BAYESIAN QUANTILE REGRESSION MODELING TO ESTIMATE EXTREME RAINFALL IN INDRAMAYU

Eko Primadi Hendri[1], AjiHamim Wigena[1*], Anik Djuraidah[1].
{ajiwigena@ymail.com}

[1]*Department of Statistics, IPB University, Bogor,16680, Indonesia*

Abstract. Quantile regression can be used to analyze symmetric or asymmetric data. Estimates of quantile regression parameters are obtained by the simplex method. Another approach is the Bayesian method based on Laplace's asymmetric distribution using MCMC. MCMC is used numerically to estimate parameters from each posterior distribution. The Bayesian quantile regression and the quantile regression can be used for statistical downscaling in extreme rainfall cases. This study used statistical downscaling to obtain relationship between global-scale data and local-scale data. The data used were monthly rainfall data in Indramayu and GCM output data. LASSO regularization was used to overcome multicollinearity problems in GCM output data. The purpose of this study was to compare Bayesian quantile regression models with quantile regression. The Bayesian quantile regression and the quantile regression couldpredict extreme rainfallmore accurate and consistent in one year ahead. The Bayesian quantile regression model is relatively better than the quantile regression.

Keywords: Bayesian Quantile Regression, LASSO, MCMC, Statistical Downscaling.

1 Introduction

Simple linear regression is used to get a functional relationship between two or more variables and can predict the effect of explanatory variables on the response variable. In asymmetric data, simple linear regression is not valid because it is sensitive to outliers. Therefore, a quantile regression method is developed. Quantile regression can analyze a number of data in the form of asymmetric and non-homogenous data [5]. Quantile regression can measure the effects of explanatory variables not only at the center of the distribution of data but also at the top and bottom of the distribution tail. This is very useful if extreme values are an important problem [4].

Parameters of quantile regression can be determined by the simplex method on linear programming [3] and the Bayesian method that is known as Bayesian quantile regression. Yu and Moyeed[12] suggested that the Bayesian quantile regression estimation uses Markov Chain Monte Carlo (MCMC) method based on Laplace asymmetric distribution. Kozumi and Kobayashi [6] suggested that the Gibbs sampling method can estimate the parameters of the Bayesian quantile regression based on Laplace asymmetric distribution.

Statistical downscaling (SD) is a technique that uses statistical methods to see the relationship between global scale data and local scale data. Global-scale data is represented by a global circulation model (GCM) data and local-scale data represented by rainfall data. GCM

ICSA 2019, August 02-03, Bogor, Indonesia

DOI 10.4108/eai.2-8-2019.2290491

output data has high dimensions that cause multicollinearity. This causes the estimated parameters of the model to be biased. This can be solved with the least absolute shrinkage and selection operator (LASSO).

SD modeling with quantile regression has been widely studied. Djuraidah and Wigena [4] used quantile regression to explore rainfall in Indramayu. Mondiana [8] researched SD modeling with quantile regression using principal component analysis to reduce the dimensional of GCM. Santri [9]discussesd quantile regression modeling at SD using penalty LASSO. Zakarina [13] discussed SD modeling using guludquantil regression. Cahyani [2] researched SD modeling using Elestic-net quantile regression.

Based on the explanation above, this study aims to develop SD modeling with Bayesian quantile regression with LASSO Penalty is used to solve multicollinearity in GCM and MCMC method to estimate extreme rainfall in Indramayu.

2 Materials

a. Describing rainfall data in Indramayu with descriptive statistical analysis

b. The Bayesian quantile regression model [6] as follows:

$$y_i = x_i^{'}\beta + pv_i + k\sqrt{\sigma v_i}u_i.$$

c. LASSO penalty is used at Q (0.75), Q (0.90), and Q (0.95) to get variables that are not multicollinearity with formulas

$$\beta_\tau^{LASSO} = \min_{\beta \in R} \sum_{i=1}^{n} \rho_\tau (y_i - x_i^{'}\beta) + \lambda \sum_{j=1}^{p} |\beta_j|.$$

d. MCMC, Gibbs sampling, is used to get the parameter estimator. The Gibbs sampling algorithm [7] are as follows:

