WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Cluster Optimization Using Cluster wise Linear Regression Method with Akaike Information Criterion to Identify Factors Influencing the Number of DHF Cases

Syarifah Desy Rahmawati[1], Sudarno[2], Yuciana Wilandari[3]

[1,2,3]*Department of Statistics, Faculty of Science and Mathematics, Diponegoro University*

A B S T R A C T

Dengue Hemorrhagic Fever (DHF) is an endemic disease that is a major health problem in Indonesia. This study aims to optimize clusters in identifying factors that affect the number of dengue cases based on districts/cities in West Java Province in 2022 using Clusterwise Linear Regression (CLR) method. CLR is a method of clustering data into clusters based on the characteristics of regression parameters. Model parameter estimation is carried out using the maximum likelihood method and the best model optimization is carried out with the minimum Akaike Information Criterion (AIC) to determine the optimal number of clusters. The results of the study stated that the best model consists of three clusters. Cluster 1 consists of 11 districts/cities which are influenced by the factors of population growth rate, number of hospitals, number of health centers; Cluster 2 consists of 8 districts/cities which are influenced by the factors of population growth rate, number of hospitals, number of flood disasters, number of health centers, percentage of decent sanitation; and cluster 3 consists of 8 districts/cities which are influenced by the factors of population growth rate, number of hospitals, number of flood disasters, percentage of decent sanitation. The value of the determination coefficient in cluster 1 was 90.95736%; cluster 2 is 71.21607%; and cluster 3 is 91.40601%. This indicates that cluster optimization using CLR with AIC can improve the model's ability to better explain data variants.

Keywords: Akaike Information Criterion, Clusterwise Linear Regression, DHF, Cluster, West Java

INTRODUCTION

Indonesia as one of the largest tropical countries in Southeast Asia is facing major health problems related to Dengue Hemorrhagic Fever (DHF) which affects all ages and spreads in various provinces. More than 80% of Indonesian children aged 10 years and above who live in urban areas have been infected with dengue (Elizabeth & Yudhastuti, 2023). WHO (2022) reports 390 million global dengue infections per year with 500,000 people requiring hospitalization. The number of dengue cases in Indonesia in 2022 was 143,184 people with the highest dengue cases in West Java reaching 36,594 people (Ministry of Health, 2022). The spread of this disease is influenced by Physical, social, and biological environmental factors (Oroh et al., 2020).

The linear regression analysis approach is often used to identify which factors affect the spread of dengue disease. The results of the analysis obtained were less than optimal because this method was not able to estimate the model that indicated the presence of clusters in the data (Putri, 2015). A method that can be applied to solve this problem is Clusterwise Linear Regression (CLR). CLR is a combination of cluster and regression techniques that can identify hidden structures in data with previously unknown clusters. CLR aims to determine the best model by identifying the hidden structures in the data that make up clusters. The best model can be demonstrated by using the minimum Akaike Information Criterion (AIC) value on each number of clusters formed. AIC is a criterion in model selection that balances the goodness of fit model based on the estimated maximum likelihood with the number of parameters used in the model (Utama & Hajarisman, 2021).

Several previous studies on clustering methods in handling various cases have been conducted. Ikbal (2024) uses the K-means Clustering algorithm to group areas with dengue case rates based on gender in West Java. The results of the study show that the region can be grouped into three clusters, namely high, medium, and low clusters with evaluation using the Davies-Bouldin Index. Meylisah et al. (2023) applying Clusterwise Linear Regression Modeling to analyze the poverty level in Indonesia. This study found three optimal clusters based on AIC and BIC (Bayesian Information Criteria) criteria. The first cluster identified the percentage of electricity users, the number of small and micro industries, and the number of tourist villages as significant factors affecting poverty; The second cluster shows the number of tourist villages as the dominant factor; The third cluster highlights the percentage of electricity users and the percentage of villages with mining and quarrying.

Based on these studies, there has been no study that identifies the factors that affect the number of DHF cases using the Clusterwise Linear Regression approach. This study presents a novelty by clustering districts/cities in West Java Province based on the factors that affect the number of DHF cases using the CLR method optimized with AIC and knowing the factors that affect the number of dengue cases in each cluster effectively and accurately.

LITERATURE REVIEW

Regression is a statistical method used to analyze the extent to which the relationship between two or more variables affects each other. This method determines the strength of relationships, interaction patterns, and the direction of influence between dependent variable and independent variable (Ghozali, 2021). In general, regression models can be expressed in the following forms:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i \quad (1)$$

with Y_i the dependent variable for observation to-i, X_{ik} the independent variable to-k for observation to-i, β_k the regression coefficient value to-k, and ε_i the observation error to-i.

Based on the regression model in Equation (1), if there are n many observations, then the regression modeling for each observation to-i is as follows:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \quad (2)$$

with **Y** vector of dependent variable, **β** vector of regression parameter, **X** matrix of observations for independent variable, and **ε** vector of error which are assumed to be identical, independent, and normally distributed with mean 0 and constant variance σ^2.

Classical assumption testing includes normality test, homokedasticity test, non-autocorrelation test, and non-multicollinearity test.

1. Normality Test

The normality test is used in regression analysis to determine whether the error value follows the normal distribution (Ghozali, 2021).

Hypothesis: H_0: $F_n(x) = F_0(x)$ error data are normally distributed

H_1 : $F_n(x) \neq F_0(x)$ error data are not normally distributed

Significance Level: α = 5%

Test Statistics:

$$D = sup(|F_n(x) - F_0(x)|) \quad (3)$$

with D maximum value for all values of x, $F_n(x)$ cumulative distribution function of the data sample, $F_0(x)$ cumulative distribution function of theoretical data.

Test criteria: if $D > D_{(n,\alpha)}$ or significance value $P_{value} < \alpha$ then H_0 rejected.

2. Homoskedasticity Test

According to Ghozali (2021), homoskedasticity occurs if there is the same variance from an observation error to another observatin, namely $E(\varepsilon_i) = \sigma^2$ for i = 1, 2, ..., n.

