INFECTION ECOLOGY & EPIDEMIOLOGY
2023, VOL. 13, 2218207
https://doi.org/10.1080/20008686.2023.2218207

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

OPEN ACCESS Check for updates

Predicting potential areas at risk of the Dengue Hemorrhagic Fever in Jakarta, Indonesia—analyzing the accuracy of predictive hot spot analysis in the absence of small geographical area data

Valentino Prasetya, Valentino Vito, Ivan N. Tanawi, Dipo Aldila and Gatot F. Hertono

Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Indonesia, Depok, Indonesia

ABSTRACT

Dengue Hemorrhagic Fever (DHF), a more severe form of dengue, is one of the most rapidly spreading mosquito-borne diseases in the world. This study is motivated by the rising DHF incidence in Jakarta, the capital city of Indonesia. We mainly utilized hot spot analysis, which employs spatial statistics to find at-risk areas for DHF outbreaks in Jakarta's five municipalities. However, producing informative results from hot spot analysis requires a complete set of data on each of Jakarta's 42 districts, which is not available. We thus propose the idea of using small area estimation (SAE) and machine learning to make up for the lack of data. To evaluate whether this proposed method is effective, we compare the hot spot results from the estimation with the actual data of each district. The results show that the estimated hot spot map is similar to the hot spot map from the actual data. This implies that it is possible to find potential at-risk areas of dengue fever without a complete dataset in every small geographic area. We expect that this research can increase the performance of DHF control measures at the district level, even in the absence of small area data.

ARTICLE HISTORY
Received 21 January 2021
Accepted 22 May 2023

KEYWORDS
Getis-Ord Gi; hot spot analysis; machine learning; small area estimation; support vector regression

Introduction

Dengue is a serious disease whose primary vector is the *Aedes aegypti* mosquito. Dengue Hemorrhagic Fever (DHF), a more severe form of dengue, is one of the most rapidly spreading mosquito-borne viral diseases worldwide [1]. According to the World Health Organization (WHO) [1,2], the incidence of dengue has increased over eightfold in the last two decades, from around 500 thousand cases in 2000 to over 4.2 million cases in 2019. The reported death toll has also increased significantly, from 960 in 2000 to 4,032 in 2015. While the disease is endemic to more than 100 countries, Asia is still the most affected region, with around 70% of DHF incidents happening in the globe.

A variety of recent research and case studies on dengue has been done in areas like China [3], Malaysia [4], Spain [5], and Africa [6]. In this paper, we focus on Jakarta, the capital city of Indonesia. Although some programs have been initiated and efforts to control DHF have been made in Indonesia, the DHF incidence and case fatality rate are still high and not showing any significant changes. This is due to several challenges, such as the dense population in the city, lack of community awareness, and lack of access to health centers [7]. As one of the most populated cities in the country, Jakarta has a significant public health burden due to its DHF outbreaks. According to Indonesia's Ministry of Health, there were 3,352 cases of DHF in Jakarta throughout 2017, with 32.41 citizens infected for every 100,000 citizens [8].

There are several factors as to why dengue fever is difficult to control. One such factor is difficulty in finding vector breeding sites. However, several techniques of spatial analysis have proven to be helpful for tackling this problem. Additionally, Guo et al. [3] employed machine learning techniques to develop prediction models for future incidence, including support vector regression (SVR), which we discuss in this paper. We followed this computational approach and applied various machine learning algorithms to our problem.

Hot spot analysis is a spatial statistical technique that is being employed more frequently as time goes on. It is a method for determining clusters of interest in the region of study. This technique has various applications in epidemiology, where it can be used to find at-risk areas for potential outbreaks [9,10]. It also has several applications in other fields of research [11,12].

Despite their utility, hot spot analysis requires a sufficient number of geographical areas to be able

CONTACT Gatot F. Hertono gatot-f1@ui.ac.id Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Indonesia, Depok 16424, Indonesia

to produce reliable and useful information. The incidence of DHF and meteorological datasets involving Jakarta, provided by Indonesia's Meteorology, Climatology, and Geophysical Agency and the Jakarta Health Department, are mostly only available at a municipal level consisting of only five geographical areas. As such, hot spot analysis will not produce a useful result if applied directly.

To solve this data-related limitation, we attempt to estimate the incidence at the district level (which consists of 42 geographical areas) using the data at the municipal level with the help of some demographical information around the districts. This idea is inspired by the method of small area estimation (SAE), which has been known to solve similar problems in different fields of research [13–15]. Machine learning methods, especially supervised learning, are then used to improve this idea and obtain a more accurate result in our estimation.

