Quantile Regression with Partial Least Squares in Statistical Downscaling for Estimation of Extreme Rainfall

Esther Ria Matulessy
Department of Statistics, Faculty of Mathematics and Natural Science
Bogor Agricultural University, Indonesia

Aji Hamim Wigena and Anik Djuraidah
Department of Statistics, Faculty of Mathematics and Natural Science
Bogor Agricultural University, Indonesia

Abstract

Extreme rainfall can lead to flooding and causes various disadvantages, such as crop failure in the field of agriculture. An analysis to examine extreme events is required to minimize these adverse impacts. Extreme rainfall can be analyzed by statistical downscaling (SD) model which is the functional relationship between local scale variable as the response variable (rainfall) and global scale variables as the explanatory variables (precipitation of global circulation). The purpose of this research is to develop SD model with quantile regression to predict extreme rainfall and partial least square to reduce dimension of explanatory variables. Extreme rainfall modeling is developed using linear, quadratic and cubic quantile regression respectively at 75th, 90th, and 95th quantiles. However, these models were not give good results. The models were improved by dummy variables. It was shown that cubic quantile regression model with the addition of dummy variables can predict extreme rainfall especially at 95th quantile in February, which was closed to the actual rainfall and consistent up to two years ahead.

Keywords: global circulation model, partial least square regression, quantile regression, extreme rainfall
1 Introduction

Climate plays a major role in people's environment such as in agriculture, transportation, telecommunication, and tourism. Rainfall as a part of the highest climate fluctuation and characterizing the most dominant climate in Indonesia is strongly influenced by global climate change, such as extreme rainfall. Extreme rainfall can lead to flooding and cause various disadvantages such as crop failure in the field of agriculture. The analysis that examines the extreme events are needed to minimize the bad impact due to the extreme rainfall. Extreme rainfall can be analyzed by statistical downscaling (SD). SD is the process of transforming information from a large-scale (global) variable to small-scale (local) variable. Downscaling generally utilizes global circulation model (GCM) output as global scale data to estimate rainfall as local scale variable. GCM produces data output as a primary source of information for assessing climate change.

Problems in GCM output are high dimension and multicollinearity. SD modeling generally uses principal component analysis (PCA) to reduce the dimension and solve the problem of multicollinearity. The problems can also be overcome by functional principal component analysis (FPCA) or partial least squares method (PLS). PCA and FPCA focuses on variance in the explanatory variables, whereas PLS focuses on the variance of the explanatory variables and the response variable [10].

SD models require a strong correlation between the precipitation of GCM output and the rainfall to obtain more accurate estimate. Strong correlation indicates the same pattern between two variables. The time lag precipitation of GCM output can result in different patterns with rainfall data. The time lag of precipitation GCM output is required through the highest correlation between rainfall and the precipitation GCM output by cross correlation function (CCF). Then dimension reduction using PCA and PLS method is applied to the precipitation GCM output with time lag [8]. The results indicated that PCA models with dummy variables gave higher $R^2$ and correlation than PLS. The addition of dummy variables in the PCA models produced a model with more homogeneous error [8].

Another method to estimate the extreme rainfall in SD modeling is quantile regression, which is an expansion from the median regression on a variety of quantile values. The model formed in quantile regression can be used to measure the effect of explanatory variables in the centre, the right or the left tail of the data distribution. The SD modeling using quantile regression to estimate extreme rainfall with PCA as the dimension reduction method can be found in [7], and the results show that quantile regression model gave the accurate prediction at 90th quantile in February. Similar SD model using quantile regression but with FPCA as dimension reduction method can be found in [9], and the results were more accurate and consistent than that with PCA. SD model with FPCA can predict the actual rainfall in February 2008 (439 mm/month) more precise than that with PCA. Both studies do not use the precipitation GCM output with time lag. This paper uses quantile regression to estimate extreme rainfall using PLS as dimension reduction method and the precipitation GCM output with time lag.
2 Literature Reviews

2.1 Global Circulation Model

GCM is an important tool in the study of climate variability and climate change. This model describes a number of subsystems of the climate on Earth, such as the processes in the atmosphere, oceans, land, and simulates global climate conditions [15]. Although GCM can properly simulate global climate variables, GCM cannot simulate well the local-scale climate variables [6]. The differences in scale between the explanatory variables and the response variable could be overcome using statistical downscaling [2].

