Semiparametric Regression Estimator of Fourier Series For Categorical Data

Muhammad Zulfadhli, I Nyoman Budiantara, Vita Ratnasari, and Afiqah Saffa Suriaslan

***Abstract*— In recent years, a relationship between predictor and response variables has often been observed, particularly when the response variable consists of categorical data. Several methods have been developed to address cases involving qualitative response data. However, many of these approaches are limited in scope. This paper proposes a novel method for estimating semiparametric models with categorical response variables. The proposed method combines two estimators: a nonparametric Fourier series and a parametric linear function. To demonstrate the application of this method, we utilize data on the poverty gap status in East Java for the year 2023. The study selects the optimal model based on the smallest deviance value, the highest accuracy, and the largest Press's Q value. The results demonstrate that the semiparametric Fourier series regression provides significantly better estimation compared to binary logistic regression and nonparametric Fourier series regression.**

***Index Terms*— categorical data, Fourier Series, maximum likelihood estimation, semiparametric regression**

I. Introduction

Semiparametric regression can be used to ascertain the relationship between predictor and response variables. particularly when the regression curve's functional form is unknown, and some predictors exhibit a linear trend. It is expected that the semiparametric regression curve is smooth, as it lies within a specific function space. Importantly, the data are expected to estimate the curve independently, without being influenced by the researcher's subjective judgment. This characteristic enhances the flexibility of the semiparametric regression method. Using smoothing methods based on observed data, semiparametric regression can be applied. Fourier Series [1], Kernel [2], Spline [3], Wavelet [4], and Local Polynomial [5] are among the various smoothing techniques that are accessible. Spline estimators are used for data that exhibits different patterns and are depend on knot points [3]. The local polynomial estimator has been used to minimize the estimator's asymptotic varianceand bias characteristics in nonparametric regression with many response variables [6]. For observations of noise-contaminated signals, an additive wavelet estimator with a Gaussian distribution has been applied [7]. Patterned data with recurrent trends are estimated using the Fourier Series estimator [1]. The Fourier Series approach is the most effective of these estimators when there is a recurring pattern with a specific trend in the relationship between the predictor and responder variables [8]. In the context of additive nonparametric regression models, the Fourier Series estimator achieves an optimal balance between computational complexity and accuracy [9]. These methods are applicable to both univariable (single predictor) and multivariable (multiple predictors) scenarios [1, 7, 10, 11].

The Fourier Series estimator was initially put up by [1] and further studied by [7]. Additionally, [12] applied the Fourier Series estimator in the context of semiparametric regression. Further developments include the creation of a biresponse Fourier Series semiparametric regression by [13]. It then [14, 15, 16] developed into a mixture estimator of Fourier Series nonparametric regression and [17] developed a mixture estimator Fourier Series semiparametric regression. However, earlier studies, such as those by [18], focused exclusively on models using quantitative data. In practice, however, there are often a relationship between the predictor and response variables, where the response variable consists of categorical data. Because of this, researchers' Fourier Series semiparametric regression model for quantitative data is insufficient to handle problems requiring qualitative (categorical) answers.

In recent developments, the modeling of categorical response variables has gained significant attention. Using categorical data, researchers have created estimators for nonparametric regression in recent years. For instance, [19] introduced a Local Likelihood Logit Estimation approach and [20] employed a Spline *Truncated* function. Other researchers have also contributed to this field: [21] proposed a Fourier Series-based nonparametric regression estimator for categorical data. To date, no semiparametric regression estimator using a Fourier Series function has been developed for categorical response data. This study addresses this gap by proposing a Fourier Series semiparametric regression estimator specifically designed for categorical response variables. We apply this method to analyze data on the poverty gap status in East Java for the year 2023, as it provides a highly relevant and practical demonstration of the effectiveness and applicability of the proposed Fourier Series semiparametric regression estimator in handling categorical

Manuscript received September 4, 2024; revised April 5, 2025.

Through the "Master Education Program Towards Doctor for Excellent Scholars" scholarship, the Directorate General of Science and Technology Resources and Higher Education, Kemristekdikti, provided funding for this study.

Muhammad Zulfadhli is a postgraduate student in Statistics Department, Faculty of Science and Data Analytics, Institut Teknologi Sepuluh Nopember, Surabaya, 60111, Indonesia. (e-mail: 7003242004@student.its.ac.id).

I Nyoman Budiantara is a professor in Statistics Department, Faculty of Science and Data Analytics, Institut Teknologi Sepuluh Nopember, Surabaya, 60111, Indonesia (corresponding author to provide e-mail: i_nyoman_b@statistika.its.ac.id).

Vita Ratnasari is a professor in Statistics Department, Faculty of Science and Data Analytics, Institut Teknologi Sepuluh Nopember, Surabaya, 60111, Indonesia (e-mail: vita_ratna@its.ac.id).

Afiqah Saffa Suriaslan is a postgraduate student in Statistics Department, Faculty of Science and Data Analytics, Institut Teknologi Sepuluh Nopember, Surabaya, 60111, Indonesia. (e-mail: 7003242003@student.its.ac.id).

response data, particularly in cases where the relationship between the response and predictor variables exhibits both linear trends and recurring or periodic patterns.

II. Material & Methods

There are several steps involved in developing a multivariable Fourier Series semiparametric regression estimator for categorical data. First, a Fourier Series semiparametric regression model is constructed, incorporating both parametric and nonparametric regression curve functions. After that, each model parameter's log-likelihood function is calculated and derived. Finally, carries out numerical iterations using the Newton–Raphson method.

