Spatial Clustered Regression Analysis of 2017 Getis Score
Indonesian Malaria Prevalence Data

Suci Sukmawati, Anik Djuraidah*, and Aji Hamim Wigena
Department of Statistics, IPB University, Bogor, 16680, Indonesia

*E-mail : anikdjuraidah@gmail.com

Abstract. Malaria is a disease that is still endemic to Indonesia. In the national scale, morbidity rate related to malaria continued to decline from 2009 to 2017. Eastern areas of Indonesia (namely Papua, West Papua, and East Nusa Tenggara) still have high endemic levels of malaria, while western areas (Islands Sumatera and Java) mostly have low endemic levels or even free of malaria. This indicates that malaria has a clustering pattern. This research used the Annual Parasite Incidence (API) data as the dependent variable and Slide Positivity Rate (SPR), Annual Blood Examination Rate (ABER), number of positive cases on children aged 1-14 years, and the number of Plasmodium falciparum cases and Plasmodium vivax cases data as the independent variables. Since they contain spatial autocorrelation, Getis Ord G_i were employed in the clustered regression analysis. Finite Mixture Model was employed as the clustering method. This study resulted in the forming of 3 clusters. Cities/regencies belonging to Cluster-1 were neither hotspots nor coldspots. Elements of Cluster-2 are SPR hotspots. Elements of Cluster-3 are hotspot areas of API, ABER, number of Plasmodium falciparum cases, number of Plasmodium vivax cases, and number of cases on children aged 1-14 years. Cluster-1 consists of 360 cities/regencies located in Islands Kalimantan, Sulawesi, Sumatera, and Java, and North Maluku Province. Cluster-2 consists of 88 cities/regencies, mostly located in provinces South Sumatera, Bengkulu, Lampung, Banten, DKI Jakarta, and West Java. Cluster-3 consists of 66 cities/regencies located in the Papua Island and East Nusa Tenggara Province

1. Introduction
Malaria is an infectious disease caused by Plasmodium and transmitted by the bite of a female Anopheles mosquito. The transmission process by using the Anopheles mosquito as a vector causes adjacent areas to have a great chance of being infected. Indonesia is a malaria endemic area with the sixth largest number of malaria cases in the world in 2018 [11].

Malaria morbidity in an area is determined by Annual Parasite Incidence (API), which is the number of positive cases of malaria per 1000 population within one year [9]. The endemicity status of a city/regency is determined based on the number of API. According to PUSDATIN [9], there are six levels of malaria endemicity status in Indonesia, namely elimination or free of malaria (API = 0), low cumulative incidence (LIC, 0 <API <1), medium cumulative incidence (MCI, 1 ≤ API <5), high cumulative incidence I (HCI I, 5 ≤ API <50), HCI II (50 ≤ API <100), and HCI III (API ≥ 100).

Indonesia succeeded in reducing the number of malaria cases during 2009 to 2017, from 1.8 per 1000 population to 0.99 per 1000 population [6]. Malaria endemicity data in 2017 shows that cities/regencies with high endemic status come from eastern areas of Indonesia, namely Papua, West Papua and East...
Nusa Tenggara Provinces. Cities/regencies of western areas of Indonesia, namely islands of Java and Sumatra are dominated by low endemic status and even elimination. The distribution of malaria tends to clustered. According to Anselin [2], the clustered pattern indicates positive spatial autocorrelation. One method that can be used to group malaria data is the Finite Mixture Model (FMM) method. The FMM method can be used to perform regression analysis and cluster analysis simultaneously. Malaria data have spatial autocorrelation, so Getis Ord G\textsuperscript{*} is used in clustered regression analysis to determine which areas are hotspots or coldspots.

1. Objectives
This study aims to group cities/regencies in Indonesia based on malaria prevalence data in 2017 using Finite Mixture Model (FMM) method on the Getis-Ord G\textsuperscript{*} local spatial autocorrelation value and determine the factors that influence malaria morbidity in Indonesia in 2017 for each group.

2. Data
This research used malaria Passive Case Detection (PCD) data from health centers in all cities/regencies in 2017. PCD results data were obtained from the Sub-Directorate of Malaria, Ministry of Health, Republic of Indonesia. The variables used in this study are presented in Table 1.

