IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 6 Issue 3, March 2019
ISSN (Online) 2348 – 7968
www.ijiset.com

Multivariate Model Based Clustering with Spatial Constrain: Spatial Clustering of Dengue Disease in Bogor, Indonesia

I Gede Nyoman Mindra Jaya[1*], Bertho Tantular[2] and Budi Nurani Ruchjana[3]

[1] Department Statistics, Universitas Padjadjaran, Indonesia

[2] Department Statistics, Universitas Padjadjaran, Indonesia

[3] Department Mathematics, Universitas Padjadjaran, Indonesia

*email: mindra@unpad.ac.id

Abstract

Considering spatial autocorrelation in spatial clustering is vital to obtain valid clusters. Standardized Getis Ord Gi statistics is combined with multivariate model-based clustering to identify the spatiotemporal distribution of dengue disease in Bogor, Indonesia including multivariate dengue disease variables. This method is constructed in two steps, i.e., creating the spatial clustering variables by means Getis Ord Gi and used multivariate model-based clustering to identify the spatial clusters. Model-based clustering with spatial constraint provides a more informative clustering result. In case of dengue disease clustering, it can be used to identify high-risk regions and the significant risk factors (i.e., larvae free rate and population density)

Keywords: *Dengue Hemorrhagic Fever, Spatial clustering, Model Based Clustering, Getis –Ord Statistic, mixture likelihood function, Bogor.*

1. Introduction

Dengue hemorrhagic fever (DD) is an endemic disease common in tropical and subtropical regions. It is spread by the mosquito species Aedes aegypti and causes death within a short period. Indonesia is a tropical country plagued by DD. The incidence rate tends to increase and to spread widely, in line with the density and mobility of the population [1];[2]. Indonesia Health Ministry (2010) [1] shows that the total number of DD cases increased by 12.65% in Indonesia in 2009: from 137,569 in 2008 to 154,855; the number of patients who died during that period grew from 1187 to 1384 or by 16.60%. Hence, the case fatality rate (CFR) reached 1.384% in 2009. Due to this high rate, Indonesia ranks highest in ASEAN. West Java is one of the provinces in Indonesia with a high Incidence Rate (IR) rate and the highest Case Fatality Rate (CFR). See Figure 1.

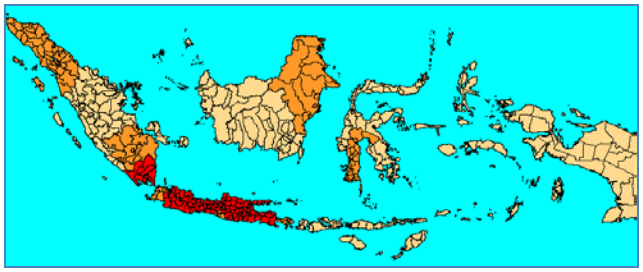


Fig. 1 Distribution of Dengue disease in Indonesia

That is, there were 35,453 cases (IR of 0.84 per 100,000 inhabitants) and 287 casualties (CFR 0.81%) in West Java in 2009 (Director of Sourced Animals for Disease Control, 2010). Bogor is a city in West Java with a relatively high IR. In 2009 there were 1505 cases (IR 172.7 per 100,000 inhabitants) and 4 casualties CFR 0.26%).

DD is an epidemiological disease with spatial characteristics which may be used to determine its distribution. Monitoring the spatial spread of a disease, particularly for a disease like DD with a very high diffusion rate, is required to identify areas that have great potential to become endemic. That is, mapping the distribution of DD cases may help guide prevention and reduction of further incidence.

In response to public health care needs, spatial analysis of health data has developed rapidly during the last decade. Substantial attention has been paid to disease clustering ([3]; [4]). Cluster analysis of diseases is important to evaluate whether a disease is randomly distributed or tends to concentrate as clusters over time and space. Analysis of confounding factors may provide clues to the etiology of the disease. Spatial clustering examines the question of whether cases tend to be located close to each other in space [5]. The pattern of spatial spread of disease reflects the combination of heterogeneity in vector distribution, human-vector contact, and human host factors. (A vector is any agent (person, animal or microorganism) that carries and transmits an infectious pathogen into another living organism). For DD the major vector is the mosquito Aedes aegypti.