   1) Suppose the initiation value for $\beta, v_i, \sigma$are $\beta^{(0)}, v_i^{(0)}, \sigma^{(0)}$.
   2) For the first iteration, do
      a) Generate $\beta^{(1)}$, $\beta^{(1)} \sim \pi(\beta | v_i^{(0)}, \sigma^{(0)}, y)$
      b) Generate $v_i^{(1)}$, $v_i^{(0)} \sim \pi(v_i | \beta^{(0)}, \sigma^{(0)}, y)$
      c) Generate $\sigma^{(1)}$, $\sigma^{(0)} \sim \pi(\sigma | \beta^{(0)}, v_i^{(0)}, y)$
   3) Repeat steps 4.ii as many as m iterations.
   4) Prepare examples that have a joint posterior distribution $\pi(\beta, v_i, \sigma | y)$

e. The model is evaluated based on the root means square error of prediction (RMSEP)

$$RMSEP = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y)^2}.$$

f. and correlation between actual rainfall and estimated rainfall

$$r_{y\hat{y}} = \frac{n\sum_{i=1}^{n} y_i\hat{y}_i - (\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} \hat{y}_i)}{\sqrt{[n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2][n\sum_{i=1}^{n} \hat{y}_i^2 - (\sum_{i=1}^{n} \hat{y}_i)^2]}}.$$

g. Validation and consistency of model.

3 Materials

The research data is secondary data from 1981 to 2009. Monthly rainfall data as the dependent variable is obtained from BMKG. Monthly rainfall data is the average rainfall from nine stations in Indramayu. GCM output data is monthly precipitation data of the Climate Forecast System Reanalysis (CFSR) with a grid size of $2.5^0 \times 2.5^0$ from a domain 4×8 grid. The GCM output data as independent variable. The GCM output data are obtained from the National Centers for Environmental Prediction (https://rda.ucar.edu).

4 Result and Discussion

4.1 Data Description

4.1.1 Rainfall in Indramayu

The Descriptive statistical analysis of monthly rainfall data is done as initial information to determine the characteristics and patterns of rainfall, which are used for the next analysis. Table 1 presents information about monthly rainfall data in Indramayu. The rainy season has an average intensity of monthly rainfall greater than 150mm/month and the dry season has an average intensity of monthly rainfall of less than 150mm/month [10]. Based on Table 1, the rainy season occurs in December-January and the dry season occurs in March-November.

Extreme rainfall occurs if the intensity of monthly rainfall is more than 400mm/month [1]. Extreme rainfall occurs in January and February because it has a monthly rainfall intensity greater than 400mm/month. The highest standard deviation occurs in February, which is 123.44 mm/month. The highest standard deviation value indicates that rainfall in February are most diverse from the other month. The value of the skew coefficient for each month is greater than zero. This shows that the distribution of observation data are skewed to the right.

Table 1. Description of Indramayu rainfall data in 1981-2009

	Mean	Standard Deviation	Maximum	Minimum	coefficient of skewness
Jan	283.73	121.59	530.67	83.78	0.20
Feb	225.52	123.44	509.00	76.67	0.90
Mar	159.23	61.85	303.89	65.00	0.97
Apr	134.33	49.40	213.78	33.11	-0.52
May	73.74	46.74	188.22	0.00	0.72
Jun	54.44	36.05	118.11	0.00	0.34
Jul	26.63	31.72	136.00	0.00	1.80
Aug	11.86	16.75	59.78	0.00	1.58
Sep	20.79	32.06	134.78	0.00	2.26
Oct	70.66	67.30	241.78	0.00	0.74
Nov	142.13	82.04	302.89	12.56	0.39
Dec	175.84	70.25	397.22	48.67	1.34

4.1.2 GCM Output Data

The GCM output data are a high dimension so it need to be checked formulticollinearity based on the method of variance inflation factors (VIF). Based on Table 2, there are grids have VIF > 10. This shows that in GCM output data have multicollinearity problem.