Probability Distribution

Given $x_{1i}, \ , x_{2i}, \ \ldots, \ x_{pi}$ and $z_{1i}, z_{2i}, \ \ldots \ , z_{qi}$; $i = 1, 2, \ldots, n$, are as many as the number of parametric predictor variables and nonparametric predictor variables q. Additionally, the variable Y_i has a probability distribution of and is a random Bernoulli distribution variable.

$$Y_i \sim B\big(1, \pi(\boldsymbol{x_i}, \boldsymbol{z_i})\big),$$

$$\pi(\boldsymbol{x_i}, \boldsymbol{z_i}) = \pi\big(x_{1i}, \ x_{2i}, \ \ldots \ , \ x_{pi}, z_{1i}, \ z_{2i}, \ \ldots \ , \ z_{qi}\big),$$

$$i = 1,2, \ldots, n$$

where the success probability

$$P(Y_i = 1) = \pi(\boldsymbol{x_i}, \boldsymbol{z_i})$$

and the unsuccessful probability

$$P(Y_i = 0) = 1 - \pi(\boldsymbol{x_i}, \boldsymbol{z_i})$$

Where $\pi(\boldsymbol{x_i}, \boldsymbol{z_i})$ is described as follows in the probability distribution function $P(Y_i = y_i)$.

$$P(Y_i = y_i) = \pi(\boldsymbol{x_i}, \boldsymbol{z_i})^{y_i}\big(1 - \pi(\boldsymbol{x_i}, \boldsymbol{z_i})\big)^{1-y_i} = \left(\frac{\pi(\boldsymbol{x_i}, \boldsymbol{z_i})}{1-\pi(\boldsymbol{x_i}, \boldsymbol{z_i})}\right)^{y_i}\big(1 - \pi(\boldsymbol{x_i}, \boldsymbol{z_i})\big) \quad (1)$$

Logit Function (Link Function)

Then, a natural logarithmic function (ln) can be used to express equation (1).

$$\ln P(Y_i = y_i) = y_i \ln\left(\frac{\pi(\boldsymbol{x_i}, \boldsymbol{z_i})}{1-\pi(\boldsymbol{x_i}, \boldsymbol{z_i})}\right) + \ln(1 - \pi(\boldsymbol{x_i}, \boldsymbol{z_i})) \quad (2)$$

The ln function (2) creates the following exponential family distribution function when expressed in exponential form.

$$\exp(\ln P(Y_i = y_i)) = \exp(y_i \ln\left(\frac{\pi(\boldsymbol{x_i}, \boldsymbol{z_i})}{1 - \pi(\boldsymbol{x_i}, \boldsymbol{z_i})}\right) + \ln(1 - \pi(\boldsymbol{x_i}, \boldsymbol{z_i}))) \quad (3)$$

Equation (3) provides the definition of the exponential family distribution function.

$$f(y_i, \theta) = \exp\left(\frac{y_i\,\theta - b(\theta)}{a(\emptyset)} + c(\theta, \emptyset)\right) \quad (4)$$

As a result, its probability distribution function is a member of the exponential distribution function family.

$$P(Y_i = y_i) = \exp\left(\frac{y_i \ln\left(\frac{\pi(\boldsymbol{x_i}, \boldsymbol{z_i})}{1-\pi(\boldsymbol{x_i}, \boldsymbol{z_i})}\right)}{1} + \ln(1 - \pi(\boldsymbol{x_i}, \boldsymbol{z_i}))\right) \quad (5)$$

where,

$$\theta = \ln\left(\frac{\pi(\boldsymbol{x_i}, \boldsymbol{z_i})}{1-\pi(\boldsymbol{x_i}, \boldsymbol{z_i})}\right) \qquad a(\emptyset) = 1$$

$$b(\theta) = \ln(1 - \pi(\boldsymbol{x_i}, \boldsymbol{z_i})) \qquad c(\theta, \emptyset) = 0$$

Since θ in function (5) is a logit function, the regression's logit function is

$$\theta = \ln\left(\frac{\pi(\boldsymbol{x_i}, \boldsymbol{z_i})}{1-\pi(\boldsymbol{x_i}, \boldsymbol{z_i})}\right) \quad (6)$$

As a link function, the logit function makes parameter estimation easier and streamlines intricate regression models. A logit transformation is used to do this.

Logit Transformation Model

The following is the definition of the logit transformation model.

$$\ln\left(\frac{\pi(\boldsymbol{x_i}, \boldsymbol{z_i})}{1-\pi(\boldsymbol{x_i}, \boldsymbol{z_i})}\right) = f\big(x_{1i}, \ \ldots \ , \ x_{pi}\big) + g\big(z_{1i}, \ \ldots \ , \ z_{pi}\big) \quad (7)$$

where f and g are an additive model-following regression equation, regression function, or regression curve. Based on equation (7), $f(x_{1i}, \ \ldots \ , \ x_{pi})$ is approximated by a parametric function using Linear model and $g(z_{1i}, \ \ldots \ , \ z_{pi})$ is approximated by a nonparametric function using Fourier Series model. Consequently, the logit transformation model looks like this.

$$\ln\left(\frac{\pi(\boldsymbol{x_i}, \boldsymbol{z_i})}{1-\pi(\boldsymbol{x_i}, \boldsymbol{z_i})}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ji} + \sum_{l=1}^{q}\Big(b_l z_{li} + \frac{1}{2}a_{0l} + \sum_{k=1}^{K} a_{kl} \cos k z_{li}\Big); \ i = 1,2, \ldots n \quad (8)$$

The Fourier Series semiparametric regression model for categorical data is produced as follows by the function (8).

$$\pi(\boldsymbol{x_i}) = \frac{e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_{ji} + \sum_{l=1}^{q}\left(b_l z_{li} + \frac{1}{2}a_{0l} + \sum_{k=1}^{K} a_{kl} \cos k z_{li}\right)}}{1 + e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_{ji} + \sum_{l=1}^{q}\left(b_l z_{li} + \frac{1}{2}a_{0l} + \sum_{k=1}^{K} a_{kl} \cos k z_{li}\right)}} \quad (9)$$

where, $i = 1,2, \ldots, n$. β_0 and β_j , $j = 1,2, \ldots, p$ are the parametric function's model parameters. Meanwhile, b_l, a_{0l} and a_{kl} , $l = 1,2, \ldots, q$, $k = 1,2, \ldots, K$ are the nonparametric function's model parameters.

Likelihood Function $l(\boldsymbol{\theta})$

By applying the Maximum Likelihood Estimation (MLE) approach, the form of the likelihood function is obtained as $l(\boldsymbol{\theta})$.

where,

$$\boldsymbol{\theta} = \begin{pmatrix} \beta_0 & \beta_1 & \cdots & \beta_p & \cdots & b_q & a_{0\mathrm{q}} & a_{Kq} \end{pmatrix}$$

so,

$$l(\boldsymbol{\theta}) = \prod_{i=1}^{n} P(Y_i = y_i) = \pi(\boldsymbol{x_i}, \boldsymbol{z_i})^{\sum_{i=1}^{n} y_i}(1 - \pi(\boldsymbol{x_i}, \boldsymbol{z_i}))^{n - \sum_{i=1}^{n} y_i} \quad (10)$$

By maximizing the log likelihood function's first derivative, the MLE approach can be used to estimate parameters in logistic regression. It is simple to maximize the likelihood function (10) as follows $\ln l(\boldsymbol{\theta})$.