Clustering algorithm is mainly based on heuristic methods, such as the partitioning method and the hierarchical method. A typical partitioning method is the K-Means method. A partitioning method is based on dissimilarity between observations. The distance criteria, which is used to measure dissimilarity, includes Manhattan distance and Euclidean distance [6]. Partitioning methods are often used in combination with hierarchical clustering methods to determine how many clusters are needed. A typical hierarchical method is Ward's method. In general clustering algorithms are based on the assumption that the observations are independent. The main characteristic of spatial data, however, is that observations are correlated, i.e., an observation at one location is influenced by and influences observations at other locations. Popular approaches to analyze spatial clustering are univariate techniques based on statistics like the Local Getis Ord statistic [7].

The spread of diseases like DD is multidimensional, i.e. it is influenced by environmental and socioeconomic factors. DD is an endemic viral disease transmitted by the Aedes mosquitoes vector. Its incidence is strongly influenced by environmental factors, particularly rainfall, social behavior as hygiene and population density [1]. So, understanding the incidence of DD requires spatial clustering techniques that can accommodate the multivariate nature of its distribution.

During the last few decades, clustering techniques based on mixed probability distributions have been developed to account for the multivariate nature of the clustering objects. One such method is model-based clustering which assumes that the data come from some distribution function. Hence, the rationality to divide data into G groups is that the data come from a mixture of G different probability models. Fraley and Raftery (1998) [8] show that model-based clustering approaches provide more accurate clustering results than a classical method like Single Link based on terms of classification error, (Nearest neighbor) and K-Means algorithms.

The purpose of this paper is to develop a clustering approach based on the combination of the local Getis Ord statistic and Model-Based Clustering. We call the method is the multivariate model based clustering with spatial constraint.

The remainder of the paper is organized as follows. Section 2 method, explain the data and multivariate model based clustering with spatial constrain. Section 3 presents result and discussion, while section 4 summarizes and concludes

2. Method

2.1. Data

For the Bogor case study, we analyze health data obtained from the Bogor City Health Department. The data set includes the DD IR, the Larva-Free Rate and Population Density by 68 districts (denoted villages). The latter two variables are important covariates of DD incidence rate. [1].

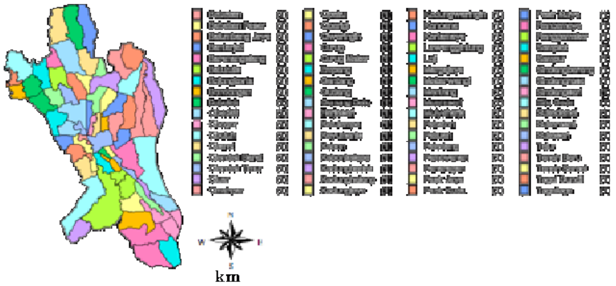


Fig. 2 Bogor City and its 68 villages.

2.2. Model Based Clustering

Model based clustering algorithms are based on probability models, such as the finite mixture model of probability densities. In this context, *model* relates to the type of constraints on and geometric properties of the covariance matrices [9]. In the model based clustering approach, the data are viewed as coming from a mixture of probability distribution, each of which represents different clustering. In other words, in model based clustering, it is assumed that the data are generated by a mixture of probability distribution in which each density function or component represent a different cluster.

Let $\boldsymbol{x} = \{\mathbf{x}_1, \mathbf{x}_2, . . . , \mathbf{x}_n\}$; $\mathbf{x}\in \mathrm{R}^p$, be a set of multivariate observations in n locations with p attributes and let $f_k(\mathbf{x}_i|\boldsymbol{\theta}_k)$ be the density or density probability function of an observation $\mathbf{x}_i$ for the kth componentwith parameters $\theta_{k.}$. For example, if we assume that the multivariate data come from a mixture of Gaussian distributions, then the parameters $\boldsymbol{\theta}_k$ consist of a mean vector μ_k and a covariance matrix $\boldsymbol{\Sigma}_k$, and the density has the form :

