Ecological Questions 30 (2019) 2: 69–77 http://dx.doi.org/10.12775/EQ.2019.010

Performance of Bayesian quantile regression and its application to eutrophication modelling in Sutami Reservoir, East Java, Indonesia

Evellin Dewi Lusiana[1,2*], Sulastri Arsad[1,2], Kusriani[1], Nanik Retno Buwono[1,2], Istiani Ratna Putri[1]

[1]Faculty of Fisheries and Marine Science, Brawijaya University, Malang, Indonesia
[2]AquaRES Research Group, Brawijaya University, Malang, Indonesia,
[*]e-mail: evellinlusiana@ub.ac.id

Received: 3 December 2018 / Accepted: 15 February 2019

Abstract. Phytoplankton has an important role in aquatic ecosystem as the primary natural feed for another aquatic biota. However, the density of phytoplankton must be controlled at desirable level in order to prevent eutrophication. Because eutrophication can damage the ecosystem and as the consequence will led to mass mortalities of fish. There are several factors that affecting phytoplankton density such as light availability, temperature, pH, and nutrient content. Nutrient content is composed by nitrate and phosphate. The relationship between nutrient content and phytoplankton density commonly performed by using simple linear regression. But, this method cannot give an overall description of the data since it is worked at conditional mean. Moreover, simple linear regression has several limitation like highly influence by outlier and need the fulfillment of classical assumption. Thus, the aims of this research are to offer an alternative method namely Bayesian quantile regression and provide its performance compared to simple linear regression under various data condition. Also to apply such model to the relationship between nutrients content and phytoplankton density in Sutami Reservoir. The results indicate that Bayesian quantile regression performs better than simple linear regression when the outlier exists. Unfortunately, the data of phytoplankton density and nutrients content in Sutami Reservoir contains outlier according to Cook's distance criteria. It means that Bayesian quantile regression should be used. The obtained model showed that the parameter values of regression model between nutrients content and phytoplankton density vary, which are depended on the analyzed quantile

Key words: simulation, phytoplankton density, nitrate, phosphate, nutrient limitation.

1. Introduction

Reservoir is an artificial lentic water which has been built for several purposes comprised avoiding flood, electricity generating power, water supply for agricultural activity, fisheries activity, and or tourism. One of this reservoir is Sutami reservoir where located in Malang regency is the largest reservoir in East Java province which provided a catchment area of 2050 km^2 and 343,000,000 m^3 of capacity storage. The establishment of Sutami reservoir were aim as flood prevention, water supply for irrigation, land fisheries, and tourism. So that the functions of the reservoir preserve, monitoring of the water quality is needed. If the reservoir provides with good water quality, it is reflecting that the waters have a low pollution and fertile so that they can work optimally. Biomonitoring or the presence of certain organisms also determines the water quality. One of organism that plays a crucial role is plankton. Plankton consists of two kind of flora and fauna namely phytoplankton and zooplankton, respectively. These organisms are floating on the water body and following the currents. Phytoplankton acts as a first-level feed as primary producer (Nybakken, 1992).

Phytoplankton are able to perform photosynthesis in order to convert inorganic materials into organic matter with the help of sunlight. Phytoplankton relay on photosynthesis to produce carbohydrate for energy. On the other side, they need other nutrients to grow and reproduce. These nutrients are typically phosphorus and nitrogen (Kadim & Arsad, 2016). However, if the concentration of such nutrients is 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) (Sugiura et al., 2004). Therefore, it is very important to be able in analyzing the relationship between nitrate and phosphates concentrations on phytoplankton density to anticipate eutrophication.

Most of the research to identify relationship between nutrients and phytoplankton density are using simple linear regression as the tools Lv et al. (2011), Donald et al. (2013), Paiki and Kalor (2017) also Putri et al. (2014). But this analysis requires some classic assumptions (normality, non-multicollinearity, non-heteroscedasticity, and non-autocorrelation) and very sensitive in the presence of outlier. Hence, the one of the objective of this research is to apply Bayesian quantile regression as a method in analysing the relationship of nutrients and phytoplankton density in Sutami reservoir. Since the method does not need data distribution assumption and robust to outlier (Yu & Stander, 2007; Yue & Hong, 2012). Moreover, the used of Bayesian quantile regression is expected to be able to estimate nutrient limitation of nitrates and phosphates for high level density of phytoplankton (Xu et al., 2015).

