Journal of Theoretical and Applied Information Technology
15th August 2023. Vol.101. No 15

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

INTEGRATION OF NON-HIERARCHY CLUSTER ANALYSIS WITH SEMIPARAMETRIC TRUNCATED SPLINE MULTIRESPONSE REGRESSION

EVITARI GALU ERWINDA[1*], SOLIMUN[2], ADJI ACHMAD RINALDO FERNANDES[3], RAHMA FITRIANI[4], ATIEK IRIANY[5]

[1,2,3,4,5]Brawijaya University, Faculty of Mathematics and Natural Sciences, Department of Statistics, 65154, Indonesia

Corresponding author e-mail: evitariwinda@gmail.com[1*)]

ABSTRACT

Cluster analysis is an approach used to find similarities in data and place them in groups. Cluster analysis is divided into two methods, namely hierarchical and non-hierarchical methods. The non-hierarchical method of several methods, namely K-Means, K-Harmonic Means, and K-Medoids, used in this study is K-Means. Then, the integration of cluster analysis with semiparametric truncated spline multiresponse regression will be carried out. This study aimed to identify and develop non-hierarchical cluster integration and semiparametric multiresponse regression using the Truncated Spline approach in the case of Farmer Inclusiveness data in West Nusa Tenggara. Modeling factors affecting the inclusiveness of farmers in NTB was conducted using cluster analysis modeling with multiresponse regression which was developed by making integration of non-hierarchical cluster analysis with semiparametric truncated spline multiresponse regression to help give Comprehensive information about farmer inclusiveness in NTB. The sample used in this study was 100 farmer groups in NTB. The results in this study produce three clusters, which have low, medium, and high inclusiveness categories. Cluster 1 is a farmer with a low inclusiveness category consisting of 42 farmer groups, cluster 2 is a farmer with a medium inclusiveness category consisting of 21 farmer groups, and Cluster 3 is a farmer with a high inclusivity category consisting of 37 farmer groups. The cluster with the highest R-squared on modeling integration of non-hierarchical cluster analysis with semiparametric truncated spline multiresponse regression is cluster 1 with R^2 value of 0.205. Thus, modeling using cluster integration with semiparametric truncated spline multiresponse regression shows a change in the nature of the data that has a variable relationship pattern and can help the government and farmers to determine the level of inclusiveness of farmers and examine the factors that influence farmer inclusiveness. This research's originality is combining cluster analysis with semiparametric truncated spline multiresponse regression with integration between the two methods.

Keywords: *Integration Clusters, Multiresponse Regression Analysis Semiparametric, Truncated Spline, Farmer Inclusivity*.

1. INTRODUCTION

Cluster analysis is an approach to find similarities in data and place them into groups [1]. Cluster analysis is divided into two methods, *hierarchical* and *non-hierarchical* methods. The *non-hierarchical* method is a grouping method that begins by first selecting the number of initial clusters that are adjusted to the number the researcher wants, after which the objects will be combined into clusters [2]. The non-hierarchical method itself consists of several methods: *K-Means, K-Harmonic Means*, and *K-Medoids*.

This study was conducted using regression analysis which is one of the data analysis in statistics on structural modeling. Regression analysis is an analysis carried out to examine, explain, and find out the form of the relationship between response variables and predictor variables [3]. In the regression analysis, there are several approaches, such as parametric, semiparametric, and nonparametric approaches.

Estimation methods in semiparametric regression that are commonly used are *spline*, kernel, and fourier. This study used the *spline family approach*. One of the estimates that are often

used in semiparametric regression is *Truncated Spline*. *Splines* are polynomial pieces that have segmented and continuous properties, the *spline approach* for semiparametric regression has high flexibility and can handle the patterns of data relationships with changing behavior at any given interval. To determine the optimal knot point, GCV is used which is better compared to other methods, which have asymptotic optimal properties, invariant to transformation and in its calculating, the population variance does not need to be known.

Regression can be distinguished from the number of variables used. Multiresponse regression is a regression model that has two or more response variables that are correlated with each other. By using multiresponses from the correlation test, it can be seen what kind of relationship exists between responses. From the statistical methods previously described, integration can be carried out, which combines or adjusts between non-hierarchical cluster analysis and regression analysis, which is the *truncated spline semiparametric multiresponse regression*. So, it will form a new statistical method that can be applied in solving solutions in other fields, including in the research that will be carried out in the agricultural sector, specifically farmer inclusiveness.

Previous research, namely [4] and [5] have not combined the use of cluster analysis with multiresponse regression. then, previous studies have not integrated non-hierarchical cluster analysis with semiparametric multiresponse regression using a truncated spline. so that in this study integration will be carried out between cluster analysis and truncated spline semiparametric multiresponse regression. This method can be used to develop smallholders with an inclusive business model approach that can integrate smallholders into markets and agricultural value chains because farmers' inclusiveness is still low.

The modeling in this study will examine how the truncated spline semiparametric multiresponse regression model, where in the truncated spline which is known flexible to data both secondary and generated data will follow the shape of a semiparametric approach. The purpose of this study is to identify and develop non-hierarchical cluster integration and semiparametric multiresponse regression using the truncated spline approach.

2. LITERATURE REVIEW

2.1. Cluster Analysis

Cluster analysis is an analytical method that aims to cluster objects based on their similar characteristics and obtained a homogeneous objects according to the desired factors for clustering [6]. The cluster formation process consists of two methods, namely hierarchical and non-hierarchical methods. In the non-hierarchical method, the procedure begins by selecting a number of values according to the desired number and then the objects of observation will be combined into clusters. In the non-hierarchical method, there are several cluster methods, and the commonly used methods are K-Means and K-Medoids .

2.2. K-Means Cluster Analysis

The K-Means algorithm is a non-hierarchical method that will initially take a portion of the many components of the population to serve as the initial cluster center. The steps of the K-Means algorithm are as follows.

