Clustering Regions Based on Socio-Economic Factors Which Affected the Number of COVID-19 Cases in Java Island

Septian Rahardiantoro*, Wataru Sakamoto
Department of Human Ecology, Graduate School of Environmental and Life Science, Okayama University, Okayama, 700-8530, Japan

*E-mail : rahardiantoro.stk@gmail.com

Abstract. Around 60% of COVID-19 positive cases in Indonesia have occurred in Java Island. This study provides clustering adjacent regions (cities and regencies) in Java Island into some groups based on some socio-economic factors that are suspected to affect the COVID-19 infection rates (positive cases per 100,000 residents), which could be useful for decision making by government. The factors involved in this study are poverty percentage, Human Development Index (HDI), average of expenditure per month, and open unemployment rate. There are two steps in our data analysis: first, we determined the factors that affected the infection rate significantly by using lasso, and then we estimated region-specific effects of each significant factor by using generalized lasso. In the generalized lasso, two types of spatial structure were considered, namely, regions divided by province, and neighbourhood regions based on k-means clustering and Voronoi tessellation. The tuning parameter in both lasso and generalized lasso was selected by 5-folds cross-validation. Based on the first step, three variables were found to affect the infection rate significantly. Then in the second step, the three variables had spatially varying coefficients in the generalized lasso using regions divided by provinces. On the other hand, HDI provided spatially varying coefficient in the generalized lasso using region based on k-means clustering and Voronoi tessellation.

1. Introduction
COVID-19 is the novel corona virus that infects humans and has already spread out globally. Recently, the number of COVID-19 positive cases in Indonesia is greater than 300,000 cases (as of October 5th, 2020), since the first case was announced by the government at the beginning of February 2020. In Indonesia, more than 60% of COVID-19 cases have occurred in Java Island, where the capital city, Jakarta, is located.

Our study was inspired by Akil and Ahmad [1], which investigated the association between the increase in body mass index (BMI) and socio-economic factors. Moreover, there are many related works for COVID-19 cases in Indonesia, such as SIRV modeling [2], and modified SEIR modeling [3]. In a different point of view, Djalante et al [4] published a review and analysis of current responses to COVID-19 in Indonesia. Therefore, we consider revealing the relationship between the socio-economic factors that affect the number of COVID-19 positive cases by using a statistical modeling approach.

In the beginning of this study, we apply the lasso approach for identifying significant predictor variables. Then, in the second step we apply the generalized lasso method [5][6] using these significant predictor variables to fit the model with spatially varying regression coefficients. We are interested in using generalized lasso method because of its flexibility to incorporate neighborhood structure. We can...
deal with a two-dimensional graph representing neighborhood structure of regions in Java Island. All of this data analysis was handled by using \textit{genlasso} package of R software.

2. Data Description

Our target variable is the COVID-19 infection rates, the number of cases per 100,000 residents, in each regency in Java Island, Indonesia. The number of COVID-19 cases (as of September 29) was collected based on the monthly report of each province governments in Java Islands [7-12]. The predictor variables that we included in this study are poverty percentage, Human Development Index (HDI), average of expenditure per month, and open unemployment rate. These predictor variables are the recent aggregate calculation based on the statistics office of each province in Java Island [13-18].

Our data set contains 119 cities and regencies in 6 provinces of Java Islands: Jakarta Province, West Java (Jawa Barat) Province, Central Java (Jawa Tengah) Province, East Java (Jawa Timur) Province, Yogyakarta Province, and Banten Province.

![Figure 1. Distribution of log COVID-19 infection rates in Java Island](image)

Figure 1 shows the distribution pattern of log COVID-19 infection rates in each regency in Java Island. The high number of cases occurred in Jakarta and surrounding regencies, Semarang (the capital city of East Java), and also Surabaya (the capital city of East Java).