2. Study area

The research was conducted in Sutami Reservoir located in Malang Regency in East Java Island (Fig. 1). The reservoir has 3 sources of flow, namely the Brantas River, Metro River and Sengguruh River and obtains input from Lahor Reservoir.

The data were collected in February 2017 until August 2018. The sampling location consists of three sites. Site 1or inlet is the upstream area is the entry of water from the Brantas River and the Sengguruh River. Site 2 or center

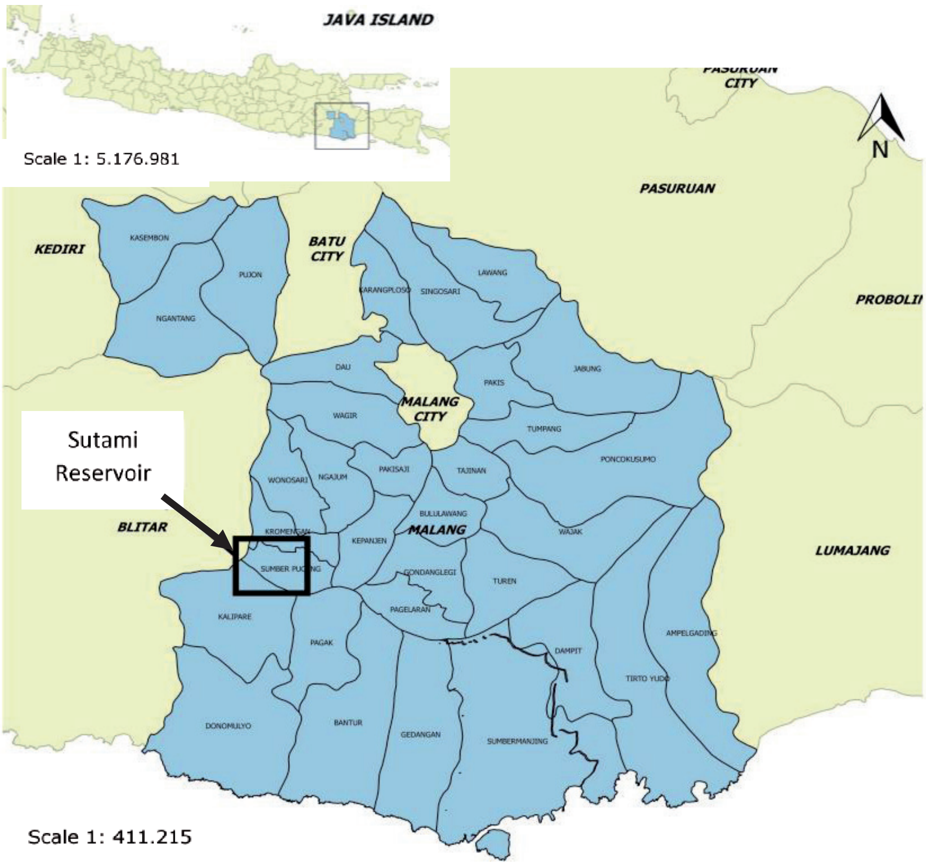


Figure 1. The Location of Sutami Reservoir at Malang Regency

is the inner part of the reservoir. Site 3 is downstream area adjacent to the water exit that will be flowed for the purpose of generating electricity. Those three stations can be seen in Figure 2.

3. Material and methods

The materials of this research are phytoplankton (genera, composition, and density). The samples were collected horizontally by using plankton net (mesh size 45 µm) in the water column. Identification process of samples was carried out by using light microscope (Olympus CX-21LED; magnification 400x), then morphological characteristic was determined appropriately by using Prescot's book Identification (Prescott, 1978). Furthermore, phytoplankton density was measured by using *Lackey drop* (microtransect) method (APHA, 1989) with the following formula:

$$D = \frac{C.A_t}{A_s.S.V} ,$$

where
D = phytoplankton desnity (cell/ml),
C = number of organisms counted,
A_t = area of cover slip (mm²),
A_s = area of one strip (mm²),
S = number of strips counted,
V = volume of sample under the cover slip (ml).