1) Determining the number of clusters (k) and set the cluster center randomly.
2) Calculating the distance of each data to the cluster center, size distance using the size equation distance *Euclidean* by using the following equation.

$$d_{jl} = \left\{ \left(x_l - x_j\right)' \left(x_l - x_j\right) \right\}^{\frac{1}{2}} = \sqrt{\sum_{k=1}^{i} \left(x_{lk} - x_{jk}\right)^2} \quad (1)$$

$$i = 1,2,...,p \ , j = 1,2,...,n \text{ and } l = 1,2,...,n$$

3) Grouping data into clusters with the shortest distance
4) Calculating the new cluster center
5) Repeating step 2 to step 4 so that no more data is moved to another cluster.

The clustering carried out in this study used a *dummy variable* in order to represent the nature of qualitative data to become quantitative. If there are two clusters, a $(k = 2)$ *dummy* function will be obtained as follows.

$$D_{1i} = \begin{cases} 1, \text{for cluster } 1 \\ 0, \text{and others} \end{cases} \quad (2)$$

Then, if there are three clusters $(k = 3)$, then the *dummy function is obtained* as follows.

$$D_{1i} = \begin{cases} 1, \text{for cluster 2} \\ 0, \text{and others} \end{cases} \quad (3)$$

$$D_{2i} = \begin{cases} 1, \text{for cluster 3} \\ 0, \text{and others} \end{cases}$$

2.3. Semiparametric Regression

Semiparametric regression is a statistical method used to determine the pattern of the relationship between the response variable and the predictor variable, with some part are known and another part are unknown. The general model of semiparametric regression is as follows [7].

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + f(z_i) + \varepsilon_i$$

$$i = 1,2,...,n$$

Where

y_i : response variable

$x_{i1}, x_{i2}, ..., x_{ik}$: predictor variable

$\beta_0, \beta_1, ..., \beta_k$: unknown parameters

$f(z_i)$:regression function of unknown form

ε_i : random errors are assumed to be identical, independent and normally distributed with zero mean and variance σ^2

2.4. Multiresponse Regression Truncated Spline Semiparametric

According to Fernandes [8] in regression can be differentiated based on the number of responses involved, such as the regression model with single response and multiresponse. The multiresponse model consists of several models which assume a correlation between responses and have the goal of obtaining a better model than the single response model, considering the effect of predictors on responses and the relationship between responses. Regression models that have one predictor with more than two responses and involve *N* observations are as follows:

$$y_{ki} = f_k(x_i) + \varepsilon_{ki}, i = 1,2,...,n \quad (5)$$

Where

y_{ki} : The k*th* response *to* the i-*th observation*

x_i : The i- *th predictor*

f_k : The regression function that connects the predictor with the kth *response*

ε_{ki} : residual in the kth response *in* the i- *th observation*

n : the number of observations

m : many predictors

Modeling a semiparametric regression with *a truncated spline* which has more than two response variables (multiresponse) for n observations, the equation can be formulated as follows [9]:

$$\underset{\sim}{y} = \mathbf{X}_1 \underset{\sim}{\gamma} + \mathbf{X}_2[K] \underset{\sim}{\beta} + \underset{\sim}{\varepsilon} \quad (6)$$

Where:

$$\underset{\sim}{y} = (\underset{\sim}{y}_1, \underset{\sim}{y}_2, ..., \underset{\sim}{y}_m)^T$$

$$\underset{\sim}{y}_1 = (y_{11}, ..., y_{1n})^T$$

$$\underset{\sim}{y}_2 = (y_{21}, ..., y_{2n})^T \quad (4)$$

$$\underset{\sim}{y}_m = (y_{m1}, ..., y_{mn})^T$$

$$\underset{\sim}{\gamma} = (\underset{\sim}{\gamma}_1, \underset{\sim}{\gamma}_2, ..., \underset{\sim}{\gamma}_m)^T$$

$$\underset{\sim}{\gamma}_i^T = (\gamma_{i0}, ..., \gamma_{im})$$

$$\underset{\sim}{\beta} = (\underset{\sim}{\beta}_1, \underset{\sim}{\beta}_2, ..., \underset{\sim}{\beta}_m)^T$$

$$\underset{\sim}{\beta}_i^T = (\beta_{i0}, ..., \beta_{im})$$

$$\underset{\sim}{\varepsilon} = (\varepsilon_{11}, ..., \varepsilon_{1n}, \varepsilon_{21}, ..., \varepsilon_{2n}, \varepsilon_{m1}, ..., \varepsilon_{mn})^T$$

With m : the m*th* response variable where m=1,2,…,m and n : observations where n=1,2,…,n.

$$\mathbf{X}_1 = \begin{bmatrix} 1 & x_{1111} & x_{1211} & \cdots & x_{1q11} \\ 1 & x_{1112} & x_{1212} & \cdots & x_{1q12} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{111n} & x_{121n} & \cdots & x_{1q1n} \\ 1 & x_{1121} & x_{1221} & \cdots & x_{1q21} \\ 1 & x_{1122} & x_{1222} & \cdots & x_{1q22} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{112n} & x_{122n} & \cdots & x_{1q2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{11m1} & x_{12m1} & \cdots & x_{1qm1} \\ 1 & x_{11m2} & x_{12m2} & \cdots & x_{1qm2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{11mn} & x_{12mn} & \cdots & x_{1qmn} \end{bmatrix} \quad (7)$$

$$\mathbf{X}_2[K] = \begin{bmatrix} \mathbf{X}_{21}[K] & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_{22}[K] & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{X}_{2m}[K] \end{bmatrix} \quad (8)$$

