Science Journal of Applied Mathematics and Statistics
2014; 2(3): 66-70
Published online July 20, 2014 (http://www.sciencepublishinggroup.com/j/sjams)
doi: 10.11648/j.sjams.20140203.12

Quantile regression in statistical downscaling to estimate extreme monthly rainfall

Aji Hamim Wigena, Anik Djuraidah

Department of Statistics, Bogor Agricultural University, Bogor, Indonesia

Email address:
ajiwigena@ymail.com (A. H. Wigena), anikdjuraidah@gmail.com (A. Djuraidah)

To cite this article:
Aji Hamim Wigena, Anik Djuraidah. Quantile Regression in Statistical Downscaling to Estimate Extreme Monthly Rainfall. *Science Journal of Applied Mathematics and Statistics.* Vol. 2, No. 3, 2014, pp. 66-70. doi: 10.11648/j.sjams.20140203.12

Abstract: Extreme rainfall events have been great interest in statistical downscaling. This paper concerns with developing model of statistical downscaling using quantile regression to estimate extreme monthly rainfall. Statistical downscaling relates functionally local scale response variable and global scale predictor variables. The response variable is monthly rainfall from 1979 to 2008 at station Bangkir Indonesia and the predictor variables are monthly precipitation of 64 grid of Global Circulation Model output in the same period. Principal Component Analysis is used to reduce dimension of predictors. A number of components for developing quantile regression model are determined based on Quantile Verification Skill Score. The results show that at 95th quantile the pattern of forecasted rainfall in January to December 2008 is similar to actual rainfall with correlation 0.98 and the forecasted rainfall (843 mm) in February 2008 is considered as the extreme rainfall which confirms well to the highest actual rainfall (727 mm) with probability 0.99.

Keywords: Collinear, Extreme, Principal Component Analysis, Statistical Downscaling, Quantile Regression, Logistic Regression

1. Introduction

Extremely high rainfall events frequently result in bad impact to the environment and human life especially for agricultural production. Extreme or heavy daily rainfall one or two consecutive days usually trigger flood. In agriculture this event will be the risk factor in production process so the extreme event potentially threatens food security and environment damage. Therefore, the identification and quantification of extreme events of certain magnitudes and its probabilities are primary importance due to their severe consequences for human society and ecological system.

Frequency and intensity of extreme rainfall changes have occurred at monsoon area in Asia and other tropical areas. Rainfall intensity globally has projected to increase especially in tropical areas and areas with high altitude. Extreme rainfall increases more than its average in many tropical areas [1]. Information about the possibility of extreme rainfall is necessary to accommodate bad risk of the extremes. Accurate information will be important and useful to decrease the impact of extreme rainfall. It is more strategic to overcome or decrease losses due to extreme events.

This information is related to developing statistical model. Statistical downscaling is a technique using the model to forecast monthly rainfall [2]. This technique develops functional relationship between GCM (Global Circulation Model) output as global scale data and local or ground data as small scale data. The purpose of the model is to get information from GCM output to forecast future local value. However, this model does not accommodate extreme events and is sensitive to outlier data. Quantile regression can overcome this drawback.

Compared to peak over threshold or block maxima method in extreme value theory to estimate an extreme value, quantile regression is the simpler method. The quantile regression model can explore and identify an outlier including extreme value based on a quantile [3] and also is insensitive to outliers. Quantile regression model has been more efficient than multiple linear regression model to predict both high and low extreme values [4] and has been used to estimate daily rainfall based on NWP (Numerical Weather Prediction) model output [5]. Censored quantile regression has been used in statistical downscaling to explore extreme daily rainfall [6].

Statistical downscaling usually uses principal component regression based on principal component analysis (PCA) to overcome the problem of multi collinear in the predictor

variables. This paper discusses the use of quantile regression with PCA in statistical downscaling to forecast extreme monthly rainfall. The probability of this extreme is estimated using logistic regression.