SD used a transfer function that describes the functional relationship of GCM output with elements of the local climate. The general equation of SD model is as follows [11]:

\[ y = f(X) \]  

with \( y \) is local variable (e.g. rainfall), \( X (t \times p) \) are global variables (e.g. precipitation), \( t \) is the amount of time (such as daily or monthly), \( p \) is the number of grid GCM grids or explanatory variables.

2.2 Partial Least Square Regression (PLSR)

PLSR can be used to reduce the dimension and to overcome the problem of multicollinearity in the explanatory variables. This is done with the aim to predict a response variable \( y \) based on explanatory variables \( X \) [12]. PLSR method can be applied to one response as well as multiple responses. Let \( X (t \times p) \), where \( t \) is the number of observations and \( p \) is the number of explanatory variables, consisting of vectors \( x_{ij} \), \( j = 1,2,...,p \), and \( y \) size \( t \times q \), with \( q \) is the number of response variables consisting of vectors \( y_k \), \( k = 1,2,...,q \). PLSR method produces a number of new components that will model the \( X \) to \( y \) to obtain the relationship between \( X \) and \( y \). New components are referred to as a score of \( X \) and can be written as \( t_a \) with \( a = 1,2,...,A \). Each score \( t_a \) produced mutually orthogonal so PLSR can overcome the problem of multicollinearity in the explanatory variables.

Score \( X \) is a linear combination of the variables origin \( x_j \) with weighting coefficients \( w_{ja} \). The process can be formulated as [14]:

\[ t_{ia} = \sum_j w_{ja}x_{ij}, \quad i = 1,2,...,t \]

\[ T = XW \]  

Score \( X \), \( t_a \) is used as an explanatory variable for the response \( y \) and models of \( X \).

2.3 Quantile Regression

Quantile regression can be used to estimate a quantile if the data distribution is not homogeneous or the curve is not a standard or non-symmetrical shape and there is a tail on the distribution. Quantile regressions provide estimates more accurate and not sensitive to outliers [3]. This method can be used to measure the
effect of explanatory variables, not just in the center of the distribution, but also on right or left tail of the distribution. This is very useful in the application, particularly for estimating extreme values [4].

For a random variable $Y$ given distribution function as follows:

$$ F(y) = P(Y \leq y) $$

the quantile to $\tau$. from $Y$ to $0 < \tau < 1$, denoted by:

$$ Q(\tau) = \inf \{ y : F(y) \geq \tau \} \quad (3) $$

3 Data and Methodology

3.1 Data

The data used in this research is the precipitation (mm/month) of GCM output from climate models precipitation intercomparison project (CMIP5) and rainfall data in Indramayu district from 1979 to 2008. Precipitation data are used as explanatory variables and rainfall data in Indramayu as response variable. GCM domain used in this study is $8 \times 8$ grids ($2.5^\circ \times 2.5^\circ$ for each grid) at $98.75^\circ$ to $116.25^\circ$ East longitude and $1.25$ until $16.25^\circ$ North latitude above the area of Indramayu. The number of explanatory variables is 64. The domain size $8 \times 8$ grids over the area of Indramayu provide a more stable or consistent and not sensitive to outliers [11].

3.2 Methodology

The steps of this study is:

1. Identify the presence of extreme rainfall with boxplot.
2. Divide data into two groups i.e., for modeling (from 1979 to 2007) and for validation (2008).
3. Identify multicollinearity of precipitation data based on variance inflation factors (VIF).
4. Reduce the dimensions of explanatory variables using the PLS. The number of components based on prediction residual sum of squares (PRESS).
5. Develop quantile regression models of $75^{th}$, $90^{th}$ and $95^{th}$ quantiles with the components selected in step 4.
6. Predict rainfall data using the validation data and evaluate the prediction by correlation and root mean squared error of prediction (RMSEP).
7. Test the consistency of the model. The consistency of the model is based on the standard deviation and correlation at each year estimate. The small standard deviation indicates the model more consistent [11].