Assuming the correlation between residuals of each response relationship, the calculation of the correlation of the residuals in each response has the form of a correlation and weighting matrix of

various sizes $MN \times MN$ on the matrix and sub-matrix as follows:

$$\hat{\rho}_{ij} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y}_i)(Y_j - \bar{Y}_j)}{\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2 \sum_{i=1}^{n}(Y_j - \bar{Y}_j)^2}} \quad (9)$$

$$\hat{\sigma}_i = \sqrt{\frac{\sum_{l=1}^{n}(Y_{il} - \bar{Y}_i)^2}{n-1}} \quad (10)$$

$$\Sigma = \begin{bmatrix} \hat{\sigma}_1^2 & 0 & \cdots & 0 & \hat{\rho}_{12}\hat{\sigma}_1\hat{\sigma}_2 & 0 & \cdots & 0 & \cdots & \hat{\rho}_{1m}\hat{\sigma}_1\hat{\sigma}_m & 0 & \cdots & 0 \\ 0 & \hat{\sigma}_1^2 & \cdots & 0 & 0 & \hat{\rho}_{12}\hat{\sigma}_1\hat{\sigma}_2 & \cdots & 0 & \cdots & 0 & \hat{\rho}_{1m}\hat{\sigma}_1\hat{\sigma}_m & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{\sigma}_1^2 & 0 & 0 & \cdots & \hat{\rho}_{12}\hat{\sigma}_1\hat{\sigma}_2 & \cdots & 0 & 0 & \cdots & \hat{\rho}_{1m}\hat{\sigma}_1\hat{\sigma}_m \\ \hat{\rho}_{12}\hat{\sigma}_1\hat{\sigma}_2 & 0 & \cdots & 0 & \hat{\sigma}_2^2 & 0 & \cdots & 0 & \cdots & \hat{\rho}_{2m}\hat{\sigma}_2\hat{\sigma}_m & 0 & \cdots & 0 \\ 0 & \hat{\rho}_{12}\hat{\sigma}_1\hat{\sigma}_2 & \cdots & 0 & 0 & \hat{\sigma}_2^2 & \cdots & 0 & \cdots & 0 & \hat{\rho}_{2m}\hat{\sigma}_2\hat{\sigma}_m & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{\rho}_{12}\hat{\sigma}_1\hat{\sigma}_2 & 0 & 0 & \cdots & \hat{\sigma}_2^2 & \cdots & 0 & 0 & \cdots & \hat{\rho}_{2m}\hat{\sigma}_2\hat{\sigma}_m \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ \hat{\rho}_{1m}\hat{\sigma}_1\hat{\sigma}_m & 0 & \cdots & 0 & \hat{\rho}_{2m}\hat{\sigma}_2\hat{\sigma}_m & 0 & \cdots & 0 & \cdots & \hat{\sigma}_m^2 & 0 & \cdots & 0 \\ 0 & \hat{\rho}_{1m}\hat{\sigma}_1\hat{\sigma}_m & \cdots & 0 & 0 & \hat{\rho}_{2m}\hat{\sigma}_2\hat{\sigma}_m & \cdots & 0 & \cdots & 0 & \hat{\sigma}_m^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{\rho}_{1m}\hat{\sigma}_1\hat{\sigma}_m & 0 & 0 & \cdots & \hat{\rho}_{2m}\hat{\sigma}_2\hat{\sigma}_m & \cdots & 0 & 0 & \cdots & \hat{\sigma}_m^2 \end{bmatrix}_{MN \times MN} \quad (11)$$

The matrix form in equation (12) can be simplified as follows:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} & \cdots & \boldsymbol{\Sigma}_{1m} \\ \boldsymbol{\Sigma}_{12} & \boldsymbol{\Sigma}_{22} & \cdots & \boldsymbol{\Sigma}_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}_{1m} & \boldsymbol{\Sigma}_{2m} & \cdots & \boldsymbol{\Sigma}_{mm} \end{bmatrix}_{MN \times MN} \quad (12)$$

The parameter estimators and functions from semiparametric regression can be obtained by looking at the weights, $\boldsymbol{\Sigma}^{-1}$ then solving the *Weighted Least Square* (WLS) optimization by completing the parametric component, and doing partial derivatives by assuming the function as follows.

$$\frac{\partial\left(\underset{\sim}{\varepsilon}^T \hat{\boldsymbol{\Sigma}}^{-1} \underset{\sim}{\varepsilon}\right)}{\partial \underset{\sim}{\gamma}} = 0$$

$$\frac{\partial\left(\underset{\sim}{y}^T \hat{\boldsymbol{\Sigma}}^{-1} \underset{\sim}{y} - 2\underset{\sim}{\gamma}^T \mathbf{X}_1^T \hat{\boldsymbol{\Sigma}}^{-1} \underset{\sim}{y} + \underset{\sim}{\gamma}^T \mathbf{X}_1^T \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{X}_1 \underset{\sim}{\gamma}\right)}{\partial \underset{\sim}{\gamma}} = 0 \quad (13)$$

$$-2\mathbf{X}_1^T \hat{\boldsymbol{\Sigma}}^{-1} \underset{\sim}{y} + 2\mathbf{X}_1^T \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{X}_1 \underset{\sim}{\gamma} = 0$$

$$-\mathbf{X}_1^T \hat{\boldsymbol{\Sigma}}^{-1} \underset{\sim}{y} + \mathbf{X}_1^T \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{X}_1 \underset{\sim}{\gamma} = 0$$

So that the value of the estimator will be obtained $\hat{\underset{\sim}{\gamma}}$ with the following results.

$$\hat{\underset{\sim}{\gamma}} = (\mathbf{X}_1^T \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{X}_1)^{-1} \mathbf{X}_1^T \hat{\boldsymbol{\Sigma}}^{-1} \underset{\sim}{y} \quad (14)$$

Next, the solution for the nonparametric components is carried out using the WLS where the residuals are assumed normal, identical, and independent errors with zero mean and variance σ^2 and to estimate the value $\underset{\sim}{\beta}$ can be estimated by minimizing $\underset{\sim}{\varepsilon}^T \underset{\sim}{\varepsilon}$, so partial derivatives of $\underset{\sim}{\beta}$.