Log-Likelihood Function $L(\boldsymbol{\theta})$

$$\ln\left[l(\boldsymbol{\theta})\right] = L(\boldsymbol{\theta}) = \sum_{i=1}^{n} y_i \ln[\pi(\boldsymbol{x_i}, \boldsymbol{z_i})] + \sum_{i=1}^{n} (1 - y_i) \ln[1 - \pi(\boldsymbol{x_i}, \boldsymbol{z_i})]$$

$$= \sum_{i=1}^{n}\{y_i\Big(f\big(x_{1i}, \ \ldots \ , \ x_{pi}\big) + g\big(z_{1i}, \ \ldots \ , \ z_{qi}\big)\Big) + - \ln\big[1 + \exp(f\big(x_{1i}, \ \ldots \ , \ x_{pi}\big) + g\big(z_{1i}, \ \ldots \ , \ z_{qi}\big))\big]\} \quad (11)$$

By partially deriving the function (11) in relation to $\beta_0, \beta_j, b_l, a_{0l}$, and a_{kl}, then equating to 0, the estimator $\widehat{\boldsymbol{\theta}}$ is produced.

$$\frac{\partial L(\boldsymbol{\theta})}{\partial \beta_0} = 0$$

$$\frac{\partial L(\boldsymbol{\theta})}{\partial \beta_j} = 0 \; ; \; j = 1,2, \dots, p$$

$$\frac{\partial L(\boldsymbol{\theta})}{\partial b_l} = 0 \; ; \; l = 1,2, \dots, q$$

$$\frac{\partial L(\boldsymbol{\theta})}{\partial a_{0l}} = 0 \; ; \; l = 1,2, \dots, q$$

$$\frac{\partial L(\boldsymbol{\theta})}{\partial a_{kl}} = 0 \; ; \; k = 1,2, \dots, K \; ; \; l = 1,2, \dots, q$$

Equation (12) will be used to determine the estimator $\widehat{\beta_0}$.

$$\sum_{i=1}^{n} \{ (y_i - \pi(\boldsymbol{x_i}, \boldsymbol{z_i})) \} = 0 \quad (12)$$

Equation (13) will be used to determine the estimator $\widehat{\boldsymbol{\beta}}$.

$$\sum_{i=1}^{n} \{ (y_i - \pi(\boldsymbol{x_i}, \boldsymbol{z_i})) x_{ji} \} = 0 \quad (13)$$

Equation (14) will be used to determine the estimator $\widehat{\boldsymbol{b}}$.

$$\sum_{i=1}^{n} \{ (y_i - \pi(\boldsymbol{x_i}, \boldsymbol{z_i})) z_{li} \} = 0 \quad (14)$$

Equation (15) will be used to determine the estimator $\widehat{\boldsymbol{a_0}}$.

$$\sum_{i=1}^{n} \left\{ \frac{1}{2} (y_i - \pi(\boldsymbol{x_i}, \boldsymbol{z_i})) \right\} = 0 \quad (15)$$

Equation (16) will be used to determine the estimator $\widehat{\boldsymbol{a_k}}$.

$$\sum_{i=1}^{n} \{ \sum_{k=1}^{K} \cos k z_{li} \, (y_i - \pi(\boldsymbol{x_i}, \boldsymbol{z_i})) \} = 0 \quad (16)$$

Newton-Raphson Iteration

The results of the derivative of $L(\boldsymbol{\theta})$ (11) against $\beta_0, \beta_j, b_l, a_{0l}$, and a_{kl} not in closed form, which were made in the implicit equation, hence the Newton–Raphson method of numerical iteration must be used.

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \left(H(\boldsymbol{\theta})^{(t)} \right)^{-1} g(\boldsymbol{\theta})^{(t)} \quad (17)$$

where $\boldsymbol{\theta}^{(t)}$ is the $\boldsymbol{\theta}$ of the t-th iteration, *t=1,2,...*, converged.

$$\boldsymbol{\theta}^{(t)} = \left({\beta_0}^{(t)} \quad {\beta_1}^{(t)} \quad \cdots \quad {\beta_p}^{(t)} \quad \cdots \quad {b_{\mathrm{q}}}^{(t)} \quad {a_{0\mathrm{q}}}^{(t)} \quad {a_{\mathrm{Kq}}}^{(t)} \right)$$

while $g(\boldsymbol{\theta})$ is the gradient vector of $\boldsymbol{\theta}$ in function (18) and $H(\boldsymbol{\theta})$ is the Hessian matrix of $\boldsymbol{\theta}$ in function (19), with the following equation.

$$g(\boldsymbol{\theta}) = \left(\frac{\partial L(\boldsymbol{\theta})}{\partial \beta_0}, \frac{\partial L(\boldsymbol{\theta})}{\partial \beta_1}, \cdots, \frac{\partial L(\boldsymbol{\theta})}{\partial \beta_p}, \cdots, \frac{\partial L(\boldsymbol{\theta})}{\partial b_q}, \frac{\partial L(\boldsymbol{\theta})}{\partial a_{0q}}, \frac{\partial L(\boldsymbol{\theta})}{\partial a_{Kq}} \right)^T \quad (18)$$

$$H(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial^2 L(\boldsymbol{\theta})}{\partial {\beta_0}^2} & \frac{\partial^2 L(\boldsymbol{\theta})}{\partial \beta_0 \partial \beta_1} & \cdots & \frac{\partial^2 L(\boldsymbol{\theta})}{\partial \beta_0 \partial a_{Kq}} \\ \frac{\partial^2 L(\boldsymbol{\theta})}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2 L(\boldsymbol{\theta})}{\partial {\beta_1}^2} & \cdots & \frac{\partial^2 L(\boldsymbol{\theta})}{\partial \beta_1 \partial a_{Kq}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 L(\boldsymbol{\theta})}{\partial a_{Kq} \partial \beta_0} & \frac{\partial^2 L(\boldsymbol{\theta})}{\partial a_{Kq} \partial \beta_1} & \cdots & \frac{\partial^2 L(\boldsymbol{\theta})}{\partial {a_{Kq}}^2} \end{bmatrix} \quad (19)$$

The first derivative of function $L(\boldsymbol{\theta})$ with respect to $\beta_0, \beta_j, b_l, a_{0l}$, and a_{kl}, yields the elements of vector $g(\boldsymbol{\theta})$ (18), whereas the second derivative of function $L(\boldsymbol{\theta})$ with respect to $\beta_0, \beta_u, b_v, a_{0v}$, and a_{kv} yields the elements of matrix $H(\boldsymbol{\theta})$ (19).