For more details, Table 1 provides the summary statistics of each variable of the data set. The highest number of COVID-19 positive rates, 954.5, was observed in Central Jakarta City. It seems that this variable has right tailed distribution. Furthermore, the summary statistics of predictor variables are also listed in this table.

| Statistics | COVID-19 Infection Rates per 100,000 residents | Poverty (%) | HDI       | Average Expenditure per Month (IDR) | Open Unemployment Rate (%) |
|------------|-----------------------------------------------|-------------|-----------|------------------------------------|----------------------------|
| Min        | 3.364                                         | 1.680       | 61.94     | 646386                             | 0.950                      |
| 1st Qu     | 24.932                                        | 6.625       | 68.67     | 863530                             | 3.565                      |
| Median     | 50.057                                        | 9.120       | 71.75     | 987853                             | 4.480                      |
| Mean       | 89.592                                        | 9.149       | 72.61     | 1142366                            | 5.273                      |
| 3rd Qu     | 84.808                                        | 11.380      | 75.58     | 1266295                            | 7.205                      |
| Max        | 954.521                                       | 20.710      | 86.65     | 2625288                            | 10.650                     |
| Stdev      | 127.289                                       | 4.036       | 5.38      | 400718                             | 2.283                      |

We use the natural logarithm of infection rates for each regency as the response variable of this study to make the distribution of this variable more symmetric. Figure 2 shows the normal Q-Q plot for the distribution of infection rate before and after natural logarithm transformation.
Figure 2. Normal Q-Q plot for the distribution of infection rate: (a) before and (b) after natural logarithm transformation

3. Data Analysis

In the data analysis process, we have two steps: (1) identifying predictor variables that affect the response variable significantly; (2) identifying spatial structure in the effect of these significant predictor variables. Therefore, for the first step, we use lasso for independent observations. Then, for the second step, we perform the generalized lasso for the data with graph structure.

3.1. Lasso

Let \( y \in \mathbb{R}^n \) be the vector of the response variable, \( X \in \mathbb{R}^{n \times p} \) be the matrix of all predictor variables, and \( \beta \in \mathbb{R}^p \) be the vector of parameters. The lasso regression problem [19] can be expressed generally as below

Lasso regression: 
\[
\hat{\beta}_{lasso} = \arg \min_{\beta} \left\{ \frac{1}{2} \| y - X\beta \|^2 + \lambda \| \beta \|_1 \right\}
\]

where \( \|a\|_1 = \sum |a_i| \) and \( \|a\|_2 = \sqrt{\sum |a_i|^2} \); with \( a \) denoting an arbitrary vector, and \( \lambda \in [0, \infty) \) is a tuning parameter.

In the application to our data set, we have \( n = 119, p = 4 \). We also include an intercept term in the regression model but it is not penalized. Before the analysis, we standardized each column of the matrix \( X \), so that \( X^T 1 = 0 \) and \( n^{-1} x_j^T x_j = 1 \); where \( x_j \) is the vector of \( j \)-th column of \( X \). Then we applied the lasso regression with \( \lambda \) selected by 5-fold cross validation minimizing the prediction error. We selected \( \lambda = 0.043 \), which resulted in \( R^2 = 0.5286 \), and RMSE = 0.7481. The coefficients estimated by lasso are described in Table 2.

| Predictors        | Estimated Coefficient |
|-------------------|-----------------------|
| (Intercept)       | 3.889                 |
| Poverty           | -                     |
| HDI               | 3.595                 |
| Expenditure       | 4.621                 |
| Open unemployment rate | -3.402               |

Note: The symbol - shows that the estimated coefficient was shrunk to 0.
Based on this result, we estimated that the variables poverty would not affect the response variable significantly. On the other hand, we estimated that the variables on HDI and expenditure had positive relationship to the response variable, while the variable on open unemployment rate had negative relationship. Therefore, in the next step we applied the generalized lasso that include HDI, expenditure, and open unemployment rate as predictor variables.

3.2. Generalized Lasso

Tibshirani and Taylor [5] proposed the generalized lasso problem and identified the solution path of this problem. The generalized lasso regression problem can be expressed as

$$\hat{\beta}_{\text{glasso}} = \arg\min_\beta \left\{ \frac{1}{2}\|y - X\beta\|^2_2 + \lambda \|D\beta\|_1 \right\}$$

where $D \in \mathbb{R}^{m \times p}$ is a specified penalty matrix that indicates the structural or geometric property.

We analysed the COVID-19 data set using generalized lasso to fit spatially varying coefficient models [20]. We divided 119 regencies into some spatial regions in the following two ways: i) 6 regions based on provinces; and ii) 10 regions based on $k$-means clustering, in which the neighbourhood structure was specified by Voronoi tessellation approach [21]. Then, we estimated region-specific coefficients for the intercept and the predictor variables selected by lasso in Section 3.1. The details are explained in the following sections.

