Clusterwise Regression Model for Statistical Downscaling to predict Daily Rainfall using Gamma Distribution

Muthia Nadhira F, Agus M Soleh*, Anik Djuraidah
Department of Statistics, IPB University, Indonesia

Abstract. This study presents the development of clusterwise regression with gamma distribution and normal distribution. This clusterwise regression model is expected to be a solution for high variance data in order to minimize the error rate in modeling. The first study is simulation of clusterwise regression with gamma distribution and normal-gamma mixture. This simulation is divided into two-cluster and three-cluster simulations with gamma distribution, normal distribution, and mixed normal-gamma distribution with different parameters. The application of clusterwise regression was carried out at four rain stations in West Java, including Bandung, Bogor, Citeko, and Jatiwangi rain stations by first doing variable reduction using principal component analysis.

1. Introduction
Indonesia is an agricultural country where agriculture plays a role important of the national economy. Rainfall will certainly have an impact in agriculture whose productivity is influenced by weather and water availability. Very high rainfall intensity has the opportunity to fail harvests and the low rain intensity will also have the potential for drought. Therefore, information regarding the estimated rainfall is needed to anticipate or reduce the impact of disasters and support the success of agricultural production. However, geographical conditions, land, sea, atmosphere the complexity of the prediction of rainfall in the Indonesian territory. Statistical downscaling (SD) method is an approach that can handle the problem of low rainfall prediction accuracy. This method makes GCM data information from a global scale projected against local scale information at a climatological station [11].

GCM predictors contain high multicollinearity [10]. Multicollinearity will cause the model becoming unstable, and one solution to overcome it is to reduce predictor mods using Principal Component Analysis (PCA) by transforming into mutually orthogonal mods or performing projections of large-dimensional early predictors into some small-dimensional new variables called principal components. In the GCM output and local data, rainfall has a very large variety, so a solution is needed to reduce this variance. As one solution to reducing variance in order to reduce the error rate is to use clustered data or create clustering and regression processes simultaneously. One way of classification is clusterwise regression. Clusterwise regression is a multivariate statistical procedure that attempts to cluster objects with the objective of minimizing the error rate for the within-clusterwise regression models [2]. Clusterwise regression is a combination derived from two techniques, there are clustering and regression analysis. It can also estimate the number of clusters and regression parameters in each group together so that it is efficient to use on data sets that require two or more regression functions to determine the data structure [1].
There has been a development of statistical downscaling research with cluster analysis with gamma distribution and the resulting RMSE value is 6%. However, this method has not been able to make predictions as measured by the RMSEP value [8]. Estimation of rainfall using clusterwise regression, principal component regression (PCR) and partial least squares regression (PLSR) concluded cluster regression was better than PCR and PLSR. Estimation of rainfall using clusterwise regression, principal component regression (PCR) and partial least squares regression (PLSR) concluded cluster regression was better than PCR and PLSR [3].

In this paper, we present a simulation of clusterwise regression development with a gamma distribution using log-link function and applied clusterwise regression at four rainfall stations from Badan Meteorologi Klimatologi dan Geofisika (BMKG) stations in West Java Province, there are Bandung, Bogor, Citeko, and Jatiwangi stations. To predict the future rainfall, a prediction method based on the distance between observations and centroids of each class is used, and compared with a prediction method based on the distance between observations and observations in the clusters.

2. Research Methods

2.1. Data
In this study, two types of data were used, GCM (Generalize Circulation Model) output data and local rainfall data. GCM data is a model that describes the global interactions between land, oceans, and air on earth. The parameter used from the GCM data is the precipitation rate. The type of GCM output used is CFS (Climate Forecast System). The domain in the CFS data used in this study is 6 × 6 in size 0.5° × 0.5° grid space. The period time used is from January 2011 to December 2019. Response data is daily rain station data published by the Badan Meteorologi Klimatologi dan Geofisika (BMKG) stations in West Java province. The rain station selected are Bandung, Bogor, Citeko, and Jatiwangi rain stations with latitude and longitude points as follows:

- Bandung : (-6.883, 107.597)
- Bogor : (-6.550, 106.750)
- Citeko : (-6.700, 106.850)
- Jatiwangi : (-6.734, 108.263)

Figure 1. Map of the rainfall station

2.2. Data Analysis Procedure
2.2.1. Develop Clusterwise Regression
The initial stage of data analysis is to determine the number of clusters $J$. Then each observation is grouped into one of the $J$ group randomly. After obtaining the initial partition, a regression model was carried out for each group. Estimating of parameters using the Generalized Linear Model (GLM). GLM is a generalizing of the linear model, in this case the response variable comes from the exponential distribution family and there is a relationship function that connects the expected value with the systematic component of the linear model [6].