$$\boldsymbol{f_k(x_i|\theta_k) = \phi(x_i|\mu_k,\Sigma_k)}$$
$$= (\mathbf{2\pi})^{-\frac{\mathbf{p}}{2}}|\boldsymbol{\Sigma}_\mathbf{k}|^{-\frac{1}{2}}\mathbf{exp}\left\{-\frac{1}{2}(\mathbf{x_i}-\boldsymbol{\mu_k})^\mathbf{T}\boldsymbol{\Sigma}_\mathbf{k}^{-1}(\mathbf{x_i}-\boldsymbol{\mu_k})\right\} \quad (1)$$

here p is the dimension of the data. A finite mixture distribution is a weighted linear combination of a finite number of component distributions. Suppose the component distribution in multivariate normal in p dimensional space with mean vector $\boldsymbol{\mu_k}$ and covariance matrix Σ_k; that is :

$$f_{mixture}(\boldsymbol{x_i}|\boldsymbol{\Theta}) = \sum_{k=1}^{G} \boldsymbol{\pi}_\mathbf{k}\boldsymbol{\phi}(\boldsymbol{x_i}|\boldsymbol{\mu_k},\boldsymbol{\Sigma}_\mathbf{k}) \quad (2)$$

The parameter $\boldsymbol{\Theta}$ is written as $\{\pi_1, \dots, \pi_{G-1}, \boldsymbol{\mu_1}, \dots, \boldsymbol{\mu_G}, \boldsymbol{\Sigma_1}, \dots, \boldsymbol{\Sigma_G}\}$; G is the number of components. The component probability π_k represent the probability that an observation 1 relates to the *k*th component, and so lies in between 0 and 1 and $\sum_{k=1}^{G} \pi_k = 1$, $f_k(\boldsymbol{x_i}|\boldsymbol{\theta_k})$ is the kth component distribution function and θ_k denotes the kth component parameter set [6]. The covariance matrix $\boldsymbol{\Sigma}_k$ can be parameterized as in Table 1 [9].

Table 1 .The geometric interpretation the various parameterizations $\boldsymbol{\Sigma}_k$

Identifier	Model	Distribution	Volume	Shape	Orientation
E	-	Univariate	Equal	-	-
V	-	Univariate	Variable	-	-
EII	λI	Spherical	Equal	Equal	None
VII	$\lambda_k I$	Spherical	Variable	Equal	None
EEI	λA	Diagonal	Equal	Equal	Coordinate Axes
VEI	$\lambda_k A$	Diagonal	Variable	Equal	Coordinate Axes
EVI	λA_k	Diagonal	Equal	Variable	Coordinate Axes
VVI	$\lambda_k A_k$	Diagonal	Variable	Variable	Coordinate Axes
EEE	λDAD^T	Ellipsoidal	Equal	Equal	Equal
EEV	$\lambda D_k A D_k^T$	Ellipsoidal	Equal	Equal	Variable
VEV	$\lambda_k D_k A D_k^T$	Ellipsoidal	Variable	Equal	Variable
VVV	$\lambda_k D_k A_k D_k^T$	Ellipsoidal	Variable	Variable	Variable

In the mixture likelihood approach, the EM algorithm is the most widely used method for estimating the parameters of a finite mixture probability density.

2.3. EM Algorithms for Clustering

The EM algorithm was introduced by Dempster *et.al* (1977) [10] to estimate missing data by maximizing the likelihood function. In the case of clustering, the EM algorithm is used to estimate the parameters of mixture models by iterative relocation to obtain the maximum of the likelihood function. In EM for cluster analysis, the "Complete" data are considered to be $y_i=(x_i, z_i)$, where $z_i = (z_{i1} \dots . z_{iG})$ with

$$z_{ik} = \begin{cases} 1, if\ and\ only\ if\ x_i \in Group_k \\ 0, otherwise \end{cases} \tag{3}$$

Note that conditioned on X, z_{ik} is a Bernoulli random variable with probability τ_{ik} for $z_{ik} = 1$, $E_{Z|X}(z_{ik}) = 1xPr(z_{ik} = 1) + 0xPr(z_{ik} = 1) = \tau_{ik}$, $\sum_{k=1}^{G} z_{ik} = 1$ and so z_i follows a multinomial distribution consisting of one draw from G categories with probabilities $\pi_1, \dots, \pi_G$; that is

$$z_i \sim f(z_i) = Mult(1, \pi) = \binom{1}{z_1, \dots, z_G} \prod_{k=1}^{G} (\pi_k)^{z_{ik}}$$
$$= \prod_{k=1}^{G} (\pi_k)^{z_{ik}} \tag{4}$$