Hypothesis: H_0 : $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \cdots = \sigma_n^2 = \sigma^2$ (homokedasticity occurs)

H_1 : there is at least one $\sigma_i^2 \neq \sigma^2$(no homocedasticity)

Significance Level: α = 5%

Test Statistics:

$$LM = nR^2 \quad (4)$$

with R^2 coefficient determination, LM following Chi Square distribution ($\chi^2_{\alpha,k}$).

Test criteria: if $LM > \chi^2_{\alpha,k}$ and $P_{value} < \alpha$ then H_0 rejected.

3. Non-Autocorrelation Test

Autocorrelation refers to the correlation between observations in one interrelated variable. According to Ghozali (2021), explained that this test is very suitable for use in resecarch involving data with a total of ≤ 100.

Hypothesis: $H_0 : \rho = 0$ (no autocorrelation)

$H_1 : \rho \neq 0$ (autocorrelation)

Significance Level: α = 5%

Test Statistics:

$$DW = \frac{\sum_{i=2}^{n}(\varepsilon_i - \varepsilon_{i-1})^2}{\sum_{i=1}^{n}\varepsilon_i^2} \quad (5)$$

Test criteria:

0 < DW < dL (H_0 rejected, there is positive autocorrelation) and $P_{value} < \alpha$

dL < DW < dU (No conclusion)

dU < DW < 4-dU (H_0 accepted, there is no autocorrelation) and $P_{value} > \alpha$

4-dU < DW < 4-dL (No conclusion)

4-dL < DW < 4 (H_0 rejected, there is negative autocorrelation) and $P_{value} < \alpha$

4. Non-Multicollinearity Test

The Multicollinearity test was performed to check the correlation between the independent variables in the regression model. In an ideal model, there should be no significant correlation between independent variables (Ghozali, 2021).

Hypothesis: $H_0 : VIF_j = 0$ (no multicollinearity)

$H_1 : VIF_j \neq 0$ (multicollinearity)

Test Statistics:

$$VIF_j = \frac{1}{1-R_j^2} \quad (6)$$

with VIF_j Variance Inflation Factor of the independent variable to-j, R_j^2coefficient of determination of the independent variable to-j, and $j = 1,2,\dots,k$ (j order of the independent variable and k total independent variable).

Test criteria: if $VIF_j > 10$ then H_0 rejected.

Hypothesis testing is used to assess the accuracy of regression models in estimating actual values. This hypothesis test consists of two types, namely the F test and the t test.

- **F Test**

The F test aims to test whether the equation of the regression model can simultaneously explain the effect of the independent variable on the dependent variable (Ghozali, 2021).

Hypothesis: $H_0 : \beta_j = 0$ (linear regression model does not fit)

$H_1 : \beta_j \neq 0$ for at least one, $j = 1,2,\dots,k$ (regression model fit)

Significance Level: α = 5%

Test Statistics:

$$F_{hitung} = \frac{KTR}{KTG} \quad (7)$$

Test criteria: if $F_{hitung} > F_{tabel}\left(F_{(\alpha,k,n-k-1)}\right)$ or $P_{value} < \alpha$ then H_0 is rejected.

- **T Test**

According to Ghozali (2021), the t-test functions to test the significance of each parameter coefficient separately to determine the extent of the influence of each independent variable on dependent variable individually.

Hypothesis: $H_0 : \beta_j = 0, j = 1,2,\dots,k$ (parameter coefficients are not significant)

$H_1 : \beta_j \neq 0, \ j = 1,2,\dots,k$ (parameter coefficients are significant)

Significance Level: α = 5%

Test Statistics:

$$t_j = \frac{\hat{\beta}_j}{Se\,(\hat{\beta}_j)} \quad (8)$$

where $Se(\hat{\beta}_j) = \sqrt{var(\hat{\beta}_j)}$ or can use the P_{value}.

With $\hat{\beta}_j$ regression coefficient of β on independent variable to-j, $Se\left(\hat{\beta}_j\right)$ standard error of regression coefficient $\hat{\beta}_j$.

Test criteria: if $|t_j| > t_{tabel}$ ($t_{(\alpha/2,n-k-1)}$) or $P_{value} < \alpha$ then H_0 rejected.

The coefficient of determination is in the interval between zero and one ($0 \leq R^2 \leq 1$) which can be R^2expressed by the one obtained from the following formula:

$$R^2 = \frac{JKR}{JKT} = 1 - \frac{JKG}{JKT} \quad (9)$$

The following are the categories of R^2values according to Chin (1998).

Table 1 - Category R^2.

$\mathbf{R^2}$ **Value**	**Category**
$0{,}67 \leq R^2 \leq 1$	Strong
$\boldsymbol{0,33 \leq R^2 < 0,67}$	Moderate
$\boldsymbol{0,19 \leq R^2 < 0,33}$	Weak
$0 \leq R^2 < 0{,}19$	Very weak

Clusterwise Linear Regression (CLR) is a clustering method based on the characteristics of regression parameters in which clusters are randomly initialized to produce a regression model large enough to achieve convergence. The CLR method was first introduced by Späth through an exchange algorithm with the formation of a number of partitions as many as K and corresponding $\boldsymbol{\beta_c}$ parameters so that the number of squares of the error calculated on all clusters is minimized by:

$$Min\ Z = \sum_{c=1}^{K} ||\boldsymbol{X_c\beta_c} - \boldsymbol{Y_c}||^2 \quad (10)$$

The existence of the solution $\boldsymbol{\beta_c}$ is rank $\boldsymbol{X_c} = J$. The conditions required for this is $n_c \geq J$ which mean $n \geq KJ$, where n_c is the number of observations in cluster. The general model of the CLR is as follows:

$$\boldsymbol{Y_c} = \boldsymbol{X_c\beta_c} + \boldsymbol{\varepsilon_c} \quad (11)$$

with $\boldsymbol{Y_c}$ dependent variable vector on cluster to-c, $\boldsymbol{X_c}$ observation matrix for the independent variable on cluster to-c, $\boldsymbol{\beta_c}$ regression coefficient vector on cluster to-c, $\boldsymbol{\varepsilon_c}$ the error vector for the observation on cluster to-c, and c = 1,2, …, K (c the sequence of clusters and K the number of clusters).