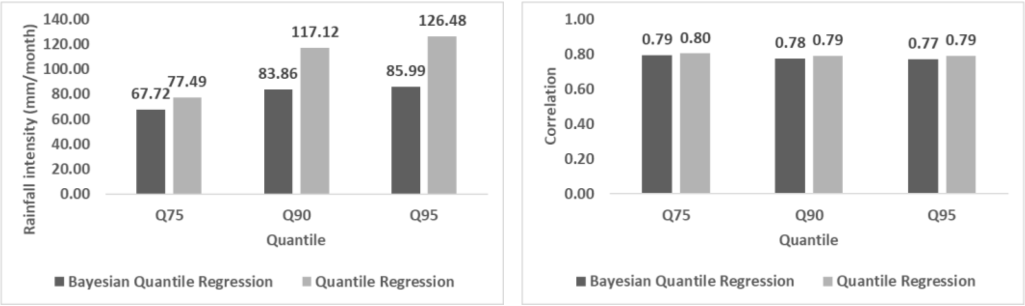
Table 2. Value of VIF

Variable	VIF	Variable	VIF	Variable	VIF	Variable	VIF
X1	3.46	X9	3.78	X17	5.20	X25	4.33
X2	4.42	X10	4.29	X18	8.12	X26	7.17
X3	4.48	X11	5.27	X19	7.72	X27	8.54
X4	4.89	X12	7.87	X20	9.39	X28	7.27
X5	4.53	X13	12.65	X21	9.86	X29	7.98
X6	3.37	X14	12.21	X22	7.45	X30	14.36
X7	3.03	X15	7.99	X23	13.21	X31	16.81
X8	4.02	X16	5.05	X24	12.04	X32	10.51

4.2Bayesian Quantile Modeling

The The domain of GCM output data used as predictor in Bayesian quantile regression model are 4×8 grids or 32 predictor. Furthermore, the LASSO Penalty is used to solve multicollinearity in Q(0.75), Q(0.90), and Q(0.95). After obtaining variables that are not multicollinearity, the MCMC method is used to obtain parameter estimators. In the next stage, the variables of the Bayesian quantile regression model in each quantile are used for quantile regression modeling. In the next step, variables of the Bayesian quantile regression model in each quantile are used for quantile regression modeling.

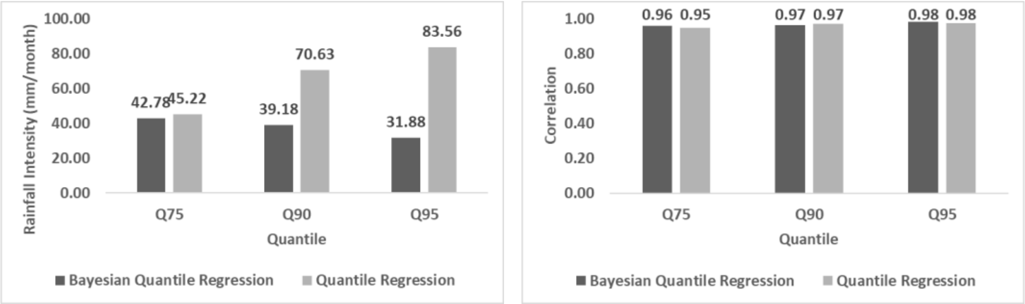
The Bayesian quantile regression model will be compared to the quantile regression model to determine the best model. The best model is based on the smallest RMSE value and the highest correlation. Based on Figure 1, the lowest RMSE value is in Bayesian quantile regression. There is no difference in the correlation value between Bayesian quantile regression with quantile regression. Therefore, the Bayesian quantile regression model is better than quantile regression.



(a) RMSE (b) Correlation

Fig 1. Value of RMSE and correlation

TheFigure 2 shows the RMSEP value and the correlation value of each model. The Bayesian quantile regression model has the lowest RMSEP value. There is no difference in the correlation value of each model. The Bayesian quantile regression model is better than quantile regression for prediction.



(a) RMSEP (b) Correlation

Fig 2. Value of RMSEP and correlation

Figure 3 shows the value of rainfall predictions and actual rainfall data. Extreme rainfall in January and February are by the Bayesian quantile regression model. Extreme rainfall in January can be predicted by model Q(0.90). Extreme rainfall in February can be predicted by the model Q(0.95).