$$\frac{\partial\left(\underset{\sim}{\varepsilon}^T \hat{\boldsymbol{\Sigma}}^{-1} \underset{\sim}{\varepsilon}\right)}{\partial \underset{\sim}{\beta}} = 0$$

$$\frac{\partial\left(\underset{\sim}{y}^T \hat{\boldsymbol{\Sigma}}^{-1} \underset{\sim}{y} - 2\underset{\sim}{\beta}\mathbf{X}_2 K^T \hat{\boldsymbol{\Sigma}}^{-1} \underset{\sim}{y} + \underset{\sim}{\beta}^T \mathbf{X}_2 K^T \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{X}_2 K_1, K\underset{\sim}{\beta}\right)}{\partial \underset{\sim}{\beta}_l} = 0 \quad (15)$$

$$-2\mathbf{X}_2 K^T \hat{\boldsymbol{\Sigma}}^{-1} \underset{\sim}{y} + 2^T \mathbf{X}_2 K^T \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{X}_2 K\underset{\sim}{\beta} = 0$$

$$-\mathbf{X}_2 K^T \hat{\boldsymbol{\Sigma}}^{-1} \underset{\sim}{y} + \mathbf{X}_2 K^T \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{X}_2 K\underset{\sim}{\beta} = 0$$

estimator of the function is obtained $\hat{\underset{\sim}{\beta}}$

$$\hat{\underset{\sim}{\beta}} = \left(\mathbf{X}_2 K^T \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{X}_2 K\right)^{-1} \left(\mathbf{X}_2 K^T \hat{\boldsymbol{\Sigma}}^{-1} \underset{\sim}{y}\right) \quad (16)$$

So that the estimator of the function will be obtained as follows.

$$\begin{aligned} \hat{\underset{\sim}{f}}(x) &= \mathbf{X}_2 [\mathbf{K}] \hat{\underset{\sim}{\beta}} \\ &= \mathbf{X}_2 [\mathbf{K}] \left(\mathbf{X}_2 [K]^T \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{X}_2 [K]\right)^{-1} \mathbf{X}_2 [K]^T \hat{\boldsymbol{\Sigma}}^{-1} \underset{\sim}{y} \end{aligned} \quad (17)$$

If the form of relationship obtained has the form of a parametric regression curve and the other is nonparametric, with the nonparametric relationship formed are polynomials linear 1 knot, linear 2 knots, quadratic 1 knot and quadratic 2 knots then the following equation will be obtained.

$$\begin{aligned} y_{1i} &= f_1(x_{1i}, x_{2i}) + \varepsilon_{1i} \\ y_{1i} &= \beta_{01} + \beta_{11} x_{1i} + \beta_{21} x_{2i} + \beta_{31}(x_{2i} - k_{11})_+ + \varepsilon_{1i} \\ y_{2i} &= f_2(x_{1i}, x_{2i}) + \varepsilon_{2i} \\ y_{2i} &= \beta_{02} + \beta_{12} x_{1i} + \beta_{22} x_{1i}^2 + \beta_{32}(x_{1i} - k_{12})_+^2 \\ &\quad + \beta_{42}(x_{1i} - k_{22})_+^2 + \beta_{52} x_{2i} + \varepsilon_{2i} \\ y_{3i} &= f_3(x_{1i}, x_{2i}) + \varepsilon_{3i} \\ y_{3i} &= \beta_{03} + \beta_{13} x_{1i} + \beta_{23}(x_{1i} - k_{13})_+ + \beta_{33}(x_{1i} - k_{23})_+ \\ &\quad + \beta_{43} x_{2i} + \beta_{53} x_{2i}^2 + \beta_{63}(x_{2i} - k_{33})_+^2 + \varepsilon_{3i} \end{aligned} \quad (18)$$

and the truncated spline function is as follows.

$$(x_{2i}-k_{11})_+ = \begin{cases} (x_{2i}-k_{11}) & ; x_{2i} \geq k_{11} \\ 0 & ; x_{2i} < k_{11} \end{cases}$$
$$(x_{1i}-k_{12})_+^2 = \begin{cases} (x_{1i}-k_{12})^2 & ; x_{1i} \geq k_{12} \\ 0 & ; x_{1i} < k_{12} \end{cases}$$
$$(x_{1i}-k_{22})_+^2 = \begin{cases} (x_{1i}-k_{22})^2 & ; x_{1i} \geq k_{22} \\ 0 & ; x_{1i} < k_{22} \end{cases}$$
$$(x_{1i}-k_{13})_+ = \begin{cases} (x_{1i}-k_{13}) & ; x_{1i} \geq k_{13} \\ 0 & ; x_{1i} < k_{13} \end{cases}$$
$$(x_{1i}-k_{23})_+ = \begin{cases} (x_{1i}-k_{23}) & ; x_{1i} \geq k_{13} \\ 0 & ; x_{1i} < k_{13} \end{cases}$$
$$(x_{2i}-k_{33})_+ = \begin{cases} (x_{2i}-k_{33})^2 & ; x_{2i} \geq k_{33} \\ 0 & ; x_{2i} < k_{33} \end{cases} \quad (19)$$

2.5. Integration of Cluster Analysis with Multiresponse Regression Truncated Spline Semiparametric

Truncated spline semiparametric multiresponse regression analysis is a development model in statistics that uses dummy variables. The dummy variable is used to make data that has qualitative characteristics become quantitative. The purposed variable is a numeric variable which can represent a variable with a binary category or more. If there are three clusters, then the cluster model with semiparametric regression analysis can be shown in the following equation.

