IOSR Journal of Mathematics (IOSR-JM)
e-ISSN: 2278-5728, p-ISSN: 2319-765X. Volume 17, Issue 5 Ser. III (Sep. – Oct. 2021), PP 35-47
www.iosrjournals.org

Parameter Estimation and Hypothesis Testing of The Modified Multivariate Adaptive Regression Spline for Modeling the Number of Diseases

Septia Devi Prihastuti Yasmirullah [1], Bambang Widjanarko Otok[2], Jerry Dwi Trijoyo Purnomo [3], Dedy Dwi Prastyo [4]

[1,2,3,4](*Departement of Statistics, Faculty of Science and Data Analytics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia)*

Abstract:
Background*: The limitations of parametric regression make nonparametric regression an alternative method that prioritizes flexibility. One of the nonparametric regression methods is the Multivariate Adaptive Regression Spline (MARS). Many researchers have used MARS to analyze the data with numerical or categorical responses. However, there is one type of numerical data that requires special attention in modeling, which is count data. The count data is often encountered, especially in the health sector, such as the number of diseases. The purpose of modeling the number of diseases is predicting, so that prevention and treatment can be carried out appropriately. However, the conventional MARS methods cannot consider count data type. The specific objective of this study was to develop the parameter estimation and the hypothesis test of the Multivariate Adaptive Poisson Regression Spline (MAPRS) model.*
Materials and Methods*: This study builds statistical models that modify the Multivariate Adaptive Regression Spline (MARS) and Poisson regression, which is the Multivariate Adaptive Poisson Regression Spline (MAPRS). Secondary data for this study comes from the East Java Health Profiles publication and the East Java Province COVID-19 Task Force website. This study focused on parameter estimation using ordinary least squares (OLS) and hypothesis testing using maximum likelihood ratio test (MLRT) and t-test.*
Results*: When the data is the count type, MAPRS model is the better than Poisson regression and conventional MARS. Thus, it can be proven that the count data type requires special handling by considering the data type. The application of the MAPRS model was used for COVID-19 modeling. The results of COVID-19 data modeling in East Java, Indonesia show that several basis functions are formed and the effect of basis function may be different in each regency. The predictor variables included in the basis function are the number of patients under surveillance, the number of lifetime migration population, the number of hospitals and public health centers, and the percentage of households with a clean and healthy lifestyle. Based on these results, each regency should take a policy according to the variables that have a significant effect in the district. Thus, COVID-19 can be carried out depending on the conditions or the character of the regency.*
Conclusion: *Statistical methods for the count data type require special attention, however, studies on parameter estimation and hypothesis testing of MAPRS methods seem a bit. Therefore, the development of the Poisson and MARS regression is potentially improved.*
Key Word*: Count Data; Multivariate Adaptive Regression Spline (MARS); Multivariate Adaptive Poisson Regression Spline (MAPRS); Poisson Regression.*

Date of Submission: 16-10-2021 Date of Acceptance: 31-10-2021

I. Introduction

One of the statistical methods often used is regression analysis, where the approaches in regression analysis are parametric, semi-parametric, and non-parametric. The parametric approach has conditions that must be fulfilled, which is knowing the shape of the regression curve. This makes the parametric approach inflexible in modeling data, especially data that has a nonlinear model and has high dimensions. One method that can be used to overcome the limitations of parametric regression is nonparametric regression. The use of nonparametric regression is an alternative method that prioritizes flexibility, where there is the possibility to search for regression curve models that have no shape or are unknown [1].

Research on nonparametric regression was carried out, i.e., the study of the propensity score stratification method with bootstrap aggregation for the analysis of classification trees [2], the study of the Multivariate Adaptive Regression Spline method (MARS) [3], the MARS method with bootstrap aggregation [4], the

development of the MARS method with Monte Carlo Simulation (MCS) [5], the development of the MARS method through the optimization of differential flower pollination (DFP) [6], parameter estimation of MARS with stepwise approach [7], the multi drug-resistant tuberculosis (mdr-tb) prevalence using MARS [8], MARS for Visible and Near-Infrared-Based Soil Organic Matter [9], assessment of pile drivability using random forest regression and MARS [10]. MARS is a method that was introduced by Friedman in 1991. The method is a nonparametric regression method that can accommodate additive effects and interaction effects between predictor variables. Additionally, MARS does not assume the functional relationship form between the response and predictor variables. Moreover, it has a flexible functional form. MARS is able to handle data whose behavior changes at certain sub-intervals, because there are knot points that can indicate changes in data behavior patterns. MARS also has the ability to obtain better predictive results or to approach the shape of the regression curve from the pattern of response relationships and predictors that are unknown [3]. Mostly, MARS modeling has been used on data with numerical or categorical responses, but there is a numerical data that attracts special attention, namely count data. The method commonly used to model count data is Poisson regression. However, there are still limitations in using the Poisson regression method. It motivated the theory development and application of the MARS method, namely the Multivariate Adaptive Poisson Regression Spline (MAPRS). This method combines the MARS method and the Poisson Regression.

In this study, the number of COVID-19 data was classified as a type of count data. COVID-19 is a contagious disease caused by a new type of coronavirus. Coronavirus is a group of viruses that can cause disease in animals or humans. Several coronaviruses can cause respiratory infections in humans, ranging from coughs and colds to more serious infections such as Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS). COVID-19 was first identified in December 2019 in Wuhan, China. Currently, COVID-19 is a pandemic that is occurring in many countries around the world, particularly in Indonesia [11]. COVID-19 is spread from person to person by droplets from the nose or mouth of an infected person when they cough, sneeze or speak [12, 13, 14]. The elderly and those with concomitant illnesses, such as high blood pressure, heart and lung problems, diabetes or cancer, are at risk of more serious illness. As a result, several governments have led to implement certain measures to control this pandemic by curbing the spread and reducing deaths from COVID-19 [15, 16, 17]. One of the identification efforts to break the chain of spread of COVID-19 is mathematical modeling. In addition, the relationship between response variables and predictor variables was explained by a mathematical model. Therefore, this study proposes to analyze COVID-19 data by the MAPRS method.