Estimation of the parameters of the model shown in the Equation (11) by using the maximum likelihood (DeSarbo & Cron, 1988). Likelihood equation for random sample consisting of n free subjects, namely:

$$L = \prod_{i=1}^{n}\left[\sum_{c=1}^{K} \lambda_c (2\pi\sigma_c^2)^{-1/2}\, exp\left[\frac{-(y_i - \boldsymbol{X_i\beta_c})^2}{2\sigma_c^2}\right]\right] \quad (12)$$

Based on Equation (12), the likelihood equation can be linearized using the ln likelihood equation as fellows:

$$ln\ L = \sum_{i=1}^{n} ln\left[\sum_{c=1}^{K} \lambda_c (2\pi\sigma_c^2)^{-1/2}\, exp\left[\frac{-(y_i - \boldsymbol{X_i\beta_c})^2}{2\sigma_c^2}\right]\right] \quad (13)$$

Equation (13) is maximized to obtain an estimate $\lambda_c, \sigma_c^2, \beta_{jc}$with constraints:

$$0 \leq \lambda_c \leq 1; \sum_{c=1}^{K} \lambda_c = 1\ \sigma_c^2 > 0 \quad (14)$$

The placement of each observation is carried out through posterior estimation on the observations to each cluster c ($\hat{p}_{ic}$) after obtaining the initial value of the estimate $\lambda_c, \sigma_c^2, \beta_{jc}$ which are calculated by:

$$\hat{p}_{ic} = \frac{\lambda_c f_{ic}(y_i|X_{ij}, \bar{\sigma}_k^2, \beta_{jc})}{\sum_{c=1}^{K} \lambda_c f_{ic}(y_i|X_{ij}, \bar{\sigma}_k^2, \beta_{jc})} \quad (15)$$

Partitions in the CLR method can be formed with the following rules:

a. Put i to a cluster c if $\hat{p}_{ic} > \hat{p}_{il}$ for all $l \neq c = 1,2, \ldots K$

b. Probability function ln with constraint in Equation (14):

$$\Phi = \sum_{i=1}^{n} ln\left[\sum_{c=1}^{K} \lambda_c f_{ic}(y_i|X_{ij}, \sigma_c^2, \beta_{jc})\right] - \mu\left(\sum_{c=1}^{K} \lambda_c - 1\right) \quad (16)$$

The maximum likelihood stationary equation is obtained by equalizing the first-order partial derivative of the ln likelihood function to a parameter $\lambda_c, \sigma_c^2, \beta_{jc}$ equal to zero, as follows:

$$\frac{\partial \Phi}{\partial \lambda_c} = \sum_{i=1}^{n} \frac{1}{\sum_{c=1}^{K} \lambda_c f_{ic}(*)} f_{ic}(*) - \mu = 0 \quad (17)$$

$$\frac{\partial \Phi}{\partial \sigma_c^2} = \sum_{i=1}^{n} \frac{1}{\sum_{c=1}^{K} \lambda_c f_{ic}(*)} \lambda_c \frac{\partial f_{ic}(*)}{\partial \sigma_c^2} = 0 \quad (18)$$

$$\frac{\partial \Phi}{\partial \beta_{jc}} = \sum_{i=1}^{n} \frac{1}{\sum_{c=1}^{K} \lambda_c f_{ic}(*)} \lambda_c \frac{\partial f_{ic}(*)}{\partial \beta_{jc}} = 0 \quad (19)$$

with $f_{ic}(*)$ that is $f_{ic}(y_i|X_{ij}, \sigma_c^2, \beta_{jc})$.

The estimation μ is done by multiplying the two sides of Equation (17) with λ_c and adding them in whole c so that the equation is:

$$\sum_{i=1}^{n} \frac{\sum_{c=1}^{K} \lambda_c f_{ic}(*)}{\sum_{c=1}^{K} \lambda_c f_{ic}(*)} - \mu \sum_{c=1}^{K} \lambda_c = 0 \quad (20)$$

$$\sum_{i=1}^{n} 1 - \mu.1 = 0$$

$n - \mu = 0$

$\hat{\mu} = n$ (21)

The estimate $\lambda_c$is obtained by multiplying the two sides of Equation (17) with λ_c so that the equation is as follows:

$\sum_{i=1}^{n} \frac{\lambda_c f_{ic}(*)}{\sum_{c=1}^{K} \lambda_c f_{ic}(*)} - \lambda_c \mu = 0$ (22)

$\sum_{i=1}^{n} \hat{p}_{ic} - \lambda_c n = 0$

$\hat{\lambda}_c = \frac{\sum_{i=1}^{n} \hat{p}_{ic}}{n}$ (23)

Estimation σ_c^2 and β_{jc} obtained from the definition $\hat{p}_{ic}$ in Equation (15) and restating Equations (18) and (19), the equation is:

$\frac{\partial \Phi}{\partial \sigma_c^2} = \sum_{i=1}^{n} \frac{1}{\sum_{c=1}^{K} \lambda_c f_{ic}(*)} \lambda_c \frac{\partial f_{ic}(*)}{\partial \sigma_c^2} = 0$ (24)

$\frac{\partial \Phi}{\partial \beta_{jc}} = \sum_{i=1}^{n} \frac{1}{\sum_{c=1}^{K} \lambda_c f_{ic}(*)} \lambda_c \frac{\partial f_{ic}(*)}{\partial \beta_{jc}} = 0$ (25)

Estimation of these parameters can be done using the two-stage E-M algorithm. E-stage is used to obtain the estimated value of λ_c and $\hat{p}_{ic}$ from Equations (15) and (23), while M-stage is used to obtain the estimated value for β_{jc} and σ_c^2. The M-stage, expansion is required on Equations (24) and (25):