Other waters quality indicators were collected in Sutami Reservoir are nitrate (NO_3, mg/l) and orthophosphate (PO_4, mg/l) and measured ex-situ (APHA, 1989). As for the methods and procedures of this study are descriptive analysis of phytoplankton composition and water quality measurement, data exploration to identify outlier using Cooks' distance criteria, comparison the result of simple linear regression and Bayesian quantile regression by simulation study, and estimation of nutrients limit of nitrate and phosphate that related to high phytoplankton density. The analysis was conducted by using Microsoft Excel version 2010 and software R version 3.4.3.

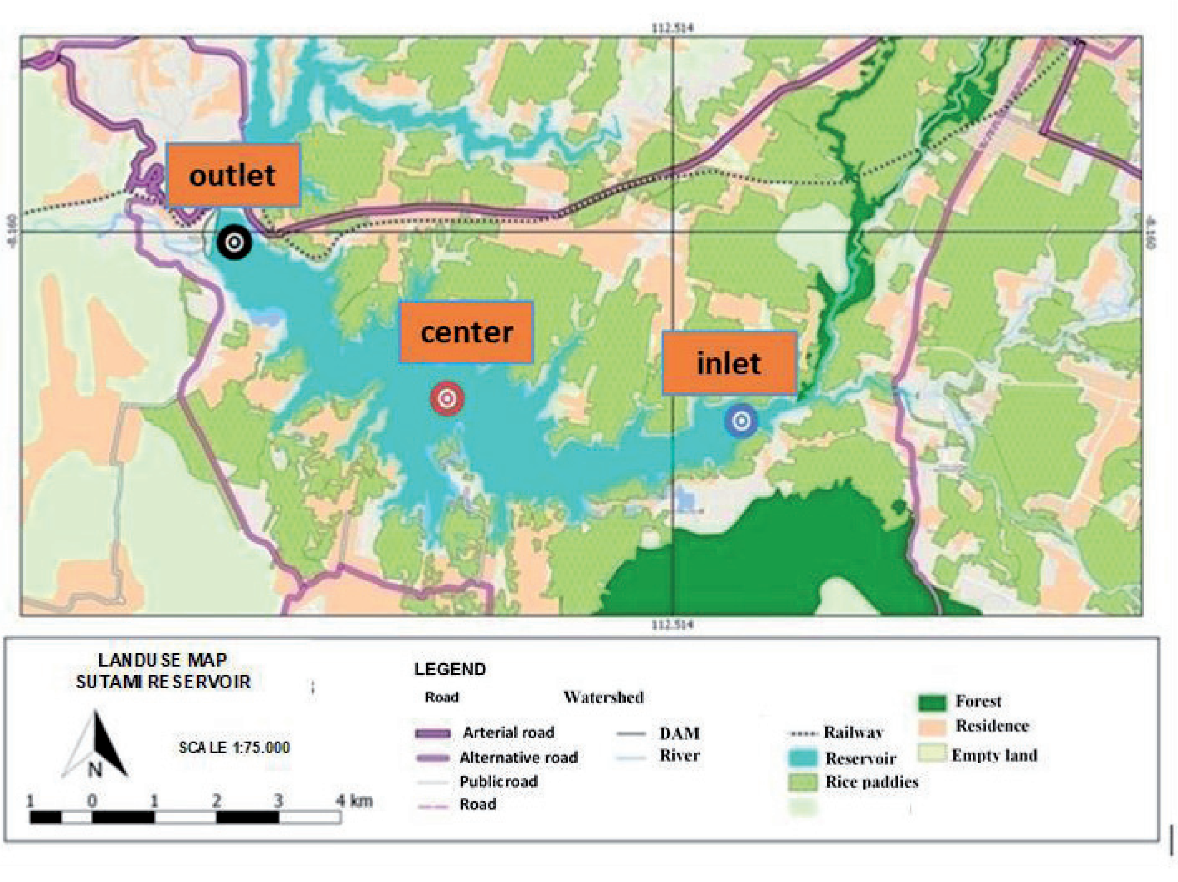


Figure 2. Sampling Location in Sutami Reservoir

3.1. Outlier detection: Cooks distance

Cooks distance is a measure that can be used to assess the presence of outlier in regression analysis. Formula presents the formulation to calculate Cooks' distance.

$$D_i = \frac{\left(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}}\right)'\mathbf{X}'\mathbf{X}\left(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}}\right)}{(k+1)s^2} \quad (1)$$

where

$\hat{\boldsymbol{\beta}}_{(i)}$ = vector of parameter estimates when the *i*-th observation is deleted,
$\hat{\boldsymbol{\beta}}$ = vector of parameter estimates using all observations,
$\mathbf{X}$ = matrix of predictors and constant,
k = number of parameters,
s^2 = variance of the fitted values.