2. Data and Methods

2.1. Data

The study uses monthly rainfall data from Bangkir station in Indramayu, West Java, Indonesia and monthly precipitation of GCM output. Both data covers the period of 30 years (from 1979 to 2008). The monthly rainfall data were from Meteorological, Climatological, and Geophysical Agency (BMKG) of Indonesia, whereas the GCM output of domain 8×8 grids was downloaded from CMAP (Climate Prediction Center Merged Analysis of Precipitation), Boulder Colorado, USA (http://www.esrl.noaa.gov.psd). The grid size is 2.5°×2.5°. The GCM domain is over the Bangkir station. Data from 1979 to 2007 are used to develop calibration model and data of 2008 are used to validate the model.

2.2. Methods

Statistical downscaling model is based on multiple regression model which shows functional relationship between monthly rainfall and monthly precipitation of GCM output. The monthly rainfall data from 1979 to 2007 are as the response variable ($y_{t\times1}$). The monthly precipitation from 1979 to 2007 of GCM output at every grid is as a predictor so the overall are 64 predictors ($X_{t\times p}$) for (8×8) grids.

Multiple regression model needs among predictors uncorrelated or not multi collinear. In statistical downscaling the predictors are usually high correlated or multi collinear, such as the monthly precipitation of GCM output at 64 grids. This is a problem in developing statistical downscaling model. Principal component analysis (PCA) is one method to overcome the problem.

Principal component analysis converts predictors into orthogonal components and each component is a linear combination of predictors. The components are not correlated and can be used as new predictors in regression model. However, only few components are used as new predictors. The method is also called dimension reduction method.

Multiple regression estimates a response mean but the mean is usually not to predict an extreme value. A response quantile is more appropriate to predict an extreme value than the mean. The quantile can be estimated by quantile regression model.

Quantile regression, introduced by Koenker & Bassett in 1978, can be used to analysis any part of distribution based on a value of quantile [8]. This method is robust to outliers and does not need a normal distribution assumption. The effect of predictor to the response can be measured not only in the center but also in the tails of distribution. Thus, the quantile regression can be used especially to predict an extreme value usually in the tails of distribution [3]. Quantile regression are used to estimate an extreme rainfall [6;9] and to analyze wind speed of tropical cyclone [10].

Quantile regression is based on the following multiple regression model:

$$y = Z'\beta + \varepsilon \tag{1}$$

with $y = (y_1, y_2, \dots, y_t)'$ is response of size $(t \times 1)$, $Z = (z_1, z_1, \dots, z_k)'$ is predictors of size $(t \times k)$, $\beta = (\beta_1, \beta_2, \dots, \beta_k)'$ is parameters of size $(k \times 1)$, and $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_t)'$ is errors of size $(t \times 1)$. In quantile regression, the parameters are estimated for a quantile $\tau \in (0,1)$ by minimizing the following function (Eq.(2)):

$$\min_{\xi \in R^p} \left[\sum_{i \in \{i; y_i \geq z_i'\beta\}}^{n} \tau |y_i - z_i'\beta| + \sum_{i \in \{i; y_i < z_i'\beta\}}^{n} (1-\tau) |y_i - z_i'\beta| \right] \tag{2}$$

or

$$\min_{\xi \in R} \sum_{i=1}^{n} \rho_\tau (y_i - z_i'\beta) \tag{3}$$

with $\rho_\tau(u)$ is called as a risk function defined as

$$\rho_\tau(u) = u\big(\tau - I(u < 0)\big), \; 0 < \tau < 1 \tag{4}$$

and $I(\cdot)$ is an indicator function.

In this study, quantile regression model is applied to estimate the value of quantiles using new orthogonal predictors resulted from PCA. The probability of each quantile representing an extreme value is computed based on logistic regression model.

First, PCA is used to overcome the multi collinear problem and to reduce the dimension of predictors into k orthogonal components (k<p) with the first largest proportion of variance. These components are the new variables ($Z_{t\times k}$) in quantile regression modeling.