The primary aim of this study is to help Jakarta's public officials predict areas at risk of an outbreak using the available data as an early warning mechanism. Our secondary aim is to evaluate whether our method of hot spot analysis, predictive modeling, and small area estimation can produce an accurate hot spot map when compared to the map produced with the actual data despite some data deficiency. We expect that this study can be applied to future DHF control efforts and preventive programs so that priority can be given to high-risk districts.

Materials and methods

Data collection and analysis

The datasets used for this study consisted of daily numerical data on DHF incidence and weather (including average temperature, rainfall, and average relative humidity) of five municipalities in Jakarta (consisting of West, East, North, South, and Central Jakarta). The datasets for incidence and weather were provided by the Jakarta Health Department and Indonesia's Meteorology, Climatology, and Geophysical Agency, respectively, and were entered in Microsoft Excel. The daily data were converted into weekly data, comprising of 455 weeks from January 2009 to September 2017.

The dataset for the years 2009 to 2016 contained information on each of the municipalities taken as one whole. That is, the data did not provide details on each district of the municipalities before 2017. Hence, we would like to only use the data at the municipal level (before the data is elaborated at the district level) to predict the incidence of 2017. Using the programming language R, we used SVR to develop the required prediction model. Afterward, we employed some SAE techniques via Python to convert the predicted municipal level data so that they were at a district level. The resulting hot spot map from these data were then compared with the hot spot map obtained from the actual district-level data of 2017.

Predictive modeling

A predictive model is used to predict an outcome when information is incomplete and to extrapolate based on similar conditions [16]. In other words, it is a model that serves to give predictions of future events based on patterns observed in the past. In our case, we trained the model using the incidence and weather data obtained from 2009 to 2016 to predict the incidence in 2017 with sufficient accuracy. Of the multiple ways to construct a predictive model, we only considered a certain type of multivariate regression method. Specifically, we use SVR, as it is the most accurate method based on the case study done previously in [3]. SVR is classified as a supervised learning algorithm.

Supervised learning

Machine learning involves the study of computer algorithms that are trained through a learning process. These algorithms can be used to develop a mathematical model based on a sample of data called the training data. Supervised learning is a subset of machine learning where the training data is already labeled. For example, a regression model evaluates some labeled data to predict the labels of the other unlabeled data. Commonly used supervised learning algorithms include linear regression, decision tree, K-nearest neighbors, random forest, and extra tree regressor [17,18]. An extensive discussion on machine learning, including both supervised and unsupervised approaches, can be found in [17].

SVR

The regression model for support vector regression (SVR) is in the form

$$y(x) = \sum_i w_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b,$$

where $\boldsymbol{x}$ is a vector consisting of the predictor variables, $\boldsymbol{x}_i$ is a data point from the training dataset in the form of a vector, w_i and b are constants, and $K(\boldsymbol{x}_i, \boldsymbol{x})$ is a type of function known as the kernel function. For more details, see Bishop [17].

In this study, we chose incidence, average temperature, rainfall, and average relative humidity (all observed a few weeks previously) as our four predictor

variables. Hence, $\boldsymbol{x}$ and $\boldsymbol{x}_i$ are four-dimensional vectors. We used the linear kernel $K(\boldsymbol{x}_i, \boldsymbol{x}) = \boldsymbol{x}_i \cdot \boldsymbol{x}$, which is found to produce a more accurate result for this type of problem in [19].

SAE

The term 'small area' is commonly used to indicate a small geographical area, such as a county or district. In general, small areas can denote any domain for which direct estimates of adequate precision cannot be produced [20]. A study involving the estimation of a small area parameter is called SAE. Specifically, SAE is a statistical technique of estimating a parameter in a small sub-population when the data sampling for the parameter of interest is done in larger geographical areas. Statistical references for SAE can be found in [20,21].

An example of its application can be seen from the study by Stasny et al. [13]. They attempted to estimate wheat production at a county level from the indicator variables of the region where the country is located and other predictor variables such as acres planted, acres harvested, and previous wheat production in the county. In this study, we used the simplest SAE models, including a direct proportion model and a regression model.

Hot spot analysis

According to [22], a hot spot analysis can indicate where the clusters in our dataset are and how significant they are. The areas indicated by the significant clusters, called hot spots, are areas with a risk for dengue infections. There are a few statistics that are in common use for hot spot analysis, one of which is the Getis-Ord Gi* statistic (or simply the Gi* statistic) [23]. This statistic is defined by the following formula:

$$G_i^* = \frac{\sum_j w_{ij} x_j}{\sum_j x_j},$$

where i and j indicate two areas in the study region, G_i^* is the value of Gi* statistic in i, x_j is the parameter of interest in j (the number of incidents in our case), and w_{ij} is the weight of the relationship between i and j. The Gi* statistic measures the degree of association between areas within the study region [23]. The greater the Gi* value, the higher is the significance of an area.