Observations who have a cook's distance greater than 4 times the mean could be classified as influential or extreme values (Zakaria et al., 2014).

3.2. Bayesian quantile regression

Quantile regression is a method to analyze relationship between variables at various quantile of the response variable (Koenker & Hallock, 2001). Equation shows the general form of quantile regression model (Buhai, 2005).

$$y_i = \mathbf{x}_i^t\boldsymbol{\beta}(\theta) + \varepsilon(\theta)_i \quad 0 < \theta < 1 \quad (2)$$

where

y_i = response variable of the *i*-th observation,
$\mathbf{x}_i^t$ = $\left(1, x_{1i}, x_{2i}, \ldots, x_{pi}\right)$,
$\boldsymbol{\beta}(\theta)$ = parameter regression at the θ-th quantile,
$\varepsilon(\theta)_i$ = error/residual model of the θ-th quantile,
i = 1, 2,..., n.

According to Koenker and Basset (1978), parameter estimation of equation is the solution of minimization and

$$\min_{\boldsymbol{\beta}\in\mathbf{R}^p}\left[\sum_{i\in\{i:y_i\geq\mathbf{x}_i'\boldsymbol{\beta}\}}\theta\left|y_i - f(\mathbf{x}_i)\right| + \sum_{i\in\{i:y_i<\mathbf{x}_i'\boldsymbol{\beta}\}}(1-\theta)\left|y_i - f(\mathbf{x}_i)\right|\right] \quad (3)$$

$$\min_{\boldsymbol{\beta}\in\mathbf{R}^p}\left[\sum_{i\in\{i:y_i\geq\mathbf{x}_i'\boldsymbol{\beta}\}}\rho_\theta\left(y_i - f(\mathbf{x}_i)\right)\right] \quad (4)$$

$\rho_\theta(u) = (\theta - 1_{\{u<0\}})u$, is namely *check function.*

Yu and Moyeed (2001) stated that minimization problem is equivalent to the likelihood maximization of Laplace asymmetric function in equation

$$f_\theta(u) = \theta(1-\theta)\exp(-\rho_\theta(u)) \quad (5)$$

it is assumed that the error/residual model of quantile regression is distributed Laplace asymmetric, the equation becomes

$$f_\theta(\varepsilon_i) = \theta(1-\theta)\exp(-\rho_\theta(y_i - \mathbf{x}_i^t\boldsymbol{\beta}(\theta)) \quad (6)$$

Yu and Moyeed (2001) said 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

$$\pi(\boldsymbol{\beta}(\theta)|\mathbf{y}) \propto L(\mathbf{y}|\boldsymbol{\beta}(\theta))\pi(\boldsymbol{\beta}(\theta)) \quad (7)$$

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

$$L(y|\boldsymbol{\beta}(\theta)) = \theta^n(1-\theta)^n\exp\left[-\sum_{i=1}^n\rho_\theta\left(y_i - \mathbf{x}_i^t\boldsymbol{\beta}(\theta)\right)\right].$$

The proposed prior distribution by Yu and Moyeed (2001) is *improper* unifor. The analysis of Bayesian quantile regression is performed by using Metropolis-Hastings algorithm of MCMC method.

3.3. Simulation study

A simulation study is considered in this research to compare the performance of simple linear regression and Bayesian quantile regression analysis. The simulation was done by generate data based on:

a) Generating response y_i from model $y_i = \beta_0 + \beta_1 x + \varepsilon_i$,
b) The parameter was set to 1.00 and three types of error distribution (standard normal, heteroscedastic, and contained outlier),
c) Sample size n = 25, 50, and 100.

Each generated data was modelled using simple linear regression and Bayesia quantile regression at median. The better model was evaluated by smaller RMSE (Root Mean Square Error).