The general form of the density function of the random variable of the exponential distribution family by adding the scale parameter (constant) is $\varphi$ [5] as follows:

$$ f(y_i; \theta_i, \varphi) = e^{\frac{y_i^{\beta_i} - b(\theta_i)}{a_i(\varphi)} + c(y_i, \varphi)} $$

in this case $a(\cdot), b(\cdot)$ and $c(\cdot)$ are certain functions, $\theta$ is a canonical parameter, and $\varphi$ is a constant disperse parameter [7]. The function of the gamma distribution density is as follows:

$$ f(y; \nu, \xi) = \frac{y^{\xi-1}e^{-\nu y}}{\Gamma(\xi)} $$
In this case, $\nu$ is the rate parameter and $\xi$ is the shape parameter. To perform the most streamlined linear modeling, it is necessary to re-parameterize the parameter $\nu$ with $\nu = \frac{\xi}{\mu}$ and the parameter $\xi$ (shape) is constant [5]. The canonical link function for the gamma distribution is [5]:

$$\eta_i = g(\mu_i) = \frac{1}{\mu_i} = \Sigma_{j=1}^{p} x_{ij} \beta_j$$  \hspace{1cm} (3)$$

Since the link function uses $\log \rightarrow g(\mu_i) = \log(\mu_i)$, then:

$$\mu = \left( \log \Sigma_{j=1}^{p} x_{ij} \beta_j \right)^{-1} = e^{\Sigma_{j=1}^{p} x_{ij} \beta_j}$$  \hspace{1cm} (4)$$

The normal distribution density function is as follows:

$$f(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma^2} e^{-(y-\mu)^2/2\sigma^2}$$  \hspace{1cm} (5)$$

Where $\sigma^2$ is the variance and $\mu$ is the expected value of the random variable $Y$. The canonical link function for normal distribution is $\eta_i = g(\mu_i) = \mu_i$, then:

$$\mu = \Sigma_{j=1}^{p} x_{ij} \beta_j$$  \hspace{1cm} (6)$$

After the model in each cluster is obtained, then calculate the error for each observation on each model that is formed. Each observation will move to the cluster with the least error. This process is carried out on all observations. After all the observations have moved, a new partition will be formed then estimate a regression model was performed on each new cluster. To determine the performance of the classification using the accuracy between the actual data and data from the clusterwise regression process.

$$\text{accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$  \hspace{1cm} (7)$$

**Figure 2.** Flowchart of clusterwise regression
2.2.2. Applying clusterwise regression algorithm to rainfall data

The first stage is analyze GCM data as predictor data using Principal Component Analysis. Principal Component Analysis (PCA) was used to reduce the predictor variable. Principal Component Analysis forms a new variable which is a linear combination of all the original variables, which is called the principal component. Although it takes \( p \) components to represent the overall variation in the data, often this variation can be represented by the \( k \) principal components, with \( k \ll p \) [4]. Then combining local rainfall data as a response variable and principal component from the previous analysis.

Data exploration is carried out by looking at the data distribution patterns with descriptive statistics. Predictions of rainfall data is measured using RMSEP (root mean square error prediction). The prediction uses 2 distance methods, the first method calculates the distance of observations with the centroid of each cluster (method A), the distance method used is the euclidean distance which can be used if all the variables used are continuous [9].

\[
d_{(i,j)} = \sqrt{\sum_{i=1}^{p}(x_{ij} - \bar{x}_{ij})^2}
\]

\( x_{ij} \) is an observation on the data testing variable \( j \) on the \( i \) cluster and \( \bar{x}_{ij} \) is mean on the data training variable \( j \) on the \( i \) cluster. The second method calculates the distance of observations with each observation in each cluster (method B).

3. Results and Discussion

In this paper, the number of clusters used is two and three clusters. For the two-cluster model, are Gamma model (GG), Normal model (NN) and Mixed Gamma-Normal model (GN). The three-cluster model are Gamma model (GGG), Normal model (NNN), Gamma-Normal-Normal mixed model (GNN), and Gamma-Normal-Normal mixed model (GGN). Starting with generate 200 observational data in each Model. The cluster information formed from the cluster of data can be detected by scatter plot between the response variable (\( Y \)) and the predictor variable (\( X \)). This simulation is done repetition of 50 times.