II. Material And Methods

2.1 Multivariate Adaptive Regression Spline (MARS)

Multivariate Adaptive Regression Spline (MARS) is a combination of the truncated spline method with Recursive Partitioning Regression (RPR) to produce a continuous model at knots [3]. If $\mathbf{Y}$ is response variable with $\mathbf{X} = \left(x_1, \cdots, x_p \right)$ is predictor variables, p is the number of predictor variabel, and n is the number of observations, then MARS function, in general, can be written in the following equation [3]:

$$\begin{aligned} y_i &= f\left(x_{1i}, \cdots, x_{pi} \right) + \varepsilon_i \\ &= f\left(\mathbf{x}_i \right) + \varepsilon_i, \\ &= a_0 + \sum_{m=1}^{M} a_m \prod_{k=1}^{K_m} \left[s_{km} \left(x_{v(k,m)i} - t_{km} \right) \right]_+ + \varepsilon_i, i = 1, \cdots, n \end{aligned} \tag{1}$$

where,

a_0	:	constant parameter of basis function
a_m	:	non-constant parameter of *m*-th basis function
M	:	number of maximum basis function
K_m	:	maximum interaction of *m*-th basis function
s_{km}	:	sign of basis function in the *k*-th interaction and *m*-th basis function, where s_{km} is +1 or -1
$x_{v(k,m)i}$	:	*v*-th predictor variable, where *v* is an index of predictor variables related to *k*-th interaction and *m*-th basis function in MARS function
t_{km}	:	value of knot in *k*-th interaction and *m*-th basis function

If,

$$B_{mi}(x,t) = \prod_{k=1}^{K_m} \left[s_{km} \left(x_{v(k,m)i} - t_{km} \right) \right] \tag{2}$$

then the MARS model was formed as:

$$\begin{aligned} y_i &= a_0 + \sum_{m=1}^{M} a_m \prod_{k=1}^{K_m} \left[s_{km} \left(x_{v(k,m)i} - t_{km} \right) \right] + \varepsilon_i \\ &= a_0 + \sum_{m=1}^{M} a_m B_{mi}(x,t) + \varepsilon_i \end{aligned} \tag{3}$$

The MARS model in equation (3) can be written in matrix form.

$$\mathbf{y} = \mathbf{B}\mathbf{a} + \boldsymbol{\varepsilon} \tag{4}$$

where,

$$\mathbf{y}_{n\times 1} = \left(y_1, \ldots, y_n \right)^T,$$

$$\mathbf{a}_{(M+1)\times 1} = \left(a_0, \ldots, a_M \right)^T,$$

$$\boldsymbol{\varepsilon}_{n\times 1} = \left(\varepsilon_1, \ldots, \varepsilon_n \right)^T,$$

$$\mathbf{B}_{n\times(M+1)} = \begin{pmatrix} 1 & \prod_{k=1}^{K_1} \left[s_{k1} \left(x_{v(k,1)1} - t_{k1} \right) \right] & \cdots & \prod_{k=1}^{K_M} \left[s_{kM} \left(x_{v(k,M)1} - t_{kM} \right) \right] \\ 1 & \prod_{k=1}^{K_1} \left[s_{k1} \left(x_{v(k,1)2} - t_{k1} \right) \right] & \cdots & \prod_{k=1}^{K_M} \left[s_{kM} \left(x_{v(k,M)2} - t_{kM} \right) \right] \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \prod_{k=1}^{K_1} \left[s_{k1} \left(x_{v(k,1)n} - t_{k1} \right) \right] & \cdots & \prod_{k=1}^{K_M} \left[s_{kM} \left(x_{v(k,M)n} - t_{kM} \right) \right] \end{pmatrix}$$

The MARS model in equation (4) is estimated by Ordinary Least Square (OLS), which is to minimize the sum squares error.

$$\begin{aligned} \boldsymbol{\varepsilon} &= \mathbf{y} - \mathbf{B}\mathbf{a} \\ \left(\boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} \right) &= \left[\left(\mathbf{y} - \mathbf{B}\mathbf{a} \right)^T \left(\mathbf{y} - \mathbf{B}\mathbf{a} \right) \right] \end{aligned} \tag{5}$$

The optimization solution of equation (5) had obtained by the partial derivation of $\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}$ to $\mathbf{a}$, furthermore, the result equal to zero.

$$\hat{\mathbf{a}} = \left[\left(\mathbf{B}^T\mathbf{B} \right)^{-1} \left(\mathbf{B}^T\mathbf{y} \right) \right] \tag{6}$$

Next, to get the estimate of the MARS function, equation (6) is used.

$$\begin{aligned} \hat{\mathbf{y}} &= \mathbf{B}\mathbf{a} \\ &= \mathbf{B}\left(\mathbf{B}^T\mathbf{B} \right)^{-1} \mathbf{B}^T\mathbf{y} \\ &= \mathbf{H}(.)\mathbf{y} \end{aligned} \tag{7}$$

where $\mathbf{H}(.) = \mathbf{B}\left(\mathbf{B}^T\mathbf{B} \right)^{-1} \mathbf{B}^T$. So, $\mathbf{H}(.)$ is a matrix that depends on $\mathbf{B}$. $\mathbf{B}$ contains the optimal basis function that was selected by a stepwise procedure., i.e. forward and backward according to minimum GCV value.

2.2 Multivariate Adaptive Poisson Regression Spline (MAPRS)

Multivariate Adaptive Poisson Regression Spline (MAPRS) is a combination of Poisson Regression and MARS. The proposed model as follows:

$$
\begin{aligned}
&Y_i \sim Poisson(\mu) \\
&\ln \mu_i = f(\mathbf{x}_i) = a_0 + \sum_{m=1}^{M} a_m \prod_{k=1}^{K_m} \left[s_{km} \left(x_{v(k,m)i} - t_{km} \right) \right]_+ \\
&\qquad = a_0 + \sum_{m=1}^{M} a_m B_{mi}(\mathbf{x}_i)
\end{aligned} \tag{8}
$$

where,

$$\mu_i = \exp\left(a_0 + \sum_{m=1}^{M} a_m B_{mi}(\mathbf{x}_i) \right)$$

The method had been used to estimate the unknown parameters in the MAPRS model in equation (8), which is Ordinary Least Squares (OLS).

2.3 Individual Test of the Basis Function Coefficients for MAPRS Model

The purpose of the individual tests is to determine the significance coefficients of the basis function for the response, where the hypothesis is:

$$H_0 : a_m = 0$$
$$H_1 : a_m \neq 0,\ m = 1, \ldots, p$$

The statistical test is:

$$t = \frac{\hat{a}_m}{se(\hat{a}_m)} \tag{9}$$

If $|t| > t_{\left(\frac{\alpha}{2}, v\right)}$ or $p - value < \alpha$ then H_0 is rejected, it means that the coefficients parameter of the basis function has a significant effect on the response.