Figure 1 shows the flowchart for the implementation of our hot spot analysis. Firstly, the spatial data consisting of the incidence and geographical information of every area was entered. Secondly, a spatial weight matrix $W = (w_{ij})$ needed to be chosen. The queen contiguity matrix is a very common spatial weight matrix in various applications [24,25]. In this type of matrix, the value of w_{ij} is 1 if areas i and j are adjacent and 0 otherwise. Thirdly, the hot spot confidence level for every area in the study region was obtained by first calculating the corresponding z-score. The calculation of z_{Gi^*} makes use of the following formula:

$$z_{G_i^*} = \frac{\sum_j w_{ij} x_j - (\sum_j w_{ij})(\bar{X})}{S\sqrt{\frac{n\sum_j w_{ij}^2 - \left(\sum_j w_{ij}\right)^2}{n-1}}},$$

where n is the number of areas in the study region, $\bar{X} = \frac{\sum_j x_j}{n}$, and $S = \frac{\sum_j x_j^2}{n} - \bar{X}^2$. Since there are 42 districts to be analyzed in Jakarta, we have $n = 42$.

Each area was put into a color-based category based on its hot spot confidence level. We use red to denote a hot spot with a 99% confidence level, orange to denote a hot spot with a 95% confidence level, and beige to denote a hot spot with a 90% confidence level. If the confidence level is below 90%, then we color the area in white to show that it is not classified as a hot spot. The resulting hot spot map illustrates which cluster of areas requires the most urgent attention.

Results

Prediction of future DHF incidence by municipality

We used the available DHF incidence and weather data (which include data on temperature, rainfall, and humidity) from 2009 to 2016 to estimate municipality-level incidence in 2017. Figure 2 provides predictions of DHF incidence in each municipality between May and September of 2017 (for a total of 20 weeks) that are obtained from the SVR model.

We can see that the SVR prediction fits the actual graphs quite closely. For a numerical illustration of the accuracy of the SVR model, Table 1 presents the predicted and actual incidence in the third week of September 2017. While the predicted values are mostly lower than the actual values, the overall results are, nevertheless, sufficiently accurate to predict the trend of future incidence.

Estimation of future DHF incidence by district

The prediction of DHF incidence by municipality was processed along with demographical data (area and population of districts) to estimate the DHF incidence of every district. Jakarta has five municipalities, which can be broken down to 42 districts. Estimating the incidence on a district-level basis allows us to conduct a proper hot spot analysis.

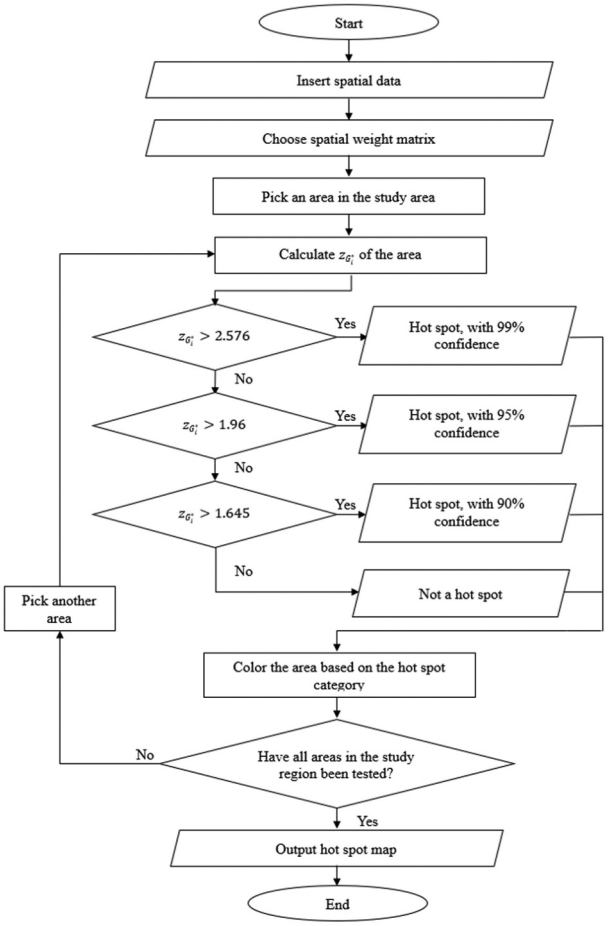

Figure 1. The hot spot analysis algorithm by the Getis-Ord Gi* statistic.

Table 2 shows the performance of several SAE models in estimating district-level incidence from data consisting of municipal-level incidence. We applied two proportion-based approaches and five supervised learning techniques. The proportion of area method estimates a district's incidence by multiplying its municipal's incidence by the ratio between the district's area and its municipal's area. In contrast, proportion of population uses the ratio between the district's population and its municipal's population. From Table 2, we find that supervised learning methods (apart from K-nearest neighbors) perform better than proportion-based methods in terms of mean squared error (MSE) and mean absolute error (MAE).