3.1. Simulation of Two Clusters

Clusterwise regression simulation was formed by generating a data set of two gamma distribution:

\[
y_1 = -2 + 0.25x_1 ; \xi = 0.5 \text{ and } y_2 = 0 + 0.5x_2 ; \xi = 15
\]

![Figure 3. Clusterwise regression simulation with gamma distribution](image)

Clusterwise regression simulation was formed by generating a data set of two normal distribution:

\[
y_1 = 5 + 4x_1 + \varepsilon \text{ and } y_2 = 10 + 10x_2 + \varepsilon
\]
Clusterwise regression simulation was formed by generating a data set of gamma-normal distribution:
\[ y_1 = 0 + 0.5.x_1 : \xi = 15 \text{ and } y_2 = 1 + 4.x_2 + \varepsilon \]

---

Table 1. Summary of accuracy of 2 cluster simulations

| no | Model | Scenario                        | Gamma Accuracy | Normal Accuracy | GN Accuracy |
|----|-------|---------------------------------|----------------|----------------|-------------|
| 1  | GG    | \( Y \sim \text{Gamma} (0.5, \nu) \) | 96.41 %        | 87.03 %        | 85.93 %     |
| 2  | NN    | \( \varepsilon \sim \text{Normal} (0.1) \) | 56.00 %        | 98.40 %        | 69.39 %     |
| 3  | GN    | \( Y \sim \text{Gamma} (15, \nu) \) | 70.14 %        | 61.15 %        | 84.90 %     |

The simulation of two-clusters, results of all models indicate that clusterwise regression is able to properly separate the data according to the actual distribution. For The simulation of two-clusters with Gamma distribution, the highest accuracy is in the Gamma cluster model, as well as in the normal data distribution and mixed data distribution. It can be concluded that the algorithm formed is able to classify data according to the characteristics of the data.

The generation data in each model has the same proportion of 50% for each distribution. In this two-cluster simulation, the accuracy value is very high because all accuracy values are above 0.90. The GG model simulation has a high average accuracy value on the gamma model of 96.41%. The simulation of the NN model has an almost perfect average value of accuracy in the normal model of 98.40%. The GN simulation has a high average accuracy value on the gamma-normal model of 84.90%. The clusterwise regression performance is very good on all models. It can be seen from Figure 3 that clusterwise regression can store data properly according to a good distribution, as well as in Figure 4 and Figure 5.
3.2. Simulation of Three Clusters
Clusterwise regression simulation was formed by generating a data set of three gamma distribution:
\[ y_1 = -2 + 0.25x_1 ; \xi = 0.5 \quad y_2 = 0.5 + 0.48x_2 ; \xi = 15 \quad y_3 = 2 + 0.4x_3 ; \xi = 15 \]

![Figure 6. Clusterwise regression simulation with gamma distribution](image)

Clusterwise regression simulation was formed by generating a data set of three normal distribution:
\[ y_1 = 0 + 4x_1 + \varepsilon \quad y_2 = 1 + 10x_2 + \varepsilon \quad y_3 = 3 + 15x_3 + \varepsilon \]

![Figure 7. Clusterwise regression simulation with normal distribution](image)

Clusterwise regression simulation was formed by generating a data set of gamma-normal-normal distribution:
\[ y_1 = 0 + 0.6x_1 ; \xi = 1 \quad y_2 = 150 + (-10).x_2 + \varepsilon \quad y_3 = 160 + 50.x_3 + \varepsilon \]

![Figure 8. Clusterwise regression simulation with normal-normal-gamma distribution](image)
Clusterwise regression simulation was formed by generating a data set of gamma-gamma-normal distribution:

\[ y_1 = -2 + 0.25x_1; \xi = 0.5, \quad y_2 = 0 + 0.48x_2; \xi = 15 \quad \text{and} \quad y_3 = 50 + (-3)x_3 + \epsilon \]

The generation data in each model has the same proportion of 33.3% for each distribution. In this three-cluster simulation, the accuracy value is not as big as the two-cluster simulation, but still produces a fairly good accuracy. The three-distribution clusterwise regression simulation consists of four models, there are GGG models, NNN models, GNN models and GGN models. The results for all models indicate that cluster regression is able to properly separate the data according to the actual distribution.

The simulation of three gamma distributions produces a fairly good average accuracy value, the GGG model has the highest average accuracy value on the gamma model of 68.05%, the NNN model has the highest average accuracy value on the normal model of 95.94%. The GNN model has the highest average accuracy value on the GNN model of 81.37%. The GGN model has the highest average accuracy value on the GGN model of 92.65%. It shows that the model of each clusterwise regression is a model that matches the generated data which also has the same distribution. It can be concluded that the algorithm formed is able to classify data according to the characteristics of the data.

