Quantile regression modeling to predict extreme precipitation

Y Q Mondiana*, A Zairina and R K Sari
Faculty of Forestry, Institut Pertanian Malang, Malang, Indonesia

*Corresponding author: yqmondiana@gmail.com

Abstract. Quantile regression is an extension of the median regression that analyzes various quantile values. This method is used to predict the relationship between the response variable (Y) and the predictor variable (X) on the conditional quantile function. Quantile regression can be used to detect extreme conditions, either extreme dry (quantile 5) or extreme wet (quantile 95). Extreme precipitation often occurs in Indonesian territory because the area is surrounded by oceans. High frequency of extreme precipitation may trigger disasters, one of which is flooding. In 2017, there were floods in Sidoarjo area with a loss of up to 2 billion. In an effort to anticipate the adverse effects of extreme precipitation, forecast information about extreme precipitation is needed, one of which is using quantile regression. The objectives of this research was to determine the best quantile regression model for predict extreme precipitation. The data of this study were secondary data obtained from BMKG with a data length of 30 years. Modeling of extreme precipitation in Sidoarjo area involved variables of humidity, temperature and air pressure for model 1. Whereas model 2 involved 3 variables in model 1 plus the month and month square variables. The addition of the month variable and the month quadratic variable in model 2 was based on the precipitation data plot which formed a quadratic trend. Based on the Pseudo $R^2$ value, it can be concluded that the best model to predict extreme precipitation in Sidoarjo is Model 2.

1. Introduction
Quantiles offer a convenient way to summarize univariate probability distributions as exemplified by Tukey’s ubiquitous boxplots. In contrast to moments, which characterize global features of the distribution and are consequently strongly influenced by tail behavior, quantiles are inherently local and are nearly impervious to small perturbations of distributional mass [1]. Quantile regression is an extension of the median regression that analyzes various quantile values. This method is used to predict the relationship between the response variable (Y) and the predictor variable (X) on the conditional quantile function. The model formed in quantile regression can be used to measure predictor variables effect not only at the center of distribution but also at the top and bottom tail of the distribution. Quantile Regression method does not require parametric assumptions and can be used to analyze a part of conditional distribution [2]. The estimate of quantile Regression is efficient if the error does not spread normally and it is robust with outliers. This method is very useful in application, especially when there are extreme values [3]. The quantile regression model has been used in research on human development indexes [4]. Quantile regression is also used in modeling pricing determinants in the hotel industry. The empirical results of quantile regression can help hoteliers in shaping investment and pricing strategies[5].
There are many extreme value cases, one of which is in precipitation case. Extreme precipitation events can be predicted by using canonical correlation analysis. But this method can not predict the extreme rainfall value [6]. Another method used in extreme precipitation prediction is multivariate regression analysis. This method can predict the value of extreme precipitation but cannot predict at different extreme level [7]. Prediction of extreme precipitation is also used a decision tree approach [8], neural network [9], kalman filter method [10], raw WRF model output [11]. Extreme precipitation often occurs in Indonesian territory because the area is surrounded by oceans. High frequency of extreme precipitation may trigger disasters, one of which is flooding. In 2017 there were floods in Sidoarjo area with a loss of up to 2 billion. In an effort to anticipate the adverse effects of extreme precipitation, forecast information about extreme precipitation is needed, one of which is using quantile regression.

2. Methods
The data of dependent variable used were monthly rainfall of the district Sidoarjo Juanda station. The method used in this research was: identifying the extreme precipitation with a box plot, then modelling the extreme precipitation in quantile 75, 90 and 95. The model was based on 3 predictor variables: humidity, temperature and air pressure for model 1 and variable of humidity, temperature, air pressure, month and month square for model 2. Then, the 2 models were compared based on pseudo $R^2$.

3. Results and Discussion
Descriptive statistics of rainfall are shown in Table 1.

| Month  | Average | Minimum | Maximum | Skewness |
|--------|---------|---------|---------|----------|
| January| 415     | 146     | 725     | 0.21     |
| February| 376    | 168     | 886     | 1.19     |
| March  | 296     | 117     | 609     | 0.62     |
| April  | 230     | 84      | 639     | 1.69     |
| May    | 167     | 2       | 393     | 0.49     |
| June   | 107     | 0       | 341     | 0.85     |
| July   | 38      | 0       | 245     | 2.12     |
| August | 10      | 0       | 141     | 2.82     |
| September| 10,1  | 0       | 129     | 3.47     |
| October| 35,8    | 0       | 502     | 4.47     |
| November| 104    | 0       | 329     | 0.85     |
| December| 232    | 0       | 427     | 0.02     |

The highest average rainfall occurred in January and the lowest in August. Information on skewness can be used to determine the tail of distribution. A positive skewness indicates tail to the right [11]. Skewness from January to December was bigger than zero. This value indicated that the distribution of monthly rainfall was not normal and stretched to the right (the average is greater than median and mode).

