Quantile Smoothing Splines Regression on The Air Relative Humidity Toward The Rubber Production in South Sumatera

Y Nalita1\* and Y Andriyana2

1Magister Student of Applied Statistics, Universitas Padjadjaran, Jalan Dipati Ukur No.35, Lebakgede, Kecamatan Coblong, Kota Bandung, Jawa Barat, Indonesia

2Department of Statistics, Universitas Padjadjaran, Jalan Raya Bandung-Sumedang KM 21, Jatianangor, Sumedang, Jawa Barat Indonesia

\*yollanda15001@mail.unpad.ac.id

**Abstract**. Rubber is one of the Indonesian plantation commodities which has an important role on a the economic activity. Rubber can grow very well in tropical area. One of the factors influencing the rubber production is the relative humidity of the air. This research is aimed at finding out the pattern showing the influence of the air relative humidity and the rubber production in South Sumatera due to the relationship pattern shows different shapes in some parts indicating a difficulty to specify parametrically, we then propose a nonparametric approach. In this case we use smoothing splines technique. However, the pattern spreads asymmetrically and the are some outliers. Hence, we also propose to use quantile objective function called quantile smoothing splines. In this research we divide the data distribution into five parts based on very low, low, moderate, high and very high characteristics. The performances of quantile smoothing splines (with , as median case) and smoothing splines (mean approach) are compared by Root Mean Square Error. As the result, it is shown that the quantile smoothing splines regression gives smaller Root Mean Square Error compared to mean smoothing splines regression.

1. Introduction

Agriculture, forestry and fishery sectors are holding important role in the economic activity. Rubber is one of the plantation commodities which has quite important role in the economic activity [1]. Generally, there are many factors influencing crop production and these include soil, climate, and diseases among others with rainfall, one of the dominant variables in tropical agriculture is relative humidity that can determine global plant distribution and productivity [11]. The production pattern of rubber plantation in Indonesia is influenced by several factors from time to time, such as the climate that influences the deciduous pattern of leaves and the pattern of annual production. The relative humidity are critical factors when selecting sites in rubber zone [11].

The high rainfall will inflence the humidity in which it will make the humidity in high level because it deals with the radiation of the sun. The problem of deciduous leaves and branches disease has close relationship to biological environment factor which impact on the productivity of rubber plant and this problem depends not only on the plants level of resistance on the disease but also the air humidity in the platation area. A very high humidity will cause the plants to be susceptible to a certain disease. Therefore, air humidity is said to have high influence on the rubber production. On the other hand, humidity also gives good effect on the soil moisture which will help the well growth of the plant. Hence, a good range of humidity is needed in rubber plantation. The development of rubber commodity fluctuates from year to year, where the biggest production of dry rubber in 2018 from South Sumatera was 30,83 percent of the total national dry rubber production [12] . South Sumatera is located on 50 10’ - 10 20’ SL and 1010 40’ - 1060 30’ EL on the world map and this geographical location makes South Sumatera becomes one of the tropical area with high humidity [11].

The above factors raise questions about the relationship pattern of the air relative humidity toward the ruber production in South Sumatera from 2013 – 2018 using regression model. In this research, the relationship pattern shows different shapes in some parts indicating a difficulty to specify parametrically,therefore nonparametric approach is used with smoothing splines regression method. Based on the data of the rubber production and the variation of air relative himidity in South Sumatera has the pattern spreads asymmetrically and the are some outliers, then use of mean smoothing spline regression becomes less optimal because it cannot represent some parts of the response distribution, therefore regression lines from several quantile are needed so more information can be obtained to find out the relationship between these two variables. Quantile regression is also robust to the outlier [5]. Hence, this research apllies the quantile smoothing splines regression method introduced by Koenker et al.(1994) using packages “quantreq” with function “rqss” on Software R [7]. Then, to know performance this method, we will compare between the smoothing spline regression by mean and quantile which as median case.