$\frac{\partial \Phi}{\partial \boldsymbol{\beta_c}} = \sum_{i=1}^{n} \frac{1}{\sum_{c=1}^{K} \lambda_c f_{ic}(*)} \lambda_c (2\pi\sigma_c^2)^{-1/2} \times exp\left[\frac{-(y_i - \boldsymbol{X_i \beta_c})^2}{2\sigma_c^2}\right] . \frac{2(y_i - \boldsymbol{X_i \beta_c}) \boldsymbol{X_i}}{2\sigma_c^2} = 0$

$\frac{\partial \Phi}{\partial \boldsymbol{\beta_c}} = \sum_{i=1}^{n} \hat{p}_{ic} (y_i - \boldsymbol{X_i \beta_c}) \boldsymbol{X_i} = 0$ (26)

Based on Equation (26), the following estimates of $\boldsymbol{\beta_c}$ is obtained:

$\widehat{\boldsymbol{\beta}}_{\boldsymbol{c}} = (\sum_{i=1}^{n} \hat{p}_{ic} \boldsymbol{X_i X_i'})^{-1} (\sum_{i=1}^{n} \hat{p}_{ic} y_i \boldsymbol{X_i})$ (27)

The estimate σ_c^2 is obtained after the previous steps are taken, namely:

$\frac{\partial \Phi}{\partial \sigma_c^2} = \sum_{i=1}^{n} \frac{1}{\sum_{c=1}^{K} \lambda_c f_{ic}(*)} \left[\lambda_c exp\left[\frac{-(y_i - \boldsymbol{X_i \beta_c})^2}{2\sigma_c^2}\right] (-1/2(2\pi\sigma_c^2)^{-3/2} 2\pi) + \lambda_c (2\pi\sigma_c^2)^{-1/2} exp\left[\frac{-(y_i - \boldsymbol{X_i \beta_c})^2}{2\sigma_c^2}\right] \frac{1/2(y_i - \boldsymbol{X_i \beta_c})^2}{\sigma_c^4}\right] = 0$

$\frac{\partial \Phi}{\partial \sigma_c^2} = \sum_{i=1}^{n} \hat{p}_{ic} \left[\frac{-1}{2\sigma_c^2} + \frac{(y_i - \boldsymbol{X_i \beta_c})^2}{2\sigma_c^4}\right] = 0$ 28)

From Equation (28), thus obtained:

$\hat{o}_c^2 = \frac{\sum_{i=1}^{n} \hat{p}_{ic} (y_i - \boldsymbol{X_i \beta_c})^2}{\sum_{i=1}^{n} \hat{p}_{ic}}$ (29)

A model is said to be good if it meets the requirements, namely having minimum AIC value. AIC values can be formulated as follows (Rizalul et al., 2017):

$AIC(K) = 2n(K) - 2ln[maxL(K)]$ (30)

with $n(K)$ number of effective parameters where $n(K) = JK + 2K - 1$, $L(K)$ maximum likelihood, J number of regression parameters (k total independent variables and Knumber of clusters).

CLR hypothesis testing uses maximum likelihood a follow:

Hypothesis: H_0: $\beta_{jc} = 0$

H_1: $\beta_{jc} \neq 0$, with =1,2, ..., and =1,2, ..., Kjkc

Significance Level: α = 5%

Test Statistics:

$Z = \frac{\beta_K - B_K}{\sqrt{f_{KK}^{-1}}}$ (31)

with β_K are the estimated coefficients of maximum likelihood in K cluster, B_K are value of the actual population parameter in K cluster, and f_{KK}^{-1} element of the asymptotic covariance matrix in K cluster.

The test criteria used is that if $|Z| > Z_{tabel}$ $(Z_{\alpha/2})$ or $P_{value} < \alpha$ then H_0 is rejected.

RESEARCH METHOD

The data source used in this study came from the Central Statistics Agency of West Java Province. The observation units in this study are 27 districts or cities in West Java Province in 2022. To process the data, this study utilizes RStudio and QGIS software. The variables used were the number of DHF cases as variable Y (people), population growth rate as variable X_1 (% per year), number of hospitals as variable X_2 (unit) number of flood disasters as variable X_3 (time), number of health centers as variable X_4 (unit), and percentage of proper sanitation as variable X_5 (%). To process the data, this study utilizes RStudio and QGIS software. The cluster variable design analyzed in this study is presented in Table 2.

Table 2 - Cluster Variable Design

District/City	$\boldsymbol{Y}$	$\boldsymbol{X_1}$	$\boldsymbol{X_2}$	$\boldsymbol{X_3}$	$\boldsymbol{X_4}$	$\boldsymbol{X_5}$
1.	Y_1	$X_{1,1}$	$X_{1,2}$	$X_{1,3}$	$X_{1,4}$	$X_{1,5}$
2.	$\boldsymbol{Y_2}$	$\boldsymbol{X_{2,1}}$	$\boldsymbol{X_{2,2}}$	$\boldsymbol{X_{2,3}}$	$\boldsymbol{X_{2,4}}$	$\boldsymbol{X_{2,5}}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
27.	Y_{27}	$X_{27,1}$	$X_{27,2}$	$X_{27,3}$	$X_{27,4}$	$X_{27,5}$

The research procedures were carried out through the following steps:

1. Input variable data for each district/city in West Java Province.
2. Describe the data using descriptive statistics.
3. Perform regression model specifications:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$

4. Predict the parameters of the regression model.
5. Perform a classic assumption test.
6. Conducting hypothesis testing in the form of F test and t test.
7. Evaluate the model with the coefficient of determination. If the regression model produced is not good, it is indicated that there are groups in the data, so that Clusterwise Linear Regression (CLR) analysis is then carried out.
8. Standardize data because variables have different units.
9. Specifies the number of clusters () that are possible.K
10. Retrieve the initial value for the estimation $\hat{\lambda}_c$, $\hat{\beta}_{jc}$, $\hat{\sigma}_c^2$.
11. Calculate $\hat{p}_{ic}$ using Equation (15).
12. Calculate estimates $\hat{\lambda}_c$ using Equation (23), $\hat{\beta}_{jc}$ using Equation (27), $\hat{\sigma}_c^2$ with Equation (29).
13. Calculating the ln likelihood value of probability with Equation (16).
14. Determine if the iteration results have converged. If not, proceed to the next iteration and repeat steps 10 through 13 until they converge.
15. Determine the optimal number of clusters using the minimum AIC value.
16. Define the members and model assumptions in each cluster.
17. Perform hypothesis testing and evaluation of the formed model.
18. Perform cluster profiling.