The number of observations within cluster k can be obtained by summing overall the indicators variables z_{ik} . that is $n_k = \sum_{i=1}^{n} z_{ik}$ and $\sum_{i=1}^{n} n_k = n$

Similarity, the density $f(x|z)$ is $f_k(x_i)$ when $z_{ik} = 1$ or simply $\prod_{k=1}^{G} [f_k(x)]^{z_{ik}}$. And $f_k(z)$ is π_k when $z_k = 1$ or simply $\prod_{k=1}^{G} [\pi_k]^{z_{ik}}$. The joint density of (X, Z) is therefore

$f_k(x, z) = f(x|z) = \prod_{k=1}^{G} [f_k(x)]^{z_{ik}} \prod_{k=1}^{G} [\pi_k]^{z_{ik}}$

The EM algorithm for a mixture of multivariate normal is as follows [6]:

1. Initialize $\Theta^{(0)}$

   Take $\pi^{(0)} = \left(\pi_1^{(0)}, \dots ., \pi_G^{(0)}\right) = \left(\frac{1}{g}, \dots ., \frac{1}{g}\right)$; the kth means vector $\mu_k^{(0)} = \bar{x}$, the overall sample mean for all groups. The kth covariance matrix $\Sigma_k^{(0)} = S$ the overall sample covariance for all group.

2. E-Step. A conditional expectation of the group membership for each observation can be evaluated by calculating

$$\tau_{ik}^{(t)} = \frac{\pi_k^{(t-1)} \phi\left(x_i | \mu_k^{(t-1)}, \Sigma_k^{(t-1)}\right)}{\sum_{k=1}^{G} \pi_k^{(t-1)} \phi\left(x_i | \mu_k^{(t-1)}, \Sigma_k^{(t-1)}\right)} \tag{5}$$

3. M-Step Compute sufficient statistics by

$$T_{k1}^{(t)} = \sum_{i=1}^{n} \tau_{ik}^{(t)}, \quad T_{k2}^{(t)} = \sum_{i=1}^{n} \tau_{ik}^{(t)} x_i \text{ , and}$$
$$T_{k3}^{(t)} = \sum_{i=1}^{n} \tau_{ik}^{(t)} x_i\, x_i^T \tag{6}$$

4. Get the parameter estimate $\widehat{\Theta}^{(0)}$ by

$$\hat{\pi}_k^{(t)} = \frac{T_{k1}^{(t)}}{n}, \mu_k^{(t)} = \frac{T_{k2}^{(t)}}{T_{k1}^{(t)}} \text{ and}$$

$$\hat{\Sigma}_k^{(t)} = \left\{T_{k3}^{(t)} - {T_{k1}^{(t)}}^{-1} T_{k2}^{(t)} {T_{k2}^{(t)}}^{T}\right\} / T_{k1}^{(t)} \quad (7)$$

5. Go back to the E-Step until the following convergence criteria are met:

$$\begin{cases} \hat{\pi}_k^{(t)} - \hat{\pi}_k^{(t-1)} \leq Treshold\ for\ k = 1,..,G \\ \hat{\mu}_k^{(t)} - \hat{\mu}_k^{(t-1)} \leq Treshold\ , k = 1,..,G \\ \hat{\Sigma}_{ij}^{(t)} - \hat{\Sigma}_{ij}^{(t-1)} \leq \text{treshold for } k = 1, \ldots G \text{ and any ij combination} \end{cases}$$

2.4. Determine the Number of Clusters

The model-based clustering framework provides a way to deal with several problems, such as the number of clusters, initial values of the parameters, and distributions of the component densities (e.g., Gaussian). The number of clusters and the distribution of the component densities can be considered as producing different statistical models for the data. The final model can be determined by the Bayesian information criterion (BIC). The model with the highest BIC value is chosen as the best model ([11];[8]). BIC is defined as

$$BIC \equiv 2loglik_M(y, \theta_k^*) - (\#parameters)_M \log(n) \quad (8)$$

where $2loglik_M(y, \theta_k^*)$ is the maximized loglikelihood for model and data, $(\#parameters)_M$ is the number of independent parameters to be estimated in the model M, and n is the number of observations in the data.

