KOMPUTASI: JURNAL ILMIAH ILMU KOMPUTER DAN MATEMATIKA
VOL. 20 (1) (2023), 44-50 p-ISSN: 1693-7554, e-ISSN:2654-3990

Analysis of Regency and City Pneumonia Clusters in West Java 2020

Yusma Yanti [1], **Septian Rahardiantoro** [2]

[1] Department of Computer Science, Faculty of Mathematics and Natural Science, Pakuan University, Bogor, West Java, 16143, Indonesia
[4] Department of Statistics, Faculty of Mathematics and Natural Science, IPB University, West Java, 16680, Indonesia

Abstract

Pneumonia is an infection of the respiratory tract caused by bacteria, viruses, or fungi. The number of pneumonia cases in West Java is relatively high, therefore, it is necessary to identify some group of regencies/cities in which have a common characteristic, to make it easier to handle. The data used is data on the number of pneumonia cases in 27 regencies/cities in West Java in 2020. In this study, chi-squared test was applied to determine the characteristics of pneumonia spread in West Java. Then, a regression-based analysis by using the Irregular Graph Fused LASSO method was used to provide the cluster of regencies/cities, by considering the adjacent locations of regencies/cities as a penalty matrix. The results obtained that the cases spread unevenly. The number of cases for every 1000 people in each regency/city was relatively high in the eastern part of West Java. There were 6 clusters obtained from 27 regencies/cities, with Pangandaran regency as the location with the highest cases occurred. Depok City and Bekasi City were locations with the lowest number of cases even though they have relatively high population numbers.

Keywords: *Chi-squared test; location clustering; Irregular Graph Fused LASSO; pneumonia; regression analysis*

1. Introduction

Pneumonia is a type of disease related to breathing. Bacteria, viruses, or fungi are the common factors that causes pneumonia by contributing infection of the respiratory tract [1]. The response of this disease affected individuals varies widely, ranging from mild to severe. The spread of this disease directly often occurs through splashes of saliva that are released from the mouth into the air like sneezing. While, for indirect spread, it is common occurred by hand contact with saliva adhering to an object such as tissue, unsterilized ventilators and so on.

West Java Province is the area with the highest number of pneumonia cases in Indonesia from 2013 to 2017 [2]. Even though the pneumonia cases have decreased starting in 2019, the pneumonia prevention by government might see inefficient because the decrease in the number of cases from year to year is relatively small. This is because the spread of this disease does not spread evenly in all districts and cities in the province of West Java. Therefore, this study will analyze the distribution of pneumonia cases in regency/city based on 2020 data using chi-squared test analysis.

[*]**Corresponding author.** *E-mail address: yusma.yanti@unpak.ac.id*
Received: 21 November 2023, Accepted: 12 January 2023 and available online 30 January 2023
https://doi.org/10.33751/komputasi.v20i1.6412

Then, in this study, we applied Irregular Graph Fused LASSO [3] to identify the cluster of regency/city based on the number of pneumonia cases for every 1000 people in each regency/city. The Irregular Graph Fused LASSO is a method with a penalty (shrinkage) which is a linear function of the regression parameters based on regency/city adjacencies. In this case, the penalty is defined as the difference between the response values that are close to each other in adjacent regencies/cities. This method is considered very effective in clustering adjacent or neighboring objects as mentioned in [4], as in the application of clustering on burglaries per household between 2005 and 2009 in Chicago, IL.

2. Methods

The data used is data on the number of pneumonia cases in 27 regencies/cities in West Java in 2020 [5]. We also used number of populations for each regency/city, to obtain the target variable (Y) as the number of pneumonia cases for every 1000 people in each regency/city. The steps of analysis used are as follows:

a) Data description
This step is used to find out whether pneumonia cases spread evenly in all regencies/cities in West Java province, with the hypothesis:

b) Calculation of the Chi-squared Goodness of fit test
This step is used to find out whether pneumonia cases spread evenly in all regencies/cities in West Java province, with the hypothesis:
H0: pneumonia cases spread uniformly (evenly) for each district/city
H1: pneumonia cases do not spread uniformly (evenly) for each district/city

c) Defining target definition variables
In this case, the target variable is obtained according to the formula below.

$$Y_i = \frac{C_i}{P_i} \times 1000 \tag{1}$$

where C_i is number of pneumonia cases in i-th regency/city, P_i is number of populations in i-th regency/city, and Y_i is number of pneumonia cases per 1000 residents in i-th regency/city, for $i = 1, 2, ..., 27$.

d) Clustering regency/city using the Irregular Graph Fused LASSO
One application of cluster analysis is for mapping objects which have similar characteristics [6]. We applied the Irregular Graph Fused LASSO to the target variable (Y), after the data is relatively symmetrical. The results of this method are in the form of 2-dimensional images that can be interpreted descriptively in terms of the distribution of cases spread. In general, this method can be described as a form of multiple linear regression model as follows:

$$Y = X\beta + \varepsilon \tag{2}$$

where Y is $n \times 1$ target variable vector, X is $n \times p$ predictors matrix, β is $p \times 1$ coefficient parameter vector, and ε is $n \times 1$ random vector.