Table 1. List of Variables

| Symbol | Variables                                      |
|--------|------------------------------------------------|
| Y      | Annual Parasite Incidence (API)                |
| X1     | Slide Positivity Rate (SPR)                    |
| X2     | Annual Blood Examination (ABER)                |
| X3     | Number of \textit{Plasmodium falciparum} cases |
| X4     | Number of \textit{Plasmodium vivax} cases      |
| X5     | Number of cases on children aged 1-14 years    |

3. Methods
Data analysis for this research was performed using Microsoft Excel 2013 and R 3.6.2 software. The data analysis steps carried out in this study are as follows:

1. Data exploratory to find an overview of the data.
2. Create a matrix \( D \) which the elements are the distances between cities/regencies. The distance is determined based on Great Circle Distance (GCD). The GCD calculation is written as follows:
   \[
   d_{ij} = R \cos^{-1}(\cos|x_i - x_j| \cos|y_i| \cos|y_j| + \sin|y_i| \sin|y_j|)
   \]
   where \( R \) is the radius of the earth, which is 6371 km.
3. Compute the optimum distance (\( d_{opt} \)) based on the standard global Getis score of variables \( Y \), \( X1 \), \( X2 \), \( X3 \), \( X4 \), dan \( X5 \).
   a. Set minimum distance and maximum distance.
   b. Compute the standard global Getis score (\( Z(G) \)) from minimum distance to maximum distance.
   c. Choose the optimum distance based on the largest \( Z(G) \).
4. Compute the standard local Getis score. Getis-Ord G\textsuperscript{*} statistics according to Getis and Ord [5] are as follows:
   a. Create spatial weighted matrix \( W(d) \) with the following conditions:
      \[
      w(d)_{ij} = \begin{cases} 
      1, & d_{ij} < d_{opt} \\
      0, & \text{else} 
      \end{cases}
      \]
      where \( w(d)_{ij} \) is the element of \( W(d) \), \( d_{ij} \) is the element of \( D \), and \( d_{opt} \) is the optimum distance obtained in step 3.
   b. Compute local Getis score for each location
\[ G_i^*(d) = \sum_{j=1}^{n} w_{ij}(d)x_j / \sum_{j=1}^{n} x_j; \quad i \neq j \]

c. Compute expected value of local Getis Score.
\[ E(G_i^*(d)) = \frac{W_i^*}{(n-1)} \]

where, \( W_i^* = \sum_{j=1}^{n} w_{ij}(d) \).

d. Compute variance of local Getis score.
\[ \text{Var}(G_i^*(d)) = \frac{W_i^*(n-1 - W_i^*)y_{i2}^*}{(n-1)^2(n-2)y_{i1}^*} \]

Where, \( y_{i1}^* = \frac{\sum_{j=1}^{n} x_j}{(n-1)} \) and \( y_{i2}^* = \frac{\sum_{j=1}^{n} x_j^2}{(n-1)} - (y_{i1}^*)^2 \).

e. Compute the standard local Getis score
\[ Z(G_i^*) = \frac{G_i^*(d) - EG_i^*(d)}{\sqrt{\text{Var}(G_i^*(d))}} \]

A location categorized as a hotspot if \( Z(G_i^*) > Z_\alpha \). Hotspot is a condition when high values of an observation are gathered. Otherwise, if \( Z(G_i^*) < Z_\alpha \) then a location is categorized as a coldspot, a condition that low values of an observation gathered.

f. Repeating steps a to e for each variable with different optimum distances.

5. Determine the best cluster regression based on \( Z(G_i^*) \) obtained in steps 4. Clustered regression analysis was carried out using Finite Mixture Model (FMM). According to Leisch [7], the equation for FMM with \( K \) clusters is written as follows:
\[ h(y|x, \psi) = \sum_{j=1}^{K} \pi_j f(y|x, b_j, \sigma_j^2) \]

where \( \pi_j \geq 0; \sum_{j=1}^{K} \pi_j = 1 \).

The parameters of the models is estimated by Expectation-Maximization (EM) algorithm [3].

a. Expectation step:
   i. Determine the initial value of the number of cluster \( K \), where \( K > 1 \).
   ii. Initialize randomly \( \pi_j, b_j, \sigma_j \) value.
   iii. Estimate the posterior probability \( p_{ij} \) based on the initialization of \( \pi_j, b_j, \sigma_j \). The posterior probability for each observation are listed as follows:
\[ \hat{p}_{ik} = P(k|x_i, y_i, \hat{\psi}) = \frac{\pi_j f(y|x, b_j, \sigma_j^2)}{\sum_{j=1}^{K} \pi_j f(y|x, b_j, \sigma_j^2)} \]

[8].

iv. Compute the log-likelihood value.