$$\begin{aligned} y_{1i} &= f_1(x_{1i},x_{2i}) + \varepsilon_{1i} \\ &= \beta_{01} + \beta_{11}x_{1i} + \beta_{21}x_{2i} + \beta_{31}(x_{2i}-k_{11})_+ + \beta_{41}D_{1i} \\ &\quad + \beta_{51}D_{2i} + \beta_{61}x_{1i}D_{1i} + \beta_{71}x_{1i}D_{2i} + \beta_{81}x_{2i}D_{1i} + \beta_{91}x_{2i}D_{2i} \\ &\quad + \beta_{101}(x_{2i}-k_{11})_+ D_{1i} + \beta_{111}(x_{2i}-k_{11})_+ D_{2i} + \varepsilon_{1i} \\ y_{2i} &= f_2(x_{1i},x_{2i}) + \varepsilon_{2i} \\ &= \beta_{02} + \beta_{12}x_{1i} + \beta_{22}(x_{1i}-k_{12})_+ + \beta_{32}x_{2i} + \beta_{42}D_{1i} \\ &\quad + \beta_{52}D_{2i} + \beta_{62}x_{1i}D_{1i} + \beta_{72}x_{1i}D_{2i} + \beta_{82}(x_{1i}-k_{12})_+ D_{1i} \\ &\quad + \beta_{92}(x_{1i}-k_{12})_+ D_{2i} + \beta_{102}x_{2i}D_{1i} + \beta_{112}x_{2i}D_{2i} + \varepsilon_{2i} \\ y_{3i} &= f_3(x_{1i},x_{2i}) + \varepsilon_{3i} \\ &= \beta_{03} + \beta_{13}x_{1i} + \beta_{23}x_{1i}^2 + \beta_{33}(x_{1i}-k_{13})_+^2 + \beta_{43}x_{2i} \\ &\quad + \beta_{53}D_{1i} + \beta_{63}D_{2i} + \beta_{73}x_{1i}D_{1i} + \beta_{83}x_{1i}D_{2i} + \beta_{93}x_{1i}^2D_{1i} \\ &\quad + \beta_{103}x_{1i}^2D_{2i} + \beta_{113}(x_{1i}-k_{13})_+^2 D_{1i} + \beta_{123}(x_{1i}-k_{13})_+^2 D_{2i} \\ &\quad + \beta_{133}x_{2i}D_{1i} + \beta_{143}x_{2i}D_{2i} + \varepsilon_{3i} \end{aligned} \quad (20)$$

Cluster 1 ($D_1 = 0, D_2 = 0$)

$$\begin{aligned} y_{1i} &= \beta_{01} + \beta_{11}x_{1i} + \beta_{21}x_{2i} + \beta_{31}(x_{2i}-k_{11})_+ + \varepsilon_{1i} \\ y_{2i} &= \beta_{02} + \beta_{12}x_{1i} + \beta_{22}(x_{1i}-k_{12})_+ + \beta_{32}x_{2i} + \varepsilon_{2i} \\ y_{3i} &= \beta_{03} + \beta_{13}x_{1i} + \beta_{23}x_{1i}^2 + \beta_{33}(x_{1i}-k_{13})_+^2 + \beta_{43}x_{2i} + \varepsilon_{3i} \end{aligned}$$

Cluster 2 ($D_1 = 1, D_2 = 0$)

$$\begin{aligned} y_{1i} &= \beta_{01} + \beta_{11}x_{1i} + \beta_{21}x_{2i} + \beta_{31}(x_{2i}-k_{11})_+ + \beta_{41}D_{1i} \\ &\quad + \beta_{61}x_{1i}D_{1i} + \beta_{81}x_{2i}D_{1i} + \beta_{101}(x_{2i}-k_{11})_+ D_{1i} + \varepsilon_{1i} \\ y_{2i} &= \beta_{02} + \beta_{12}x_{1i} + \beta_{22}(x_{1i}-k_{12})_+ + \beta_{32}x_{2i} + \beta_{42}D_{1i} \\ &\quad + \beta_{62}x_{1i}D_{1i} + \beta_{82}(x_{1i}-k_{12})_+ D_{1i} + \beta_{102}x_{2i}D_{1i} + \varepsilon_{2i} \\ y_{3i} &= \beta_{03} + \beta_{13}x_{1i} + \beta_{23}x_{1i}^2 + \beta_{33}(x_{1i}-k_{13})_+^2 + \beta_{43}x_{2i} \\ &\quad + \beta_{53}D_{1i} + \beta_{73}x_{1i}D_{1i} + \beta_{93}x_{1i}^2D_{1i} + \beta_{113}(x_{1i}-k_{13})_+^2 D_{1i} \\ &\quad + \beta_{133}x_{2i}D_{1i} + \varepsilon_{3i} \end{aligned}$$

Cluster 3 ($D_1 = 0, D_2 = 1$)

$$\begin{aligned} y_{1i} &= \beta_{01} + \beta_{11}x_{1i} + \beta_{21}x_{2i} + \beta_{31}(x_{2i}-k_{11})_+ \\ &\quad + \beta_{51}D_{2i} + \beta_{71}x_{1i}D_{2i} + \beta_{91}x_{2i}D_{2i} + \beta_{111}(x_{2i}-k_{11})_+ D_{2i} + \varepsilon_{1i} \\ y_{2i} &= \beta_{02} + \beta_{12}x_{1i} + \beta_{22}(x_{1i}-k_{12})_+ + \beta_{32}x_{2i} \\ &\quad + \beta_{52}D_{2i} + \beta_{72}x_{1i}D_{2i} + \beta_{92}(x_{1i}-k_{12})_+ D_{2i} + \beta_{112}x_{2i}D_{2i} + \varepsilon_{2i} \\ y_{3i} &= \beta_{03} + \beta_{13}x_{1i} + \beta_{23}x_{1i}^2 + \beta_{33}(x_{1i}-k_{13})_+^2 + \beta_{43}x_{2i} \\ &\quad + \beta_{63}D_{2i} + \beta_{83}x_{1i}D_{2i} + \beta_{103}x_{1i}^2D_{2i} + \beta_{123}(x_{1i}-k_{13})_+^2 D_{2i} \\ &\quad + \beta_{143}x_{2i}D_{2i} + \varepsilon_{3i} \end{aligned}$$

2.6. Generalized Cross Validation

One of the methods for selecting optimal knot points is *Generalized Cross Validation* (GCV). The optimal knot point selected by looking at the minimum GCV value. The selection of knot points using the GCV method is defined as follows [10]:

$$GCV(K_1,K_2,...,K_r) = \frac{MSE(K_1,K_2,...,K_r)}{\left(n^{-1}trace\left[I-A(K_1,K_2,...,K_r)\right]\right)^2} \quad (21)$$

with

$$MSE(K_1,K_2,...,K_r) = n^{-1}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2$$

where $K_1, K_2, ..., K_r$ is the point of the first knot to the rth knot.