The next step, quantile regression model is applied using variety numbers of components. One of the models is selected according to the value of QVSS (Quantile Verification Skill Score) [6]. The model with the largest QVSS is the best model used to predict the extreme value.

QVSS in Eq.(5) is computed to assess the model [6,7]:

$$QVSS = 1 - \frac{\sum_{i=1}^{n} \rho_\tau |y_i - \hat{\boldsymbol{\beta}}_\tau^T \boldsymbol{z}_i|}{\sum_{i=1}^{n} \rho_\tau |y_i - Q_\tau(y)|} \tag{5}$$

and

$$\rho_\tau(u) = \begin{cases} \tau u & ; u \geq 0 \\ (\tau - 1)u & ; u < 0 \end{cases} \tag{6}$$

where $Q_\tau(y)$ is the quantile τ of y and $\hat{\beta}_\tau^T z_i$ is the estimate of quantile regression model.

The best model estimates the values of 75th, 90th, and 95th quantiles using data in 2008. The 90th and 95th quantiles are expected to be the extreme values. The estimates of extreme rainfall according to each quantile are compared to the actual rainfall by comparing the patterns of both estimates and actual rainfall using validation data in 2008.

Besides forecasting extreme values, the probabilities of each extreme event are also estimated. The probabilities of

extreme values at 75th, 90th, and 95th quantiles are estimated using logistic regression model. Before applying logistic regression, the response (y) is discretized first into the following two categories such as in Eq.(7):

$$y_i = \begin{cases} 1, & y_i \geq Q_\tau(y) \\ 0, & y_i < Q_\tau(y) \end{cases} \quad (7)$$

So, the response (monthly rainfall data) in logistic regression modeling is binary variable with 0 and 1 values. For example, in order to estimate the probability of extreme rainfall at 95th quantile, the actual rainfall data (y) is discretized such that $y_i = 1$ if y_i is greater than or equal to the value of 95th quantile and $y_i = 0$ if y_i is less than the value of 95th quantile of actual rainfall data. The same procedure is implemented to estimate the probability of extreme rainfall at 75th and 90th quantiles.

3. Results

The predictors (monthly precipitation of GCM output) are explored first to know whether or not these predictors were highly correlated or multi collinear. The data exploration shows that the correlations of many predictors are high which indicates that there is multi collinear problem in predictors.

PCA is applied to overcome this problem by reducing the dimension of predictors to smaller dimension of new orthogonal or uncorrelated predictors. The number of new predictors potentially used in developing statistical downscaling model corresponds to the number of components resulted from PCA.

According to the result of dimension reduction, the statistical downscaling model with quantile regression uses five new orthogonal predictors. These new predictors are based on the first five components which represent 82% of total variance. However, in this study the number of components is not based on the proportion of variance but based on the maximum value of QVSS of any quantile regression model [6]. The larger QVSS indicates the stronger relationship between response and predictors. The number of new predictors with largest QVSS is selected for developing quantile regression model.

The values of QVSS are computed for every model with j components (j = 1,2, … ,k) for the 75th, 90th, and 95th quantiles. Fig. 1 shows the plot of QVSS and the number of components (k = 30) for each quantile. The lowest values of QVSS are for the models with one component and the QVSS increases as the number of components increases until 22 components. The other QVSSs are smaller. The largest values of QVSS for 22 components are chosen. These QVSS values are 0.56, 0.76, and 0.77 respectively for 75th, 90th, and 95th quantiles. The quantile regression models are developed and expected to be the best models to forecast extreme values corresponding to the estimates of 75th, 90th, and 95th quantiles, especially the estimate of 95th quantile.

Forecasting the monthly rainfall at the 75th, 90th, and 95th quantiles for one year ahead in 2008 uses the best models with 22 components. The forecasted monthly rainfall is validated with the actual monthly rainfall in the same year. Fig. 2 shows the forecasted and actual monthly rainfall. The forecasted monthly rainfall is presented by box and whisker diagram for every month. The up whisker is as long as the value of 95th quantile and the box shows the different between the value of 90th and 70th quantiles. The top of box represents the value of 90th quantile and the bottom of box represents the value of 75th quantile. The actual monthly rainfall is presented by dot symbols.