Then, the LASSO is defined by adding the l_1 penalty term on the coefficient parameter of formula (2) as stated in [7]. The coefficient estimator of LASSO can be written as:

$$\widetilde{\beta} = argmin\{|\mathbf{Y} - \mathbf{X}\beta|\frac{2}{2} + \lambda|\beta|_1\} \tag{3}$$

where $|a|_2 = \sqrt{\sum a_i^2}, |a|_1 = \sum |a_i|$, for arbitrary vector a, and $\lambda > 0$ is a tuning parameter. Moreover, the Irregular Graph Fused LASSO considers the coefficient adjacencies which notated in the $m \times n$ penalty matrix D in the penalty terms [8]. Each row of penalty matrix D contains 1 and -1 for adjacent districts/cities. Consider n locations in the data and let the i-th and j-th locations be adjacent. Therefore, the row of penalty matrix D can be stated as:

$$(0, \ldots., -1, \ldots..,1, \ldots,0)$$
$$\uparrow \quad \uparrow$$
$$i \quad j$$

The coefficient estimator of Irregular Graph Fused LASSO can be written as:

$$\widetilde{\beta} = argmin \left\{ |Y - X\beta|_2^2 + \lambda |D\beta|_1 \right\} \tag{4}$$

When $D = I$, then formula (4) is equivalent to formula (3). Practically, to apply the Irregular Graph Fused LASSO for spatial clustering, we define the predictor matrix X is identical, so that $X = I$. Therefore, the formula (5) can be simplified as:

$$\widetilde{\beta} = argmin \left\{ |Y - \beta|_2^2 + \lambda |D\beta|_1 \right\} \tag{5}$$

In this case, to determine the optimal tuning parameter λ, we used the ALOCV approach [9], [10]. The application of ALOCV in the Irregular Graph Fused LASSO has small error and good performance for spatial clustering [11].

3. Result and Discussion

The distribution of number of pneumonia cases in 27 regencies/cities in West Java 2020 can be seen in Figure 1. We can see that the histogram is skewed to the right with an outlier which has large number of pneumonia cases. It is indicated that majority of regencies/cities in West Java occurred relatively small number of pneumonia cases, while, in several regencies/cities the number of pneumonia cases relatively higher compared others.

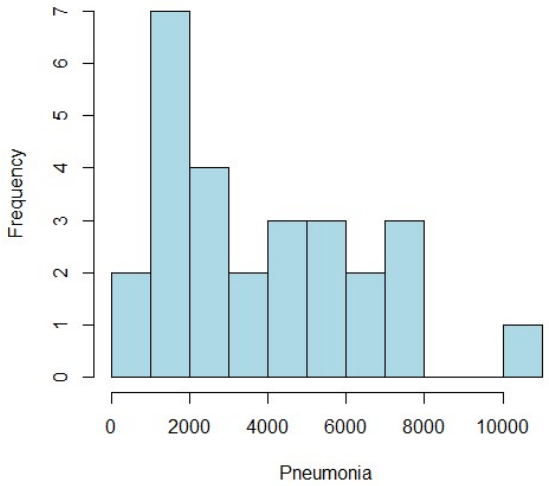

Figure 1. Description of the spread of pneumonia in West Java in 2020

Then, the chi-squared test was applied to see the characteristics of the spread of pneumonia cases in West Java, in which the results is summarizes in Table 1. Based on these results, it can be concluded that reject H_0, or in other words, the spread of pneumonia has an uneven distribution point in each location. There are some groups of locations have a different pneumonia spread cases compared to other locations. Therefore, it is reasonable to identify regency/city cluster based on the spread of pneumonia cases in West Java 2020.

Table 1. Summary of chi-squared test for pneumonia cases spread

Chi-squared test statistics	Degree of Freedom	p-Value
43445	26	$< 2.2 \times 10^{-16}$

The next step is to define the target variable (Y) for regencies/cities clustering as the number of pneumonia cases per 1000 residents (equation (1)). The distribution of Y is presented as histogram in Figure 2. The distribution is also relatively skewed to the right a not symmetrical.