4 Results and Discussion

4.1 Data Exploration

The average monthly rainfall is relatively high in January, February, March, November and December. The range of monthly rainfall in rainy season is from 148.2 to 308.8 mm/month and that of dry season is from 14.6 to 141.2 mm/month.
In general, the lowest rainfall is 0 mm/month while the highest rainfall occurred in January 1981, which is 583 mm/month. According to Meteorology, Climatology and Geophysics Agency (BMKG), the intensity of extreme is greater than 400 mm/month [1]. The largest standard deviation of rainfall is 126.3 mm/month in January and the lowest is 16.5 mm/month in August. This facts show that rainfall in January in the period of 1979-2008 are very diverse.

Slope coefficients for all months are more than zero. The highest in is 2.00 in July and the lowest is 0.23 in May. Slope coefficient more than zero indicates the abnormal distribution of data and in the right part the average value is greater than the median and mode which is extreme rainfall. Boxplot of monthly rainfall is presented in Figure 1.

![Boxplot of Monthly Rainfall](image)

### 4.2 Quantile Regression

The RMSEP value of the linear quantile regression model at 75\textsuperscript{th} quantile is 74.48, at 90\textsuperscript{th} quantile is 101.57, and at 95\textsuperscript{th} quantile is 129.52, and the correlation (r) values for all quantiles are 0.90. Linear quantile regression model cannot pursue the actual pattern, so it must be added dummy variables to the model. The dummy variables are used for a better estimate. Dummy variables are determined based on the results of PLS method. A Plot between $y$ and $X$ scores shows five groups of rainfall data which indicate that four dummy variables are necessary to be used in the model.

The addition of the dummy variables to the linear quantile regression model results in the pattern of predicted rainfall similar to that of actual rainfall, but the model is only able to estimate rainfall at 90\textsuperscript{th} quantile. The RMSEP value of linear quantile regression models with dummy variables at 75\textsuperscript{th} quantile is 34.77, at 90\textsuperscript{th} quantile is 52.41, and at 95\textsuperscript{th} quantile is 61.34. The correlation (r) for all quantiles is 0.98. The RMSEPs of linear quantile regression model with dummy variables are about 48%-53% less than that without dummy variables, and the correlations of quantile regression models with dummy variables are about 9% larger than that without dummy variables (Table 1).
The relation pattern between rainfall and PLS component scores tend to be not linear so that quadratic or cubic quantile regression are also developed besides the linear quantile regression. However, the quadratic and cubic quantile regressions also produce the patterns not like the actual patterns. The addition of dummy variables is necessary to include in the models. A quadratic and cubic quantile regression with dummy variables produce smaller RMSEP values than the linear regression even though the linear regression results in larger correlation than the quadratic and cubic quantile regression without dummy variables. The RMSEP values of quadratic quantile regression models with dummy variable decrease about 45% - 51% from the RMSEP value of quadratic quantile regression model without the dummy variables, and the correlation values of quadratic quantile regression models with dummy variables increase about 8% from quantile regression model without the dummy variables. For the cubic quantile regression models with dummy variables, RMSEP values decrease approximately 50% - 58% from the RMSEP values of cubic quantile regression models without a dummy variable and the correlations of cubic quantile regression models with dummy variables increase about 11% from the cubic quantile regression model without the dummy variables.

Table 1 Comparison of RMSEP value and the correlation of linear, quadratic, cubic quantile regression model with dummy variable or without a dummy variable