The $L(\boldsymbol{\theta})$ Function's Second Derivative in Relation to β_0

$$\frac{\partial^2 L(\boldsymbol{\theta})}{\partial {\beta_0}^2} = -\sum_{i=1}^{n} \pi(\boldsymbol{x_i}, \boldsymbol{z_i}) \left(1 - \pi(\boldsymbol{x_i}, \boldsymbol{z_i}) \right) \quad (20)$$

The $L(\boldsymbol{\theta})$ Function's Second Derivative in Relation to β_u

$$\frac{\partial^2 L(\boldsymbol{\theta})}{\partial \beta_u \partial \beta_j} = -\sum_{i=1}^{n} x_{ji} x_{ui} \, \pi(\boldsymbol{x_i}, \boldsymbol{z_i}) \left(1 - \pi(\boldsymbol{x_i}, \boldsymbol{z_i}) \right) \quad (21)$$

The $L(\boldsymbol{\theta})$ Function's Second Derivative in Relation to b_u

$$\frac{\partial^2 L(\boldsymbol{\theta})}{\partial b_v \partial b_l} = -\sum_{i=1}^{n} x_{li} x_{vi} \, \pi(\boldsymbol{x_i}, \boldsymbol{z_i}) \left(1 - \pi(\boldsymbol{x_i}, \boldsymbol{z_i}) \right) \quad (22)$$

The second derivative of the parameter combinations is determined in the same way as in (22).

$$\frac{\partial^2 L(\boldsymbol{\theta})}{\partial a_{kv} \partial b_l} = -\sum_{i=1}^{n} \sum_{k=1}^{K} \pi(\boldsymbol{x_i}, \boldsymbol{z_i}) \left(1 - \pi(\boldsymbol{x_i}, \boldsymbol{z_i}) \right) x_{li} \cos k x_{vi} \quad (23)$$

The $L(\boldsymbol{\theta})$ Function's Second Derivative in Relation to a_{0u}

$$\frac{\partial^2 L(\boldsymbol{\theta})}{\partial a_{0v} \partial a_{0l}} = -\frac{1}{4} \sum_{i=1}^{n} \pi(\boldsymbol{x_i}, \boldsymbol{z_i}) \left(1 - \pi(\boldsymbol{x_i}, \boldsymbol{z_i}) \right) \quad (24)$$

The second derivative of the parameter combinations can be found in the same way (24).

$$\frac{\partial^2 L(\boldsymbol{\theta})}{\partial a_{kv} \partial a_{0l}} = -\frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{K} \pi(\boldsymbol{x_i}, \boldsymbol{z_i}) \left(1 - \pi(\boldsymbol{x_i}, \boldsymbol{z_i}) \right) \cos k x_{vi} \quad (25)$$

The $L(\boldsymbol{\theta})$ Function's Second Derivative in Relation to a_{ku}

$$\frac{\partial^2 L(\boldsymbol{\theta})}{\partial a_{kv} \partial a_{kl}} = -\sum_{i=1}^{n} \sum_{k=1}^{K} \cos k x_{li} \sum_{k=1}^{K} \cos k x_{vi} \, \pi(\boldsymbol{x_i}, \boldsymbol{z_i}) \left(1 - \pi(\boldsymbol{x_i}, \boldsymbol{z_i}) \right) \quad (26)$$

The second derivative of the parameter combinations is derived in the same way as in (26).

$$\frac{\partial^2 L(\boldsymbol{\theta})}{\partial a_{0v} \partial a_{kl}} = -\frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{K} \pi(\boldsymbol{x_i}, \boldsymbol{z_i}) \left(1 - \pi(\boldsymbol{x_i}, \boldsymbol{z_i}) \right) \cos k x_{li} \quad (27)$$

Estimator $\widehat{\boldsymbol{\theta}}$

$\widehat{\boldsymbol{\theta}}$ will be derived from the Newton-Raphson iteration equation when

$$\left| \boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)} \right| < \varepsilon, \; \varepsilon = 0{,}000001 \quad (28)$$

Thus, the estimator $\widehat{\boldsymbol{\theta}}$ obtained when (28) is

$$\widehat{\boldsymbol{\theta}} = \left(\hat{\beta}_0 \quad \hat{\beta}_1 \quad \cdots \quad \hat{\beta}_p \quad \cdots \quad \hat{b}_{\mathrm{q}} \quad \hat{a}_{0\mathrm{q}} \quad \hat{a}_{\mathrm{Kq}} \right)$$

The Fourier Series semiparametric regression model for categorical data can be expressed using the estimator $\widehat{\boldsymbol{\theta}}$ result:

$$\hat{\pi}(\boldsymbol{x_i}, \boldsymbol{z_i}) = \frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_p x_{pi} + \cdots + \hat{b}_q z_{qi} + \frac{1}{2} \hat{a}_{0q} + \hat{a}_{Kq} \cos K z_{qi} \right)}{1 + \exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_p x_{pi} + \cdots + \hat{b}_q z_{qi} + \frac{1}{2} \hat{a}_{0q} + \hat{a}_{Kq} \cos K z_{qi} \right)} \quad (29)$$

where $i = 1,2, \dots, n$. The nonparametric function's estimate model estimator for the predictor variable x_p is represented by $\hat{\beta}_0$ and $\hat{\beta}_p$, and the estimator model of the nonparametric function for predictor variable z_q is represented by $\hat{b}_q, \hat{a}_{0q}$ and $\hat{a}_{kq}$. In this case, p and q are the number of predictor variables, and k is the number of oscillation parameters.

III. Data Application

We apply the Fourier Series semiparametric regression approach for categorical data using the poverty gap status in East Java Province for the year 2023 as the application dataset. The data were obtained from secondary sources, including dynamic tables from the BPS (Badan Pusat Statistik) website and provincial publications from East Java in figures. The dataset consists of four predictor variables (x) and one response variable (y).

