The application of Bayesian quantile regression to analyse the relationship between nutrients content and phytoplankton abundance in Sutami reservoir

E D Lusiana, S Arsad, Kusriani, N R Buwono, I R Putri

Department of Water Resources Management, Faculty of Fisheries and Marine Science, Universitas Brawijaya, Malang, Indonesia

E-mail: evellinusiana@ub.ac.id

Abstract. Phytoplankton plays a significant role in aquatic ecosystem as the main feed for another aquatic biota. However, the abundance of phytoplankton must be controlled. This is due to the exaggeration of phytoplankton abundance can lead to eutrophication and the mass mortalities of fish. One of the dependent factors for phytoplankton abundance is the nutrients (nitrate and phosphate) in the relevant aquatic ecosystem. This study aims to analyze the relationship between nutrients and phytoplankton abundance, also to determine the maximum limit of the nutrient content in order to prevent eutrophication in Sutami Reservoir. The relationship usually analyzed by simple linear regression. Unfortunately, the data of phytoplankton abundance and nutrients content in Sutami Reservoir contains outlier according to Cook’s distance criteria. It means that simple linear regression cannot be used. Thus, the alternative method which is Bayesian quantile regression considered. The result of analysis indicates that the parameter value of regression model between nutrients content and phytoplankton abundance is varying, which are depended on the analyzed quantile.

1. Introduction
Reservoir is an artificial fresh water through damming certain rivers with various objectives, such as a prevention of flooding, electricity generation, water supply for agricultural irrigation needs, for fisheries activities, and even for tourism activities [1]. Thus the existence of the reservoir has provided its own benefits for the surrounding community. Sutami reservoir in Malang regency is the largest reservoir in East Java province which has a catchment area of 2050 km2 and storage capacity of 343,000,000 m3. The purposes of the reservoir are as flood prevention, irrigation, land fisheries, and tourism. In order to keep the role and function [2].

In order to maintain the role and function of the reservoir, monitoring of the condition of reservoir water quality is needed. Reservoirs with good water quality indicate that these waters have a low degree of pollution and a high level of fertility so that they can function optimally. The degree of pollution and water quality can be determined based on the presence of organisms and water quality parameters. One organism that plays a crucial role is plankton. Plankton are floating organisms whose movements depend on the flow of water [3]. Plankton consists of two types namely phytoplankton and zooplankton. Phytoplankton ranks top trophic level, or in other words it acts as a first-level feed source or primary producer [4].
Phytoplankton can carry out photosynthesis to convert inorganic materials into organic matter with the help of sunlight. Photosynthesis results from producers will be used for themselves and other organisms [1]. In addition to the light that plays a role in photosynthesis, phytoplankton are also strongly influenced by the presence of nutrients, especially nitrates and phosphates [5]. However, if the concentrations of such nutrients are too high, it will lead to algae blooming or eutrophication and water polluted instead. This condition will result to a decrease in water quality, low dissolved oxygen, arising gases and toxic substances (cyanotoxin) [6]. Therefore, it is very important to be able in analysing the relationship between nitrate and phosphates concentrations on phytoplankton abundance to anticipate eutrophication.

Most of the research to identify relationship between nutrients and phytoplankton abundance are using simple linear regression as the tools such as Lv et al. [7]. But, this analysis requires some classic assumptions (normality, non-multicollinearity, non-heteroscedasticity, and non-autocorrelation) and very sensitive in the presence of outlier. Hence, in this research, we consider Bayesian quantile regression to analyse relationship of nutrients and phytoplankton abundance in Sutami reservoir. Since this analysis does not need data distribution assumption and robust to outlier [8, 9]. Moreover, the used of Bayesian quantile regression is expected to be able to estimate nutrient limitation of nitrates and phosphates for high level abundance of phytoplankton [10].

2. Materials and methods
The objects of this research are nutrients content (mg/L) and phytoplankton abundance (ind/L) data in Sutami Reservoir that collected in February-March 2017. The sampling location is displayed in Figure 1.

In general, procedures of the study are data exploration to identify outlier using Cooks’ distance criteria, comparison the result of simple linear regression and Bayesian quantile regression, and estimation of nutrients limit related to high phytoplankton abundance. The analysis was conducted by using R version 3.4.3.
2.1. Cooks distance
Cooks distance is a measure that can be utilized to examine the presence of outlier in regression analysis. It is displaying the effect of the observation for fitted values. The measure can be calculated using formula (1)

\[ D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})'X'X(\hat{\beta}_{(i)} - \hat{\beta})}{(k+1)s^2} \]  

(1)

Where:
- \( \hat{\beta}_{(i)} \) = vector of parameter estimate when the ith observation is deleted
- \( \hat{\beta} \) = vector of parameter estimate using all observations
- \( X \) = matrix of predictors and constant
- \( k \) = number of parameters
- \( s^2 \) = variance of the fitted values

Those observations that have a Cook’s distance greater than 4 times the mean may be classified as influential or had extreme values [11].