2.5. Multivariate model based spatial clustering with spatial constraint

Several methods exist for taking spatial information into account in a clustering process. First, modify existing clustering algorithms by specifying which objects are neighbors and allowing an object to be assigned to a class if and only if this class already contains a geographical neighbor (Franz et al, 2002). This approach has the drawback of producing classes which are necessarily geographically connected. Secondly, use spatial autocorrelation statistics as data input into clustering algorithms. For instance, Luca (2005) [7] used the Getis Ord (G). Statistic as data input in K-means clustering. In this paper on model based clustering we use the local G statistic which is defined as:

$$G_i(d) = \frac{\sum_{i=1}^{n} w_{ij} x_j}{\sum_{i=1}^{n} x_j} \quad (9)$$

Where x_j is the measured attribute of interest at location j and w_{ij} a weight indexing the location of i relative to j (equal to 1 if locations i and j are within an a priori defined distance from each other and 0 if not). A group of features with high values indicates a cluster with high attribute values, or a "*hot spot*." On the other hand, a group of features with low values indicates a "cold spot." Finally, a value near 0 indicates that there is no concentration of either high or low values surrounding the target feature. This will occur when the surrounding values are all near the mean or when the target feature is surrounded by a mix of high and low values.

G_i is asymptotically normally distributed as the number of neighbours of i increases. For not very skewed distributions of X, a number of 8 neighbours or more is enough to ensure a sufficient approximation. Therefore, inference can be drawn on the basis of standardized scores computed from the following moments:

$$E(Gi) = \frac{w_i}{n} \quad (10)$$

$$Var(Gi) = \frac{w_i(n - w_i)s^2}{n^2(n-1)\bar{x}^2} \quad (11)$$

Where $w_i = \sum_j w_{ij}, \bar{x} = \sum_i x_i / n$ and $s^2 = \sum_i (xi - \bar{x}^2)^2 / n$. It can be shown that the standardized local Getis Ord Statistics is given by

$$Z(Gi) = \frac{\sum_{j=1}^{n} w_{ij} x_j - \bar{x} w_i}{\sqrt{\frac{s^2}{n-1}\left(n\left(\sum_{j=1}^{n} w_{ij}^2\right) - w_i^2\right)}} \quad (12)$$

The interpretation z(Gi) is straightforward: positive significant value indicate clusters of high value around the i-th location, while negative significant value indicates clusters of low value around the i-th location ([7];[12]).

We proposes the following model based clustering procedure to detect patterns of the spread of DD based on the standardized Local Getis Ord Statistic. Create a spatial weights matrix **W.** In this study we use a spatial weights matrix based on distance. The maximum distance for locations to be neighbors is 1840.59 m. So w_{ij} = 1 if the distance of a give site to other sites is less than or equal to m and w_{ij} = 0 if it is more 1840.59 *m*. The resulting spatial matrix is equivalent to a queen contiguity matrix

$$w_{ij} = \begin{cases} 1 if\ d \leq 1840.59m \\ 0 if\ d > 1840.59m \end{cases}$$

a. For each variable calculate the Standardized Local Getis Ord Statistics (Equation 9)
b. Perform spatial clustering using the Standardized Local Ord Statistics by means of Model-Based Clustering based on the assumption that the data is from

Multivariate Gaussian Distribution with density function (1)

c. Choose the model with largest BIC (Equation 8)

3. Result and Discussion

3.1. Data description

Dengue disease data in Bogor, Indonesia (2009) is used to apply the model based clustering method with spatial constraint. There are three variables considered: incidence rate, larvae free rate and population density. The distribution of those variable are presented in Figure 3.

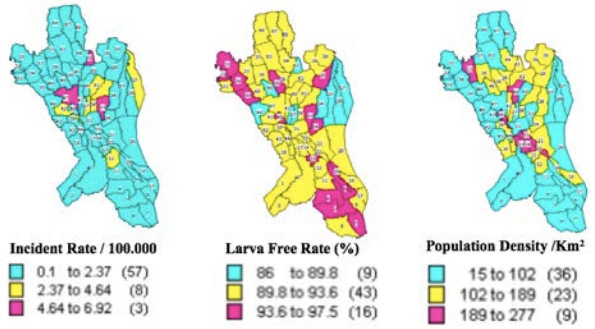


Fig. 3 Spatial distribution incident rate, larva free rate, and population density

Figure 3 shows the incidence rate has negative correlation with larvae free rate and positive correlation with population density. The regions with low incidence rate have high larvae free rate and small population density.