3.2.1. Generalized Lasso with Regions Based on Provinces. Figure 3 shows the 6 provinces. To estimate region-specific coefficients for the intercept and predictor variables, the generalized lasso problem can be expressed as follows:

$$\min_{\beta} \left\{ \sum_{l} (y_{lk} - \beta_{l0} - X_{lk1}\beta_{l1} - X_{lk2}\beta_{l2} - X_{lk3}\beta_{l3})^2 + \lambda \|D\beta\|_1 \right\}$$

where $X_{lk1}, X_{lk2}$ and $X_{lk3}$ represents the predictor variables for HDI, expenditure, and open unemployment rate, respectively, $l = 1, 2, ... 6$ represents 6 provinces, $k = 1, 2, ...$ denotes the $k$th regency at each province. We considered intercept and coefficients for predictors at each province, so that $X \in \mathbb{R}^{119 \times 24}$ and $D \in \mathbb{R}^{24 \times 24}$. The matrix $D$ is constructed based on adjacencies between provinces, that is,

$$\|D\beta\|_1 = \sum_l (|\beta_{l0} - \beta_{m0}| + |\beta_{l1} - \beta_{m1}| + |\beta_{l2} - \beta_{m2}| + |\beta_{l3} - \beta_{m3}|)$$

where the summation is taken on the edges of the graph, which corresponds to the adjacent regions $l$ and $m$, so that there are 6 connections that are represented by each row in the matrix. Four $6 \times 6$ adjacency matrices of the same form are arranged diagonally to form a block diagonal matrix $D$, thus $D \in \mathbb{R}^{24 \times 24}$.
Figure 4. Plot of spatially varying coefficient estimates using generalized lasso with regions based on provinces, (a) intercept; (b) HDI; (c) average of monthly expenditure; (d) open unemployment rate
Then we applied the generalized lasso regression, and selected $\lambda = 0.0807$ by 5-folds cross validation minimizing the prediction error. We obtained $R^2 = 0.7772$, and RMSE = 0.5130. Figure 4 shows the spatially varying coefficient estimates of the intercept and the variables HDI, expenditure, and open unemployment rate.

The result provided the different estimates of intercept for each province, of which the highest intercept belonged to Jakarta Province, followed by East Java Provinces. The lowest intercept occurred in the West Java Province. On the other hand, 3 clusters were constructed on the estimated coefficient for HDI: the lowest coefficient occurred in Yogyakarta Province; the second cluster included Central Java and East Java Province; and the cluster with the highest coefficient included Banten Province, Jakarta Province, and West Java Province. Three clusters were constructed on the estimated coefficient for the monthly average of expenditure: Jakarta Province; Banten Province; and the cluster of West Java, Central Java, Yogyakarta, and East Java Province; in increasing order of the estimated coefficient. On the estimated coefficient for open unemployment rate, Central Java and Yogyakarta Province had negative relationship to response variable, while the rest provinces had positive relationship to the response variable.

3.2.2. Generalized Lasso with Regions Based on k-means Clustering and Voronoi Tessellation. The regions were constructed by using k-means method with $k = 10$ based on Euclidian distance between the centre of each regency in longitude and latitude coordinates. Each region was composed of regencies close to each other. Based on the centroids of the constructed regions, a Voronoi tessellation was used to determine the neighbourhood structure, as is shown in Figure 5.

![Figure 5](image_url)

**Figure 5.** The neighbourhood structure determined by Voronoi tessellation for the regions constructed by the $k$-means clustering ($k = 10$)

We applied the generalized lasso problem (4) again, where we have $l = 1, 2, \ldots, 10$ for 10 regions. To identify cluster effects in the coefficients for the intercept and predictor variables, we construct $X \in \mathbb{R}^{19 \times 40}$ and $D \in \mathbb{R}^{48 \times 40}$. Then we selected the $\lambda = 0.4661$ by 5-folds cross validation, and obtained $R^2 = 0.7205$ and RMSE = 0.5743. Figure 6 shows spatially varying coefficient estimates of the intercept and HDI variables based on this analysis. We obtained common estimates of coefficients for the other variables, average of monthly expenditure and open unemployment rate, over all regions.

The estimated intercept was separated into seven clusters, of which we obtained the highest estimate in the cluster around East Java Province and some part of Central Java Province, while the lowest estimate in the east region of West Java and its surroundings. For HDI, the estimated coefficient was separated into two clusters, of which the higher estimate occurred in Banten, Jakarta, West Java
Province, and some western part of Central Java Province. The result of the analysis with 10 regions was quite similar with the result of analysis by generalized lasso with regions based on provinces, in which the higher effect of HDI was suggested in Jakarta, Banten, West Java and its surroundings.