Table 2 further shows that the random forest model has the least number of errors compared to other models that we used. The random forest algorithm produced an estimation with an MSE of 1.10 and an MAE of 0.71, indicating that the overall estimation deviated from the actual data by about one case. A comparison between the estimation results of the random forest model and the actual data is provided in Table 3.

Hot spot analysis

After the incidence estimates of each of Jakarta's 42 districts were compiled, the last step was to conduct a hot spot analysis. This analysis is done on both the estimation and actual data for evaluative purposes, and Figure 3 illustrates the hot spot map for both.

Discussion

Overall, the SVR predictive model gave reasonably accurate municipality-level incidence estimates, which is in accordance with the findings of [3,19]. As observed in Figure 2, while the model tends to underestimate the number of cases, the overall trends and patterns of future incidence are reliably captured. This is a crucial first step to ensure that the estimation performed at the district level and the resulting hot spot map produce sufficiently accurate results.

The performance results of Table 2 show that simply dividing the incidence of each municipality to its districts in terms of area resulted in a poor estimation at the district level. The MSE and MAE of the estimates are 3.55 and 1.38, respectively, which are higher than most other models. Dividing in terms of population resulted in a markedly better estimation, but it still paled in comparison to machine learning algorithms, such as the decision tree and the extra tree regressor. The K-nearest neighbors algorithm gave the worst MSE out of all the models, which is in line with its poor accuracy in some past studies [26–28]. In general, however, the advanced machine learning methods produced a better estimation than merely calculating the proportions of the area and population of each district.

The notable accuracy of the random forest model compared to other models in Table 2 is due to its high performance as a machine learning algorithm in general [29–32]. We note, however, that the DHF incidence of each district ranges between 0 to 6; thus, the district-level incidence estimation error is still somewhat significant. However, considering that we applied the SVR and the random forest models in succession, each of which decreased the accuracy of the estimation, we feel that the results are satisfactory.

As observed in Table 3, there are slight errors in the incidence estimation. However, we were still able to capture the high and low patterns of the data. This is important so that hot spot areas are not designated as non-hot-spot areas by mistake or vice versa. To further check the quality of the model, we can calculate the coefficient of determination R^2 of the estimation results [33]. The coefficient of determination measures the correlation between the estimated values and the actual values. An R^2 value greater than 0.50 means that