3. RESEARCH METHODOLOGY

3.1. Research Data

The secondary data used on Strengthening Rice Farmers' Inclusivity in Supporting National Food Security in West Nusa Tenggara Province in 2022. The data was collected from questionnaires that was directly distributed, and the scale used in this study was the *Likert scale*. The samples of this

Journal of Theoretical and Applied Information Technology
15th August 2023. Vol.101. No 15

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

study are 100 farmers and the variables used were Production Volatility (X_1), Price Volatility (X_2), Agricultural Market Access (Y_1), Farmer Income (Y_2), and Farmer Welfare (Y_3).

3.2. Research Stages

The stages of this research are as follows:

1) Performing the cluster analysis
2) linearity testing using Ramsey's Reset Test
3) Parameter estimating for parametric and nonparametric components
4) Selecting the best model based on R^2
5) Obtaining the integration model of cluster analysis with semiparametric truncated spline multiresponse regression
6) Interpretating the results

4. RESULTS AND DISCUSSION

4.1. K-Means Cluster

Cluster analysis aims to classify as well as set objects with various characteristics based on a certain characteristics. For that reason, using non-hierarchical cluster analysis will making a classification on farmers in the province of NTB using the K-Means method with size Euclidean distance. The size Euclidean distance can be obtained using equation (1) and Then grouped with K-Means. The centroid calculation results for each cluster are shown in Table 1.

Table 1: K-Means Cluster Analysis Center

Variable	Cluster 1	Cluster 2	Cluster 3
Production Volatility (X_1)	2.736	3.539	3.644
Price Volatility (X_2)	2.776	2.717	3.651
Agricultural Market Access (Y_1)	2.704	4.027	3.071
Farmer Income (Y_2)	2.784	4.259	3.153
Farmer Welfare (Y_3)	2.944	4.313	2.970

K-Means method from three cluster presented in Figure 1.

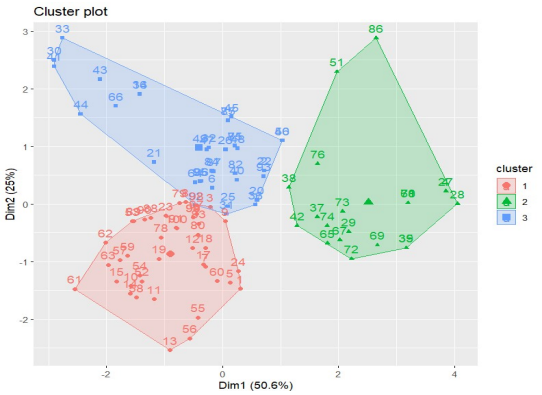

Figure 1: Results K-means cluster

Figure 1 shows that by using K-Means method and distance used is *Euclidean Distance*, farmer inclusiveness divided into 3 clusters. Cluster one is displayed in red, cluster two is displayed in green, and cluster three is displayed in blue. The details of member on each clusters can be seen on Table 2.

Table 2: Total Member from Three K-Means Cluster

Cluster	Farmer Group Code	Amount Member Cluster
Cluster 1	1, 3, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 17, 18, 19, 23, 24, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 78, 79, 80, 83, 88, 89, 90, 91, 94, 97, 98, 99, 100	42
Cluster 2	4, 27, 28, 29, 35,37, 38, 39, 42, 51, 65, 67, 68, 69, 70, 71, 72, 73, 74, 78, 86	21
Cluster 3	2, 6, 16, 20, 21, 22, 25, 26, 30, 31, 32, 33, 34, 36, 40, 41, 43, 44, 45, 46, 47, 48, 49, 50, 64, 66, 75, 77, 81, 82, 84, 85, 87, 92, 93, 95, 96	37

4.2. Linearity Test

In modeling statistics, need to know form connection between variable to be used, i.e between predictor variables and response variable to choose the appropriate method and approach for the data. For that, linearity test using Ramsey's Reset test with the following results.

Table 3: Linearity Test Results

Connection	P-Value	Linearity
X_1 to Y_1	0.842	linear
X_2 to Y_1	0.994	Not Linear
X_1 to Y_2	0.947	Not Linear
X_2 to Y_2	0.575	linear
X_1 to Y_3	0.534	Not Linear
X_2 to Y_3	0.794	linear

based on results of linearity test presented in Table 3, it is obtained that there is three linear relationship and three non-linear relationship. More details about non-linear relationship done by choosing the knot point using a truncated spline based on the results of test linearity and form from the obtained curve as following.

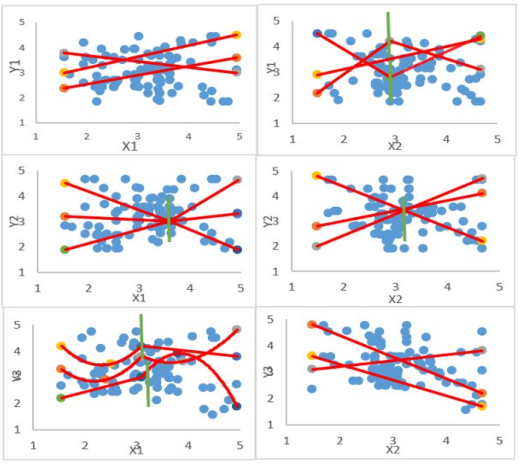

Figure 2: Scatter Plot Model

4.3. Integration of K-Means Cluster Analysis with Multiresponse Regression Truncated Spline Semiparametric

Semiparametric multiresponse regression using a truncated spline as approach is used in order to solve modeling by doing restrictions until quadratic with a maximum knot point used is 1 knots. Based on the linearity test results and choosing the knot point using GCV then obtained connection between predictor and response variables, with the formation of the model as following.