### 3.3. Clusterwise Regression application on rainfall data

Clusterwise regression applied to daily rainfall data. The response data used is daily rainfall data sourced from BMKG (Meteorology, Climatology and Geophysics Agency) starting from 2010 to 2019 at the four selected rain stations in West Java province, there are Bandung, Bogor, Citeko and Jatiwangi rain stations. The distribution pattern of rainfall for each rain station is depicted in boxplot (Figure 10). Based on Figure 10, all rain stations have outlier data that sticks out to the right indicating the amount of high rainfall or extreme rainfall. The location of boxplot tends to be on the left, meaning that the data tends to gather below, rainfall data that is in the range of 0 to 20 mm.
The rain station with the highest variance is at the Bogor rain station with a standard deviation of 20.60 and the lowest is at the Bandung rain station with a standard deviation of 14.33. The area with the highest average rainfall is Bogor at 8.3 mm/day. The highest maximum rainfall for the last 10 years was at Citeko station at 192.8 mm/day while the lowest maximum rainfall for the last 10 years was at Bandung station at 122.9 mm/day. After that, perform predictor data reduction using principal component analysis.

**Table 3.** Statistic descriptive of data

| Stasiun   | Minimum | Median | Mean  | Maximum | Std. Deviasi |
|-----------|---------|--------|-------|---------|--------------|
| Bandung   | 0       | 1.6    | 8.306 | 122.9   | 14.33169     |
| Bogor     | 0       | 4.3    | 13.17 | 169.1   | 20.60330     |
| Citeko    | 0       | 2.5    | 10.07 | 192.8   | 17.12798     |
| Jatiwangi | 0       | 0.0    | 7.863 | 187.8   | 17.28319     |

To determine the number of principal components using eigenvalue and cumulative proportion of variety. From the scree plot graph, it is determined from eigenvalue of more than 1, based on the proportion of the cumulative variety if the value is 75%. To make it easier to see these determinations, they are presented in a scree plot and a cumulative table.

**Table 4.** Principal components

| Rain station | Com.1 | Com.2 | Com.3 | Com.4 | Com.5 | Com.6 | Com.7 | Com.8 |
|--------------|-------|-------|-------|-------|-------|-------|-------|-------|
| Bandung      | 0.358 | 0.486 | 0.578 | 0.655 | 0.708 | 0.751 | **0.790** | 0.820 |
| Bogor        | 0.370 | 0.499 | 0.602 | 0.663 | 0.719 | **0.762** | 0.797 | 0.825 |
| Citeko       | 0.370 | 0.499 | 0.602 | 0.663 | 0.719 | **0.762** | 0.797 | 0.825 |
| Jatiwangi    | 0.395 | 0.532 | 0.620 | 0.682 | 0.729 | **0.770** | 0.803 | 0.833 |
Determining the main component is seen from the eigenvalue, provided that the selected component is a component with an eigenvalue > 1. Based on table 4, for Bandung rain station, 7 principal components were selected based on the screeplot and also the cumulative proportion value of 79%. For Bogor, Citeko and Jatiwangi rain station, 6 principal components were selected based on the screeplot and also the value of the cumulative proportion of 76%. For Bogor and Citeko have the same latitude and longitude, so the selected components will be the same, and proportion of diversity of 77% for Jatiwangi rain station. After obtaining the main components, the data will be analyzed using clusterwise regression.

**Table 5. Prediction of rainfall**

| Rain station | Cluster |      |      |      |      |      |      |      |      |      |      |      |
|--------------|---------|------|------|------|------|------|------|------|------|------|------|------|
|              |         | 1 A  | 2 (GN) | B  | 3 (GGN) | A | B | 3 (GNN) | A | B | A | B |
| Bandung      | 20.430  | 13.512 | 13.445 | 13.760 | **13.406** | 13.790 | 13.666 |
| Bogor        | 21.352  | 18.901 | 18.871 | **18.596** | 18.717 | 18.675 | 18.855 |
| Citeko       | 37.152  | 15.339 | 15.457 | 15.205 | **15.122** | 15.295 | 15.202 |
| Jatiwangi    | 38.808  | 15.404 | 15.357 | **15.345** | 15.668 | 15.736 | 15.784 |