4. Results and discussion

4.1. Composition and density of phytoplankton

Based on the results of identification of the genera and density of phytoplankton (Fig. 3), phytoplankton genera from four class and 22 genera were found. The classification is *Chlorophyceae* (10 genera), *Bacillariophyceae* (7 genera), *Cyanophyceae* (3 genera), and *Dynophyceae* (2 genera) at Site 1, 2 and 3. The types of phytoplankton identified from the *Chlorophyceae* class include *Ankis-*

trodesmus, *Chlorella*, *Cosmarium*, *Gloeocystis*, *Mougeotia*, *Pediastrum*, *Scenedesmus*, *Selenastrum*, *Staurastrum*, and *Tetraedron*. Moreover, in the *Bacillaryophyceae*, the genera of phytoplankton identified included *Cyclotella*, *Cymbella*, *Navicula*, *Neidium*, *Nitzschia*, *Pinnularia*, and *Synedra*. The *Cyanophyceae* only found three genera, namely *Merismopedia*, *Microcystis*, and *Spirulina*. Nonetheless, the density of four class is most commonly found in all sampling Site. On the other hand, in the *Dinophyceae*, the genera and density found were the least, namely *Ceratium* and *Peridinium*.

The density of species from the *Cyanophyceae* has the highest percentage compared to other class at all stations. While the genera of phytoplankton from the *Dinophyceae* was found the fewest in the waters of the Sutami Reservoir. The most found and dominating types of phytoplankton are *Microcystis* sp. Sachlan (1972) said that *Microcystis* sp. small round shape and colony life. Richmond (2005) stated the density of the *Cyanophyceae* because this phytoplankton is able to adapt to unfavorable conditions (low CO_2, low or too high temperatures, and low light). Furthermore, the height of *Cyanophyceae* is caused when sampling is done when the intensity of sunlight is not too high. Ac-

Figure 3. Phytoplankton composition (a) Site 1; (b) Site 2; (c) Site 3

cording to Goldman and Horne (1994), during the morning *Cyanophyceae* will float to the surface of the water. The vertical movement of *Cyanophyceae* is because it has vacuole gas.

The phytoplankton density at Site 1 is on average 4,750 cells/ml, at Site 2 – 4,389 cells/ml, and at Site 3 – 4,673 cells/ml. According to Landner (1978), if the density of phytoplankton ranges from 2000 to 15,000 cells/ml, the waters are classified as mesotrophic. It can therefore be concluded that the waters in the Sutami Reservoir are entirely mesotrophic. This means that they have a moderate level of fertility. The high proportion of *Cyanophyceae* in the species composition at all sites indicates, however, that the Sutami flow-through reservoir located in tropical zone may be at risk of eutrophication.

4.2. Simulation results

The RMSE values of simulation study can be seen in following figures:

According to Figure 3(a) and 3(b), estimator of simple linear regression has smaller RMSE than Bayesian quantile regression. However, when the outlier exists as in Figure 3(c), the latter performs better. Thus, outlier detection is needed before implementing the regression analysis.