Cluster 1 ($D_1 = 0, D_2 = 0$)

$$y_{1i} = \beta_{01} + \beta_{11}x_{1i} + \beta_{21}x_{2i} + \beta_{31}\left(x_{2i} - k_{11}\right)_+ + \varepsilon_{1i}$$

$$y_{2i} = \beta_{02} + \beta_{12}x_{1i} + \beta_{22}\left(x_{1i} - k_{12}\right)_+ + \beta_{32}x_{2i} + \varepsilon_{2i}$$

$$y_{3i} = \beta_{03} + \beta_{13}x_{1i} + \beta_{23}x_{1i}^{\ 2} + \beta_{33}\left(x_{1i} - k_{13}\right)_+^2 + \beta_{43}x_{2i} + \varepsilon_{3i}$$

Cluster 2 ($D_1 = 1, D_2 = 0$)

$$y_{1i} = \beta_{01} + \beta_{11}x_{1i} + \beta_{21}x_{2i} + \beta_{31}\left(x_{2i} - k_{11}\right)_+ + \beta_{41}D_{1i} + \beta_{61}x_{1i}D_{1i} + \beta_{81}x_{2i}D_{1i} + \beta_{101}\left(x_{2i} - k_{11}\right)_+ D_{1i} + \varepsilon_{1i}$$

$$y_{2i} = \beta_{02} + \beta_{12}x_{1i} + \beta_{22}\left(x_{1i} - k_{12}\right)_+ + \beta_{32}x_{2i} + \beta_{42}D_{1i} + \beta_{62}x_{1i}D_{1i} + \beta_{82}\left(x_{1i} - k_{12}\right)_+ D_{1i} + \beta_{102}x_{2i}D_{1i} + \varepsilon_{2i}$$

$$y_{3i} = \beta_{03} + \beta_{13}x_{1i} + \beta_{23}x_{1i}^{\ 2} + \beta_{33}\left(x_{1i} - k_{13}\right)_+^2 + \beta_{43}x_{2i} + \beta_{53}D_{1i} + \beta_{73}x_{1i}D_{1i} + \beta_{93}x_{1i}^{\ 2}D_{1i} + \beta_{113}\left(x_{1i} - k_{13}\right)_+^2 D_{1i} + \beta_{133}x_{2i}D_{1i} + \varepsilon_{3i}$$

Cluster 3 ($D_1 = 0, D_2 = 1$)

$$y_{1i} = \beta_{01} + \beta_{11}x_{1i} + \beta_{21}x_{2i} + \beta_{31}\left(x_{2i} - k_{11}\right)_+ + \beta_{51}D_{2i} + \beta_{71}x_{1i}D_{2i} + \beta_{91}x_{2i}D_{2i} + \beta_{111}\left(x_{2i} - k_{11}\right)_+ D_{2i} + \varepsilon_{1i}$$

$$y_{2i} = \beta_{02} + \beta_{12}x_{1i} + \beta_{22}\left(x_{1i} - k_{12}\right)_+ + \beta_{32}x_{2i} + \beta_{52}D_{2i} + \beta_{72}x_{1i}D_{2i} + \beta_{92}\left(x_{1i} - k_{12}\right)_+ D_{2i} + \beta_{112}x_{2i}D_{2i} + \varepsilon_{2i}$$

$$y_{3i} = \beta_{03} + \beta_{13}x_{1i} + \beta_{23}x_{1i}^{\ 2} + \beta_{33}\left(x_{1i} - k_{13}\right)_+^2 + \beta_{43}x_{2i} + \beta_{63}D_{2i} + \beta_{83}x_{1i}D_{2i} + \beta_{103}x_{1i}^{\ 2}D_{2i} + \beta_{123}\left(x_{1i} - k_{13}\right)_+^2 D_{2i} + \beta_{143}x_{2i}D_{2i} + \varepsilon_{3i}$$

Integration cluster with semiparametric multiresponse regression using Weighted Least Square and cluster analysis using K-Means with the Euclidean distance that has been obtained give 3 cluster as the results. Next, cluster analysis integrated with semiparametric truncated spline multiresponse regression. The formed models are as follows.

Low Clusters ($D_1 = 0, D_2 = 0$)

$$\hat{f}_1(x_1, x_2) = 3.2091 + 0.1891x_{1i} - 0.2319x_{2i} + 12.5796062x_{2i}\left(x_{2i} - 4.5723\right)_+$$

$$\hat{f}_2(x_1, x_2) = 2.8400 + 0.4363x_{1i} - 2.8229x_{1i}\left(x_{1i} - 4.5366\right)_+ - 0.3254x_{2i}$$

$$\hat{f}_3(x_1, x_2) = 3.443 + 0.4349x_{1i} - 0.0247x_{1i}^{\ 2} - 5.2270x_{1i}\left(x_{1i} - 4.4607\right)_+^2 - 0.4277x_{2i}$$

Medium Clusters ($D_1 = 1, D_2 = 0$)

$$\hat{f}_1(x_1,x_2) = 3.9237 + 0.7734D_{1i} - 0.2943x_{1i} + 0.4118x_{1i}D_{1i} - 0.1411x_{2i} + 0.0302x_{2i}D_{1i} - 0.8612(x_{2i} - 1.4905)_+ + 0.5024(x_{2i} - 1.4905)_+ D_{1i}$$

$$\hat{f}_2(x_1,x_2) = 4.0013 + 0.7799D_{1i} - 0.0446x_{1i} - 0.0135x_{1i}D_{1i} + 0.4259(x_{1i} - 2.1718)_+ - 0.2333(x_{1i} - 2.1718)_+ D_{1i} - 0.4114x_{2i} + 0.5580x_{2i}D_{1i}$$

$$\hat{f}_3(x_1,x_2) = 3.3356 + 1.2900D_{1i} + 0.3691x_{1i} + 0.0798x_{1i}D_{1i} - 0.0510x_{1i}^2 - 0.0674x_{1i}^2D_{1i} + 0.2292(x_{1i} - 2.2021)_+^2 - 0.4871(x_{1i} - 2.2021)_+^2 D_{1i} - 0.3673x_{2i} + 0.3647x_{2i}D_{1i}$$