RESULTS & DISCUSSION

Descriptive statistics in this study were carried out to find out the initial picture of the condition of each research variable, which is presented in Table 3.

Table 3 - Descriptive Statistics.

Variabel	Min	Mean	Max	Skewness	Kurtosis	Standard Deviation
Y	15,00	92,41	310,00	1,467	4,624	74,766
$\boldsymbol{X_1}$	0,410	1,301	1,860	-0,408	2,599	0,378
$\boldsymbol{X_2}$	1,00	14,63	53,00	1,526	4,471	13,425
X_3	0,000	7,185	30,000	1,501	5,284	7,000
X_4	10,00	40,78	101,00	0,833	3,685	21,393
X_5	45,80	73,58	96,21	-0,443	1,918	14,963

The linear regression model formed according to Equation (1) is:

$Y = 450{,}7847 - 63{,}3853X_1 + 2{,}1844X_2 + 2{,}1616X_3 - 2{,}5810X_4 - 2{,}9653X_5 + \varepsilon$

Classical assumption testing must be met before performing linear regression analysis and hypothesis testing. This aims to ensure that the regression model used is free of assumption deviations and produces valid estimates.

1. Normality Test

Hypothesis: $H_0: F_n(x) = F_0(x)$ error data are normally distributed

$H_1 : F_n(x) \neq F_0(x)$ error data are not normally distributed

Significance Level: α = 5%

Test Statistics:

$D = \sup(|F_{27}(x) - F_0(x)|) = 0{,}097204$; $P_{value} = 0{,}7376$

The value of $D_{(n,\alpha)}$ is obtained from the Kolmogorov-Smirnov (K-S) table with α = 5% and $n = 27$ so that the value $D_{(27,0.05)}$ is 0.254.

Test criteria: if or $D > D_{(n,a)}$ significance value $P_{value} < \alpha$ then H_0 rejected.

Decision and Conclusion: H_0 accepted because the value $D < D_{(27,0.05)} = 0{,}097204 < 0{,}254$ or the value $P_{value} > \alpha = 0{,}7376 > 0{,}05$ so the error data is normally distributed.

2. Non-Multicollinearity Test

Hypothesis: $H_0 : VIF_j = 0$ (no multicollinearity)

$H_1 : VIF_j \neq 0$ (multicollinearity)

Test Statistics:

$VIF_1 = \frac{1}{1-{R_1}^2} = 1.243473$; $VIF_2 = \frac{1}{1-{R_2}^2} = 1.602682$; $VIF_3 = \frac{1}{1-{R_3}^2} = 2.527059$; $VIF_4 = \frac{1}{1-{R_4}^2} = 3{,}284812$; $VIF_5 = \frac{1}{1-{R_5}^2} = 1.317664$

Test criteria: if $VIF_j > 10$ then H_0 rejected.

Results and Conclusions: H_0 accepted because the value VIF_1 for the variable X_1 is $1.243473 < 10$; VIF_2 for the variable X_2 is $1.602682 < 10$; VIF_3 for the variable X_3 is $2.527059 < 10$; VIF_4 for the variable X_4 is $3.284812 < 10$; VIF_5 for the variable X_5 is $1.317664 < 10$ so there is no multicollinearity.

3. Homoskedasticity Test

Hypothesis: $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \cdots = \sigma_n^2 = \sigma^2$ (homokedasticity occurs)

H_1 : there is at least one $\sigma_i^2 \neq \sigma^2$(no homocedasticity)

Significance Level: α = 5%

Test Statistics:

$LM = nR^2 = 10{,}801$; $P_{value} = 0.05548$

From the table Chi Square ($\chi^2_{\alpha,k}$) with α = 5% and $k = 5$, the value $\chi^2_{0.05,5} = 11{,}070$.

Test criteria: if $LM > \chi^2_{\alpha,k}$ and $P_{value} < \alpha$ then H_0 rejected.

Decision and Conclusion: H_0 accepted because the value $LM < \chi^2_{0.05,5} = 10.801 < 11.070$ or the value $P_{value} > \alpha = 0{,}05548 > 0{,}05$ so there is homokedasticity occurs.

4. Non-Autocorrelation Test

Hypothesis: $H_0 : \rho = 0$ (no autocorrelation)

$H_1 : \rho \neq 0$ (autocorrelation)

Significance Level: α = 5%

Test Statistics:

$DW = \frac{\sum_{i=2}^{n}(\varepsilon_i - \varepsilon_{i-1})^2}{\sum_{i=1}^{n}{\varepsilon_i}^2} = 2{,}1349$; $P_{value} = 0{,}5565$

From Durbin Watson's table with α = 5%, $n = 27$, and $k = 5$. The values dL = 1,0042 and dU = 1,8608.

Test Criteria:

$0 < DW < dL$ (H_0 rejected, there is positive autocorrelation) and $P_{value} < \alpha$

dL < DW < dU (No conclusion)

dU < DW < 4-dU (H_0 accepted, there is no autocorrelation) and $P_{value} > \alpha$

4-dU < DW < 4-dL (No conclusion)

4-dL < DW < 4 (H_0 rejected, there is negative autocorrelation) and $P_{value} < \alpha$

Results and Conclusions: H_0 accepted because the value of dU (1.8608) < DW (2.1349) < 4-dU (2.1392) or value $P_{value} > \alpha$ = 0,5565 > 0,05 so there is no autocorrelation.