2.4 Generalized Cross-Validation (GCV)

Generalized cross-validation (GCV) is a criterion for selecting the best MAPRS model, the model that has the lowest GCV value among the other models is the best [3].

$$GCV(M) = \frac{\frac{1}{N}\sum_{i=1}^{N} \left[y_i - \hat{f}_M(x_i) \right]^2}{\left[1 - \frac{\left(trace\left[\mathbf{B}\left(\mathbf{B}^T\mathbf{B} \right)^{-1} \mathbf{B}^T \right] + 1 \right) + dM}{N} \right]^2} \tag{10}$$

where,

y_i : response variable

x_i : predictor variable

N : number of observation

$\hat{f}_M(x_i)$: estimation value of response variable on M basis function at x_i

M : maximum number of basis function

$\mathbf{B}$: matrix of M basis function

d : value when each basis function reaches optimization ($2 \leq d \leq 4$)

2.5 Research Variables

This study used secondary data from the East Java Health Profiles publication and the East Java Province COVID-19 Task Force website (http://infocovid19.jatimprov.go.id/). There is one limitation to this study, which is the data period from March 17, 2020 to May 6, 2020. The research variables for this study have been presented in Table 1.

Table 1. Research Variables

Code	Variables	Data Scale
Y	The Number of COVID-19 Positive Patients (people)	Ratio
X_1	The Number of Patients Under Surveillance (Bahasa : Pasien Dalam Pengawasan or PDP) (people)	Ratio
X_2	Lifetime Migration Population (people)	Ratio
X_3	Population Density (people per km^2)	Ratio
X_4	The Number of Health Workers (people)	Ratio
X_5	The Number of Poor Population (people)	Ratio
X_6	The Number of Hospitals & Public Health Centers (unit)	Ratio
X_7	Households have a clean and healthy lifestyle (Bahasa : Perilaku Hidup Bersih dan Sehat or PHBS) (%)	Ratio

2.6 Steps of Analysis

The steps of analysis in this study are:

1. Exploring and analyzing descriptive statistics on the research variables.
2. Estimate the model parameters using Equation (5).
3. Estimate the statistical test using Equation (9).
4. Modeling of the number of COVID-19 positive patients using the MAPRS method with the R code attached in Appendix 1.
   a. Determine the maximum possible number of Basis Function (BF).
   b. Determine the maximum number of interactions.
   c. Determine the minimum observation between knots by trial and error.
   d. Determine the best model according to the minimum GCV value.
5. Interpret the best model.
6. Test the significance parameter of the best model.
7. Draw the conclusions and suggestions

2.7. COVID-19

There are several specific terms in the mention of patients who are infected or suspected to be infected with COVID-19 in Indonesia, which are ODP, PDP, and Confirm [18].

a. ODP stands for people under observation ("Bahasa: Orang dalam Pengawasan or ODP"). A person who has a fever (≥ 38 ° C) or a history of fever; or symptoms of respiratory system disorders such as a runny nose, sore throat, cough and the last 14 days before the onset of symptoms having a history of travel or living abroad or in a transmission area local in Indonesia.
b. PDP means a patient under surveillance ("Bahasa: Pasien dalam Pengawasan or PDP"). A person having a fever (≥ 38 ° C), when accompanied by any of the symptoms of respiratory illnesses, such as cough, shortness of breath, sore throat, runny nose, mild to severe pneumonia. Then the last 14 days before symptoms appear have a history of travel or stay abroad or in a local transmission area in Indonesia.
c. A patient with Confirm status is someone infected with COVID-19 with a positive laboratory test result.

III. Result

3.1 Parameter Estimation of Multivariate Adaptive Poisson Regression Spline (MAPRS)

Multivariate Adaptive Poisson Regression (MAPRS) spline is a combination of Poisson regression and MARS. The purpose of method development is to accommodate the counts data type. The count data are often encountered, especially in the health sector. Therefore, this method must be developed.

Lemma 1: If the MARS model can accommodate the count data type, then the response variable assumes as a Poisson distribution, $Y_i \sim Poisson(\lambda)$ with $E[Y] = Var[Y] = \mu$. Then, the proposed model can be estimated by the Ordinary Least Squares (OLS) method. Then, the parameter estimation of the MAPRS model is:

$$\hat{f}(\mathbf{X}) = \ln \mu_i = \mathbf{B}\hat{a} = \ln\left(\left(\mathbf{B}^T \exp\left(\hat{\mathbf{a}}^T \mathbf{B}^T\right)\right)^{-1} \mathbf{B}^T \exp\left(\hat{\mathbf{a}}^T \mathbf{B}^T\right) \underset{\sim}{y}\right)$$

If Lemma 1 is satisfied, then the Multivariate Adaptive Poisson Regression Spline (MAPRS) model proposed as follows:

$$y_i = f(\mathbf{x}_i) + \varepsilon_i, i = 1, \cdots, n$$

where,

$$\ln \mu_i = f(\mathbf{x}_i) = a_0 + \sum_{m=1}^{M} a_m \prod_{k=1}^{K_m} \left[s_{km} \left(x_{v(k,m)\mathrm{i}} - t_{km} \right) \right]_+$$

$$= a_0 + \sum_{m=1}^{M} a_m B_{mi}(\mathbf{x}_i)$$

$$\mu_i = \exp\left(a_0 + \sum_{m=1}^{M} a_m B_{mi}(\mathbf{x}_i) \right) = \exp(\mathbf{Ba})$$

The estimation of the parameters of the MAPRS model uses the ordinary least squares (OLS) method, which minimizes the following functions:

$$\underset{\sim}{\varepsilon}^T \underset{\sim}{\varepsilon} = \left(\underset{\sim}{y} - \exp(\mathbf{Ba}) \right)^T \left(\underset{\sim}{y} - \exp(\mathbf{Ba}) \right)$$

$$= \underset{\sim}{y}^T \underset{\sim}{y} - \underset{\sim}{y}^T \exp(\mathbf{Ba}) - \exp\left(\mathbf{a}^T \mathbf{B}^T\right) \underset{\sim}{y} + \exp\left(\mathbf{a}^T \mathbf{B}^T\right) \exp(\mathbf{Ba})$$

$$= \underset{\sim}{y}^T \underset{\sim}{y} - \exp\left(\mathbf{a}^T \mathbf{B}^T\right) \underset{\sim}{y} - \exp\left(\mathbf{a}^T \mathbf{B}^T\right) \underset{\sim}{y} + \exp\left(\mathbf{a}^T \mathbf{B}^T\right) \exp(\mathbf{Ba})$$

$$= \underset{\sim}{y}^T \underset{\sim}{y} - 2\exp\left(\mathbf{a}^T \mathbf{B}^T\right) \underset{\sim}{y} + \exp\left(\mathbf{a}^T \mathbf{B}^T\right) \exp(\mathbf{Ba})$$

