.

[8]BPS, *Statistics Indonesia*. *"Surabaya Municipality in Figures 2017"*. *surabayakota.bps.go.id. Archived from the original on 2019-04-01. Retrieved 2019-04-01.*
[9] Indonesia, *"Indonesia: Java (Regencies, Cities and Districts) – Population Statistics, Charts and Map"*.
[10]Kompas, https://regional.kompas.com/read/2019/01/30/21522801/ada-2660-kasus-demam-berdarah-di-jatim-46-penderita-meninggal.
[11]Kelanakota, https://kelanakota.suarasurabaya.net/news/2019/216123-Puncak-Wabah-DBD-Diperkirakan-April-dan-Maret,-Pemkot-Lakukan-Cegah-Dini.
[17] Kassambara, A., *Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning*, STHDA,2017, http://www.sthda.com
[18] Kaufman L., and Rousseeuw, P.J., *Finding Group in Data: An Introduction to Cluster Analysis*, Wiley, New York, 1990.
[19] Anselin, L., Local Indicators of Spatial Association-LISA, *Geographical Analysis,* 27, 1995, pp.930115.
[20] ver Hoef, J.M., Peterson, E.E., Hooten, M.B., Hanks, E.M., and Fortin, M.J., Spatial Autoregressive Models for Statistical Inference from Ecological Data, *Ecological Society of America*, 88(1), 2018, pp. 36-59.
[21] Fotheringham, A.S., Brunsdon, C.C., and Charlton, M.E., *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*, Wiley, Chichester, 2002.
[22] Gollini, I., Lu, B., Charlton, M., Brunsdon, C., and Harris, P., GWmodel: an R Package for Exploring Spatial Heterogeneity using Geographically Weighted Models, *Journal of Statistical Software*, 63(17), 2015, pp.
[23]Jumantik, https://surabaya.kompas.com/read/2018/11/01/16393191/ gerakan-1-rumah-1-jumantik-langkah-risma-agar-surabaya-bebas-dbd
[24]Surabaya dalam Angka, Biro Pusat Statistika Surabaya, https://surabayakota.bps.go.id/publication/2018/08/21/35de76f19338e3ecd445b838/kota-surabaya-dalam-angka-2018.html.
[25] Rousseeuw, P. J., Silhouttes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis, *Computational and Applied Mathematics*, 20, 1987, pp. 53-65

Statistical Learning for Predicting Dengue Fever Rate in Surabaya

ORIGINALITY REPORT

15%	13%	11%	11%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

MATCH ALL SOURCES (ONLY SELECTED SOURCE PRINTED)

4%

★ www.scopus.com

Internet Source

Exclude quotes	Off	Exclude matches	< 1%
Exclude bibliography	Off		