4.3 Outlier detection

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

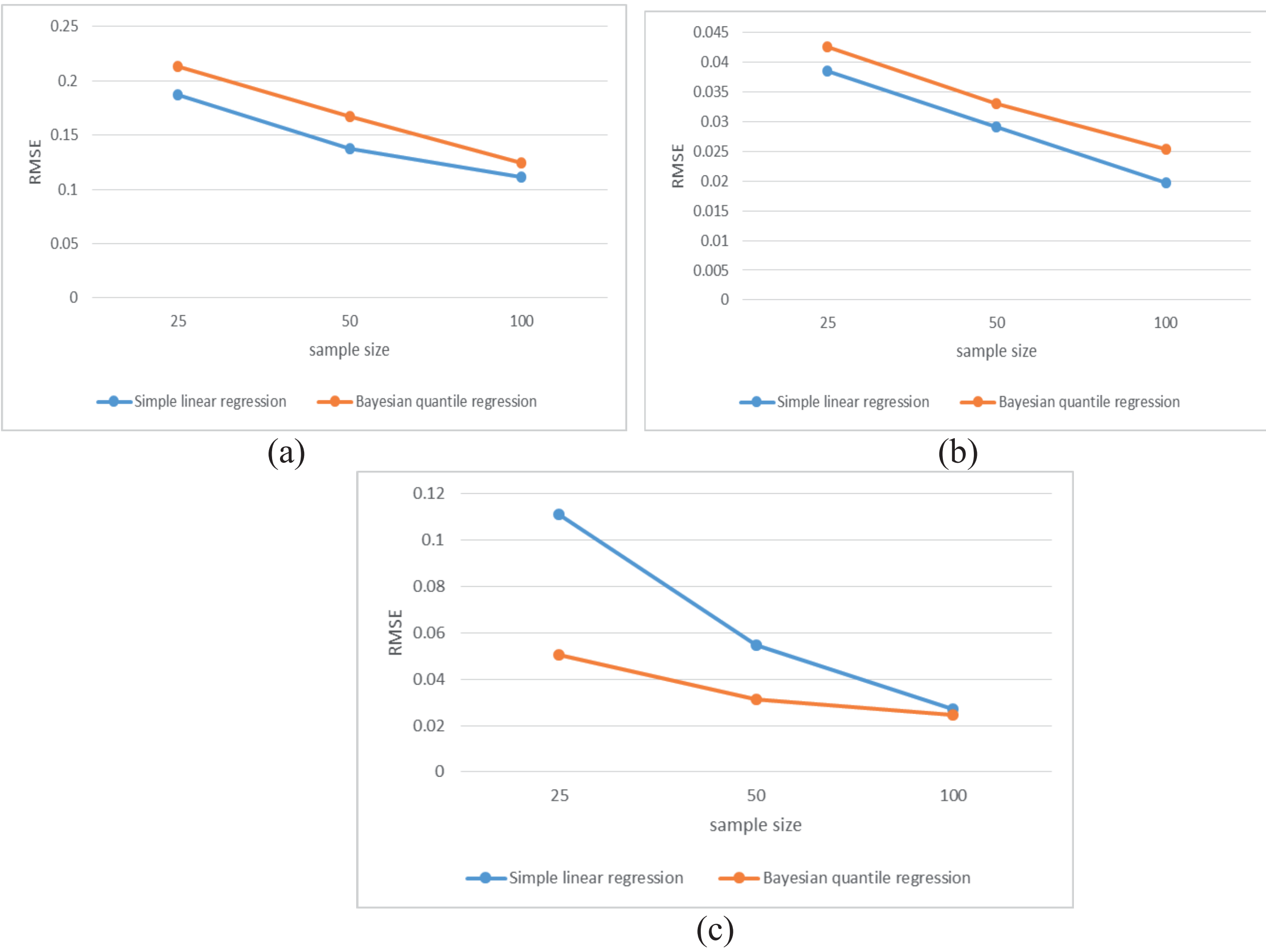
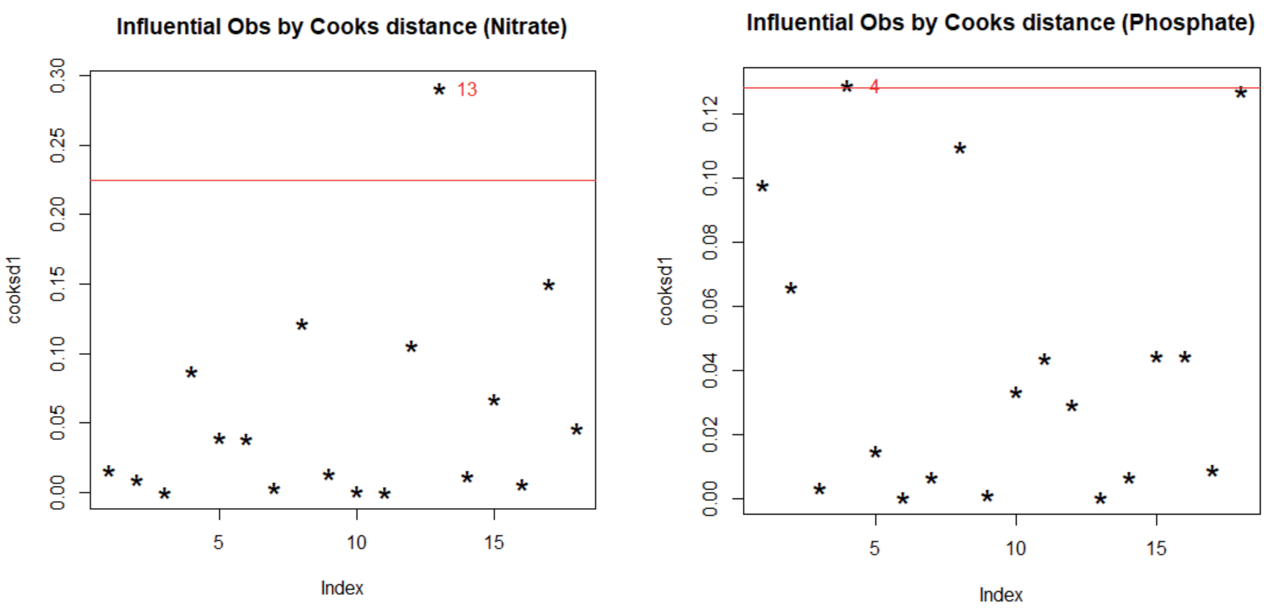