<table>
<thead>
<tr>
<th></th>
<th>Linear quantile regression</th>
<th>Quadratic Quantile regression</th>
<th>Cubic Quantile regression</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Without dummy</td>
<td>With dummy</td>
<td>Without dummy</td>
</tr>
<tr>
<td>75th Quantile</td>
<td>74.48</td>
<td>34.77</td>
<td>75.76</td>
</tr>
<tr>
<td>90th Quantile</td>
<td>101.57</td>
<td>52.41</td>
<td>95.44</td>
</tr>
<tr>
<td>95th Quantile</td>
<td>129.52</td>
<td>61.34</td>
<td>127.75</td>
</tr>
<tr>
<td>75th Quantile</td>
<td>0.90</td>
<td>0.98</td>
<td>0.91</td>
</tr>
<tr>
<td>90th Quantile</td>
<td>0.90</td>
<td>0.98</td>
<td>0.91</td>
</tr>
<tr>
<td>95th Quantile</td>
<td>0.90</td>
<td>0.98</td>
<td>0.91</td>
</tr>
</tbody>
</table>

4.3 Prediction

The RMSEP value of the quadratic quantile regression models with dummy variables at 75th quantile is 40.18, at 90th quantile is 52.06, and at 95th quantile is 61.55, and the correlation (r) is 0.98 for all quantiles. For the cubic quantile regression models with dummy variables, the RMSEP value at 75th quantile is 40.09, at 90th quantile is 51.29, and at 95th quantile is 59.33, and the correlation for
all quantiles is around 0.98. The cubic quantile regression models with dummy variables produce the better estimate than the quadratic quantile regression models with dummy variable.

Table 2 The Estimation of rainfall in 2008

<table>
<thead>
<tr>
<th>Month</th>
<th>75th quantile</th>
<th>90th quantile</th>
<th>95th quantile</th>
<th>Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>298.30</td>
<td>359.11</td>
<td>381.87</td>
<td>351.46</td>
</tr>
<tr>
<td>2</td>
<td>427.15</td>
<td>438.49</td>
<td>439.66</td>
<td>439.33</td>
</tr>
<tr>
<td>3</td>
<td>305.74</td>
<td>342.42</td>
<td>344.06</td>
<td>260.86</td>
</tr>
<tr>
<td>4</td>
<td>104.83</td>
<td>109.65</td>
<td>109.99</td>
<td>97.13</td>
</tr>
<tr>
<td>5</td>
<td>91.48</td>
<td>103.17</td>
<td>106.91</td>
<td>19.73</td>
</tr>
<tr>
<td>6</td>
<td>70.29</td>
<td>87.23</td>
<td>97.03</td>
<td>23.46</td>
</tr>
<tr>
<td>7</td>
<td>41.11</td>
<td>59.78</td>
<td>76.00</td>
<td>0.00</td>
</tr>
<tr>
<td>8</td>
<td>32.03</td>
<td>50.35</td>
<td>68.22</td>
<td>7.33</td>
</tr>
<tr>
<td>9</td>
<td>41.70</td>
<td>60.38</td>
<td>76.49</td>
<td>2.20</td>
</tr>
<tr>
<td>10</td>
<td>69.44</td>
<td>86.51</td>
<td>96.53</td>
<td>68.00</td>
</tr>
<tr>
<td>11</td>
<td>191.16</td>
<td>202.50</td>
<td>205.41</td>
<td>136.46</td>
</tr>
<tr>
<td>12</td>
<td>192.21</td>
<td>210.80</td>
<td>223.93</td>
<td>198.40</td>
</tr>
</tbody>
</table>

Figure 2. Estimates of monthly rainfall in 2008 using cubic quantile regression models with dummy variables at 75th, 90th and 95th quantiles
Figure 2 shows that the cubic quantile regression models with dummy variables can estimate the intensity of monthly rainfall well. The estimate of rainfall can follow the pattern of actual rainfall properly, especially the extreme rainfall. The highest rainfall intensity occurred in February 2008 is 439.33 mm/month. This value is estimated well by the estimate at 95th quantile around 439.66 mm/month (Table 2). In general, for the months that are in the dry season (April-September), the estimates at 75th, 90th, and 95th quantiles are higher than the actual value, but the patterns are similar to the actual pattern.

5 Conclusion

Linear quantile regression models with dummy variables are capable to follow the actual pattern but the estimate value has not been able to follow the extreme value. The best quantile regression SD model is the cubic quantile regression with dummy variables which gives the higher correlation values, the lower RMSEP values and able to estimate the extreme values properly.

References


Received: March 30, 2015; Published: June 23, 2015