3.2. Global Getis Ord G Statistics

Global Getis Ord Statistics is used to identify the global spatial correlation. Table 1 presents the Global Getis Ord Statistics for all variables.

Table 2 Global Getis Ord G Statistics

Statistics	DD Incidence Rate (Case/100.000)	Larva Free Rate (%)	Population Density (people/km^2)
Getis-Ord General (G)	0.281015	0.199928	0.280781
Expected (E(G))	0.200615	0.200615	0.200615
Different G-E(G)	0.080400	-0.000687	0.080167
SE(G)	0.026737	0.001087	0.021541
Normality Significant (G)	3.007110	-0.632137	3.721542
p-value	0.010000	Not-significant	0.001000

Table 2 indicates that DD and Population Density have significant G statistics with Z (G)> 2.00. For larvae-free rate the statistic is not significant. This results suggest that we need to include the spatial dependence in the clustering process.

3.3. Local Getis Ord Gi Statistics

The next step, calculate the Local Getis Ord G Statistics which is used as an input in model based clustering. The Local Getis Ord G Statistics is presented in Figure 4.

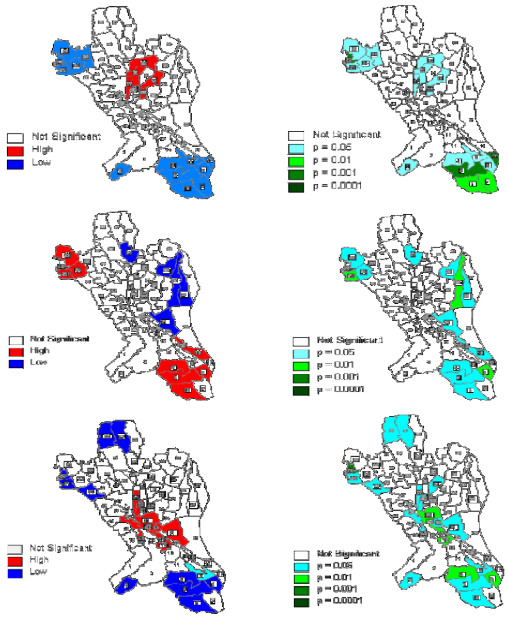


Fig 4 Visualization of Local Getis Ord Gi Statistics

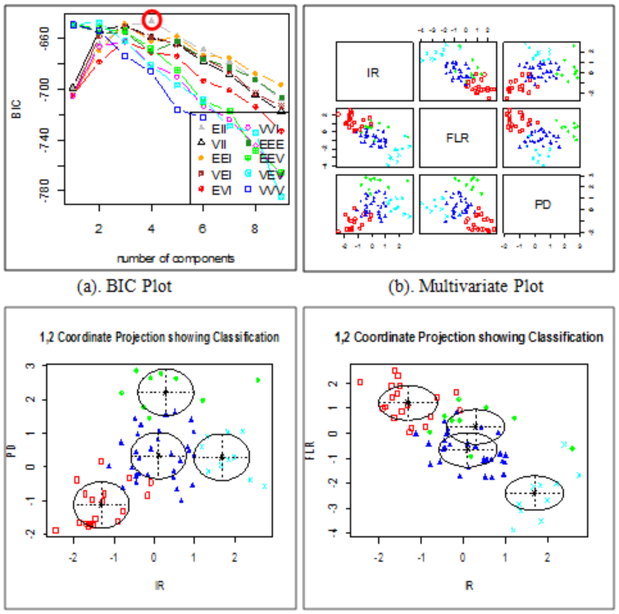
Using local Getis Ord Gi Statistics, we can identify the univariate spatial cluster for each dengue disease variable which can be categorized as high (hot spot) and low cluster (cold spot). Based on incident rate of DD, the location including in hotspot area are Bantarjati Village (23), Sempur (36), Ciwaringin (41), Kebon Pedes (60) and Kedung Rhino (62). While the location is the Village Margajaya including cold spot (51), cockscomb Beaver (52), Situ Gede (53), Bubulak (54), Semplak (55) and the village is located at the southern end of Bogor city

Based on Larva Free Rate, the location including in HotSpot Area are Genteng Village (4), Kertamaya (5), Rancamaya (6), Harjasari (8), Sidang Raya (18), and Katulampa (19).