High Clusters ($D_1 = 0, D_2 = 1$)

$$\hat{f}_1(x_1,x_2) = 3.8804 + 0.7786D_{2i} - 0.3180x_{1i} + 0.4427x_{1i}D_{2i} - 0.1022x_{2i} - 0.0021x_{2i}D_{2i} - 1.0516(x_{2i} - 2.7649)_+ + 0.7464(x_{2i} - 2.7649)_+ D_{2i}$$

$$\hat{f}_2(x_1,x_2) = 4.0558 + 0.8374D_{2i} - 0.0310x_{1i} - 0.1207x_{1i}D_{2i} + 0.5913(x_{1i} - 2.1955)_+ - 0.6137(x_{1i} - 2.1955)_+ D_{2i} - 0.4442x_{2i} + 0.6708x_{2i}D_{2i} + \varepsilon_{2i}$$

$$\hat{f}_3(x_1,x_2) = 3.1249 + 1.1334D_{2i} + 0.5015x_{1i} + 0.0508x_{1i}D_{2i} - 0.0764x_{1i}^2 - 0.0521x_{1i}^2D_{2i} + 0.3605(x_{1i} - 2.3245)_+^2 - 0.7787(x_{1i} - 2.3245)_+^2 D_{2i} - 0.3524x_{2i} + 0.3995x_{2i}D_{2i}$$

4.4. Selecting the Best Cluster from Integration Model Validity Clusters with Multiresponse Regression Truncated Spline Semiparametric

Selecting the best cluster and model validity was carried out by choosing a model that has greatest total R^2 value, at this research has the largest total R^2 value is Cluster one called Low Cluster with a suitability level of 0.205. It means that the variable volatility production And volatility price can explain diversity variable access market agriculture, income farmer And well-being farmer by 20.5%, meanwhile the remaining 79.5% is not explained completely.

5. CONCLUSION AND SUGGESTIONS

5.1. Conclusion

This research was conducted to develop a form of integration of non-hierarchical cluster analysis with semiparametric multiresponse regression using the Truncated spline approach in the case of Farmer Inclusivity and to obtain an estimate of its function. In contrast to previous studies which did not combine cluster analysis with semiparametric truncated spline multiresponse regression.

Based on the results of the analysis and discussion that has been carried out, it can be concluded as follows:

1) By using the Euclidean distance in the K-means cluster, 3 clusters were obtained, where cluster 1 had the low farmer inclusiveness consisting of 42 farmers, cluster 2 has a medium farmer inclusiveness consisting 21 farmers, and cluster 3 has a high farmer inclusiveness consisting 37 farmers.
2) The integration application of cluster analysis with semiparametric truncated spline multiresponse regression using Euclidean distance yields obtained 3 clusters, and each of them has a different members which formed dummy variable become different and it affects R^2 value. From the three clusters obtained, on cluster 1 has the highest R^2 value by 20.5%.
3) Modeling using the integration cluster with semiparametric truncated spline multiresponse regression showed the changing data properties that have the changing pattern in their relationship. In addition, this modeling can contribute to the government and farmers to know level of inclusiveness of farmers as well as study factors affecting farmer inclusiveness.

5.2. Suggestions

Some suggestions that can be given based on the results of this study are as follows:

1) Future research can compare distance measures such as euclidean, manhattan and mahalanobis.
2) Future research can also use methods other than truncated splines such as smoothing splines and kernels.

REFERENCES:

[1] Afira, N., & Wijayanto, AW, Cluster Analysis Using the Partitioning and Hierarchical Methods on Provincial Poverty Information Data in Indonesia in 2019, *Komputika : Journal of Computer Systems*, *10* (2), 101–109, 2021, https://doi.org/10.34010/komputika.v10i2.4317.

[2] Musfiani, M, Cluster Analysis Using the Partition Method for Contraceptive Users in West Kalimantan. *Bimaster : Scientific Bulletin of Mathematics, Statistics and Its Applications* , *8* (4), 893–902, 2019, https://doi.org/10.26418/bbimst.v8i4.36584

[3] Hasna, & Achmad, AI, Binary Probit Regression Method for Modeling Factors Influencing Heart Disease Diagnosis, *Journal of Statistical Research*, 28–34, 2022, https://doi.org/10.29313/jrs.vi.721.

[4] Hablum, R. J., Khairan, A., & Rosihan, R. (2019). Clustering Hasil Tangkap Ikan Di Pelabuhan Perikanan Nusantara (Ppn) Ternate Menggunakan Algoritma K-Means. *Jurnal Informatika dan Komputer*, *2*(1), 26-33.

[5] Hidayati, L., Chamidah, N., & Budiantara, I. N. Estimasi Selang Kepercayaan Nilai Ujian Nasional Berbasis Kompetensi Berdasarkan Model Regresi Semiparametrik Multirespon Truncated Spline. *Media Statistika*, *13*(1), 92-103.

[6] Hikmah, H., Fardinah, F., Qadrini, L., & Tande, E, Cluster Analysis of District Grouping in West Sulawesi Based on Education Indicators. *Scientific*, *8* (2), 188–196, 2022, https://doi.org/10.31605/saintfik.v8i2.383.

[7] Ningrum, MN, Satyahadewi, N., & Debataraja, NN, Modeling the Factors Influencing the Human Development Index in Indonesia Using Spline Semiparametric Regression, *Bimaster : Scientific Bulletin of Mathematics, Statistics and Its Applications*, *9* (1), 57–64, 2020, https://doi.org/10.26418/bbimst.v9i1.38583.

[8] Fernandes, AAR, Budiantara, IN, Otok, BW, & Suhartono, S, *Dissertation Exam (Closed) Spline Estimator in Nonparametric Biresponse Regression for Longitudinal Data (Case Study in Patients with Pulmonary TB in Malang), 9,* 2016.

[9] Kaltsum, YES, *Development of Truncated Spline Semiparametric Multi-Response Regression for Modeling Defects in Cowhide,* Brawijaya University, 2023.

[10] Bintariningrum, MF, & Budiantara, IN, Truncated Spline Nonparametric Regression Modeling and Its Application to Crude Birth Rate in Surabaya. *Journal of Science and Arts ITS* , *3* (1), D7–D12, 2014, http://ejurnal.its.ac.id/index.php/sains_seni/article/view/6098%0Ahttps://ejurnal.its.ac.id.