b. Maximization step:
   i. Estimate new \( \pi_j, b_j, \sigma_j \) based on estimate of \( p_{ij} \).
\[ \hat{\pi}_j = \frac{1}{n} \sum_{i=1}^{n} p_{ij} \]
\[ \hat{b}_j = (X'WX)^{-1} (XW_jy) \]
\[ \hat{\sigma}_j = \frac{\sum_{i=1}^{n} p_{ij} (y_i - X_i b_j)^2}{\sum_{i=1}^{n} p_{ij}} \]
where 

\[
W_j = \begin{pmatrix}
p_{1j} & \cdots & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & p_{nj}
\end{pmatrix}
\]

ii. Compute the new log-likelihood value.

iii. Repeating steps a point ii to steps b point iii until it converges.

c. Compute Akaike’s Information Criterion (AIC) = \(-2\ln(L) + 2g\) [1]; where L is log-likelihood function and g is the number of parameters.

d. Repeating steps a to c with different K and optimum distance.

g. Chose the best clustered regression based on the smallest AIC value.

6. Exploration of clustering result using a map of standar local Getis score.

4. Result and Discussion

4.1. Data Exploration

In general, Indonesia has a low morbidity rate of malaria. Figure 1 shows that most cities/regencies in Islands Sumatera, Java, Sulawesi and Kalimantan are with low endemic status and some cities/regencies are free of malaria. The distribution of API numbers in 2017 shows that the more to the eastern areas of Indonesia, the greater the API numbers. Papua Island is the eastern most area of Indonesia with cities/regencies that have high endemic status and even some cities/regencies have API numbers above 100.

Figure 1. Indonesian malaria endemicity map in 2017

4.2. Global Getis Score

The standard global Getis score for each variable is calculated from the minimum distance which is 293 km to the maximum distance 2673 km. The minimum distance is determined based on the distance between cities/regencies with each city/regency location point having at least 1 neighbor and maximum distance is determined from the farthest distance between cities/regencies. The optimum distance that chosen is the distance when it reaches the largest standard global Getis score and exceeds the threshold used, namely \(Z_{0.025} = 1.96\) which is rounded to 2.
Figure 2. Global Getis score every variables based on distance

The peak point of the $Z(G)$ value in Figure 2 for each variable exceeds the threshold, which means there are global spatial autocorrelation at that distance. The API ($Y$) variable get the largest $Z(G)$ value when the distance used is 493 km. The optimum distance between cities/regencies obtained from the standard global Getis score of SPR ($X_1$), ABER ($X_2$) variables, the number of Plasmodium falciparum ($X_3$) cases, the number of Plasmodium vivax cases ($X_4$), and the number of cases of children aged 1-14 years ($X_5$) is 383 km, 573 km, 533 km, 553 km, and 553 km.

4.3. Local Getis Score

The map of the standard local Getis score is in Figure 3. Papua Island in red means all variables have a standard local Getis score exceeding number 2, so that it is categorized as a hotspot. The mapping of API ($Y$) standard local Getis score in Figure 3 (b) shows that most cities/regencies in Indonesia are green and orange, which means they are neither hotspot nor coldspot ($-2 \leq Z(G_i^*) \leq 2$), but some cities/regencies in Java Island are blue which shows that coldspot areas with $Z(G_i^*)$ less than -2. While the standard local Getis score of SPR variable ($X_1$) in Figure 3 (a), Java Island is a hotspot and several cities/regencies in Islands Sumatera, Kalimantan and Sulawesi are coldspots. Based on the standard local Getis score number of cases of Plasmodium falciparum ($X_3$) and Plasmodium vivax ($X_4$) in Figure 3 (c) and Figure 3 (d), most cities/regencies in Indonesia are neither hotspot nor coldspot, but most of Java and several cities/regencies on the Sumatera Island including to coldspots areas. The mapping of ABER ($X_2$) and number of cases on children aged 1-14 years ($X_5$) standard local Getis score in Figure 3 (e) and Figure 3 (f) shows that most of Java and Sumatra are coldspots.
Gambar 3. Mapping of local Getis score of (a) SPR (383 km), (b) API (493 km), (c) *P. falciparum* (533 km), (d) *P. vivax* (553 km), (e) number of cases aged 1-14 years (553 km), (f) ABER (573 km)

4.4. Clustered Regression Analysis

The results of clustered regression analysis with the number of groups 2 and 3 will be compared based on Akaike's Information Criterion (AIC). The best cluster regression equation is selected based on the smallest AIC value. Based on Table 2, the smallest AIC value is obtained when the optimum distance is 383 km so that the best cluster regression equation with 3 groups is chosen.