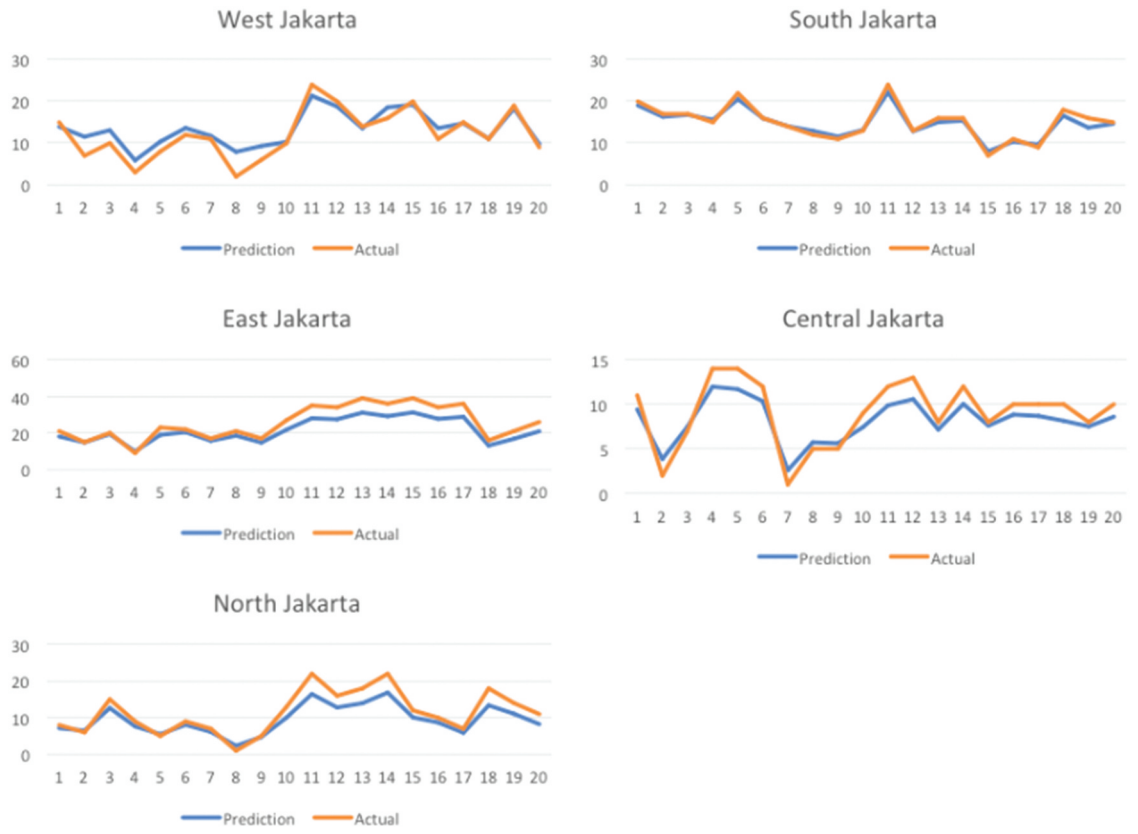

Figure 2. A week-incidence graph illustrating the performance of the SVR model over 20 weeks in 2017. The blue graph represents the number of incidents obtained by SVR, while the orange graph denotes the actual number of incidents.

Table 1. Results of the SVR model versus the actual data in the third week of September 2017.

Municipality	Prediction	Actual Data
West Jakarta	9.816595	9
East Jakarta	20.930816	26
North Jakarta	8.253456	11
South Jakarta	14.664965	15
Central Jakarta	8.600496	10

Table 2. Performance of SAE models for estimating DHF incidence measured by mean squared error (MSE) and mean absolute error (MAE).

Model	MSE	MAE
Proportion of area	3.55	1.38
Proportion of population	2.14	1.06
Linear regression	2.07	1.05
Decision tree	1.59	0.95
K-nearest neighbors	3.90	1.31
Random forest	1.10	0.71
Extra tree regressor	1.56	0.94

the correlation level is strong [34]. A score of $R^2 = 1$ leads to an estimated hot spot map that is identical to the actual hot spot map. We calculated that the results of the random forest algorithm produced a score of $R^2 = 0.8018$. Therefore, we expect that the hot spot maps generated by the estimation and actual data have some degree of similarity.

As expected, we can see from Figure 3 that the hot spot map constructed from the estimated values was quite similar to the one constructed using actual data. For example, the districts of Koja, Kelapa Gading, Tanjung Priok, Cilincing, and Cakung were all hot spots in the two maps, with Koja and Kelapa Gading having the same confidence level in both maps. We note that four of these five districts are located in North Jakarta, a municipality that has become a central interest for a study on DHF in the past [35]. On the other hand, the estimated map contained slight deviations from the actual map. For example, the district of Duren Sawit was designated as a hot spot in the actual hot spot map, while it was not so according to the estimated hot spot map. Moreover, the confidence levels of some hot spots were different compared to the actual hot spot map, namely Tanjung Priok, Cilincing, and Cakung. Despite this, the overall estimation was still able to denote the cluster of areas with a high risk of dengue outbreak.

We note that from Table 3, there are 24 out of 42 districts whose DHF estimates do not match with the actual data. This contrasts with the accuracy of the hot spot map which correctly classifies the hot spot category of all but only four districts. Notable examples include the districts of Kalideres and Penjaringan (top left part of the map). Although the estimates for these districts are both equal to 2 instead of the actual 0, neither the districts nor their adjacent districts are mistakenly

Table 3. Actual and estimated DHF incidence by the random forest model compared side by side. The estimated values are rounded to the nearest integer.

District	Actual DHF incidence	Estimated DHF incidence
Cakung	3	3
Cempaka Putih	1	1
Cengkareng	1	2
Cilandak	0	1
Cilincing	4	6
Cipayung	2	2
Ciracas	2	3
Duren Sawit	3	3
Gambir	0	0
Grogol Petamburan	1	1
Jagakarsa	4	3
Jatinegara	2	2
Johar Baru	0	1
Kalideres	0	2
Kebayoran Baru	1	2
Kebayoran Lama	4	3
Kebon Jeruk	2	1
Kelapa Gading	2	3
Kemayoran	4	2
Kembangan	1	1
Koja	6	4
Kramat Jati	1	2
Makasar	2	2
Mampang Prapatan	2	2
Matraman	2	2
Menteng	0	1
Pademangan	3	1
Palmerah	1	1
Pancoran	1	2
Pasar Minggu	4	3
Pasar Rebo	2	2
Penjaringan	0	2
Pesanggrahan	1	1
Pulo Gadung	5	4
Sawah Besar	0	1
Senen	0	1
Setiabudi	1	1
Taman Sari	0	0
Tambora	6	5
Tanah Abang	5	2
Tanjong Priok	5	5
Tebet	2	2

assigned as a hot spot. This suggests that hot spot analysis is also quite robust to these estimation errors.