There are two types of hypothesis tests, namely simultaneous hypothesis tests (F test) and partial hypothesis tests (t test).

- **F Test**

Hypothesis: H_0 : β_j= 0 (linear regression model does not fit)

H_1 : $\beta_j \neq 0$ for at least one, $j = 1,2,\ldots,k$ (regression model fit)

Significance Level: α = 5%

Test Statistics:

$F_{hitung} = \frac{KTR}{KTG}$ = 3,372; P_{value} = 0.02162

$F_{tabel}(F_{(\alpha,k,n-k-1)})$ obtained from the distribution table F with α = 5%, $n = 27$, and $k = 5$ so that the value of $F_{(0.05,5,21)}$ = 2,68.

Test criteria: if $F_{hitung} > F_{tabel}(F_{(\alpha,k,n-k-1)})$ or $P_{value} < \alpha$ then H_0 is rejected.

Decision and Conclusion: H_0 rejected because the value of $F_{hitung} > F_{tabel}$ ($F_{(0.05,5,21)}$) = 3,372 > 2,68 or value $P_{value} < \alpha$ = 0,02162 < 0,05 can be concluded that the linear regression model fits so that the independent variables simultaneously have an influence on the dependent variables.

- **T Test**

Hypothesis: H_0 : β_j= 0, $j = 1,2,..,k$ (parameter coefficients are not significant)

H_1 : $\beta_j \neq 0$, $j = 1,2,\ldots,k$ (parameter coefficients are significant)

Significance Level: α = 5%

Test Statistics: Based on Table 4, the results of Equation (8) are obtained as follows:

Table 4 - Partial Parameter Significance Test Results.

Variabel	Estimation	Std. Error	t_i	P_{value}
X_1	-63,3853	35,8635	-1,767	0,091691
$\boldsymbol{X_2}$	2,1844	1,1458	1,906	0,070371
$\boldsymbol{X_3}$	2,1616	2,7594	0,783	0,442162
X_4	-2,5810	1,0295	-2,507	0,020458
X_5	-2,9653	0,9322	-3,181	0,004497

t_{tabel} ($t_{(\alpha/2,n-k-1)}$ obtained from the distribution table t with α = 5%, $n = 27$, and $k = 5$ so that the value is $t_{(0.025,21)}$ = 2,07961.

Test Criteria: if $|t_j| > t_{tabel}$ ($t_{(\alpha/2,n-k-1)}$) or $P_{value} < \alpha$ then H_0 rejected.

Decision and Conclusion: H_0 rejected for the independent variable X_4 and X_5 because $|t_j| > t_{tabel}$ ($t_{(0.025,21)}$) so that partially there is influence on the dependent variable Y.

The Adjusted R-squared value shows that R^2 is 0.3133 or 31.33% which means that 31.33% of the variance in the Y variable can be explained by the X variable, while the remaining 68.67% is influenced by other factors outside this research model. According to Chin (1998), this value of R^2 is classified as a weak category. The increase in values was analyzed using the Clusterwise Linear Regression method assuming that the observation data formed clusters that must be separated so that the best regression model is obtained with R^2 which is close to 1.

Further analysis need standardization using Z-score values because the data used has different units. The CLR method is used to generate the best model when the observational data forms clusters. The conditions required to confirm the existence of the Clusterwise Linear Regression method are $n \geq KJ = 27 \geq K \times 6 = K \leq 4{,}5$ so that the number of possible clusters is $K = 2{,}3{,}4$. The best model is indicated by the minimum AIC value at each number of clusters formed. The value of the Akaike Information Criterion can be presented in Table 5.

Table 5 - Akaike Information Criterion (AIC).

K	AIC
2	35,69484
3	-6,88821
4	13,76395

From Table 5, the best model is obtained in cluster $K = 3$ with a minimum AIC value of -6.88821 so that there are 3 optimal number of clusters.

The results of the Clusterwise Linear Regression analysis show that:

1. Cluster 1 has a cluster size of 11 districts/cities which include Bogor, Sukabumi, Cianjur, Garut, Tasikmalaya, Majalengka, Subang, Bogor City, Cirebon City, Depok City, and Cimahi City. The models formed are:

$$Z_Y = -0{,}25914452 - 0{,}20649973Z_1 + 0{,}65275533Z_2 + 0{,}02853121Z_3 - 0{,}43474849Z_4 - 0{,}04465134Z_5 + \varepsilon$$

Hypothesis: H_0: $\beta_{jc} = 0$

H_1 : $\beta_{jc} \neq 0$, with j =1,2, …, k and c = 1,2, …, K

Significance Level: α = 5%

Test Statistics: Based on Table 6, the results of Equation (31) are obtained as follows:

Table 6 - Results of Cluster 1 Parameter Significance Test.

Variabel	Estimation	Std. Error	Z	P_{value}	Results
X_1	-0,207343	0,039274	-5,2794	$1{,}296 \times 10^{-7}$	H_0 rejected
$\boldsymbol{X_2}$	0,654259	0,063664	10,2767	$< 2.2\times \mathbf{10^{-6}}$	$\mathbf{H_0}$ rejected
$\boldsymbol{X_3}$	0,027990	0,052860	0,5295	0,5964	H_0 Accepted
X_4	-0,433882	0,054377	-7,9792	$1{,}473\times 10^{-5}$	H_0 rejected
X_5	-0,045363	0,047073	-0,9637	0,3352	H_0 Accepted

Test criteria: if $|Z| > Z_{tabel}$ $(Z_{\alpha/2})$ atau the value $P_{value} < \alpha$ then H_0 rejected.