The next step is to find the first derivative of the function $\underset{\sim}{\varepsilon}^T \underset{\sim}{\varepsilon}$ with respect to $\mathbf{a}$, then equalize zero.

$$\frac{\partial \left(\underset{\sim}{\varepsilon}^T \underset{\sim}{\varepsilon} \right)}{\partial (\mathbf{a})} = \frac{\partial \left[\underset{\sim}{y}^T \underset{\sim}{y} - 2\exp\left(\mathbf{a}^T \mathbf{B}^T\right) \underset{\sim}{y} + \exp\left(\mathbf{a}^T \mathbf{B}^T\right) \exp(\mathbf{Ba}) \right]}{\partial (\mathbf{a})}$$

$$0 = -2\mathbf{B}^T \exp\left(\mathbf{a}^T \mathbf{B}^T\right) \underset{\sim}{y} + 2\mathbf{B}^T \exp\left(\mathbf{a}^T \mathbf{B}^T\right) \exp(\mathbf{Ba})$$

Then,

$$2\mathbf{B}^T \exp\left(\mathbf{a}^T \mathbf{B}^T\right) \exp(\mathbf{Ba}) = 2\mathbf{B}^T \exp\left(\mathbf{a}^T \mathbf{B}^T\right) \underset{\sim}{y}$$

$$\mathbf{B}^T \exp\left(\mathbf{a}^T \mathbf{B}^T\right) \exp(\mathbf{Ba}) = \mathbf{B}^T \exp\left(\mathbf{a}^T \mathbf{B}^T\right) \underset{\sim}{y}$$

$$\exp(\mathbf{Ba}) = \left(\mathbf{B}^T \exp\left(\mathbf{a}^T \mathbf{B}^T\right) \right)^{-1} \mathbf{B}^T \exp\left(\mathbf{a}^T \mathbf{B}^T\right) \underset{\sim}{y}$$

$$\mathbf{Ba} = \ln\left(\mathbf{B}^T \exp\left(\mathbf{a}^T \mathbf{B}^T\right) \right)^{-1} \mathbf{B}^T \exp\left(\mathbf{a}^T \mathbf{B}^T\right) \underset{\sim}{y}$$

$$\hat{\mathbf{a}} = (\mathbf{B})^{-1} \ln\left(\mathbf{B}^T \exp\left(\hat{\mathbf{a}}^T \mathbf{B}^T\right) \right)^{-1} \mathbf{B}^T \exp\left(\hat{\mathbf{a}}^T \mathbf{B}^T\right) \underset{\sim}{y}$$

The estimator of MAPRS model as follows:

$$\hat{\mathbf{a}}_{OLS} = (\mathbf{B})^{-1} \ln\left(\left(\mathbf{B}^T \exp\left(\hat{\mathbf{a}}^T \mathbf{B}^T\right) \right)^{-1} \mathbf{B}^T \exp\left(\hat{\mathbf{a}}^T \mathbf{B}^T\right) \underset{\sim}{y} \right)$$

In addition, the parameter estimation of the MAPRS model as follows:

$$\hat{f}(\mathbf{X}) = \ln \mu_i = \mathbf{B}\hat{a}$$

$$= \mathbf{B}(\mathbf{B})^{-1} \ln\left(\left(\mathbf{B}^T \exp\left(\hat{\mathbf{a}}^T \mathbf{B}^T\right) \right)^{-1} \mathbf{B}^T \exp\left(\hat{\mathbf{a}}^T \mathbf{B}^T\right) \underset{\sim}{y} \right)$$

$$= \ln\left(\left(\mathbf{B}^T \exp\left(\hat{\mathbf{a}}^T \mathbf{B}^T\right) \right)^{-1} \mathbf{B}^T \exp\left(\hat{\mathbf{a}}^T \mathbf{B}^T\right) \underset{\sim}{y} \right)$$

Therefore, the lemma 1 was fulfilled.

3.2 Hypothesis Testing of Multivariate Adaptive Poisson Regression Spline (MAPRS)

The Maximum Likelihood Ratio Test (MLRT) is a simultaneous parameter test method. A particular advantage of this test statistic is that it immediately extends to the case of the multiparameter, because the definition of this statistic does not refer to the dimension of the parameter space [19].

Lemma 2: If the Multivariate Adaptive Poisson Regression Spline (MAPRS) parameter model is tested by Maximum Likelihood Ratio Test (MLRT), then the hypothesis is $\mathrm{H}_0 : a_1 = a_2 = \cdots = a_M = 0$, and $\mathrm{H}_1 :$ at least one $a_m \neq 0$ where $m = 1, 2, \cdots, M$. Next, the test statistics is $G^2 = -2 \ln \Lambda = 2\left(\ln L\left(\hat{\Omega}\right) - \ln L\left(\hat{\omega}\right)\right)$.

If Lemma 2 is satisfied, then the parameter under the population can be defined as,

$$\Omega = \left\{a_0, \cdots, a_m\right\}$$

where $n(\Omega) = m + 1$. Next, the parameter under the H_0 can be defined as,

$$\omega = \left\{a_0\right\}$$

where $n(\omega) = 1$. Furthermore, the ln-likelihood under population as follows:

$$L\left(\hat{\Omega}\right) = \prod_{i=1}^{n} \left(\frac{\exp\left(-\exp\left(\mathbf{B}\hat{\mathbf{a}}\right)\right)\exp\left(\mathbf{B}\hat{\mathbf{a}}\right)^{y_i}}{y!} \right)$$

$$= \frac{\left(\exp\left(-\sum_{i=1}^{n}\exp\left(\mathbf{B}\hat{\mathbf{a}}\right)\right)\right)\left(\prod_{i=1}^{n}\exp\left(\mathbf{B}\hat{\mathbf{a}}\right)^{y_i}\right)}{\prod_{i=1}^{n} y_i!}$$

$$\ln L\left(\hat{\Omega}\right) = \ln\left(\frac{\left(\exp\left(-\sum_{i=1}^{n}\exp\left(\mathbf{B}\hat{\mathbf{a}}\right)\right)\right)\left(\prod_{i=1}^{n}\exp\left(\mathbf{B}\hat{\mathbf{a}}\right)^{y_i}\right)}{\prod_{i=1}^{n} y_i!} \right)$$

$$\ln L\left(\hat{\Omega}\right) = \sum_{i=1}^{n}\left(-\exp\left(\mathbf{B}\hat{\mathbf{a}}\right) + y_i\mathbf{B}\hat{\mathbf{a}} - \ln y_i!\right)$$

$$= \sum_{i=1}^{n}\left(y_i\mathbf{B}\hat{\mathbf{a}} - \exp\left(\mathbf{B}\hat{\mathbf{a}}\right) - \ln y_i!\right)$$