The main limitation of this study lies in the accuracy of its predictive model. Although it is able to forecast future outbreaks, SVR has a tendency to underpredict incidence, especially during spikes of cases [19]. This prediction error carries over to the SAE and hot spot analysis performed afterward, which causes some discordance between estimated and actual data, as observed in Table 3 and Figure 3. Therefore, it is necessary to consider how to improve the performance of the predictive model. One option is to consider models other than regression in future studies, such as ensemble and deep learning models [36,37]. Furthermore, the addition of predictor variables besides the weather and previous incidence could be beneficial for the performance of the model.

Another limitation of this study is the lack of mosquito time series data in Jakarta. It is well known that dengue is a vector-borne disease, and the dynamics of mosquitoes determines the rapid spread of dengue in the community. In some references, the authors try to accommodate the dynamics of mosquitoes into their prediction method [38,39] using systems of ordinary differential equations. Further improvements of our method can be done in the direction of using vector dynamics to obtain better results.

We believe that determining hot spots at the district level is important, as it then becomes possible to minimize the burden of the disease by focusing on the areas at risk. From this research, we urge local public officials to focus on areas surrounding North Jakarta to prevent future DHF outbreaks. We hope that the SAE methods

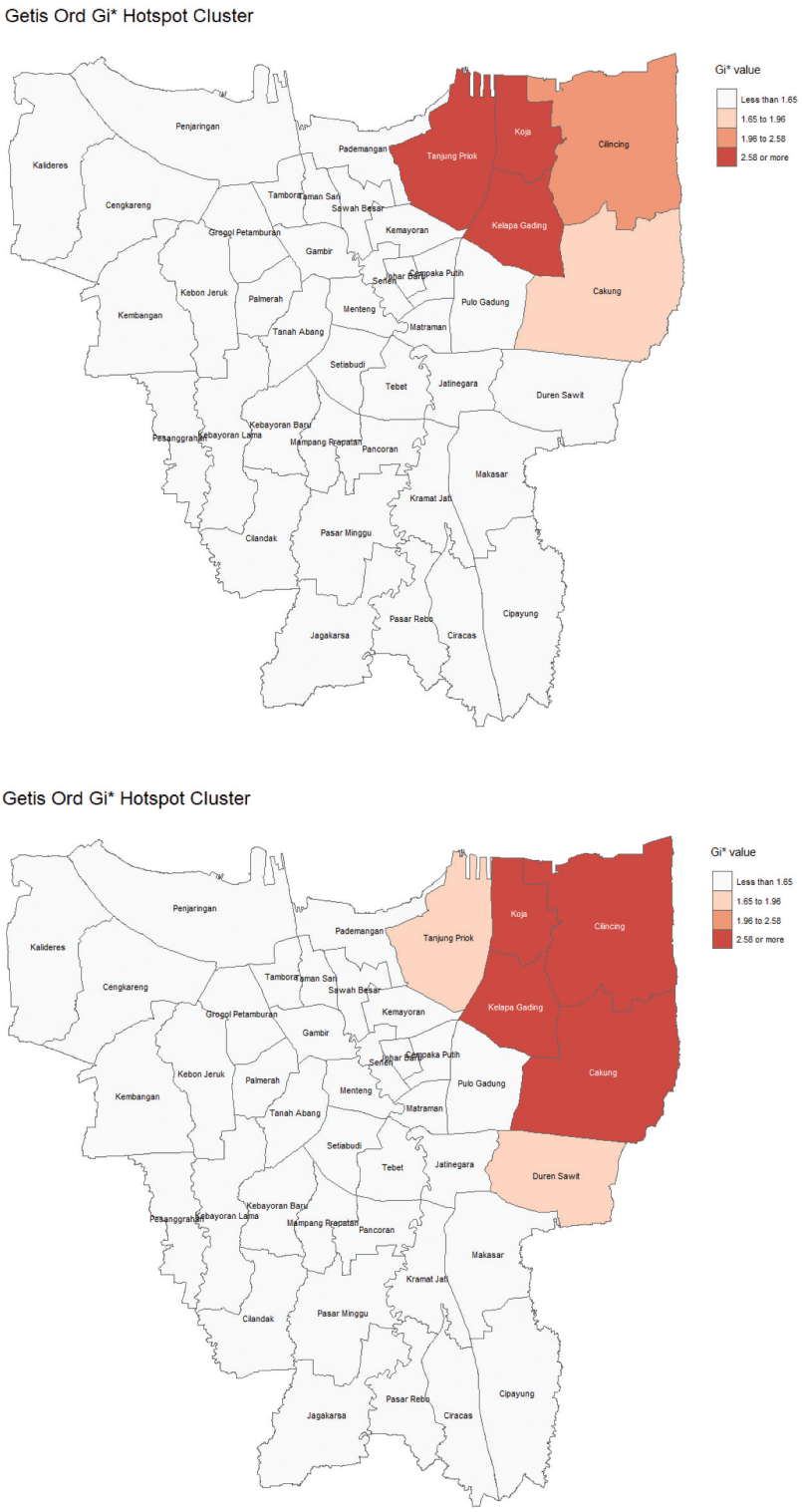

Figure 3. Hot spot map constructed from the DHF incidence estimation (top) and from the actual data (bottom).

performed in this study can be adopted in other regions where local data is a limited resource [40,41] as an early warning mechanism for future dengue outbreaks.