We can see that location in cold spot area based on incident rate indicator is hot spot based on larva free rate

3.3. Model Based Clustering

The Local Spatial Autocorrelation Getis Ord Gi Statistics provides partial information about the pattern of spread of dengue disease and its covariates in the city of Bogor, but not in simultaneously. For that purpose we apply model-based clustering with spatial constrain. We use the R-package Mclust. The optimal number of clusters is determined by means of BIC (Bayesian Information Criterion). See Figure 5



(a). BIC Plot

(b). Multivariate Plot

(c) Classification

Fig. 5 Visualization BIC Model Based Clustering and Classification

Figure 5a an Table 3 show that the model with the largest BIC is model EII, which consists of four clusters with spherical distribution, and equal shape and volume.

Table 3 The optimal Number of Cluster Based On BIC

```
Classification Table:
 1  2  3  4
30 19  9 10

Best BIC Values:
    EII,4     VEV,2     EEI,3
-646.5185 -647.6214 -648.7231
```

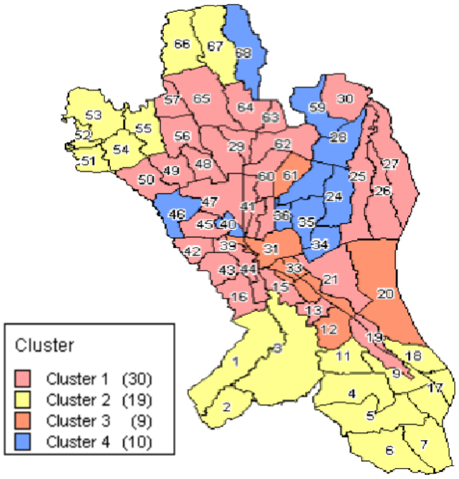


Fig. 6 Spatial clustering of dengue disease

The above map shows the spatial distribution of dengue disease under three an indicator Incident Rate, larvae Free Rate and Population Density. This grouping of information in line with Local Getis-Ord Statistics with the characteristics of each cluster are presented in Table 4.

Table 4 Characteristics each cluster

Cluster	Incident Rate		Free Larva Rate		Population Density	
	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
Cluster 1	2.053	1.379	91.4300	2.3669	112	56
Cluster 2	.711	.590	93.5605	1.8349	65	52
Cluster 3	2.042	1.224	91.7889	1.0671	175	86
Cluster 4	2.399	1.321	91.6700	2.7825	133	58

Cluster 2 is the group village with the possibility of dengue endemic areas is very low. While the cluster is cluster 4 with the possibility of a dengue endemic areas is very high because Incident Rate detected high, with low larva-free rate and high population density.

Based on the results of this analysis, Bogor City Health Department should further enhance the implementation of the Surveillance Epidemiology of cases of Dengue Hemorrhagic Fever (DD) in the villages included in cluster 4 because these villages are likely to become endemic area of DD disease, facilitating the subsequent availability of facilities and countermeasures DD, such as insecticides, Larvasida, reagents (limited), and vector control equipment.

Table 5. List of Village in Bogor City

Cluster 1		Cluster 2		Cluster 3		Cluster 4	
No.	Village	No.	Village	No.	Village	No.	Village
10	Pakuan	1	Mulyaharja	12	Lawang Gintung	23	Bantarjati
13	Batu Tulis	2	Pamoyanan	14	Bondongan	24	Tegal Gundil
15	Empang	3	Ranggamekar	20	Katulampa	28	Cibuluh
16	Cikaret	4	Genteng	22	Sukasari	34	Tegal Lega
19	Tajur	5	Kertamaya	31	Paledang	35	Babakan
21	Baranangsiang	6	Rancamaya	32	Gudang	36	Sempur
25	Tanah Baru	7	Bojongkerta	33	Babakan Pasar	40	Kebon Kalapa
26	Cimahpar	8	Harjasari	39	Panaragan	46	Loji
27	Ciluar	9	Muarasari	61	Tanah Sareal	59	Kedung Jaya
29	Kedunghalang	11	Cipaku			68	Kencana
30	Ciparigi	17	Sindang Sari				
37	Pabaton	18	Sindang Rasa				
38	Cibogor	51	Margajaya				
41	Ciwaringin	52	Balungbang Jaya				
42	Pasir Mulya	53	Situ Gede				
43	Pasir Kuda	54	Bubulak				
44	Pasir Jaya	55	Semplak				
45	Gunung Batu	66	Kayu Manis				
47	Menteng	67	Mekarwangi				
48	Cilendek Timur						
49	Cilendek Barat						
50	Sindang Barang						
56	Curug Mekar						
57	Curug						
58	Kedungwaringin						
60	Kebon Pedes						
62	Kedung Badak						
63	Sukaresmi						
64	Sukadamai						
65	Cibadak						
Total :	30		19		9		10