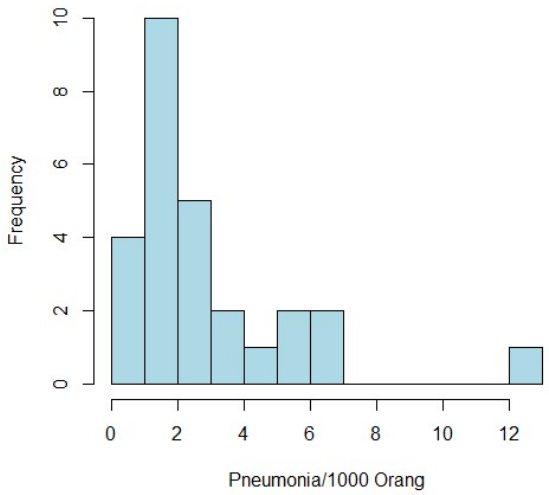

Figure 2. The number of cases for each region

Since the Irregular Graph Fused LASSO assumes that the distribution of the target variable should be symmetrical [8], we transformed the target variable (Y) by using the natural logarithm transformation. Therefore, a more symmetrical distribution is obtained, as shown in histogram in Figure 3.

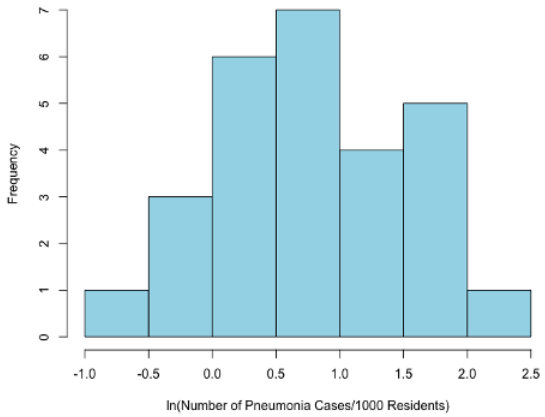

Figure 3. Histogram of the natural logarithm of number of pneumonia cases per 1000 residents (ln(Y))

The next step is performing cluster analysis by using the Irregular Graph Fused LASSO based

on ln(Y). Figure 4 presented the spatial distribution of ln(Y) to see the actual distribution before the clustering processing with the green color indicates the lower value, yellow color indicates the medium value, and red color indicates the higher value. There are several locations with a relatively low number of cases: Bandung City, Bandung Regency, Bekasi City, and Tasikmalaya City. While the location with the highest cases occurred in Pangandaran Regency. Purwakarta Regency also has a high number of cases, while the surrounding locations are relatively low

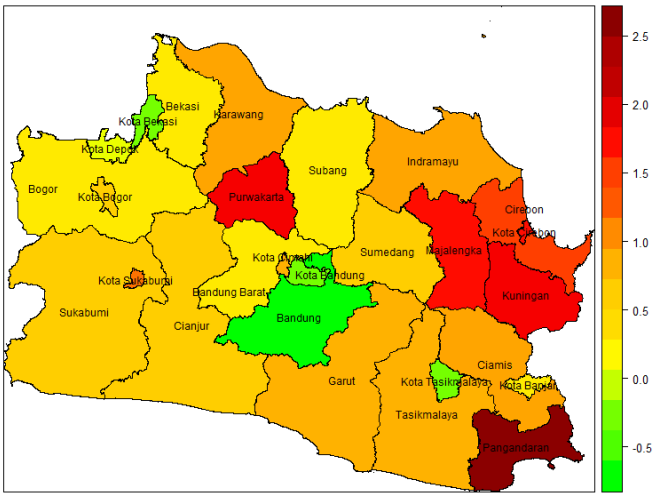

Figure 4. Spatial plot of ln Y

Then, the Irregular Graph Fused LASSO was applied to identify the clusters of regencies/cities based on ln(Y). In this case, we have $n = 27$ regencies/cities with $m = 53$ adjacent locations. Therefore, the dimension of the penalty matrix D is 53×27. We applied the formula (6) and ALOCV [9] to obtain the optimal tuning parameter λ. As a result, the optimal tuning parameter λ was 0.3019 with the mean squared errors (MSE) was 0.2441. The spatial plot of the clustering results is presented in Figure 5.

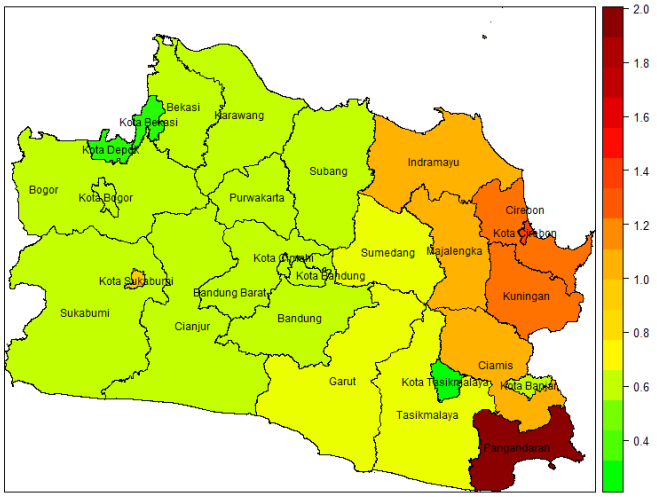

Figure 5. Spatial plot of regencies/cities clusters of ln(Y) based on the Irregular Graph Fused LASSO method