Figure 3. RMSEs comparison of simple linear regression and Bayesian quantile regression (θ=0.50) estimator under error distribution (a) standard normal; (b) heteroscedastic; (c) outlier



(a) (b)

Figure 4. Cooks distance of simple linear regression analysis between (a) nitrate and phytoplankton density; (b) phosphate and phytoplankton density

It can be seen from Figure 4(a) and 4(b) that there are outliers in the result of simple linear regression model that indicated by a data point upper the red-line. It means that simple linear regression is not suitable for analyzing relationship of nutrient content and phytoplankton density, so Bayesian quantile regression needs to be considered.

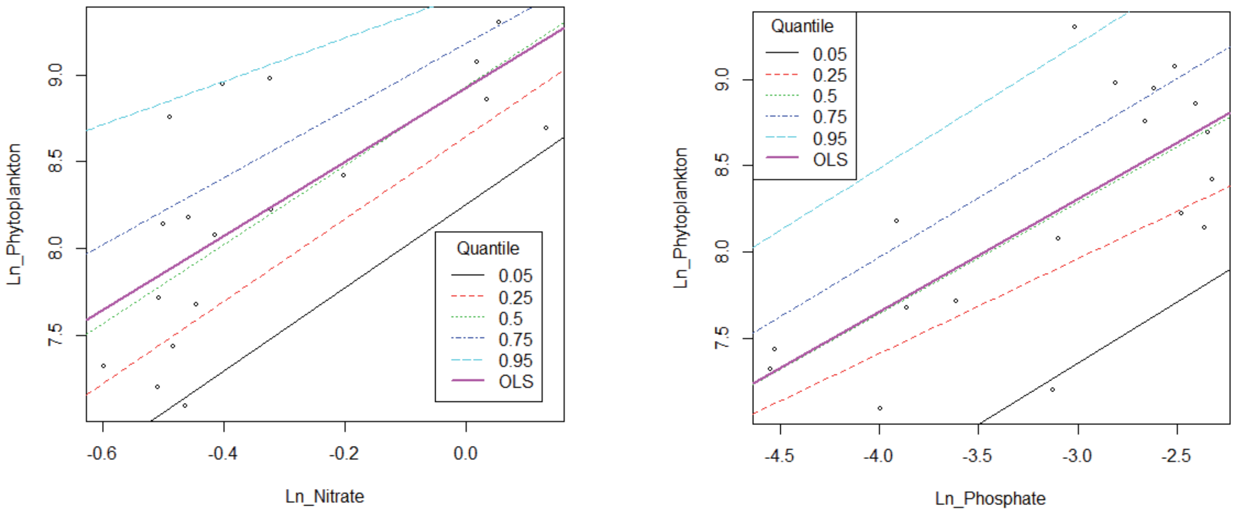
4.4. Bayesian quantile regression and nutrient limit estimation

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

Table 1 shows that the coefficients from Bayesian quantile regression are varying among quantile. Meanwhile, simple linear regression model only produce one

Table 1. Parameter estimates of simple linear and Bayesian quantile regression

Model	Distribution	Phytoplankton-abundance (ind/l)	Parameter estimates	
			Nitrate	Phosphate
Simple linear regression	Mean	3,738.71	2.140	0.655
Bayesian quantile regression	0.05	1,315.99	2.334	0.639
	0.25	2,177.75	2.385	0.568
	0.50	3,638.23	2.292	0.655
	0.75	6,845.84	1.870	0.707
	0.95	9,020.94	1.140	0.706



(a) (b)

Figure 5. Scatterplot and fitted line Bayesian quantile regression and simple linear regression of (a) nitrate regressor; (b) phosphate regressor

model so that it cannot represents the dataset well, especially when outlier is exist. This matter also supported by fitted line of Bayesian quantile regression and simple linear regression as in Figure 5.