This methodology allows us to explore the relationship between the predictors and the poverty gap status while

accommodating the complexities inherent in categorical data. By leveraging the flexibility of the Fourier Series semiparametric regression model, we aim to provide a more accurate model of the factors influencing poverty in East Java. A detailed description of the variables are provided in Table I.

TABLE I
VARIABLE DESCIPTION

Variable	Description	Unit	Scale
y	Status of Poverty Gap	0 = Low 1 = High	Nominal
x_1	Percentage of Mean Years of Schooling	Percent	Rasio
x_2	GRDP per Capita	Million rupiah	Rasio
x_3	Percentage of Households with Proper Sanitation	Percent	Rasio
x_4	The Open Unemployment Rate	Percent	Rasio

8 district exist in East Java and shown in Table II below.

TABLE II
LIST OF POVERTY GAP IN EAST JAVA 2023 ACCORDING TO NATIONAL AVERAGE

Status	District	Total
Low ($y = 0$)	Pasuruan City, Mojokerto City, Madiun City, Ponorogo, Blitar, Kediri, Malang, Lumajang, Trenggalek, Tulungagung, Jember, Mojokerto, Jombang, Nganjuk, Madiun, Magetan, Pamekasan, Kediri City, Blitar City, Banyuwangi, Pasuruan, Sidoarjo, Surabaya City, Batu City, Malang City, Probolinggo City	26 Districts
High ($y = 1$)	Situbondo, Probolinggo, Pacitan, Bondowoso, Ngawi, Bojonegoro, Tuban, Bangkalan, Sampang, Sumenep, Lamongan, Gresik	12 Districts

Descriptive Analytics

Table III shows the characteristics of the data for each predictor variable as determined by descriptive analysis.

TABLE III
DESCRIPTIVE STATISTICS OF PREDICTOR VARIABLES

Variable	Mean	Variance	Min	Max
x_1	8.38	-0.41	5.07	11.82
x_2	7.06	-0.78	2.38	54.11
x_3	84.59	-2.65	50.30	98.18
x_4	4.66	-0.25	1.71	8.05

The features of the variables, which include the poverty gap status in 38 East Java regencies and cities in 2023, are detailed in Table III. Furthermore, there are no multicollinearity amongst predictor variables and that none of the variables contain missing values. Poverty is a multifaceted issue encompassing education, economic conditions, infrastructure, and labor factors [22]. Poverty and education are closely interrelated. For instance, [23] employed spatial data regression techniques to examine the relationship between education levels and the degree of poverty. Their findings revealed that the poverty depth index is negatively influenced by the average number of years of education. Fig. 1 shows the conceptual predictor variable that was employed in this investigation.

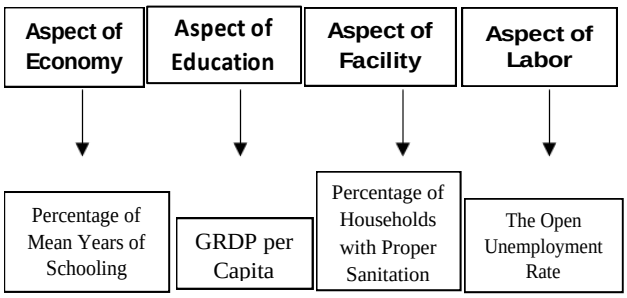

Fig. 1. Conceptual Diagram of Variables

As illustrated in Fig. 1, economic factors also play a significant role in poverty alleviation efforts. This is supported by numerous studies, including one by [24], which found that GDP per capita significantly impacts the poverty depth index.

Furthermore, poverty is linked to infrastructure. [25] utilized robust regression to analyze the factors influencing the poverty depth index. Their results indicated that the proportion of households with access to adequate and sustainable sanitation facilities negatively affects the poverty depth index.

Additionally, poverty is associated with labor factors. [24] applied panel data regression to study the poverty depth index, revealing that the proportion of impoverished individuals working in the informal sector positively impacts the poverty depth index.

Binary Logistic Regression Model

The following is the binary logistic regression model.

$$\pi(\boldsymbol{x_i}) = \frac{\exp(\beta_0+\beta_1 x_{1i}+\cdots+\beta_p x_{pi})}{1+\exp(\beta_0+\beta_1 x_{1i}+\cdots+\beta_p x_{pi})}, i = 1,2,\ldots,n \quad (30)$$

Where p is the number of predictor parametric variable.

Parameter Estimation

The results of parameter estimate in the model for data on the poverty gap condition in East Java in 2023 are based on the binary logistic regression model (30).

$$\hat{\pi}(\boldsymbol{x_i}) = \frac{\exp(8.384-1.0429x_{1i}+0.0354x_{2i}-0.003x_{3i}-0.189x_{4i})}{1+\exp(8.384-1.0429x_{1i}+0.0354x_{2i}-0.003x_{3i}-0.189x_{4i})} \quad (31)$$

Additional information is shown in the Table IV.

TABLE IV
PARAMETER ESTIMATION IN BINARY LOGISTIC REGRESSION MODEL

Parameters	Estimations
β_0	8.3840
β_1	-1.0429
β_2	0.0354
β_3	-0.0030
β_4	-0.1890

Fourier Series Nonparametric Regression Model

The following is the Fourier Series nonparametric regression model.

$$\pi(\boldsymbol{x_i}) = \frac{\exp\left(\sum_{j=1}^{p}\left(b_j x_{ji}+\frac{1}{2}a_{0j}+\sum_{k=1}^{K} a_{kj}\cos kx_{ji}\right)\right)}{1+\exp\left(\sum_{j=1}^{p}\left(b_j x_{ji}+\frac{1}{2}a_{0j}+\sum_{k=1}^{K} a_{kj}\cos kx_{ji}\right)\right)} \quad (32)$$

Where p is the number of predictor variable.

Parameter Estimation

Using the Fourier Series nonparametric regression model (32), the parameter estimation results following nonparametric regression model offers the best oscillation parameter combinations for data on the poverty gap status in East Java in 2023: $x_1 = 1$, $x_2 = 2$, $x_3 = 1$, and $x_4 = 2$.

$$\hat{\pi}(\boldsymbol{x_i}) = \frac{\exp(21.2734+\cdots-1.7470x_{4i}+1.4472\cos x_{4i}+2.8371\cos 2x_{4i})}{1+\exp(21.2734+\cdots-1.7470x_{4i}+1.4472\cos x_{4i}+2.8371\cos 2x_{4i})} \quad (33)$$

Additional information is shown in Table V.