Next, the ln-likelihood under H_0 as follows:

$$\ln L\left(\hat{\omega}\right) = \ln\left(\frac{\left(\exp\left(-\sum_{i=1}^{n}\exp\left(\hat{a}_0\right)\right)\right)\left(\prod_{i=1}^{n}\exp\left(\hat{a}_0\right)^{y_i}\right)}{\prod_{i=1}^{n} y_i!} \right)$$

$$\ln L\left(\hat{\omega}\right) = \sum_{i=1}^{n}\left(-\exp\left(\hat{a}_0\right) + y_i\hat{a}_0 - \ln y_i!\right)$$

$$= \sum_{i=1}^{n}\left(y_i\hat{a}_0 - \exp\left(\hat{a}_0\right) - \ln y_i!\right)$$

Furthermore,

$$G^2 = 2\left[\ln\left(L\left(\hat{\Omega}\right)\right) - \ln\left(L\left(\hat{\omega}\right)\right)\right]$$

$$= 2\left[\sum_{i=1}^{n}\left(y_i\mathbf{B}\hat{\mathbf{a}} - \exp\left(\mathbf{B}\hat{\mathbf{a}}\right) - \ln y_i!\right) - \sum_{i=1}^{n}\left(y_i\hat{a}_0 - \exp\left(\hat{a}_0\right) - \ln y_i!\right)\right]$$

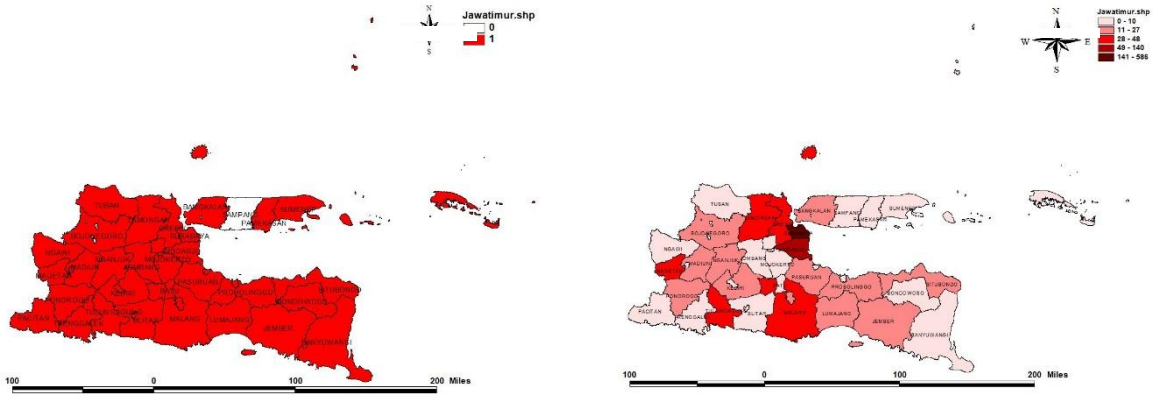
$$= 2\left[\sum_{i=1}^{n}\left(y_i\mathbf{B}\hat{\mathbf{a}} - \exp\left(\mathbf{B}\hat{\mathbf{a}}\right)\right) - \sum_{i=1}^{n}\left(y_i\hat{a}_0 - \exp\left(\hat{a}_0\right)\right)\right]$$

Therefore, the lemma 2 was fulfilled.

3.3 Application of Multivariate Adaptive Poisson Regression Spline (MAPRS)

3.3.1 Characteristics of Research Variables

East Java is one of the provinces of Indonesia affected by COVID-19. As of May 6, 2020, there were 1,220 confirmed positive cases, with 3,645 patients under surveillance (PDP) and 20,608 people under observation (ODP). The spread of COVID-19 patients in each regency in East Java has been observed in Figure 1.



(a) (b)

Figure 1. The spread of COVID-19 patients in East Java

The red color in Figure 1 (a) shows the presence of a patient confirmed positive for COVID-19 in a regency. According to Figure 1, almost all regencies in East Java have COVID-19 cases. Meanwhile, the white color in Figure 1 (a) shows that there are no patients confirmed positive for COVID-19. It shows that there is only one regency has white color, which is Sampang. It means Sampang has not been any positive cases of COVID-19.

Figure 1 (b) shows the number of COVID-19 patients per regency. The lightest color shows the regency with the lowest number of COVID-19 patients, while the darkest color shows the highest number. Surabaya is the city with the darkest color, so the highest number of COVID-19 patients for East Java Province is in Surabaya. Then, followed by Sidoarjo and Lamongan.

Furthermore, descriptive statistics of each variable are shown in Table 2.

Table 2. Descriptive Statistics of Research Variables

Variables	Mean	St.Dev	Variance	Min	Median	Max
The Number of Positive COVID-19 Patients	32.1	95.4	9093	0	10.5	586
The Number of Patients Under Surveillance	95.9	225.9	51020.8	0	31	1354
Lifetime Migration Population	114070	201344	4.054E+10	17602	54978	1092204
Population Density	1912	2167	4696658	278	890	8233
The Number of Health Workers	2551	2001	4002233	650	2081	12142
The Number of Poor Population	114	70.2	4923.2	7	116.3	268.5
The Number of Hospitals & Public Health Centers	35.37	20.47	419.05	9	31.5	122
Households have a clean and healthy lifestyle	51.92	15.62	243.84	18.2	50.05	83.2

The average number of positive patients with COVID-19 in each regency in East Java is 32 people. But, there is still one regency that does not have positive patients with COVID-19, which is Sampang. Meanwhile, most of the COVID-19 patients were in the city of Surabaya, namely 586 people. Furthermore, the average number of the patient under surveillance is 96 people, with the highest number of the patient under surveillance in Surabaya, namely 1354 people. East Java has an average lifetime migration population of 114070 people, with an average population density of 1912 people/km2. The average number of health workers in East Java is 2551 people, with an average number of hospitals and public health centers of 35 units. The average number of poor people in East Java is 114 people. Then there are 51.92% of households in East Java implementing a clean and healthy lifestyle.