The parameter estimates for nitrate and phosphate are positive, which mean that if each the nitrate and phosphate content is increasing, it will cause the high density of phytoplankton. Furthermore, by using Bayesian quantile regression method, the nutrient limiting to anticipate eutrophication can be simply calculated based on the upper quantile (q=0.95). The nitrate limit is 0.765 mg/l and limiting content for phosphate is 0.043 mg/l.

5. Conclusions

Bayesian quantile regression estimator is outperformed the simple linear regression when the outlier exists. According to outlier detection, there are found outliers in the relationship between nutrient content and phytoplankton density in Sutami reservoir, so that Bayesian quantile regression should be used. The result showed that the coefficients regression are varying among quantile. Based on the model, the nitrate and phosphate limit to anticipate eutrophication are 0.765 mg/l and 0.043 mg/l respectively.

References

APHA, 1989, Standard Methods for the Examination of Water and Wastewater, 17th edition. Prepared and published jointly by American Public Health Association, American Water Work Association, Water Pollution Control Federation (joint editorial board A.E. Greenberg, L.S. Cleseri, A.D. Eaton). Washington DC.

Buhai I.S., 2005, Quantile Regression: Overview and Selected Applications. Ad Astra 4, 1-17.

Donald D.B., Bogard M., Finlay K., Bunting L. & Leavitt P., 2013, Phytoplankton-Specific Response to Enrichment of Phosphorus-Rich Surface Waters with Ammonium, Nitrate, and Urea. PlLoS ONE 8: 1-14.

Goldman C.R. & Horne A.J., 1994, Limnology. McGraw-Hill, New York.

Kadim M.K. & Arsad S., 2016, Distribution and Density of Microalgae Based on Coastal Characteristic and Ecology in Bone Bolango Coastal Region. Asian Jr. of Microbiol. Biotech. Env. Sc. 18: 395-401.

Koenker R. & Basset G., 1978, Regression Quantiles. Econometrica 46: 33-50.

Koenker R. & Hallock K., 2001, Quantile Regressions. Journal of Economic Perspectives 15: 143-156.

Landner L, 1978, Eutrophication of lakes. Analysis Water and Air Pollution Research Laboratory Stockholm, Stockholm.

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

Nybakken J.W., 1992. Marine Biology. PT. Gramedia, Jakarta.

Paiki K. & Kalor J.D., 2017, Distribusi Nitrat dan Fosfat Terhadap Kelimpahan Fitoplankton di Perairan Pesisir

Yapen Timur. Journal of Fisheries and Marine Science 1: 65-71.

Prescott G.W., 1978, The Freshwater Algae. WM. C. Brown Company Publishers., Iowa.

Putri F.D.M., Widyastuti E. & Christiani, 2014, Hubungan Perbandingan Total Nitrogen dan Total Fosfor dengan Kelimpahan Fitoplankton di Perairan Waduk Panglima Besar Soedirman, Banjarnegara. Scripta Biologica 1: 96-101.

Richmond A., 2005, Microalgal Culture, Biotechnology and Applied Phycology. Blackwell Publishing, New Jersey.

Sachlan M., 1972, Planktonologi. Direktorat jenderal Perikanan, Departemen Pertanian, Jakarta.

Sugiura N., Utsumi M., Wei B., Iwami N., Okano K., Kawauchi Y. & Maekawa T., 2004, Assessment for tthe Complicated Occurence of Nuisance Odours from Phytoplankton and Environmental Factors in a Eutrophic Lake. Lake & Reservoirs: Res. and Man 9, 195-201.

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

Yu K. & Moyeed R.A., 2001. Bayesian Quantile Regression. Statistics and Probability Letters 54: 437-447.

Yu K. & Stander J., 2007, Bayesian Analysis of Tobit Quantile Regression Model. Journal of Econometrics 137: 260-276.

Yue Y.R. & Hong H.G., 2012, Bayesian Tobit Quantile Regression Model for Medical Expenditure Panel Survey Data. Statistical Modelling 12: 323-346.

Zakaria A., Howard N.K. & Nkansah B.K., 2014, On the detection of influential outliers in linear regression analysis. American Journal of Theoretical and Applied Statistics 3: 100-106.