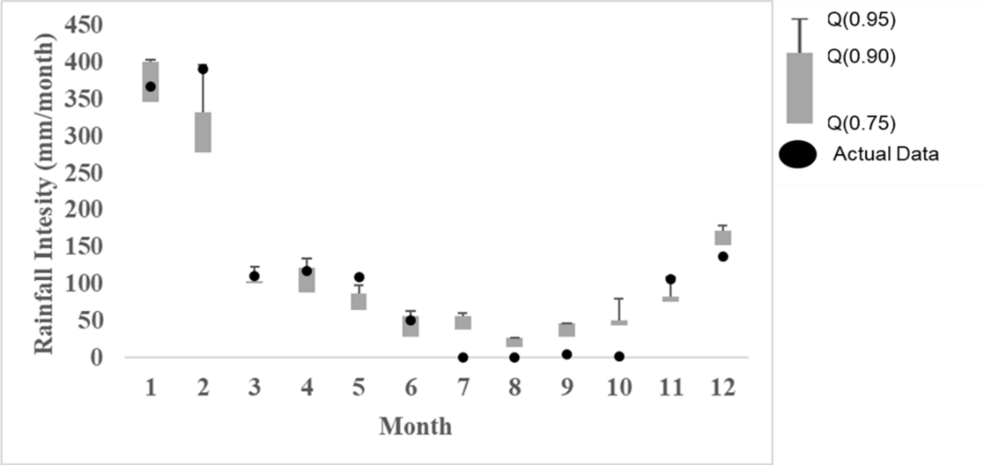


Fig 3. Plot of actual data and rainfall prediction using Bayesian quantile regression model

Figure 4 shows the value of rainfall predictions and actual rainfall data. Extreme rainfall in February is by the quantile regression model. Extreme rainfall in February can be predicted by model Q(0.90).

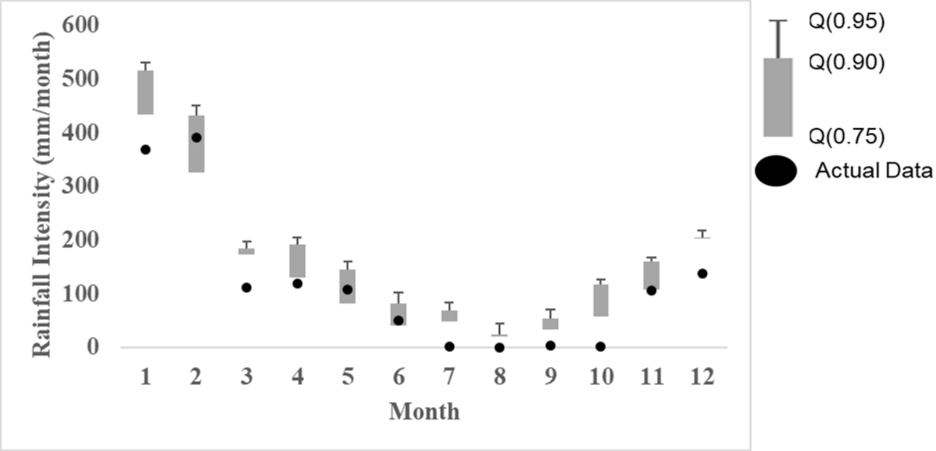


Fig 4. Plot of actual data and rainfall prediction using quantile regression model

4.3 Validation and Consistency of Model

Validation is a step to see the accuracy of model predictions. Table 3 shows that the Bayesian quantile regression model and quantile regression models are better used to predict extreme rainfall for the next 1 year. This is because the model has the largest correlation value and the lowest RMSEP value.

Table 3. Value of RMSEP and correlation of each model

Data Training	Data Testing	Quantile	Bayesian Quantile Regression		Quantile Regression	
			RMSEP	Correlation	RMSEP	Correlation
1981-2008	2009	0.75	42.78	0.96	45.22	0.95
		0.90	39.18	0.97	70.63	0.97
		0.95	31.88	0.98	83.56	0.98
1981-2007	2008-2009	0.75	55.63	0.94	59.80	0.92
		0.90	66.90	0.93	68.58	0.93
		0.95	56.54	0.94	77.05	0.93
1981-2006	2007-2009	0.75	60.69	0.94	60.34	0.91
		0.90	70.08	0.89	83.02	0.89
		0.95	57.22	0.93	79.77	0.94
1981-2005	2006-2009	0.75	77.48	0.88	68.89	0.87
		0.90	65.34	0.88	86.83	0.88
		0.95	88.87	0.83	103.27	0.88