The relationship pattern between the number of positive COVID-19 patients and other research variables have been seen in Figure 2.

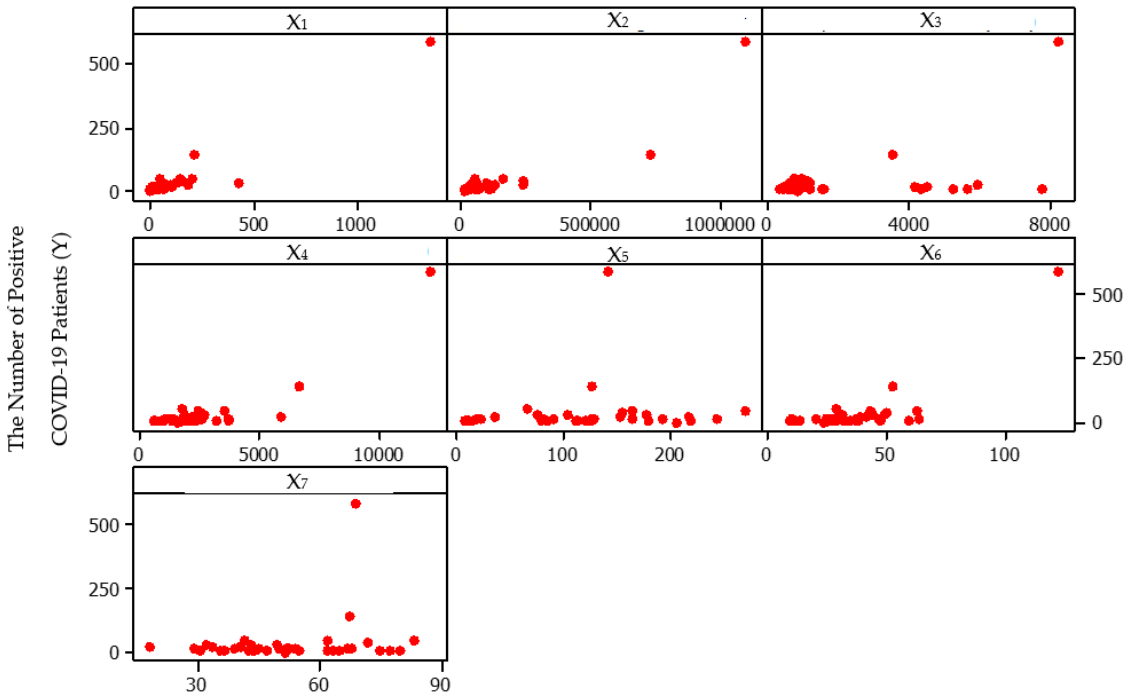


Figure 2. The Relationships Patterns between Research Variables

Figure 2 shows the relationship pattern between the response variable, which is the number of positive COVID-19 patients, with the predictor variables. The plot pattern has been randomly distributed, and cannot be known, so the appropriate approach to use is the non-parametric regression approach. One of the non-parametric regression approaches is the Multivariate Adaptive Poisson Regression Spline (MAPRS) [20].

3.3.2 Multivariate Adaptive Poisson Regression Spline (MAPRS) Modeling

Multivariate Adaptive Poisson Regression Spline (MAPRS) is modeling that is carried out by combining the maximum number of basis functions (BF), Maximum Interaction (MI), and Minimum Observation (MO) between knots. So, it obtains an optimal model with a minimum GCV value. In the study of the number of COVID-19 patients in East Java used seven predictor variables, so that the BF was 14, 21, and 28. The results of the combination of BF, MI, and MO in this study has shown in Table 3.

Table 3. Combination of BF, MI and MO

BF	MI	MO	GCV	R2	BF	MI	MO	GCV	R2	BF	MI	MO	GCV	R2
14	1	0	184.3381	0.97918	21	1	0	111.1532	0.987446	28	1	0	111.1532	0.987446
14	1	1	46.97232	0.994695	21	1	1	46.97232	0.994695	28	1	1	46.97232	0.994695
14	1	2	86.10249	0.990275	21	1	2	86.10249	0.990275	28	1	2	86.10249	0.990275
14	1	3	101.0263	0.988589	21	1	3	101.0263	0.988589	28	1	3	101.0263	0.988589
14	1	4	108.2552	0.987773	21	1	4	108.2552	0.987773	28	1	4	108.2552	0.987773
14	1	5	117.0626	0.986778	21	1	5	117.0626	0.986778	28	1	5	117.0626	0.986778
14	2	0	68.07068	0.992312	21	2	0	55.71253	0.993707	28	2	0	55.71253	0.993707
14	2	1	46.30097	0.99477	21	2	1	46.30097	0.99477	28	2	1	46.30097	0.99477
14	2	2	61.59636	0.993043	21	2	2	61.59636	0.993043	28	2	2	61.59636	0.993043
14	2	3	52.5676	0.994063	21	2	3	52.5676	0.994063	28	2	3	52.5676	0.994063
14	2	4	62.91915	0.992893	21	2	4	62.91915	0.992893	28	2	4	62.91915	0.992893
14	2	5	43.28971	0.995111	21	2	5	43.28971	0.995111	28	2	5	43.28971	0.995111
14	3	0	68.07068	0.992312	21	3	0	55.71253	0.993707	28	3	0	55.71253	0.993707
14	3	1	46.30097	0.99477	21	3	1	46.30097	0.99477	28	3	1	46.30097	0.99477
14	3	2	43.3597	0.995103	21	3	2	43.3597	0.995103	28	3	2	43.3597	0.995103
14	3	3	52.5676	0.994063	21	3	3	52.5676	0.994063	28	3	3	52.5676	0.994063

BF	MI	MO	GCV	R2	BF	MI	MO	GCV	R2	BF	MI	MO	GCV	R2
14	3	4	61.15003	0.993093	21	3	4	61.15003	0.993093	28	3	4	61.15003	0.993093
14	3	5	43.28971	0.995111	21	3	5	43.28971	0.995111	28	3	5	43.28971*	0.995111

Note : * GCV minimum value

The results of the combination of BF, MI, and MO in Table 2 show that the best MAPRS model has BF = 28, MI = 3, and MO = 5 with a GCV value of 43.28971 and R^2 value 0.9951. Furthermore, it can be written as the best MAPRS models:

$$\hat{f}(x) = 3.0844 - 0.0230BF_1 + 0.0609BF_2 + 0.00001BF_3 - 0.00002BF_4 + 0.0000001BF_5 - 0.0014BF_6 + 0.0000006BF_7$$

where,

$BF_1 = h(146 - X1)$

$BF_2 = h(X1 - 146)$

$BF_3 = h(115520 - X2)$

$BF_4 = h(X2 - 115520)$

$BF_5 = h(146 - X1) * X2$

$BF_6 = h(X1 - 146) * X7$

$BF_7 = h(X2 - 115520) * X6$

Next, interpret each basis function from the best model.