1. Method

Data used in this research is secondary data obtained from Badan Pusat Statistik and satellite Aqua -Moderate Resolution Imaging Spectroradiometer (AQUA-MODIS). The rubber production data is obtained from the statistic publication of Indonesian rubber from year to year. The data used is the result of Indonesian plantation dry rubber production monthly with the unit of thousands of tons. Beside the result of rubber production, the variables being observed is the percentage of the air relative humidity which is the satelite AQUA-MODIS with AIRS instrument [13] . The research limits the data used by only taking data of South Sumatera Province during the period of January 2013 up to December 2018. Variables which are about to be observed are one response variable (*Y*) and one predictor variable (*X*):

 : Rubber production in South Sumatera (thousands of tons)

 : Air relative humidity in South Sumatera (%)

## Regression Analysis

Regression analysis is an analysis to trace the functional relationship pattern between response variable and predictor variables, in which it is expressed in the equation which relates the response variable *Y* to one or more predictor varibles in general formulation:

 (1)

Regression model basically has three approaches among which are parametric, nonparametric and semiparametric which is the combination of both. Parametric approach is an approach of which model is already known or the regression function is known, hence the regression structure has been discovered to do a prespecification model. Generally, the regression model is focused on linear regression can be written as follows:

 (2)

With is the corresponding regression coefficients must be estimated, is model residual to*-i*, parametric approach with linear regression generally obligates to fullfiling a strict assumption. Nonparametric regression is one of statistical methods used to find out the relationship pattern between response variable and predictor variable of which the form function is difficulty to specify parametrically. If for example *X* as predictor and *Y* as the response variable with then the model which is used to see the relationship between and as in equation (1) with the regression function is unknown and is residual. Nonparametric regression is not tied to a certain assumption such as in the parametric regression and is more flexible, but nonparametric regression curve is assumed to be smooth, where the data will look for its own estimation form from the function [3]. To identify the data pattern, there are several estimation techniques commonly used in the nonparametric regression, among them are the estimation of kernel and splines.

## Splines Regression

Splines regression is one of the nonparametric regression that is built from piecewise polynomial which gives the flexibility of common polynomial so it can adapt itself effectively to the data characteristics and gives a smoother function and minimize the data variants.The common form of splines regression with the degree to *m* is [9]

 (3)

with,

 : Parameter of each linear function,

 : Parameter of polynomial function with degree to *m*,

 : Function to-*i* with knot in polynomial degree.

Knot is interpreted as a focal point in the splines function which is formed segmented on that focal point and hence makes the curve smoother. Splines is said to be linear with degree 1, quadratic splines with degree 2, and cubical splines with degree 3.

## Smoothing Splines Regression

Splines will show deficiency when it is using very high orde and too many knots which will cause it to form calculation matrix which is almost singular, therefore smoothing splines method is developed which uses a certain amount of knot with the same interval as the predictor variable [10]. In other words, we can set the knot value. The estimation with ordinary least square when the function as in equation (1) has unknown function, hence it will result a non accurate estimation. Therefore, smooth function is used to obtain the smaller sum of squared residual. One of the ways to get a quite smooth value of function is by minimizing the function of Penalized Least Square (PLS) [4] :

 (4)

with is smoothing parameter and is called the roughness penalty which is the measurement of smoothness. The parameter is so important in controlling the accuracy of the smoothness of the curve and a good result or it can also be said to minimize the value of residual. However, basically we know that the function on the nonparametric regression is unknown, so based on equation (4) we need to estimate the data and then the estimator is minimized toward in order to get a good estimator [9]. One of the methods proposed in choosing the smoothing parameter is Generalized Cross Validation (GCV) method introduced by Craven dan Wahba (1979) in the context of smoothing splines as follows [2] :

 (5)

The optimal value of is obtained by minimizing .

## Quantile Regression

Quantile regression is the generalization of median regression introduced by Koenker dan Basset in 1978 which is used to overcome the unknown form of the nonparametric regression. Quantile regression is the regression technique that explains the relationship between response variable and predictor variable in many quantiles, not only on the size of the median response variable. The quantile regression is useful when we have distribution of data such as asymmetric, fat-tailed, or outlier existed because the estimator is more efficient [9]. The use of quantile noted by gives more information on the relationship between response and predictor variables. Quantile to *-* of a distribution is defined as where the quantile is invers from the cumulative distribution function .

## Quantile Smoothing Splines Regression

The model of quantile smoothing splines regression is the modeling of quantile regression using the smoothing splines techniques. The solution of quantile smoothing splines regression is obtained by minimizing [6]:

 (6)

with is the knot point, where is the check function with and is the continue function which is the natural linear splines . The quantile objective function to- for quantile smoothing splines regression is defined as follows [9]:

 (7)

is called - loss function. The objective function of quantile smoothing splines regression or is not differentiable at zero point, which cause it to fail in obtaining an explicit solution. Therefore, the quantile solution is obtained by changing the function in the linear equation by linear programming.