2.2. Bayesian quantile regression
Quantile regression is a method to analyze relationship between variables at various quantile of the response variable [12]. Equation (2) shows the general form of quantile regression model [13].

\[ y_i = x_i'\beta(\theta) + \varepsilon(\theta), \quad 0 < \theta < 1 \]  

(2)

Where:
- \( y_i \) = response variable of the ith observation
- \( x_i' \) = \( (1, x_{i1}, x_{i2}, \ldots, x_{ip}) \)
- \( \beta(\theta) \) = parameter regression at the \( \theta \)th quantile
- \( \varepsilon(\theta) \) = error/residual model of the \( \theta \)th quantile
- \( i = 1, 2, \ldots, n \)

According to Koenker and Basset [14], parameter estimation of equation (2) is the solution of minimization (3) and (4):

\[ \min_{\beta \in \mathbb{R}^p} \left[ \sum_{i \in \{y_i \geq x_i' \beta \}} \theta|y_i - f(x_i)| + \sum_{i \in \{y_i < x_i' \beta \}} (1-\theta)|y_i - f(x_i)| \right] \]  

(3)

\[ \min_{\beta \in \mathbb{R}^p} \left[ \sum_{i \in \{y_i \geq x_i' \beta \}} \rho_\theta(y_i - f(x_i)) \right] \]  

(4)

\( \rho_\theta(u) = (\theta - 1_{u<0})u \), is called as check function.

Yu and Moyeed [15] found that minimization problem (4) is equivalent to the likelihood maximization of Laplace asymmetric function in equation (5):

\[ f_\theta(u) = \theta(1-\theta)\exp(-\rho_\theta(u)) \quad ,0 < \theta < 1 \]  

(5)

It is assumed that the error/residual model of quantile regression is distributed Laplace asymmetric, the equation (5) becomes:
\[ f_{\theta}(\epsilon_i) = \theta(1-\theta)\exp(-\rho_{\theta}(y_i - x_i^T \beta(\theta))) \quad , 0 < \theta < 1 \]  

Yu and Moyeed [15] stated that the basic principle of Bayesian modelling is to obtain posterior distribution as long as the prior distribution and likelihood function are known. The posterior distribution is presented in (7).

\[
\pi(\beta(\theta)|y) \propto L(y|\beta(\theta))\pi(\beta(\theta))
\]

\(\pi(\beta(\theta))\) is prior distribution for \(\beta(\theta)\) and \(L(y|\beta(\theta))\) is likelihood function that assuming the residual model is asymmetric Laplace distributed.

\[
L(y|\beta(\theta)) = \theta^\prime (1-\theta)^n \exp\left[-\sum_{i=1}^{n} \rho_{\theta}(y_i - x_i^T \beta(\theta))\right]
\]

The proposed prior distribution by Yu and Moyeed [15] is improper uniform. The analysis of Bayesian quantile regression is performed by using Metropolis-Hastings algorithm of MCMC method.

3. Results and Discussions

3.1. Outlier detection using Cooks distance

The outlier detection in relationship between nutrients content and phytoplankton abundance in Sutami reservoir based on simple linear regression are presented in following figures.

**Figure 2.** Cooks distance of simple linear regression analysis between nitrate and phytoplankton abundance  

**Figure 3.** Cooks distance of simple linear regression analysis between phosphate and phytoplankton abundance

It can be seen from Figure 2 and Figure 3 that there are outliers in the result of simple linear regression model that indicated by a datapoint upper the redline. It means that simple linear regression
is not suitable for analyzing relationship of nutrient content and phytoplankton abundance, so bayesian quantile regression needs to be considered.