As seen in Figure 5, there were six clusters constructed in which indicated by the common color in each regency/city. The regencies/cities cluster with the lowest cases occurred in Depok City and Bekasi City. Meanwhile, Depok City and Bekasi City are locations with relatively high population numbers, but the number of cases is relatively low. The clusters with the higher cases were constructed in the eastern part of West Java. The further east, the more constructed clusters have the higher the number of cases. Pangandaran Regency is the cluster with the highest number of cases, in which followed by the cluster of Cirebon City, and cluster of Cirebon Regency and Kuningan Regency. It can be suggested to the government that to handle the pneumonia cases which occurs in West Java, the priority regencies/cities located in the eastern part of West Java, especially in Pangandaran Regency, Cirebon City, Cirebon Regency and Kuningan Regency.

4. Conclusion

In this study it can be concluded that the incidence of pneumonia cases is not spread evenly for each regency/city in West Java. Statistically, it can be stated that the number of pneumonia cases that occur has different probabilities for each regency/city. Based on the Irregular Graph Fused LASSO method, it can be seen the clusters of the number of pneumonia cases per 1000 people. The clusters results obtained that the further east, there were the clusters with higher the number of pneumonia cases. This means that the eastern region of West Java should be the priority to handle the spread of pneumonia cases by the government.

References

[1] Sari, E F dkk. 2016. Faktor-faktor yang berhubungan dengan Diagnosis Pneumnia pada Pasien Usia Lanjut. Jurnal Penyakit Dalam Indonesia Vol 3(4). [Online]. Available: https://scholar.ui.ac.id/en/publications/faktorfaktor-yang-berhubungan-dengan-diagnosis-pneumonia-pada-pas

[2] Sari, M P dan Cahyati, W H. 2019. Tren Pneumonia Balita di Kota Semarang Tahun 2012-2018. Higeia Journal of Public Health Research and Development 3 (3) 2019. [Online]. Available: https://journal.unnes.ac.id/sju/index.php/higeia/article/view/30266

[3] Rahardiantoro, S dan Sakamoto, W. 2021. Clustering Regions Based on Socio Economic Factors Which Affected the Number of COVID-19 Cases in Java Island. J. Phys.: Conf. Ser. 1863 012014 [Online]. Available: https://iopscience.iop.org/article/10.1088/1742-6596/1863/1/012014/meta

[4] BPS. 2021. Jawa Barat Province in Figure 2020. Catalog Number: 1102001.32 [Online] Available: https://jabar.bps.go.id/publication/2020/04/27/cfab9a400cf304f800182a5f/provinsi-jawa-barat-dalam-angka-2020.html

[5] Tibshirani, Robert. 1996. Regression Shrinkage and Selection via the Lasso. J.R. Statist.Soc B (1996) 58, No1, 267-288 [Online] Available: https://www.jstor.org/stable/2346178

[6] Tibshirani R J and Taylor J. 2011. The solution path of the generalized lasso Ann. Statist. 39 (3): 1335–1371 [Online] Available: https://www.jstor.org/stable/23033600

[7] Yanti, Yusma dan Rahardiantoro, Septian.2018. Alternatif Penggerombolan Data Deret Waktu dengan Kondisi terdapat Data Kosong. Indonesia Journal of Statistics and It's Applications. Volume 2 No 1: 13-22 [Online] Available: https://journal.stats.id/index.php/ijsa/article/view/55

[8] Tibshirani, R J and Taylor, J. 2011. The solution path of the generalized lasso Ann. *Statist.* 39 (3): 1335–1371 [Online] Available: https://www.jstor.org/stable/23033600

[9] Rad, K R and Maleki, A. 2020. A scalable estimate of the extra-sample prediction error via approximate leave-one-out. *Journal of the Royal Statistical Society Series B.* 82(4): 965-996. arXiv:1801.10243

[10] Rad, K R, Zhou, W, and Maleki, A. 2020. Error bounds in estimating the out-of-sample prediction error using leave-one-out cross validation in high-dimensions. *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 108.

[11] [11] Rahardiantoro, S and Sakamoto, W. 2022. Optimum tuning parameter selection in generalized lasso for clustering with spatially varying coefficient models. *IOP Conference Series: Earth and Environmental Science*, 950(1), 012093. https://doi.org/10.1088/1755-1315/950/1/012093

.