**Table 2.** Akaike’s Information Criterion (AIC) value

| Distance (km) | k = 2   | k = 3   |
|--------------|---------|---------|
| 383          | -779.53 | **-1429.60** |
| 493          | -362.61 | -1327.37 |
| 533          | -765.75 | -627.96 |
| 553          | -379.95 | -769.58 |
| 573          | -339.59 | -903.54 |
| Combination  | -374.78 | -570.79 |

The regression coefficient estimator of each group with the optimum distance of 383 km is presented in Table 3. The independent variables that significantly affect the API (*Y*) in Cluster-1 are SPR (*X1*), ABER (*X2*), number of *Plasmodium falciparum* cases (*X3*), and number of cases on children aged 1-14 years (*X5*). All independent variables have a significant effect on the API (*Y*) in Cluster-2. Cluster-3 has 4 independent variables that significantly affect the API (*Y*), namely SPR (*X1*), ABER (*X2*), number of *Plasmodium vivax* (*X4*) cases, and number of cases on children aged 1-14 years (*X5*).

**Table 3.** Parameter estimation

| Variable       | Cluster-1  | Cluster-2  | Cluster-3  |
|----------------|------------|------------|------------|
|                | Coefficient| p-value    | Coefficient| p-value    | Coefficient| p-value    |
| Intercept      | 0.0661     | 0.0000     | -0.2033    | 0.0000     | 0.0191     | 0.9378     |
| Z(G^*_i)X1     | 0.0028     | 0.0000     | 0.0150     | 0.0000     | 0.6427     | 0.0000     |
| Z(G^*_i)X2     | -0.0265    | 0.0000     | 0.0531     | 0.0000     | 0.1234     | 0.0027     |
| Z(G^*_i)X3     | **0.6529** | 0.0000     | **1.8280** | 0.0000     | 0.1477     | 0.5885     |
| Z(G^*_i)X4     | -0.0098    | 0.1871     | 0.6723     | 0.0000     | -0.4361    | 0.0428     |
| Z(G^*_i)X5     | 0.3151     | 0.0000     | -1.5519    | 0.0000     | **0.9002** | 0.0009     |

The independent variable that has the greatest influence on the API (*Y*) for each cluster is indicated by the largest regression equation coefficient. The greatest influence on Cluster-1 and Cluster-2 was derived from the number of *Plasmodium falciparum* cases (*X3*). The independent variable that has the greatest effect on Cluster-3 is the number of cases on children aged 1-14 years (*X5*).
The map of the distribution of the standard local Getis scores in **Figure 4** is formed by 3 colors, namely blue for Cluster-1, green for Cluster-2, and red for Cluster-3. Cluster-1 consists of 360 cities/regencies, generally on the islands of Kalimantan, Sulawesi, Sumatera, Java and North Maluku Province. The number of cities/regencies that are included in the Cluster-2 are 88, generally from the provinces of South Sumatra, Lampung, Bengkulu, Bangka Belitung, DKI Jakarta, Banten, and West Java. Cluster-3 consists of 66 cities/regencies, which are generally located in Papua Island and NTT Province. Cities/regencies in Papua Island are generally members of Cluster-3, however, there are cities/regencies that are included in Cluster-1 and Cluster-2, namely Mamberamo Tengah, Paniai, Lanny Jaya, Tolikara, and Nduga. The cities/regencies that are members of Cluster-3 are surrounded by cities/regencies of Cluster-1 and Cluster-2 members, including Pesawaran, Kota Pangkal Pinang, and Bangka Tengah.

The mapping of the standard local Getis score for each variable in **Figure 4** shows that there are coldspots cities/regencies, but the average value for each cluster in **Table 4** shows there is no coldspot area. Cities/regencies in Cluster-1 are characterized as neither hotspot nor coldspot. The Cities/regencies in Cluster-2 are included in the hotspot area of the SPR variable. Meanwhile, Cluster-3 is characterized by cities/regencies that fall into the hotspot area of the variables API (Y), ABER (X1), the number of *Plasmodium falciparum* cases (X3), the number of *Plasmodium vivax* cases (X4), and the number of cases on children aged 1-14 year (X5).

**Table 4.** Local spatial autocorrelation for each cluster

| Cluster | Y | X1 | X2 | X3 | X4 | X5 |
|---------|---|----|----|----|----|----|
| 1       |   |    |    |    |    |    |
| 2       |   | +  |    |    |    |    |
| 3       | + |    | +  | +  | +  | +  |

Cities/regencies that are included in Cluster-1 are generally neither hotspots nor coldspots, but some are hotspot or coldspot. There are cities/regencies in the provinces of Central Java, DIY, and East Java which are ABER coldspots (X2), the number of *Plasmodium falciparum* cases (X3), the number of *Plasmodium vivax* cases (X4), the number of cases on children aged 1-14 years (X5) and the SPR (X2) hotspot. Other cities/regencies are in Aceh, West Kalimantan, Central Kalimantan, East Kalimantan and Central Sulawesi Provinces which are SPR coldspot (X1) and cities/regencies in North Sumatera Province which are ABER (X2) coldspot.