- $BF_1 = h(146 - X1)$

  The coefficient BF_1 will be statistically significant when the number of patients under surveillance is less than 146 patients. Each BF_1 increase by one unit (with the other is considered constant), then a decrease in the number of positive COVID-19 patients is 0.0230.

- $BF_2 = h(X1 - 146)$

  The coefficient BF_2 will be statistically significant when the number of patients under surveillance is more than 146 patients. Each BF_2 increase by one unit (with the other is considered constant), then an increase in the number of positive COVID-19 patients is 0.0609.

- $BF_3 = h(115520 - X2)$

  The coefficient BF_3 will be statistically significant when the lifetime migration population is less than 115520 people. Each BF_3 increase by one unit (with the other is considered constant), then an increase in the number of positive COVID-19 patients is 0.00001.

- $BF_4 = h(X2 - 115520)$

  The coefficient BF_4 will be statistically significant when the lifetime migration population is more than 115520 people. Each BF_4 increase by one unit (with the other is considered constant), then a decrease in the number of positive COVID-19 patients is 0.00002.

- $BF_5 = h(146 - X1) * X2$

  The coefficient BF_5 will be statistically significant when the number of patients under surveillance is less than 146 patients, and multiplied by the number of lifetime migration population. Each BF_5 increase by one unit (with the other is considered constant), then an increase in the number of positive COVID-19 patients is 0.0000001.

- $BF_6 = h(X1 - 146) * X7$

  The coefficient BF_6 will be statistically significant when the number of patients under surveillance is more than 146 patients, and multiplied by the percentage of households have a clean and healthy lifestyle. Each BF_6 increase by one unit (with the other is considered constant), then a decrease in the number of positive COVID-19 patients is 0.0014.

- $BF_7 = h(X2 - 115520) * X6$

  The coefficient BF_7 will be statistically significant when the lifetime migration population is more than 115520 people, and multiplied by the number of hospitals & public health centers. Each BF_7 increase by one unit (with the other is considered constant), then an increase in the number of positive COVID-19 patients is 0.0000006.

3.3.3 Significance Test of The Multivariate Adaptive Poisson Regression Spline (MAPRS) Model Parameters

The significance test of the modeling results using the MAPRS method has shown in table 4.

Table 4 Significance Test Results of the Best Model Parameters

Parameter	Estimate	Std. Error	Pr(>|t|)
(Intercept)	3.08E+00	5.02E-15	<2e-16 [a]
h(X1-146)	6.09E-02	7.80E-16	<2e-16 [a]
h(146-X1)	-2.31E-02	9.63E-17	<2e-16 [a]
h(X1-146)*X7	-1.47E-03	1.82E-17	<2e-16 [a]
h(X2-115520)	-2.70E-05	3.49E-19	<2e-16 [a]
h(115520-X2)	1.66E-05	1.28E-19	<2e-16 [a]
h(146-X1)*X2	1.39E-07	1.09E-21	<2e-16 [a]
h(X2-115520)*X6	6.53E-07	7.63E-21	<2e-16 [a]

Note: [a] Significant at 0.05 level

Based on the test results in Table 4, it shows all the basis function parameters of the model are significant, where the *p-value* < 0.05.

3.3.4 Comparison of Poisson Regression, MARS and MAPRS Model

The model selection can be determined based on the R^2 value. According to the criteria, the best model is a model with the highest R^2.

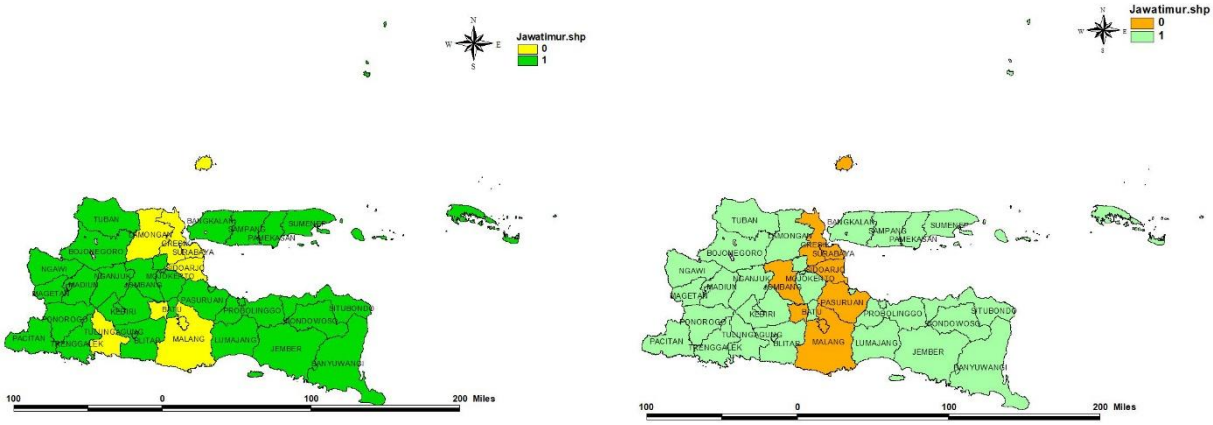
Table 5 Models Comparison

Models	R^2
Poisson Regression	0.9875
MARS	0.9947
MAPRS	**0.9951**

Table 5 shows that MAPRS is the best model than the others, where the R^2 value of MAPRS is 0.9951. Thus, it can be proven that the count data type requires special handling by considering the data type.

IV. Discussion

The results of COVID-19 data modeling in East Java using the MAPRS method show that there are several basis functions are formed. Where each basis function has affects the number of COVID-19 patients. Based on this model, it shows that the effect of the basis function can be different in each regency. One of them has been seen by the map in Figure 3.



(a) (b)

Gambar 3. The Effect of The basis function Each Regency in East Java

The green color in Figure 3 (a) is the regency that has a significant influence on BF_1, which is regency that has the number of patients under surveillance is less than 146 patients. Meanwhile, BF2 (the number of patients under surveillance is more than 146 patients) is significant in the yellow regency in Figure 3 (a). Furthermore, BF_3 (the lifetime migration population is less than 115520 people) is significant in the regency, which is colored green in Figure 3 (b). The orange color in Figure 3 (b) shows regency that has a significant influence on BF_4 (the lifetime migration population is more than 115520 people). Based on these results, each regency should take a policy according to the variables that have a significant effect in the district. Thus, COVID-19 can be carried out depending on the conditions or the character of the regency.