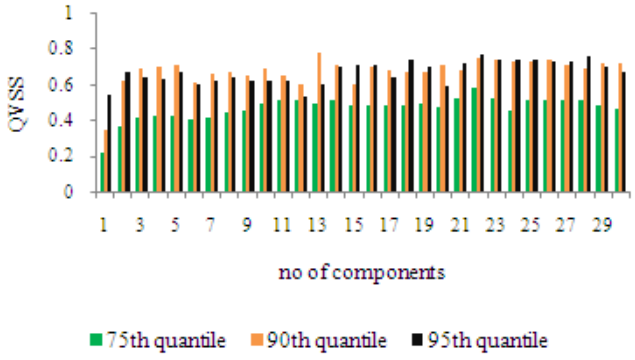

Fig. 1. *QVSS and number of components*

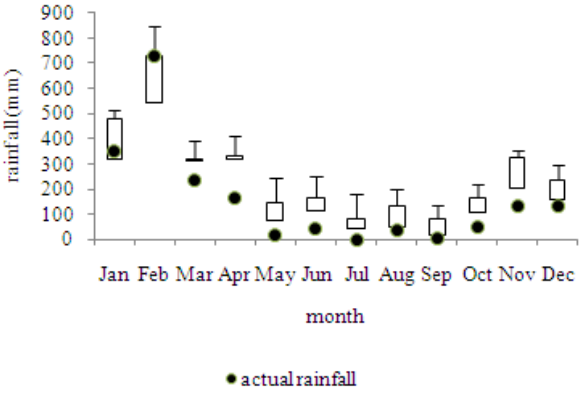

Fig. 2. *Forecasted and actual monthly rainfall 2008*

Fig.2 indicates the patterns of forecasted rainfall at 75th, 90th, and 95th quantiles and actual rainfall from January to December are relatively similar. These patterns are indicated by high correlations between the forecasted rainfall at 75th, 90th, and 95th quantiles and actual rainfall. The correlations are 0.94, 0.98, and 0.98 respectively for 75th, 90th, and 95th quantiles. In January to December 2008 the forecasted rainfall at 95th quantile are higher than the actual rainfall. In February, which is in rainy season from October to March, the actual rainfall (727 mm) is the highest which indicates the extreme event in 2008 while the forecasted rainfall at the 95th quantile (843 mm) is also the highest. This forecasted rainfall confirms well to the actual rainfall.

Both the forecasted rainfall at 75th and 90th quantiles and actual rainfall in January and March are lower than those in February. Those in other months, which is in dry season from April to September, are also the lower. These facts show that quantile regression model is appropriate to

forecast extreme rainfall.

Besides forecasting the quantities of extreme rainfall, in order to know the probability of extreme events, logistic regression model is applied to compute the probability of extreme value at each quantile. Table 1 shows the probabilities of rainfall at 75th, 90th, and 95th quantiles for every month. In January the probabilities of forecasted rainfall at 75th and 90th quantiles are 0.97 and 0.84 respectively but the probability at 95th quantile is 0.32. This fact is indicated that the extreme rainfall event represented by 95th quantile occurred with very low probability. While in February the probabilities of forecasted rainfall at 75th, 90th, and 95th quantiles are 0.99, 0.97, and 0.99 respectively. The extreme rainfall event occurred with very high probability especially at 95th quantile (843 mm) which confirms to the actual rainfall (727 mm) as the highest rainfall. These facts show that the extreme rainfall especially in rainy season, such as in February, can be forecasted quite accurate with the highest probability.