TABLE V
PARAMETER ESTIMATION IN FOURIER SERIES NONPARAMETRIC REGRESSION MODEL

Parameters	Estimations	Parameters	Estimations
a_0	21.2734	b_3	0.0854
b_1	-2.7233	$a_{1,3}$	1.3234
$a_{1,1}$	-1.1289	b_4	-1.7470
b_2	0.1264	$a_{1,4}$	1.4472
$a_{1,2}$	2.3892	$a_{2,4}$	2.8371
$a_{2,2}$	2.6540		

Fourier Series Semiparametric Regression Model

Each predictor variable that was present in multiple groups was plotted against the number of high poverty gaps using a scatterplot that we constructed ($y = 1$) in each group in order to determine the relationship that followed the Fourier Series semiparametric regression model. Fig. 2 displays the scatterplots as follows.

Fig. 2. Multiple data groups' scatterplots compared to the group's substantial poverty gap

Based on Fig. 2, The probability of a high poverty gap ($y = 1$) for variable x_1 has a linear pattern, but for variables x_2, x_3, and x_4 it has a repeating pattern and follows a downward trend line. his suggests that the relationship between these variables and the probability of a high poverty gap is more complex and non-linear compared to x_1. Therefore, a semiparametric approach is needed to capture these more flexible and intricate patterns.

Choosing the Best Oscillation Settings

The lowest AIC value was used to choose the oscillation parameters for the Fourier Series nonparametric regression model. In order to create a model that is not overly complex and produces results that are adequately relevant, the number of oscillation parameters employed in this experiment was limited. The AIC findings, as determined by the R program, are shown in Table VI.

TABLE VI
LOWEST AIC RESULTS FOR EVERY OSCILLATION PARAMETER NUMBER

The Number of Oscillation Parameter	Oscillation Parameter Combinations (K)			AIC (K)
	x_2	x_3	x_4	
K=1	1	1	1	48.8913
K=2	2	1	2	44.0591
K=3	3	3	3	42.8249

The Fourier Series model with the best oscillation parameters, according to Table VI, is the one with the minimum AIC value and a combination of oscillation parameters $x_2 = 3$, $x_3 = 3$, and $x_4 = 3$.

Parameter Estimation

These are the findings of parameter estimate in the model utilizing data on the poverty gap status in East Java in 2023, which is based on the Fourier Series semiparametric regression model (9).

$$\hat{\pi}(\boldsymbol{x_i}, \boldsymbol{z_i}) = \frac{\exp(459.01-118.11x_{1i}+\cdots+109.48\cos 3x_{4i})}{1+\exp(459.01-118.11x_{1i}+\cdots+109.48\cos 3x_{4i})}$$

Additional information is shown in Table VII.

TABLE VII
PARAMETER ESTIMATION IN FOURIER SERIES SEMIPARAMETRIC REGRESSION

Parameters	Estimations	Parameters	Estimations
β_0	459.0154	$a_{1,2}$	31.9289
β_1	-118.1128	$a_{2,2}$	106.6754
b_1	1.1825	$a_{3,2}$	-41.4193
$a_{1,1}$	78.9075	b_3	-84.8413
$a_{2,1}$	31.6283	$a_{1,3}$	69.0279
$a_{3,1}$	-17.5459	$a_{2,3}$	137.2307
b_2	10.7718	$a_{3,3}$	109.4801

Comparing Binary Logistic Regression, Fourier Series Nonparametric Regression, and Fourier Series Semiparametric Regression

In order to assess and compare the performance of the models, it is essential to evaluate how well each model fits the observed data. One of the key statistical measures used for this purpose is the deviance value.

Utilizing Deviance Value to Determine the Best Model

The regression model with the lowest deviation value is selected, as it indicates that the model with the smallest deviance offers the best fit to the observed data compared to alternative models. The results of the deviation statistical test are shown in Table VIII.

TABLE VIII
DEVIANCE VALUES COMPARISON

Methods	Deviance Values
Parametric	34.1024
Nonparametric	23.9881
Semiparametric	14.8250

As shown in Table VIII, the binary logistic regression (34.1024) and Fourier Series nonparametric regression (23.9881) exhibit higher deviance values compared to the Fourier Series semiparametric regression (14.825). Since the Fourier Series semiparametric regression model has the lowest deviance value, it is the most suitable model for analyzing the poverty gap status in East Java in 2023.

Getting the Best Classification Based on Accuracy & Press's Q Value

The best accuracy or the higher Press's Q was shown by the chosen Fourier Series semiparametric regression model. The categorization test yields the following findings, which are shown in Table IX.

TABLE IX
COMPARISON OF ACCURACY AND PRESS'S Q

Methods	Accuracy	Sensitivity	Specificity	AUC	Press's Q value
Parametric	76.32%	76.31%	88.46%	50%	10.526
Nonparametric	86.84%	88.46%	83.33%	85%	20.631
Semiparametric	89.47%	96.15%	75%	86%	23.684

As presented in Table IX, the Fourier Series semiparametric regression achieves an accuracy of 89.47%, which is higher than Fourier Series nonparametric regression (86.84%) and binary logistic regression (76.32%). A comparison of the accuracy values for the three methods across different thresholds are illustrated in Fig. 3.