![Figure 6. Plot of spatially varying coefficient estimates using generalized lasso with neighborhood regions based on $k$-means clustering and Voronoi tessellation, (a) intercept; (b) HDI](image)

### 4. Conclusion
We found that the HDI variable, average of expenditure per month, and open unemployment rate significantly affected the response variable. Moreover, the HDI variable provided spatially varying coefficient estimate with the two generalized lasso approaches. This suggests that the HDI factor would have different effect on the infection rate of COVID-19 between regions on Java Island, that is, the western part of Java Island would have relatively higher HDI effect than the eastern part of Java Island. In the future, we can add some variables to enrich the analysis such as the number of PCR test conducted, number of COVID-19 specialized hospitals, index of population mobility, and so on.

### References

[1] Akil L and Ahmad A 2011 Effects of socioeconomic factors on obesity rates in four southern states and Colorado *Ethn Dis.* 21(1): 58–62.

[2] Sugiyanto and Abrori M 2020 A mathematical model of the COVID-19 cases in Indonesia (under and without lockdown enforcement) *Biology, Medicine, & Natural Product Chemistry*, 9(1): 15-19.

[3] Rustan R and Handayani L 2020 The outbreak’s modeling of Coronavirus (COVID-19) using the modified SEIR model in Indonesia *SPEKTRA: Jurnal Fisika dan Aplikasinya*. 5(1): 61-68.

[4] Djalante R., Lassa J, Setiamarga D, Sudjatma A, Indrawan M., Haryanto B, Mahfud C., Sinapoy M S, Djalante S, Rafliana I, Gunawan L A, Surtiari G A K, and Warsilah H 2020 Review and
analysis of current responses to COVID-19 in Indonesia: Period of January to March 2020. *Progress in Disaster Science*. 6(100091): 1-9

[5] Tibshirani R J and Taylor J 2011 The solution path of the generalized lasso *Ann. Statist*. 39 (3): 1335–1371.

[6] Arnold T and Tibshirani R J 2016 Efficient implementations of the generalized lasso dual path algorithm. *Journal of Computational and Graphical Statistics*, 25(1): 1–27.

[7] Perkembangan COVID-19 Provinsi Banten. Retrieved from https://infocorona.bantenprov.go.id/maps

[8] Data terkait COVID-19 di DI Yogyakarta. Retrieved from https://corona.jogjaprov.go.id/data-statistik

[9] Open data COVID-19 Provinsi DKI Jakarta. Retrieved from https://riwayat-file-covid-19-dki-jakarta-jakartagis.hub.arcgis.com/

[10] Sebaran kasus COVID-19 di Jawa Barat. Retrieved from https://pikobar.jabarprov.go.id/distribution-case

[11] Tabel sebaran COVID-19 Jawa Tengah. Retrieved from https://corona.jatengprov.go.id/data

[12] Peta sebaran COVID-19 Jatim. Retrieved from http://infocovid19.jatimprov.go.id/

[13] BPS-Statistics of Banten Province 2020 *Banten Province in figures 2020*. CV Dharmaputra.

[14] BPS-Statistics of Daerah Istimewa Yogyakarta Province 2020 *Daerah Istimewa Yogyakarta in figures 2020*. BPS-Statistics of Daerah Istimewa Yogyakarta Province.

[15] BPS-Statistics of DKI Jakarta Province 2020 *DKI Jakarta Province in figures 2020*. BPS-Statistics of DKI Jakarta Province.

[16] BPS-Statistics of Jawa Barat Province 2020 *Jawa Barat Province in figures 2020*. BPS-Statistics of Jawa Barat Province.

[17] BPS-Statistics of Jawa Tengah Province 2020 *Jawa Tengah Province in figures 2020*. BPS-Statistics of Jawa Tengah Province.

[18] BPS-Statistics of Jawa Timur Province 2020 *Jawa Timur Province in figures 2020*. BPS-Statistics of Jawa Timur Province.

[19] Tibshirani R 1996 Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol*. 267–288.

[20] Gelfand A E, Kim H-J, Sirmans C, and Banerjee, S 2003 Spatial modeling with spatially varying coefficient processes. *J. Amer. Statist. Assoc*. 98(462): 387–396.

[21] Zhao Y and Bondell H 2019 Solution paths for the generalized lasso with applications to spatially varying coefficients regression. Computational Statistics and Data Analysis. 142(106821): 1-14.