The model without clustering is formed from the regression of the Gamma distribution. The two-cluster model is the Gamma-Normal (GN) mixed model. The three-cluster model consists of two models, the first model formed from two groups of Gamma regression and one distribution normal regression (GGN). The second model is formed from two groups of normal regression and one spread Gamma regression (GNN). The model performance is evaluated by comparing the observed and predicted rainfall using the RMSEP (Root Mean Squared Error Prediction). Repetition for each rain station is done 50 times for each cluster model. Based on table 5, the column is without clusters, which means that only Gamma regression analysis is carried out and produces high RMSEP values for all rain stations. The two-clusters model is the Gamma-Normal (GN) model and the three-clusters are Gamma-Normal-Normal (GNN) and Gamma-Gamma-Normal (GGN) models. The highest RMSEP value without cluster is at Jatiwangi station at 38.80 and the smallest is at Bandung rain station at 20.43. Two or three clusters produce lower RMSEP values than those without clustering. There was no significant difference in the RMSEP value between two groups or three groups at all rain stations.

The model without clustering produces an average RMSEP of 20.43 at the Bandung rain station. The best model for the Bandung rain station is the GNN prediction B model with an average RMSEP value of 13.40. The average value of RMSEP without clustering is 21.35 at the Bogor rain station. The best model for the Bogor rain station is the GNN model prediction A with an average RMSEP value of 18.59. The Citeko rain station modeled without clustering produces an average RMSE of 37.15. The best model for the Citeko rain station is in the GNN prediction B model with an average RMSEP value.
of 15.12. Jatiwangi rain station modeled without clustering produces an average RMSE of 38.8. The best model for the Jatiwangi rain station is in the GNN model prediction A with an average RMSEP value of 15.34.

4. Conclusion
The simulation results show that each new partition is almost similar to the actual data distribution. The results presented in this paper indicate that a cluster model simulation with a gamma distribution and a normal distribution is able to group data into a model that can be a solution for data with high variance. The application of cluster regression results in the Gamma-Normal-Normal (GNN) model which is the most suitable model for modeling rainfall in all rain stations with the smallest RMSEP compared to the Gamma-Normal (GN) and Gamma-Gamma-Normal (GGN) models. Therefore, the mixed model with Gamma and normal distribution can be said to be the best model for estimating daily rainfall when the Gamma model is not yet able to produce the best model.

References
[1] Bagirov A, Mahmood A, and Barton A 2017 Prediction of monthly rainfall in Victoria, Australia: Clusterwise linear regression approach, *Atmospheric Research* **188** 20-29
[2] Brusco MJ, Cradit JD, Steinley D and Fox GL 2008 Cautionary Remarks on the Use of Clusterwise Regression, *Multivariate Behavioral Research*, 43:1, 29-49
[3] Butar-butar VP, Soleh AM, and Wigena AH 2019 Pemodelan Clusterwise Regression Pada Statistical downscaling Untuk Pendugaan Curah Hujan Bulanan. Indonesian Journal of Statistics and Its Applications, 3(3), 236–246
[4] Jolliffe I 2002 *Principal Component Analysis 2nd Edition* (New York: Springer)
[5] Lindsey JK 1997 *Applying Generalized Linear Models* (New York: Springer)
[6] Soleh AM, Wigena AH, Djuraidah A and Saefudin A 2015 Statistical downscaling to predict monthly rainfall using linear regression with L1 regularization (LASSO) *Applied Mathematics Sciences* **9**(108) 5361-5369
[7] Soleh AM, Wigena AH, Djuraidah A and Saefuddin A 2016 Gamma distribution linear modeling with statistical downscaling to predict extreme monthly rainfall in Indramayu In 2016 12th International Conference on Mathematics, Statistics, and Their Applications (ICMSA) (pp. 134-138) IEEE
[8] Syafruddin R, Soleh AM and Wigena AH 2019 Clusterwise Regression Model Development with Gamma Distribution *Proceedings of the 1st International Conference on Statistics and Analytics, ICSA*
[9] Sumertajaya IM, Erfiani and Putri WDY 2007 Clusterwise analysis using two step cluster method. Forum Statistika dan Komputasi p: 18-23 Vol 12 No.1
[10] Wigena AH 2006 *Pemodelan Statistical Downscaling dengan Regresi Projection Pursuit untuk Peramalan Curah Hujan Bulanan*. Dissertation (ID): Institut Pertanian Bogor.
[11] Zorita E and von Storch H 1999 The analog method as a Simple statistical downscaling technique *Journal of Climate and Applied Meteorology* 12: 2474-2489