Both smoothing spline regression (mean approach) and quantile smoothing splines regression, the choosing of the smoothing parameter is also an important part in setting the estimator. Koenker et al. (1994) proposes to use Schwarz Information Criterion (SIC) which can be written as follows [6]:

 (8)

The optimal value of is obtained by minimizing [8]

1. Result and Discussion

## Model Specification

Regression analysis can be used to find out the relationship pattern between the air relative himidity toward the rubber production, where there is a good correlation between the two variables, which is 55.55%. The regression analysis is done based on the regression function itself when the regression function is unknown or the relationship pattern shows different shapes in some parts indicating a difficulty to specify parametricallythen we then propose a nonparametric approach. In this primary steps model specification is done toward the pattern of data distribution by scatter plot.

|  |
| --- |
|  |
| **Figure 1**. Scatter Plot data of air relative humidity and the rubber production of South Sumatera  |

Based on figure 1it is seen that the pattern of data distribution has a function which is not easy to be formed parametrically and also the distribution of the data tends to gather on the right side, making it difficult to make the prespesification model. Hence, visually we decided that this research will use the nonparametric approach. However, to make it sure we can also se it empirically bassed on the normality assumption testing on residual. In this data testing will be done to find out whether the residual distributes normally or not, based on the result of the analysis using Shapiro Wilk testing the *p-value* obtained is < 0.05 which means the residual does not distribute normally. Both visually and empirically giving information that the parametric approach cannot be done, so for further steps smoothing splines approach will be used to find out the relationship pattern between air relative himidity toward the rubber production in South Sumatera.

## Smoothing Splines Regression (Mean)

In this research, smoothing splines regression method is one of the ways to estimate the nonparametric regression curve. The criteria needed to be focused on is the optimal smoothing parameter optimal and deciding the knot value. Where the smoothing parameter is the control device to determine the smoothness of the curve, which can be determined based on the minimum value of GCV. This research has obtained the value of the smoothing parameter by default through Software R with GCV criteria. Table 1 shows the components on the smoothing splines regression.

**Table 1.** Components on the Smoothing Splines Regression

|  |  |  |  |
| --- | --- | --- | --- |
| Component |  | GCV | Number of Coefficient (knot+2) |
| Value | 0.021535 | 113.6747 | 57 |

By using GCV, optimal lambda was found 0.021535, then this is used will be used in smoothing Splines regression model as shown in the figure 2 below:

|  |
| --- |
|  |
| **Figure 2**. The curve Smoothing Splines Regression (Mean approach) on Data of Relative humidity and Rubber Production |

Based on the fitting curve and the estimation result with red line explaining the estimation result, the curve obtained is smooth and yet has not explained the data distribution visually due to the very flexible characteristics of smoothing splines.

## Quantile Smoothing Splines Regression Model

Quantile smoothing splines regression model which is explained in figure 2. Results curve which hasn’t been able to explain the data distribution completely. This assumed that the distribution of data assymetrically and is also identified to having outliers. Hence, the mean regression approach is considered not good to be used and that it is better to use the quantile regression approach. Based on the boxplot from the residual using smoothing splines, the description is obtained as follows :

|  |
| --- |
|  |
| **Figure 3**.Boxplot Residual Mean Smoothing Splines Regression Model |

Based on figure3 We can inform that the relationship between response and predictor variables contains outlier, so that the use of quantile regression can be a mean regression alternative in accomodating outlier and to catch information about data distibution through its quantiles. In quantile smoothing splines regression, the decision of smoothing parameter according to its quantiles uses the minimum criteria of Schwarz Information Criterion (SIC), in the following plot, the value of optimal is 1.1 is obtained.

|  |
| --- |
|  |
| **Figure 4.** SIC and in Quantile Smoothing Splines Regression |

By using packages “quantreq” with function “rqss” on Software R with smoothing parameter = 1.1 based SIC and hence quantile smoothing splines regression is obtained as follows:

|  |  |  |
| --- | --- | --- |
|  |  | Very HighHigh ModerateLowVery Low |
| **Figure 5.** The curve of Quantile Smoothing Splines Regression (Median Case) |  | **Figure 6.** The curve of Quantile Smoothing Splines Regression |