Conclusion

The aim of this research is to develop a spatial method for finding at-risk areas for DHF outbreaks using Jakarta as a case study. At the time of this research, Jakarta's meteorological data collection are available only up to the municipal level. Thus, we incorporate SAE and machine learning to estimate incidence at the district level. The results of the estimation by the SVR and random forest algorithms in succession produce relatively accurate estimates, even though we only have the area and population of each district as supporting datasets. As we compare the hot spot map generated by the estimation results, we conclude that the map has

a high degree of similarity compared to the actual hot spot map. Hence, it is possible to conduct a hot spot analysis in Jakarta without the district data with some degree of accuracy. We expect that this result will be helpful in countries lacking in small area data. Finally, we hope that this research can be an inspiration for better spatial models in the future.

Acknowledgments

We would like to thank Indonesia's Meteorology, Climatology, and Geophysical Agency and the Jakarta Health Department for their datasets and general support. Additionally, we wish to thank Cindy and Cynthia for their help in the data preprocessing stage.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by Universitas Indonesia, PUTI 2020 Grant Scheme Number: NKB-1011/UN2.RST/HKP.05.00/2020.

References

[1] World Health Organization. Global strategy for dengue prevention and control 2012–2020. Geneva: World Health Organization. 2012.
[2] World Health Organization. Dengue and severe dengue. No. WHOEM/MAC/032/E. Cairo: World Health Organization. Regional Office for the Eastern Mediterranean; 2014.
[3] Guo P, Liu T, Zhang Q, et al. Developing a dengue forecast model using machine learning: a case study in China. PLoS negl trop dis. 2017;11(10):e0005973.
[4] Murphy A, Rajahram GS, Jilip J, et al. Incidence and epidemiological features of dengue in Sabah, Malaysia. PLoS negl trop dis. 2020;14(5):e0007504.
[5] Redondo-Bravo L, Ruiz-Huerta C, Gomez-Barroso D, et al. Imported dengue in Spain: a nationwide analysis with predictive time series analyses. J Travel Med. 2019;26(8):taz072.
[6] Sintayehu DW, Tassie N, De Boer WF. Present and future climatic suitability for dengue fever in Africa. Infection Ecology & Epidemiology. 2020;10(1):1782042.
[7] Haryanto B. Indonesia dengue fever: status, vulnerability, and challenges. Current Topics In Tropical Emerging Diseases And Travel Medicine. 2018;5:81–92.
[8] Ministry of Health of the Republic of Indonesia. The situation of dengue fever in Indonesia in the year 2017. Jakarta: Ministry of Health of the Republic of Indonesia. Text in Indonesian; 2018.
[9] Sulistyawati S, Fitriani I . Risk factor and cluster analysis to identify malaria hot spot for control strategy in Samigaluh Sub-District, Kulon Progo, Indonesia. Iran J Public Health. 2019;48(9):1647.
[10] Bogale GG, Gelaye KA, Degefie DT, et al. Spatial patterns of childhood diarrhea in Ethiopia: data from Ethiopian demographic and health surveys (2000, 2005, and 2011). BMC Infect Dis. 2017;17(1):426.
[11] Hart TC, Zandbergen PA. Effects of data quality on predictive hotspot mapping, 239861. Washington (DC): National Criminal Justice Research Service; 2012.
[12] Wang X, Varady DP. Using hot-spot analysis to study the clustering of Section 8 housing voucher families. Housing Studies. 2005;20(1):29–48.
[13] Stasny EA, Goel PK, Ramsey OJ. County estimates of wheat production. Survey Methodology. 1991;17(2):211–225.
[14] Christiaensen L, Lanjouw P, Luoto J, et al. Small area estimation-based prediction methods to track poverty: validation and applications. The Journal Of Economic Inequality. 2012;10(2):267–297.
[15] Molina I, Rao J. Small area estimation of poverty indicators. Can J Stat. 2010;38(3):369–385.
[16] Costa R, Ed. Predictive modeling and risk assessment. Vol. 4 New York: Springer Science & Business Media; 2008.
[17] Bishop CM. Pattern recognition and machine learning. New York: Springer; 2006.
[18] Mishra G, Sehgal D, Valadi JK. Quantitative structure activity relationship study of the anti-hepatitis peptides employing random forests and extra-trees regressors. Bioinformation. 2017;13(3):60.
[19] Tanawi IN, Vito V, Sarwinda D, et al. Support vector regression for predicting the number of dengue incidents in DKI Jakarta. Procedia Comput Sci. 2021;179:747–753.
[20] Rao JNK, Molina I. Small area estimation. John Wiley & Sons: 2015. 10.1002/9781118735855.
[21] Rao JN. Small-area estimation. Wiley StatsRef: Statistics Reference Online. Hoboken: John Wiley & Sons; 2014. p. 1–8.
[22] Kalinic M, Krisp JM (2018). Kernel Density Estimation (KDE) vs. hot-spot analysis–detecting criminal hot spots in the city of San Francisco. *Proceeding of the 21st Conference on Geo-Information Science*, Lund.
[23] Getis A, Ord JK. 2010. The analysis of spatial association by use of distance statistics. Perspectives on Spatial Data Analysis pp. 127–145. Springer. 10.1007/978-3-642-01976-0_10.
[24] Suryowati K, Bekti RD, Faradila A. A comparison of weights matrices on computation of dengue spatial autocorrelation. IOP Conf Ser Mater Sci Eng. 2018;335(1):012052.
[25] Yan Y, Hu W. Does foreign direct investment affect tropospheric SO2 emissions? A spatial analysis in Eastern China from 2011 to 2017. Sustainability. 2020;12(7):2878.
[26] Danades A, Pratama D, Anggraini D, et al. (2016). Comparison of accuracy level K-nearest neighbor algorithm and support vector machine algorithm in classification water quality status. *2016 6th International Conference on System Engineering and Technology (ICSET)*; Bandung (pp. 137–141). IEEE.
[27] Tamatjita EN, Mahastama AW (2016). Comparison of music genre classification using Nearest Centroid Classifier and k-Nearest Neighbours. *2016 International Conference on Information Management and Technology (ICIMTech)*; Bandung (pp. 118–123). IEEE.
[28] Brown JM (2017). Predicting math test scores using k-nearest neighbor. *2017 IEEE Integrated STEM Education Conference (ISEC)*; Princeton (pp. 104–106). IEEE.