Z_{tabel} $(Z_{\alpha/2})$ obtained from the distribution table Z with α = 5%, it can be seen that the value is $Z_{\alpha/2}$ = 1,96 so that in cluster 1 for the standardized independent variable Z_1, Z_2, and Z_4 there is an influence on the model because the value of $|Z| > Z_{\frac{\alpha}{2}}$ or $P_{value} < \alpha$, while Z_3 (number of flood disasters) and Z_5 (percentage of feasible sanitation) there is no influence on the model because the value of $|Z| < Z_{\frac{\alpha}{2}}$ or $P_{value} > \alpha$.

The Adjusted R-squared value in cluster 1 shows that R^2 is 0.9226919 means that 92.26919% of the dependent variable variant can be explained by the independent variable, while the remaining 7.7308% is influenced by other factors outside the research model.

2. Cluster 2 has a cluster size of 8 districts/cities which include Cirebon, Sumedang, Purwakarta, Bekasi, West Bandung, Pangandaran, Bandung City, and Banjar City. The models formed are:

$$Z_Y = 0{,}155009675 - 0{,}723833960Z_1 + 0{,}222176937Z_2 + 0{,}412007591Z_3 - 0{,}689087285Z_4 - 0{,}561149757Z_5 + \varepsilon$$

Hypothesis: H_0: $\beta_{jc} = 0$

H_1 : $\beta_{jc} \neq 0$, with j =1,2, …, k and c = 1,2, …, K

Significance Level: α = 5%

Test Statistics: Based on Table 7, the results of Equation (31) are obtained as follows:

Table 7 - Results of Cluster 2 Parameter Significance Test.

Variabel	Estimation	Std. Error	Z	P_{value}	Results
X_1	-0,7237422	0,0032299	-224,074	$< 2.2\times 10^{-6}$	H_0 rejected
$\boldsymbol{X_2}$	0,2219725	0,0044238	50,177	$< 2.2\times \mathbf{10^{-6}}$	$\mathbf{H_0}$ rejected
$\boldsymbol{X_3}$	0,4113756	0,0087822	46,842	$< 2.2\times 10^{-6}$	H_0 rejected
X_4	-0,6887045	0,0091041	-75,648	$< 2.2\times 10^{-6}$	H_0 rejected
X_5	-0,5606493	0,0071777	-78,110	$< 2.2\times 10^{-6}$	H_0 rejected

Test criteria: if $|Z| > Z_{tabel}$ $(Z_{\alpha/2})$ atau the value $P_{value} < \alpha$ then H_0 rejected.

Conclusion: $Z_{tabel}(Z_{\alpha/2})$ obtained from the distribution table Z with $\alpha = 5\%$, it can be seen that the value of $Z_{\alpha/2} = 1{,}96$ so that cluster 2 for all standardized independent variables $(Z_1, Z_2, Z_3, Z_4, Z_5)$ there is an influence on the model because $|Z| > Z_{\frac{\alpha}{2}}$ or $P_{value} < \alpha$.

The Adjusted R-squared value in cluster 2 shows that R^2 is 0.7121607 means that 71.21607% of the dependent variable variants can be explained by independent variables, while the remaining 28.7839% is influenced by other factors outside the research model.

3. Cluster 3 has a cluster size of 8 districts/cities which include Bandung, Ciamis, Kuningan, Indramayu, Karawang, Sukabumi City, Bekasi City, and Tasikmalaya City. The models formed are:

$$Z_Y = 0{,}25620603 + 0{,}20057410Z_1 + 0{,}54183874Z_2 - 0{,}19021501Z_3 + 0{,}02441280Z_4 - 1{,}42951940Z_5 + \varepsilon$$

Hypothesis testing in cluster 3

Hypothesis: H_0: $\beta_{jc} = 0$

H_1 : $\beta_{jc} \neq 0$, with j =1,2, …, k and c = 1,2, …, K

Significance Level: α = 5%

Test Statistics: Based on Table 8, the results of Equation (31) are obtained as follows:

Table 8 - Results of Cluster 3 Parameter Significance Test.

Variabel	Estimation	Std. Error	Z	P_{value}	Results
X_1	0,200588	0,025444	7,8834	$3{,}185\times 10^{-5}$	H_0 rejected
$\boldsymbol{X_2}$	0,541986	0,015890	34,1077	$< 2.2\times \mathbf{10^{-6}}$	$\mathbf{H_0}$ rejected
$\boldsymbol{X_3}$	-0,190068	0,022634	-8,3975	$< 2.2\times 10^{-6}$	H_0 rejected
X_4	0,024408	0,042350	0,5763	0,5644	H_0 Accepted
X_5	-1,429686	0,028568	-50,0450	$< 2.2\times 10^{-6}$	H_0 rejected

Test criteria: if $|Z| > Z_{tabel}$ $(Z_{\alpha/2})$ atau the value $P_{value} < \alpha$ then H_0 rejected.

Conclusion: $Z_{tabel}\left(Z_{\frac{\alpha}{2}}\right)$ obtained from the distribution table Z with $\alpha = 5\%$, it can be seen that the value is $Z_{\frac{\alpha}{2}} = 1{,}96$ so that in cluster 3 for standardized independent variables Z_1 (population growth rate), Z_2 (number of hospitals), Z_3 (number of floods), and Z_5 (percentage of decent sanitation) there is an influence on the model because the value of $|Z| > Z_{\frac{\alpha}{2}}$ or $P_{value} < \alpha$, while Z_4 (number of health centers) has no influence on the model because of the value $|Z| < Z_{\frac{\alpha}{2}}$ or $P_{value} > \alpha$.

The Adjusted R-squared value in cluster 3 shows that R^2 is 0.9140601 means that 91.40601% of the dependent variable variants can be explained by independent variables, while the remaining 8.594% is influenced by other factors outside the research model.

Cluster profiling is represented by means of the average of each cluster that is formed.