Quantile smoothing splines regression ( **:** median case) is a quantile smoothing splines regression with . Based on figure5the fitting curve and blue line explains the estimation result, the curve produced is smooth and visually, it can be said that the curve has explained a better data distribution than the mean regression in figure 2. Moreover, to give more information we can use several quantile values so that the information of the regression lines as much as the needed is obtained. This research uses quantile which is divided into 5 parts in order to get the quantile value . According figure 6 Quantile smoothing splines regression divides the data distribution into five parts based on very low, low, moderate, high and very high characteristics. The result of rubber production is an important part in Indonesian statistics because of having a real impact on Indonesian economics. Therefore, we need to know which characteristics the production result in one period of time brings so we can estimate the production target. The Pattern formed in figure 6can be used to identify the position of the rubber production for the next time ahead. With quantile smoothing splines regression model, the quantile is obtained as follows:

 (9)

 (10)

 (11)

 (12)

 (13)

Knot which is used in this model is a unique predictor or is different from other predictors [9]

## Model Comparison

In this research, the use of quantile approach gives a better and efficient result in giving information compare to mean approach due to the pattern spreads asymmetrically and the are some outliers. Quantile approach can accomodate that oulier data. Dynamically, it can also be seen based on the Root Mean Square Error that quantile smoothing splines regression (median case) model is smaller than Root Mean Square Error of smoothing splines regression (mean approach). The following shows the comparison of the two Root Mean Square Error models:

**Table 2.** The Comparison of RMSE value

|  |  |
| --- | --- |
| Regression Model  | RMSE |
| Mean Smoothing Splines Regression | 7.1196 |
| Quantile Smoothing Splines Regression ( | 5.0877 |

Based on table 2. it is shown that the quantile smoothing splines regression gives smaller Root Mean Square Error compared to mean smoothing splines regression.

1. Conclusion

Based on the previous discussion, it can be concluded that quantile smoothing splines regression model is better to use in the case of modelling the influence of air relative humidity toward rubber production, because the analysis of this quantile regression is more robust to the outlier. This quantile regression analysis also can give additional information about the grouping of response variables or categories of response variables based on the information of the predictors we have. The quantile smoothing splines regression estimator (τ = 0.5 median case) is more efficient than the smoothing splines regression (mean approach) that can compare by smaller of Root Mean Square Error. In the case, the data distribution is divided into five characteristics; very low, low, moderate, high and very high. The growth pattern of this production result can be categorized good if there is a slow increase from very low category to very high category. But when the growth pattern decreases then the company and government need to find out the cause and evaluate the possibility of other factors that might cause the changes such as climate.

5. References

1. Badan Pusat Statistik.2018. Statistik Karet Indonesia 2018. Jakarta
2. Craven, P., & Wahba, G. 1979. Smoothing Noisy Data with Spline Functions: Estimating the Correct Degree of Smoothing by The Method of Generalized Cross-Validation. Numerische Mathematik.31,377-473
3. Eubank, R. 1999. Nonparametric Regression and Splines Smoothing, Marcel Dekker, Inc., New York.
4. Hastie, T.J. and R.J. Tibshirani. 1990. Generalized Additive Models. Chapman and Hall, London.
5. Koenker, R. and Bassett, G. W. 1978. Regression Quantiles. Econometrica, 46, 33–50
6. Koenker, R., Ng, P.T., & Portnoy, S.L. 1994. Quantile Smoothing Splines. Biometric, 81, 678-680
7. Koenker, R., Quantile Regression In R: A Vignette. CRAN [https://CRAN.R-project.org/package=quantreg](https://cran.r-project.org/package%3Dquantreg).
8. Koenker R. 2011. Additive Models for Quantile Regression: Model Selection and Confidence Bandaids. Brazillian Journal of Probability and Statistics 25. 30-50.
9. Mulyani,S.2017. Modeling the Human Development Indeks and The Percentage of Poor People Using Quantile Smoothing Splines. AIP Conference Proceedings.1827
10. Rodriguez, German. 2001. Smoothing and Non-Parametric Regression. Spring. Grodri@princeton.edu.
11. Umar,H.Y., et al.2017. Evaluation of The Impact of Climatic Factors on Latex Yield of Hevea Brasiliensis. Internasional Journal of Research in Agricultural Sciences.
12. <http://www.bps.go.id>
13. <https://giovanni.gsfc.nasa.gov/giovanni/#service=ArAvTs&starttime=&endtime=>