[29] Syaliman KU, Nababan EB, Sitompul OS. Improving the accuracy of k-nearest neighbor using local mean based and distance weight. J Phys. 2018;978(1):012047. doi:10.1088/1742-6596/978/1/012047. IOP Publishing.
[30] Caruana R, Niculescu-Mizil A (2006). An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd International Conference on Machine learning*; Pittsburgh (pp. 161–168).
[31] Caruana R, Karampatziakis N, Yessenalina A (2008, July). An empirical evaluation of supervised learning in high dimensions. *Proceedings of the 25th International Conference on Machine learning*; Helsinki (pp. 96–103).
[32] Calhoun P, Hallett MJ, Su X, et al. Random forest with acceptance–rejection trees. Computational Statistics. 2019;35:983–999.
[33] Rawlings JO, Pantula SG, Dickey DA. Applied regression analysis: a research tool. New York: Springer Science & Business Media; 2001.
[34] Cleophas TJ, Zwinderman AH. Regression analysis in medical research: for starters and 2nd levelers. New York: Springer; 2018. 10.1007/978-3-319-71937-5
[35] Sungkar S, Fadli RS, Sukmaningsih A. Trend of dengue hemorrhagic fever in North Jakarta. Journal Of The Indonesian Medical Association. 2012;61(10):394–399.
[36] Cawood P, Van Zyl T. Evaluating state-of-the-art, forecasting ensembles and meta-learning strategies for model fusion. Forecasting. 2022;4(3):732–751.
[37] Mahmoud A, Mohammed A. 2021. A survey on deep learning for time-series forecasting. Machine learning and big data analytics paradigms: analysis, applications and challengespp. 365–392. Springer; Cham:10.1007/978-3-030-59338-4_19.
[38] Wijaya KP, Aldila D, Schäfer LE. Learning the seasonality of disease incidences from empirical data. Ecol Complexity. 2019;38:83–97.
[39] Nuraini N, Fauzi IS, Fakhruddin M, et al. Climate-based dengue model in Semarang, Indonesia: predictions and descriptive analysis. Infect Dis Model. 2021;6:598–611.
[40] Faridah L, Rinawan FR, Fauziah N, et al. Evaluation of health information system (HIS) in the surveillance of dengue in Indonesia: lessons from case in Bandung, West Java. Int J Environ Res Public Health. 2020;17(5):1795.
[41] Brady OJ, Gething PW, Bhatt S, et al. Refining the global spatial limits of dengue virus transmission by evidence-based consensus. PLoS negl trop dis. 2012;6(8):e1760.