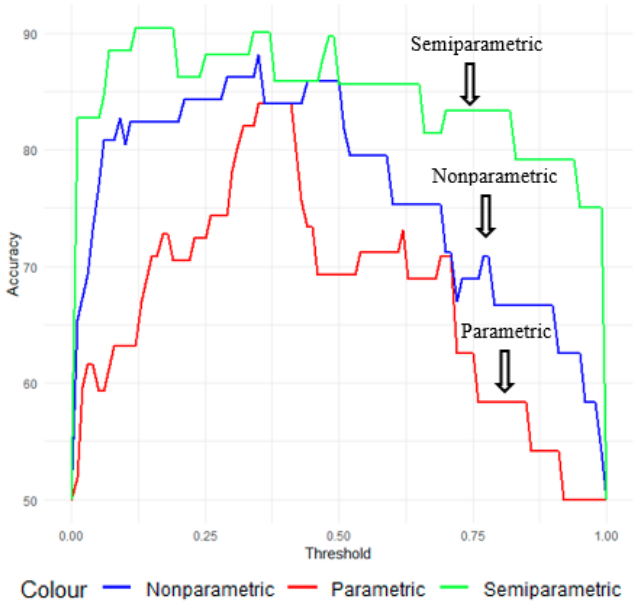

Fig. 3. Comparison of accuracy value in Binary Logistic, Fourier Series Nonparametric and Fourier Series Semiparametric regression

Furthermore, the Fourier Series semiparametric regression model demonstrates effective classification capabilities, as evidenced by its higher Press's Q value (23.6842), which exceeds the critical Chi-Square value. This indicates a greater probability of rejecting the null hypothesis (H0). A comparison of the estimation results are presented in Fig. 4.

In addition to its superior classification performance, the model also highlights the robustness of the Fourier Series approach in capturing non-linear relationships within the data. By utilizing both the flexibility of nonparametric regression and the structure of parametric models, the semiparametric model is able to provide more reliable predictions with less overfitting. These characteristics make the model particularly valuable for analyzing complex, real-world data like poverty status, where the interactions between predictor variables may not follow a simple linear pattern.

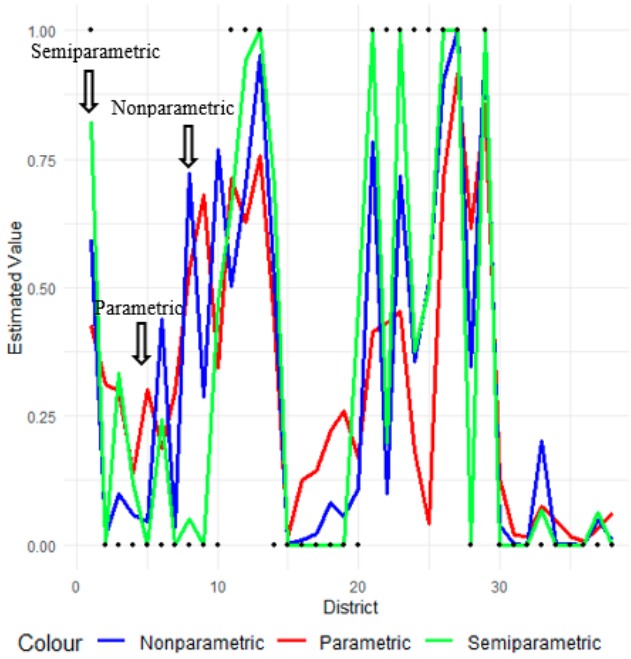

Fig. 4. Comparison of predicted value in Binary Logistic, Fourier Series Nonparametric and Fourier Series Semiparametric regression

The plots in Fig. 4 show that the predicted outcomes of the three methods vary, and no single approach consistently outperforms the others. However, in the specific case examined in this study, the Fourier Series semiparametric regression generally outperforms compared to others. The Fourier Series semiparametric regression is the optimal choice, as its probability estimates closely match the actual values in nearly all selected districts.

IV. CONCLUSION

This study's main goal was to provide a new semiparametric estimator for categorical data. Additionally, we introduced the MLE method for parameter estimation. To achieve superior estimation results, we compared the semiparametric approach with the other methods. This process was further refined to select the optimal model based on the previously proposed framework.

The optimal model for both, the case study and simulation study was selected using the Akaike Information Criterion (AIC). To identify the optimal model, various combinations of smoothing conditions were tested. One of the case study's

main conclusions is that combining ($x_2 = 3, x_3 = 3, x_4 = 3$) for every predictor variable results in the best model.

Compared to the other two models, the Fourier Series semiparametric regression model demonstrates the smallest deviance value, the highest accuracy, and the largest Press's Q value. These findings validate the hypothesis that the semiparametric estimator outperforms individual estimators when estimating the poverty gap status in East Java in 2023.

APPENDIX

TABLE RESULTS OF AIC COMPARISONS FOR EVERY OSCILLATION PARAMETER COMBINATIONS IN SEMIPARAMETRIC REGRESSION MODEL

Oscillation Parameter Combinations (K)			AIC (K)
x_2	x_3	x_4	
1	1	1	48.8913
2	1	1	48.3442
3	1	1	50.1357
1	2	1	48.1021
2	2	1	47.8624
3	2	1	49.4003
1	3	1	46.3338
2	3	1	47.1193
3	3	1	48.5888
1	1	2	46.7705
2	1	2	44.0591
3	1	2	45.1193
1	2	2	47.4183
2	2	2	44.9095
3	2	2	45.2476
1	3	2	45.2757
2	3	2	44.9268
3	3	2	45.0191
1	1	3	47.9156
2	1	3	45.8831
3	1	3	46.9094
1	2	3	48.4006
2	2	3	46.4109
3	2	3	46.6957
1	3	3	43.2803
2	3	3	43.5334
3	3	3	**42.8250**

REFERENCES

[1] M. Bilodeau, "Fourier Smoother and Additive Models," *The Canadian of Statistic*, vol. 3, pp. 257–259, 1992.

[2] H. Okumura and K. Naito, "Non-parametric Kernel Regression for Multinomial Data," *Journal of Multivariate Analysis*, vol. 97, pp. 2009–2022, 2006.

[3] R. L. Eubank, *Spline Smoothing and Nonparametric Regression*. New York: Marcel Dekker, 1998.

[4] L. Li, *Nonlinear Wavelet-Based Nonparametric Curve Estimation with Censored Data and Inference on Long Memory Processes*. ProQuest Information and Learning Company, 2002.

[5] L. Sua and A. Ullah, "Local Polynomial Estimation of Nonparametric Simultaneous Equations Models," *Journal of Econometrics*, vol. 144, pp. 193–218, 2008.

[6] A. H. Welsh and T. W. Yee, "Local Regression for Vector Responses," *Journal of Statistical Planning and Inference*, vol. 136, pp. 3007–3031, 2006.

[7] A. Antoniadis, J. Bigot, and T. Spatinas, "Wavelet Estimators in Nonparametric Regression: A Comparative Simulation Study," *Journal of Statistical Software*, vol. 6, pp. 1–83, 2001.