Table 1. *Probability of 75th, 90th, and 95th quantiles every month*

Month	Quantile		
	75th	90th	95th
Jan	0.97	0.84	0.32
Feb	0.99	0.97	0.99
Mar	0.99	0.48	0.06
Apr	0.28	0.02	0
May	0.09	0.01	0
Jun	0.12	0	0
Jul	0.08	0	0
Aug	0.05	0	0
Sep	0.01	0	0
Oct	0.11	0	0
Nov	0.83	0.18	0
Dec	0.89	0.15	0

The probabilities of forecasted rainfall in the other months are smaller than those in February. In January the probability of rainfall at 90th quantile (0.84) is higher than that at 95th quantile (0.32). In April to September the probability of rainfall at 90th and 95th quantiles are less than 0.1 except in March the probability at 90th is less than 0.5. These low probabilities are for the low forecasted rainfall. Generally, in dry season extreme rainfall is almost impossible but in rainy season the extreme rainfall is possible to happen with high probability.

4. Conclusions

Statistical downscaling technique based on GCM output can be used to forecast extreme rainfall well using quantile regression model. The pattern of forecasted and actual rainfall from January to December is almost similar even though the different between both values is large. The minimum extreme value can be equal or more than the forecasted rainfall at 95th quantile. In February the forecasted rainfall which represents the extreme value can reflect the highest actual rainfall.

Logistic regression model can be used to calculate the probability of the forecasted rainfall especially at 95th quantile. The extreme rainfall at 95th quantile can occur in February with high probability (0.99).

Quantile regression and logistic regression are potentially useful in statistical downscaling based on principal component analysis to forecast extreme value with its appropriate probability.

The different between forecasted and actual rainfall is relative large. This is probably because of linear principal component analysis. The use of functional principal component analysis, instead of linear principal component analysis, is expected to improve forecasting rainfall at 70th, 90th, and 95th quantiles in order to decrease the different.

Acknowledgement

I would like to thank to Meteorological, Climatological, and Geophysical Agency (BMKG) of Indonesia for providing rainfall data.

References

[1] IPCC. *IPCC fourth assessment report: climate change 2007, Working Group I: The Physical Science Basis*, (Cambridge University Press, 2007, p.996.

[2] A.H. Wigena. *Modeling of statistical downscaling using projection pursuit regression for forecasting monthly rainfall*, doctoral diss., Bogor Agricultural University (in Indonesian), Indonesia. 2006.

[3] A. Djuraidah, and A.H.Wigena. Quantile regression to explore rainfall pattern. *Jurnal Ilmu Dasar, 12(1)*. 2011. (in Indonesian).

[4] S.I.V. Sousa, J.C.M. Pires, F.G. Martins, M.C. Pereira, and M.C.M. Alvim-Ferraz. Potentialities of quantile regression to predict ozone concentrations, *Environmetrics, 20*, 2009, 147–158. (DOI: 10.1002/env.916).

[5] J.B. Bremnes. Probabilistic forecasts of precipitation in terms of quantiles using NWP model output, *Monthly Weather Review, 132*, 2004, pp.338-347.

[6] P. Friederichs, and A. Hense,. Statistical Downscaling of Extreme Precipitation Events Using Censored Quantile regression. *Monthly Weather Review, 135*, 2007, pp. 2365-2378.

[7] P. Friederichs. Statistical downscaling of extreme precipitation events using extreme value theory. *Extremes, 13*, 2010, pp.109–132. (DOI 10.1007/s10687-010-0107-5).

[8] I.S. Buhai. *Quantile regression: overview and selected applications. Roger Koenker's lecture notes*, 2004. (the recent Netherlands Network of Economics Workshop in Groningen 2004).

[9] S. Chamaille´-Jammesa, H. Fritz, and F. Murindagomo. Detecting climate changes of concern in highly variable environments : Quantile regressions reveal that droughts worsen in Hwange National Park, Zimbabwe. *Journal of Arid Environments, 71*, 2007, pp. 21–326.

[10] T.H. Jagger and J.B. Elsner. Modeling tropical cyclone intensity with quantile regression. *Int. J. Climatol*, 2008, Published online in Wiley Inter Science (*www.interscience.wiley.com*). (DOI: 10.1002/ joc.1804).