V. Conclusion

The city of Surabaya is an area that has the highest number of positive COVID-19 patients in East Java, which is 586 people and total patients under the surveillance of 1354 people. Furthermore, the best model from the formed MAPRS approach has a GCV value of 43.28971, with BF = 28, MI = 3, and MO = 5. The model is:

$$\hat{f}(x) = 3.0844 - 0.0230BF_1 + 0.0609BF_2 + 0.00001BF_3 - 0.00002BF_4 + 0.0000001BF_5 - 0.0014BF_6 + 0.0000006BF_7$$

Based on the results of the significance tests of the model parameters, it shows that all the basis function parameters of the model are significant. When the predictor variables included in the basis functions include the number of patients under surveillance, the number of lifetime migration population, the number of hospitals and public health centers, and the percentage of households with a clean and healthy lifestyle. In addition, each regency has a different effect of significant variables. Therefore, each regency should take a policy based on the variables that have a significant effect in the district. Thus, COVID-19 can be carried out depending on the conditions or the character of the regency.

Meanwhile, statistical methods for the count data type require special attention, however, studies on parameter estimation and hypothesis testing of MAPRS methods seem a bit. Therefore, the development of the Poisson and MARS regression is potentially improved.

Acknowledgement

We would like to express our gratitude to the Ministry of Research and Technology/National Research and Innovation Agency of Republic Indonesia (RISTEK-BRIN) for the financial support of this research, and the anonymous reviewers for suggestions for the improvement of this paper.

References

[1] Eubank, R. L. *Nonparametric Regression and Spline Smoothing*, 2nd ed.; Marcel Dekker, Inc.: New York, USA, 1999.
[2] Otok, B.W.; Musa, M.; Purhadi; Yasmirullah, S.D.P. Propensity score stratification using bootstrap aggregating classification trees analysis. *Heliyon* **2020**, 6, e04288.
[3] Friedman, J. H. Multivariate Adaptive Regression Splines. *The Annals of Statistics* **1991**, 19, 1-67.
[4] Otok, B.W.; Putra, R.Y.; Sutikno; Yasmirullah, S.D.P. Bootstrap Aggregating Multivariate Adaptive Regression Spline for Observational Studies in Diabetes Cases. *SRP* **2020**, 11, 406-413.
[5] Liu, L.; Zhang, S.; Cheng, Y. M.; Liang, L. Advanced reliability analysis of slopes in spatially variable soils using multivariate adaptive regression splines. *Geoscience Frontiers* **2019**, 10, 671–682.
[6] Tien Bui, D.; Hoang, N. D.; Samui, P. Spatial pattern analysis and prediction of forest fire using new machine learning approach of Multivariate Adaptive Regression Splines and Differential Flower Pollination optimization: A case study at Lao Cai province (Viet Nam). *Journal of Environmental Management* **2019**, *237*, 476–487.

[7] Yasmirullah, S D P *et al* 2021 *J. Phys.: Conf. Ser.* **1752** 012017
[8] Yasmirullah, S D P *et al* 2021 AIP Conference Proceedings **2329** 060019
[9] Wang, X.; Yang, C.; Zhou, M. Partial Least Squares Improved Multivariate Adaptive Regression Splines for Visible and Near-Infrared-Based Soil Organic Matter Estimation Considering Spatial Heterogeneity. Appl. Sci. **2021**, 11, 566.
[10] Wengang Zhang, Chongzhi Wu, Yongqin Li, Lin Wang & P. Samui. Assessment of pile drivability using random forest regression and multivariate adaptive regression splines, Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards **2019**, 15:1, 27-40
[11] WHO, Coronavirus disease (COVID-19). Available online: https://www.who.int/indonesia/news/novel-coronavirus (Accessed 2 September 2021).
[12] WHO, Coronavirus disease (COVID-19): How is it transmitted? Available online: https://www.who.int/news-room/q-a-detail/coronavirus-disease-covid-19-how-is-it-transmitted (Accessed 2 January 2021).
[13] Centers for Disease Control and Prevention, How COVID-19 Spreads? Available online: https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/how-covid-spreads.html (Accessed 24 July 2021).
[14] UNICEF, How does the novel coronavirus spread? Available online: https://www.unicef.org/indonesia/coronavirus/FAQ#howdoescoronavirusspread (Accessed 24 July 2021).
[15] WHO, Pertimbangan penyesuaian langkah-langkah kesehatan masyarakat dan sosial dalam konteks COVID-19. Available online: https://www.who.int/docs/default-source/searo/indonesia/covid19/who-2019-ncov-adjusting-ph-measures-2020-1-eng-indonesian.pdf?sfvrsn=63d5d4fe_2 (Accessed 2 Mei 2020).
[16] Satuan Tugas Penanganan COVID-19, Indonesia fights back the covid-19 second wave. Available online: https://covid19.go.id/p/berita/indonesia-fights-back-covid-19-second-wave (Accessed 2 Juli 2021).
[17] Keputusan Menteri Kesehatan Republik Indonesia Nomor HK.01.07/MENKES/5671/2021, Manajemen Klinis Tata Laksana Corona Virus Disease 2019 (Covid-19) di Fasilitas Pelayanan Kesehatan. Available online: https://covid19.go.id/storage/app/media/Regulasi/2021/Agustus/kmk-no-hk0107-menkes-5671-2021-ttg-manajemen-klinis-tata-laksana-covid-19-di-fasilitas-pelayanan-kesehatan-signed-1.pdf.
[18] Satuan Tugas Penanganan COVID-19 Jatim, Jatim Tanggap COVID-19. Available online: http://infocovid19.jatimprov.go.id/ (Accessed 2 Mei 2020).
[19] Maximized Likelihood Ratio Tests. Available online: http://www.stats.ox.ac.uk/~steffen/teaching/bs2siMT04/si13c.pdf (Accessed 1 December 2020).
[20] Yasmirullah, S D P *et al* 2021 *J. Phys.: Conf. Ser.* **1863** 012078

Septia Devi Prihastuti Yasmirullah, et. al. "Parameter Estimation and Hypothesis Testing of The Modified Multivariate Adaptive Regression Spline for Modeling the Number of Diseases." *IOSR Journal of Mathematics (IOSR-JM)*, 17(5), (2021): pp. 35-47.