This value also revealed that there was high rainfall in the data. A box plot of monthly rainfall from 1989 to 2018 is presented in Figure 1.
Based on Figure 1, it can be seen that in February, April, July, August and October, there was rainfall that exceeded the monthly average rainfall. The most extreme rainfall occurred in February 2006 with rainfall of 886 mm.

Model 1 which involved 3 variables humidity, temperature dan air pressure showed that variables of humidity and air pressure have a significant effect on extreme precipitation at 75th, 90th and 95th quantile. Model 2 which used the month and month square variables, and 3 variables in Model 1 indicated that month, month square, humidity and air pressure variable have an effect on extreme precipitation at 75, 90 and 95th quantile. The best model selection based on pseudo $R^2$ value is shown in Table 2.

| Quantile | 75    | 90    |
|----------|-------|-------|
| Model 1  | 0.33  | 0.295 |
| Model 2  | 0.48  | 0.49  |

Based on Table 2, the largest pseudo $R^2$ is in model 2 for the 90th quantile. This means that extreme rainfall events recorded at Juanda observation station can be modeled more accurately at 90th quantile with the addition of the number of variables. The result of the extreme precipitation prediction is presented in Table 3.

| Q75     | Q90   | Q95   | Actual Y |
|---------|-------|-------|----------|
| 726.496 | 784.907 | 819.9946 | 314      |
| 548.222 | 775.257 | 781.9412 | 553      |
| 471.196 | 501.403 | 551.6576 | 433      |
| 331.87  | 345.498 | 380.856  | 270      |
| 233.612 | 236.702 | 260.0287 | 0        |
| 161     | 161.896 | 189.8909 | 77       |
| 91.152  | 122.112 | 133.5937 | 0        |
| 86.342  | 105.256 | 170.249  | 0        |
| 73.276  | 76.22  | 85.8191  | 0        |
| 90.334  | 108.881 | 119.8534 | 0        |
| 109.3   | 192.191 | 219.3345 | 150      |
| 227.91  | 325.788 | 362.245  | 387      |
The prediction of extreme precipitation in January to December was bigger than the actual data. Inaccurate predictions in actual data indicate that the number of predictive variable must be added such as wind direction [12], wind speed and sunshine duration [13]. However, the prediction has followed the actual data pattern as shown in Figure 2.

![Figure 2. Extreme precipitation prediction at 75th, 90th and 95th quantile](image)

Figure 2 shows that the prediction in quantile 75, 90 and 95 follows the actual data pattern. Even though the prediction plot had followed the data pattern, the pseudo R$^2$ was still less than 50%. Pseudo R$^2$ measure for evaluating goodness of fit in regression model [14]. This indicates that which also affects rainfall [12].

4. Conclusion
Extreme precipitation modeling using quantile regression can predict the extreme precipitation. This can be seen in the prediction plot which followed the actual data pattern. The addition of the month and month square could improve the accuracy of the model. Based on the Pseudo R$^2$ value, it can be concluded that the best model to predict extreme precipitation in Sidoarjo is Model 2.

References
[1] Koenker R 2016 *Annu. Rev. Econ.* 9(1) 155
[2] Buhai S 2005 *Ad Astra* 4(4) 1
[3] Djuraidah A and Wigena A H 2011 *J. Ilmu dasar* 12(1) 50
[4] Mohamed M R and S S Mzee 2017 *Int. J. Econ. Manag. Account.* 25(1) 27
[5] Hung W T, Shang J K and Wang F C 2010 *Int. J. Hosp. Manag.* 29(3) 378
[6] Maldonado T, Alfaro E, Fallas-López B and Alvarado L 2013 *Adv. Geosci.* 33 41
[7] Gao T and Xie L 2014 *Adv. Meteorol.* 20 (1)
[8] Wenguang W and Zhongwei Y 2019 *Int. J. Climatol.* 40(1) 255
[9] Xu Q et al 2017 *Expert Syst. Appl.* 76(2) 129
[10] Minh P T, B T Tuyet and Thao T T T 2018 *Vietnam J. Earth Sci.* 40(1) 367
[11] Gustari I 2009 *J. Meteorol. dan Geofis.* 10(1) 29
[12] Arieska D I and N H Pusponegoro 2017 *J. Apl. Stat. dan Komputasi Stat.* 8 57
[13] Zhang C et al 2019 *Sci. Total Environ.* 653 1426
[14] Chiang S M, Chen H F and Lin C T 2013 *Glob. Finance J.* 24(1) 30
[15] Hemmert 2016 *Sage J* 27(3)