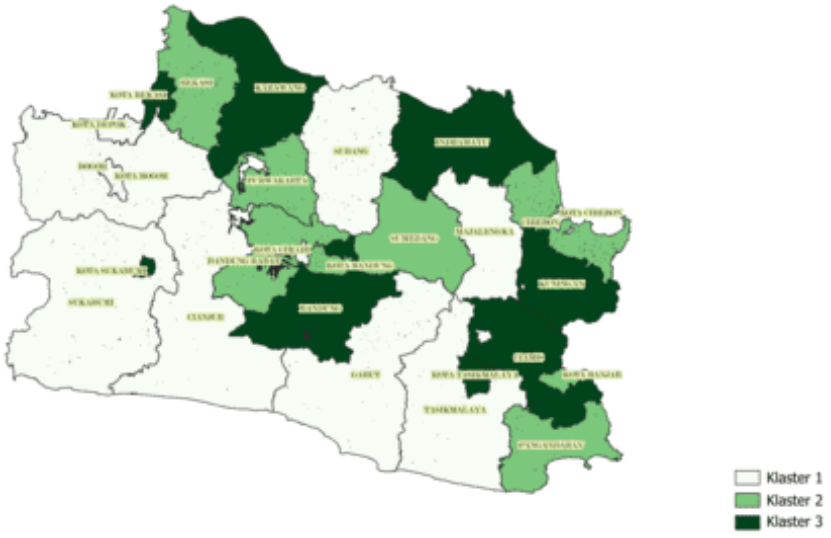

Fig. 1 – Clustering Result Map

Table 9 - Average Variables in Each Cluster

Variabel	Information	Cluster 1	Cluster 2	Cluster 3
X_1	Number of dengue cases	*60,5*	99	**130**
$\boldsymbol{X_2}$	Population growth rate	*1,29*	1,30	**1,31**
$\boldsymbol{X_3}$	Number of hospitals	*12,1*	**16,4**	**16,4**
X_4	Number of flood disasters	**8,91**	*5,62*	6,38
X_5	Number of health centers	**43,9**	*37,2*	40

Description: **Bold** means highest value, *Italic* means lowest value

Based on Table 9, cluster 3 has the highest average number of Dengue Hemorrhagic Fever (DHF) cases among other clusters. The high number of dengue cases in cluster 3 is indicated by low sanitation, a very high population growth rate, and a significant number of flood disasters. This condition makes cluster 3 a cluster that needs more attention from the Government and the Health Office.

CONCLUSION

Clustering districts/cities in West Java Province uses Clusterwise Linear Regression method with the minimum Akaike Information Criterion (AIC) value resulted three optimal clusters. Cluster 1 is 11 districts/cities which include Bogor, Sukabumi, Cianjur, Garut, Tasikmalaya, Majalengka, Subang, Bogor City, Cirebon City, Depok City, and Cimahi City. Cluster 2 is 8 districts/cities which include Cirebon, Sumedang, Purwakarta, Bekasi, West Bandung, Pangandaran, Bandung City, and Banjar City. Cluster 3 is 8 districts/cities covering Bandung, Ciamis, Kuningan, Indramayu, Karawang, Sukabumi City, Bekasi City, and Tasikmalaya City. The factors that affect the number of dengue cases in each cluster show that cluster 1 is influenced by a variables Z_1, Z_2, Z_4 with a value R^2 of 92.3%; cluster 2 is influenced by all standardized independent variables (Z_1, Z_2, Z_3, Z_4, Z_5) with a value R^2 of 71.2%; and cluster 3 is influenced by a variables Z_1, Z_2, Z_3, Z_5 with a value R^2 of 91.4%.

References

Badan Pusat Statistik Provinsi Jawa Barat. (2022). *Kasus penyakit menurut kabupaten/kota dan jenis penyakit di Provinsi Jawa Barat*. Retrieved from https://jabar.bps.go.id/id

Chin, W. W. (1998). The partial least squares approach for structural equation modeling. In G. A. Marcoulides (Ed.), *Modern methods for business research* (pp. 295–336). Mahwah, NJ: Lawrence Erlbaum Associates.

DeSarbo, W. S., & Cron, W. L. (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification*, *5*(2), 249–282. https://doi.org/10.1007/BF01897167

Elizabeth, A. H., & Yudhastuti, R. (2023). Gambaran kasus demam berdarah dengue (DBD) di Provinsi Jawa Barat tahun 2016–2020. *Media Gizi Kesmas*, *12*(1), 179–186. https://doi.org/10.20473/mgk.v12i1.2023.179-186

Ghozali, I. (2018). *Aplikasi analisis multivariate dengan program IBM SPSS* (Edisi ke-9). Semarang: Badan Penerbit Universitas Diponegoro.

Ikbal, A., Purnamasari, A. I., & Ali, I. (2024). Analisis klasterisasi untuk prediksi jumlah kasus DBD berdasarkan jenis kelamin dan kabupaten/kota di Jawa Barat. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *7*(6), 3789–3796. https://doi.org/10.36040/jati.v7i6.8296

Kementerian Kesehatan. (2022). *Profil kesehatan Indonesia 2022*. Jakarta: Kementerian Kesehatan Republik Indonesia.

Meylisah, E., Rini, D. S., Fransiska, H., Agwil, W., & Sartono, B. (2023). Modeling clusterwise linear regression on poverty rate in Indonesia. *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, *17*(3), 1653–1662. https://doi.org/10.30598/barekengvol17iss3pp1653-1662

Oroh, M. Y., Pinontoan, O. R., & Tuda, J. B. S. (2020). Faktor lingkungan, manusia dan pelayanan kesehatan yang berhubungan dengan kejadian demam berdarah dengue. *Indonesian Journal of Public Health and Community Medicine*, *1*(3), 35–46.

Putri, M. S., Sartono, B., & Susetyo, B. (2015). *Analisis regresi linear gerombol dengan algoritma pertukaran (exchange algorithm)* (Undergraduate thesis, Institut Pertanian Bogor).

World Health Organization (WHO). (2022). *Dengue and severe dengue*. Retrieved from https://www.who.int/newsroom/fact-sheets/detail/dengue-and-severe-dengue.