[8] A. Tripena and I. N. Budiantara, "Fourier Estimator in Nonparametric Regression," in *Proc. Int. Conf. On Natural Sciences and Applied Natural Sciences*, pp. 2–4, Yogyakarta, 2007.

[9] U. Amato, A. Antoniadis, and I. De Feis, "Fourier Series Approximation of Separable Models," *Journal of Computational and Applied Mathematics*, vol. 146, pp. 459–479, 2002.

[10] J. Silverberg and J. Morton, "Fourier Series of Half-range Functions by Smooth Extension," *Journal of Applied Mathematical Modelling*, vol. 33, pp. 812–821, 2007.

[11] D. De Canditiis and I. De Feis, "Pointwise Convergence of Fourier Regularization for Smoothing Data," *Journal of Computational and Applied Mathematics*, vol. 196, pp. 540–552, 2004.

[12] L. J. Asrini and I. N. Budiantara, "Fourier Series Semiparametric Regression Models (Case Study: The Production of Lowland Rice Irrigation in Central Java)," *ARPN Journal of Engineering and Applied Sciences*, vol. 9, pp. 1501–1506, 2014.

[13] I. N. Budiantara, V. Ratnasari, I. Zain, M. Ratna, and M. F. F. Mardianto, "Modeling of HDI and PQLI in East Java (Indonesia) using Biresponse Semiparametric Regression with Fourier Series Approach," *Asian Transactions on Basic and Applied Sciences Journal*, vol. 5, pp. 21–28, 2015.

[14] I. W. Sudiarsa, I. N. Budiantara, and S. W. Purnami, "Combined Estimator Fourier Series and Spline Truncated in Multivariable Nonparametric Regression," *Applied Mathematics and Sciences*, vol. 9, pp. 4997–5010, 2015.

[15] V. Ratnasari, I. N. Budiantara, I. Zain, M. Ratna, and N. P. A. M. Mariati, "Comparison Truncated Spline and Fourier Series in Multivariable Nonparametric Regression Models (Application: Data of Poverty in Papua, Indonesia)," *International Journal of Basic and Applied Sciences*, vol. 15, pp. 9–12, 2015.

[16] I. N. Budiantara, V. Ratnasari, M. Ratna, W. Wibowo, N. Afifah, D. P. Rahmawati, and M. A. D. Octavanny, "Modeling Percentage of Poor People in Indonesia Using Kernel and Fourier Series Mixed Estimator in Nonparametric Regression," *Investigacion Operacional*, vol. 40, pp. 538–551, 2019.

[17] K. Nisa, I. N. Budiantara, and A. T. Rumiati, "Multivariable Semiparametric Regression Model with Combined Estimator of Fourier Series and Kernel," in *IOP Conf. Series: Earth and Environmental Science*, pp. 012028, IOP Publishing, 2017.

[18] L. Laome, I. N. Budiantara, and V. Ratnasari, "Poverty Modelling with Spline Truncated, Fourier Series, and Mixed Estimator Geographically Weighted Nonparametric Regression," in *AIP Conf. Proc.*, AIP Publishing, 2024.

[19] S. Suliyanto, M. Rifada, and E. Tjahjono, "Estimation of Nonparametric Binary Logistic Regression Model with Local Likelihood Logit Estimation Method (Case Study of Diabetes Mellitus Patients at Surabaya Hajj General Hospital)," in *Symposium on Biomathematics 2019*, pp. 1551–7616, AIP Conference Proceedings, Bali, 2020.

[20] A.S. Suriaslan, I.N. Budiantara, and V. Ratnasari, "Nonparametric Regression Estimation Using Multivariable Truncated Splines for Binary Response Data," *MethodsX*, p. 103084, 2025.

[21] M. Zulfadhli, I.N. Budiantara, and V. Ratnasari, "Nonparametric Regression Estimator of Multivariable Fourier Series for Categorical Data," *MethodsX*, p. 102983, 2024.

[22] K. P. Utama and L. K. Sari, "Analisis Spasial Indeks Kedalaman Kemiskinan Tiga Provinsi di Pulau Jawa Tahun 2021," in *Seminar Nasional Official Statistics*, vol. 2023, no. 1, pp. 353–362, Oct. 2023.

[23] S. Chattopadhyay, A. Majumder, and H. Jaman, "Decomposition of Inter-regional Poverty Gap in India: A Spatial Approach," *Empirical Economics*, vol. 46, no. 1, pp. 65–99, 2013.

[24] N. Fajriyah and S. P. Rahayu, "Pemodelan Faktor-Faktor yang Mempengaruhi Kemiskinan Kabupaten/Kota di Jawa Timur Menggunakan Regresi Data Panel," *Jurnal Sains dan Seni ITS*, vol. 5, no. 1, pp. D45–D50, 2016.

[25] I. K. Wardani, Y. Susanti, S. Subanti, P. S. Statistika, and U. S. Maret, "Pemodelan Indeks Kedalaman Kemiskinan di Indonesia Menggunakan," in *Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST) 2021*, pp. 15–23, 2021.

Muhammad Zulfadhli, born on July 23, 2001, is a postgraduate student at the Department of Statistics, Faculty of Data Science and Analysis, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia. He earned his S.Stat degree at Universitas Negeri Makassar, Makassar, Indonesia and his M.Stat degree at Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia. Nonparametric regression using categorical data and applied statistics are areas of research interest.

I Nyoman Budiantara, born on June 3, 1965, is a Professor at the Department of Statistics, Faculty of Data Science and Analysis, Institut

Teknologi Sepuluh Nopember, Surabaya, Indonesia. He earned his Ph.D from Gadjah Mada University, Yogyakarta, Indonesia. His research focuses on Nonparametric Regression. His research has been published in a number of national journals, international proceedings, and international journals that are indexed by Scopus.

Vita Ratnasari, born on September 10, 1970, is a Professor at the Department of Statistics, Faculty of Data Science and Analysis, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia. She earned her Ph.D from Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia. Her research focuses on Categorical Data Analysis. Her research has been published in a number of national journals, international proceedings, and international journals that are indexed by Scopus.

Afiqah Saffa Suriaslan, born on November 7, 2000, is a postgraduate student at the Department of Statistics, Faculty of Data Science and Analysis, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia. She earned her S.Stat degree at Universitas Negeri Makassar, Makassar, Indonesia. Her focuses on Nonparametric Regression.