3.2. Bayesian quantile regression result
The result of bayesian quantile regression analysis and its comparison to simple linear regression (OLS) is displayed as follows.

| Model                        | Distribution | Phyto-abundance (ind/L) | Coefficient |
|------------------------------|--------------|-------------------------|-------------|
|                              |              |                         | Nitrate     | Phosphate   |
| OLS                          | Mean         | 6993.56 (8.806)         | 2.459       | -0.495      |
| Bayesian quantile regression | q=0.25       | 5972.973 (8.695)        | 1.54        | -1.03       |
|                              | q=0.50       | 7016.361(8.856)         | 2.13        | -1.3        |
|                              | q=0.75       | 7942.632 (8.980)        | 1.88        | -1.25       |

Table 1 shows that the coefficients from bayesian quantile regression are varying among quantile. Meanwhile, OLS model only produce one model so that it cannot represents the dataset well, especially when outlier is exist. Sign of coefficient for nitrate is positive or in other word if the nitrate content is high, it will cause the high abundance of phytoplankton. However, the sign of phosphate coefficient is negative, or when the phosphate content is increasing, the abundance of phytoplankton is decreasing instead. Negative values on phosphate variables can also be caused by the composition and relative abundance of the highest phytoplankton, from the Cyanophyta division, where the Cyanophyta division utilizes phosphate rather than nitrate for growth because the division is capable of nitrogen fixation. Therefore, the higher the concentration of phosphate accompanied by an increase in abundance of Cyanophyta can cause phytoplankton death due to competition in the utilization of these nutrients.

Furthermore, by using bayesian quantile regression method, the nutrient limiting to anticipate eutrophication can be simply calculated based on the upper quantile (q=0.75). The nitrate limit is 0.791 mg/L. However, the nutrient limiting for phosphate cannot be calculated since the model is not appropriate.

4. Conclusion
There are found outliers in the relationship between nutrient content and phytoplankton abundance in Sutami reservoir, so that simple linear regression or OLS method is not suitable to be used. As the alternative, bayesian quantile regression analysis is considered and the results showed that the coefficients regression are varied among quantile. Based on the model, the maximum concentration of nitrate to anticipate eutrophication is 0.791 mg/L. However, the model with phosphate as the predictor is contradicting with the theory because of the the highest abundance of phytoplankton is from Cyanophyta division which can utilizes phosphate rather than nitrate for growth.

References

[1] Apridayanti E 2008 Management Evaluation of Lahor Reservoir Malang Regency, East Java. In: Master of Environmental Science Universitas Diponegoro Semarang Indonesia. [In Indonesian]

[2] Suroso, Anwar M R, Rahmanto M C 2007 Study of the Effect of Brantas River Sedimentation on the Capacity and Age of the Malang Sutami Reservoir Plan Jurnal Rekayasa Sipil 1 33-42. [In Indonesian]
[3] Odum E P, Barrett G W 2005 Fundamentals of Ecology Thomson Brooks/Cole California USA.

[4] Nybakken J W 1992 Marine Biology PT. Gramedia Jakarta Indonesia. [In Indonesian]

[5] Suryanto A M 2011 Abundance and phytoplankton composition in Selorejo Reservoir, Ngantang District, Malang Regency Jurnal Kelautan 4 34-39. [In Indonesian]

[6] Sugiura N, Utsumi M, Wei B, Iwami N, Okano K, Kawauchi Y, Maekawa T 2004 Assessment for the complicated occurrence of nuisance odours from phytoplankton and environmental factors in a Eutrophic Lake Lake Reserv. Res. Manag. 9 195-201.

[7] Lv J, Wu H, Chen M 2011 Effects of nitrogen and phosphorus on phytoplankton composition and biomass in 15 subtropical, urban shallow lakes in Wuhan, China Limnologica 41 48-56.

[8] Yu K, Stander J 2007 Bayesian analysis of tobit quantile regression model J. Econ. 137 260-276.

[9] Yue Y R, Hong H G 2012 Bayesian Tobit quantile regression model for medical expenditure panel survey data Stat. Model. 12 323-346.

[10] Xu Y, Schroth A, Isles P, Rizzo D 2015 Quantile Regression Improves Models of Lake Eutrophication with Implications for Ecosystems-Specific Management Freshwater Biol. 60 1841-1853.

[11] Zakaria A, Howard N K, Nkansah B K 2014 On the detection of influential outliers in linear regression analysis Am. J. Theoret. Appl. Stat. 3 100-106.

[12] Koenker R, Hallock K 2001 Quantile Regressions J. Econ. Perspect. 15 143-156.

[13] Buhai I S 2005 Quantile regression: overview and selected applications Ad Astra 4 1-17.

[14] Koenker R, Basset G 1978 Regression quantiles Econometrica 46 33-50.

[15] Yu K and Moyeed R A 2001 Bayesian quantile regression Stat. Probability Lett. 54 437-147.