The consistency of the model can be seen from the results of the estimation at different times. The model will give the best results if the relationship between the independent variable and the dependent variable does not change with time changes. The consistency of the model is measured based on the standard deviation value of the correlation value in each estimation year. The smaller the standard deviation, the more consistent the model is [11]. Based on Table 4, the standard deviation of the correlation values by the Bayesian quantile regression model and the quantile regression model for each quantile has a very small value. Standard deviation value in Bayesian quantile regression are 0.06 in Q(075), 0.03 in Q(0.90), and 0.04 in Q(0.95). Standard deviation value in quantile regression are 0.08 in Q(075), 0.05 in Q(0.90), and 0.05 in Q(0.95). Therefore, the Bayesian quantile regression model and the quantile regression model are consistent in predicting extreme rainfall for the next 1 year.

Table 4. Value of correlation each model

Training Data	Testing Data	Quantile	Correlation	
			Bayesian Quantile Regression	Quantile Regression
1981-2008	2009	0.75	0.96	0.95
		0.90	0.97	0.97
		0.95	0.98	0.98
1981-2007	2008	0.75	0.93	0.91
		0.90	0.91	0.91
		0.95	0.92	0.90
1981-2006	2007	0.75	0.92	0.94
		0.90	0.90	0.92
		0.95	0.88	0.90
1981-2005	2006	0.75	0.82	0.78
		0.90	0.89	0.84
		0.95	0.90	0.86

5 Conclusion

The Bayesian quantile regression model and quantile regression model can predict rainfall in one year ahead.The Bayesian quantile regression model is relatively better than the quantile regression model.

Acknowledgements

This work is fully supported by Kemenristek DIKTI (KementrianRisetTeknologidanPendidikan Tinggi) of Indonesia.

References

[1]Badan Meteologi, Klimatologi, dan Geofisika (BMKG): Laporan Mateorologi, Klimatologi, dan Geofisikan, Jakarta (2008)

[2] Cahyani, T.B.N.: Statistical Downscaling Modelling with Ridge and Elastic-net Regularized Quantile Regression for Rainfall Prediction in Indramayu, Thesis, IPB University, Bogor (2016)

[3] Chen, C. and Wei, Y.: Computation Issues For Quantile Regression. The Indian Journal Statistics. Vol. 67(2), pp. 399-417 (2005)

[4] Djuraidah, A. and Wigena, A.H.: RegresikuantiluntukEksplorasiPola Curah Hujan di KabupatenIndramayu. JurnalIlmuDasar, vol. 12(1), pp. 50-56 (2011)

[5] Koenker, R. and Bassett, G.: Regression Quantile. *Econometrica*, vol. 46(1), pp. 33-50 (1987)

[6] Kozumi, K. and Kobayashi, G.: Gibbs Sampling Methods for Bayesian Quantile Regression. Journal of Statistical Computation and Simulation, vol. 81(11), pp. 1565-1578 (2011)

[7] Geman, S. and Gaman, D.: Stochastic Relation, Gibbs Distribution, and The Bayesian Restoration of Image. IEEE Transaction on Pattern Analisys and Machine Intelligence, vol. 6, pp. 721-741 (1984)

[8] Mondiana, Y.Q.: Statictical Downscaling Modeling with Quantile Regression to Estimate Extreme Precipitation (A Case Study in Bangkir Station, Indramayu), Thesis, IPB University, Bogor (2012)

[9] Santri, D.: Statistical Downscaling Modeling with Quantile Regression using LASSO to Estimate Extreme Rainfall, Thesis, IPB University, Bogor (2016)

[10] Pribadi, H.Y.: Variabilitas Curah HujandanPergeseranMusim di Wilayah BantenSehubungandenganVariasisuhuMukaLautPerairan Indonesia, SamudraPasifikdanSamudraHindia, Thesis, Universitas Indonesia, Depok (2012)

[11] Wigena, A.H.: Statistical Downscaling Modeling using projection Pursuit Regression to Forecash Monthly Rainfall,Dissertation, IPB University, Bogor (2006)

[12] Yu, K. and Moyeed, R.A.: Bayesian Quantile Regresion. Statistics & Probability Latter, vol. 54(4), pp. 437-447 (2001)

[13] Zakarina, H.:Lasso and Ridge Quantile Regression using Cross Validation to Estimate Extreme Rainfall, Thesis, IPB University, Bogor (2016)