5. Conclusion

Standardized Getis Ord statistics is usually used to identify spatial clustering considering the spatial autocorrelation. However, this method is developed for univariate data, while in most cases, data is multivariate. Here we proposed, model-based clustering with spatial constraint. We combine the model-based clustering and Standardized Getis Ord Statistics to identify the spatial clustering for multivariate data. The method uses a two-stage procedure.

Standardized Getis Ord statistics is used to create the clustering variables considering spatial autocorrelation and model-based clustering is used to identify spatial cluster given multivariate data.

Model-based clustering with spatial constraint provides a more informative clustering result. We applied this method to identify the spatial clustering of dengue incidence rate in Bogor, Indonesia. The clustering outcomes, consider the incidence rate, larvae free rate, and population density. Therefore, it can be used to identify high-risk regions and the significant risk factors (i.e., larvae free rate and population density)

Acknowledgments

This paper is funded by the RFU Unpad contract: 1732 d/UN6.RKT/LT/2018. The authors thank Rector Universitas Padjadjaran and to the anonymous referee whose valuable checking has improved this paper

Reference

[1] Fathi, K. Soedjajadi and U.W. Chatarina "Peran faktor lingkungan dan perilaku terhadap penularan demam berdarah dengue di Kota Mataram" Jurnal Kesehatan Lingkungan, 2, 1, 2005, pp. 1-11

[2] J. Soepardi, Profil Kesehatan Indonesia 2010, Indonesia: Kemenkes RI, 2010

[3] S Brooker "Spatial clustering of malaria and associated risk factors during an epidemic in a highland area of western Kenya", Tropical Medicine and International Health, 9, 7 , 2004, pp: 757-766

[4] L. Fang, L. Yan, L. Song, S.J.D. Vlas et. Al, "Spatial analysis of hemorrhagic fever with renal syndrome in China", BMC Infectious Diseases, 6,77, 2006, pp. 1-10

[5] T. Tango, Statistical Methods for Disease Clustering, Springer New York Dordrecht Heidelberg London,: Springer, 2010

[6] H. S. Steven, "Model Based Clustering" , A Research Paper. Presented to the University of Waterloo in fulfillment of the research paper requirement for the degree of Master of Mathematics in Statistical Computing, 2005 .

[7] S Luca, "Clustering multivariate spatial data based on local measures of spatial autocorrelation, An application to the labour market of Umbria". Dipartimento di Economia, Finanza e Statistica Universit`a degli Studi di Perugia, Italy, 2005

[8] C. Fraley and A. E. Raftery, "How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis", The Computer Journal, 41, 8, 1998, pp 578-588

[9] C. Fraley and E. Adrian, Raftery, "MCLUST Version 3 for R: Normal Mixture Modeling and Model-Based Clustering" Technical Report No. 504. Department of Statistics University of Washington, 2006

[10] A. P. Dempster, N. M. Laird and D. B. Rubin "Maximum Likelihood from Incomplete Data via the EM Algorithm", Journal of the Royal Statistical Society. Series B (Methodological), Vol. 39, No. 1. (1977), pp. 1-38.

[11] G. Goujun, M. Chaoqun, W. Jianhong, Data Clustering Theory, Algorithm, and Application, Vriginia: SIAM, 2007.

[12] L. A. Waller and C. A. Gotway, Applied Spatial Statistics For Public Health Data, New Jersey: John Wiley & Sons, 2004