In general, cities/regencies in Cluster-2 are not included in hotspot or coldspot. Some cities/regencies in provinces of South Sumatra, Bengkulu and Bangka Belitung are in SPR coldspot areas. Other cities/regencies are located in the Provinces of Lampung, Banten, DKI Jakarta and West Java, which are the ABER coldspot and SPR hotspots.

Regencies/cities in Papua Island, which are members of the Cluster-3, are colored red for all variables. This shows that the regencies/cities in Cluster-3 are hotspot for all variables. However, Mamberamo Tengah, Lanny Jaya, Tolikara, and Nduga have low endemic status in the mountainous region. According to Saputri [10], Anopheles is rarely found in areas with an altitude more than 2000-2500
meters, so that the cities/regencies were included in Cluster-1 or Cluster-2 even though the surrounding cities/regencies were included in Cluster-3.

Each cluster has a different characteristic. The character of a region can determine the malaria disease management program policy. City/regency governments in the same group can collaborate in the preparation of malaria management programs. Management of malaria disease for regencies/cities in Cluster-1 and Cluster-2 can focus on the variable with the largest regression equation coefficient, namely the number of Plasmodium falciparum cases ($X_3$) which is the most virulent species with the highest morbidity and mortality rates [4] Efforts to achieve malaria free must be made in the areas that are included in the Cluster-3, namely the provinces of Papua, West Papua and NTT. Efforts that be made can focus on preventing or checking malaria cases in children aged 1-14 years ($X_5$) because they have the largest regression estimates in Cluster-3.

5. Conclusion

There are 3 clusters formed from the local Getis score. The character of each group, namely, Cluster-1 is characterized by a local Getis score in the interval [-2,2] which is not a hotspot or coldspot area. Cluster-2 is characterized as a hotspot with the standard local Getis score over the SPR variable. Cluster-3 was characterized as a hotspot of the variable API, ABER, the number of Plasmodium falciparum cases, the number of Plasmodium vivax cases, and the number of cases on children aged 1-14 years.

References

[1] Arnold T W 2010 Uninformative parameters and model selection using Akaike’s information criterion Journal of Wildlife Management 74(6)1175-1178
[2] Baltagi B H 2001 A Companion to Theoritical Econometrics (Oxford: Blackwell Publishing Ltd)
[3] Butar-Butar V P 2019 Pemodelan clusterwise regression pada statistical downscaling untuk pendugaan curah hujan bulanan [thesis] (Bogor : IPB University)
[4] Dewi G A N Y S, Gustawan I W, Utama M G D L U, and Arhana B N P 2019 Karakteristik infeksi malaria pada anak di RSUD Dekai Papua April-Juni 2018 Medicina Journal 50(3)488-492
[5] Getis A and Ord J K 1992 The analysis of spatial association by use of distance statistics Geographical Analysis 24(3)190-206
[6] [KEMENKES] Kementerian Kesehatan Republik Indonesia 2017 Profil Kesehatan Indonesia Tahun 2017 (Jakarta : KEMENKES RI)
[7] Leisch F 2004 Flexmix: a general framework for finite mixture models and latent class regression in R Journal of Statistical Software 11(8)1-18
[8] Melnykov V and Maitra R 2009 Finite mixture models and model-based clustering Statistics Surveys 4(2010)80-116
[9] [PUSDATIN] Pusat Data dan Informasi Kementerian Kesehatan Republik Indonesia 2016 InfoDATIN Malaria (Jakarta : KEMENKES RI)
[10] Saputri T M 2012 Analisis pengaruh faktor lingkungan terhadap kerapatan larva nyamuk Anopheles sundaicus dan dampaknya terhadap kasus malaria (Kasus: Pantai Batu Kalang, Kabupaten Pesisir Selatan, Sumatera Barat) [undergraduate thesis] (Bogor : IPB University)
[11] [WHO] World Health Organization 2019 World Malaria Report 2019 (Paris: Publications of the World Health Organization)
[12] Yan X and Su X G 2009 Linear Regression Analysis: Theory and Computing (Singapore: World Scientific Publishing Co. Pte. Ltd)