FOURIER SERIES APPLICATION FOR MODELING "CHOCOLATE" KEYWORD SEARCH TRENDS IN GOOGLE TRENDS DATA

Andrea Tri Rian Dani[1*], Fachrian Bimantoro Putra[2], Muhammad Aldani Zen[3], Vita Ratnasari[4], Qonita Qurrota A'yun[5]

[1,2,3] Statistics Study Program, Department of Mathematics, Faculty of Mathematics and Natural Sciences, Mulawarman University, Samarinda, Indonesia

[4] Department of Statistics, Faculty of Science and Data Analytics, Sepuluh Nopember Institute of Technology, Surabaya, Indonesia

[5] Mathematics Study Program, Department of Mathematics, Faculty of Mathematics and Natural Sciences, Mulawarman University, Samarinda, Indonesia

*e-mail: *andreatririandani@fmipa.unmul.ac.id*

Abstract: In some cases of regression modeling, it is very common to find a repeating pattern. To model this, of course, the approach used must be in accordance with the characteristics of the data. The Fourier series is one of the proposed approaches, because it has advantages in modeling relationship patterns that tend to repeat, such as cosine sine waves. The Fourier series is a subset of nonparametric regression, which has good flexibility in modeling. In this study, the Fourier series approach was applied to model search trend data for the keyword "Chocolate" sourced from Google Trends. Generalized Cross-Validation (GCV) is used as model evaluation criteria. Based on the results of the analysis, the best Fourier series nonparametric regression model is obtained with the number of oscillations of 5, which is indicated by the minimum GCV value.

1. INTRODUCTION

The Fourier Series is one of the most widely used nonparametric regression approaches. The Fourier series was introduced by Bilodeau (1992) as a form of trigonometry polynomial which has such flexibility that it can adjust and estimate the regression curve effectively [1], [2]. The Fourier Series tends to give good estimation results when applied to a regression curve with a repeating pattern [3]–[5].

As the Fourier Series is part of nonparametric regression, it can also be used to model the relationship pattern between the dependent variable and the independent variable, whose relationship pattern is unknown [6]–[8]. This is a distinct feature of nonparametric regression [9]. In nonparametric regression, the data are expected to find their own form of estimation without being influenced by the subjectivity of the research designer [10]–[12].

In the Fourier Series, there is a smoothing parameter generally referred to as the oscillation parameter [1], [13]. This oscillation parameter is used to capture repeated changes in data patterns [14]–[16]. It becomes crucial to determine the optimal oscillation parameters [17]. If the oscillation used is too high, it will impact the number of parameters in the regression model that need to be estimated. Of course, the regression model that is formed is not parsimony [14], [18]. However, if the oscillation used is too low, then the regression curve estimation with the Fourier series will not be able to capture the local nature of the data.

Several studies that examine and develop the Fourier series approach include Dani and Adrianingsih (2020) regarding Nonparametric Regression Modeling with Truncated Spline Estimators and Fourier Series [3]. Octavanny et al. (2021) modeled children ever born in Indonesia using Fourier series nonparametric regression [5]. Furthermore, Dani et al. (2022) reviewed simulation studies and applications of the Fourier Series estimator in nonparametric regression modeling [19].

This study will apply the Fourier Series approach to Google Trends data. Google as a search engine for various information that is very popular and frequently used, of course, can be used to see trends in society. Google Trends is one of the official pages from Google that captures events in the community based on search keywords.

The search keyword that will be examined in this article is "Chocolate". Chocolate is a processed food or drink that comes from cocoa beans, and can be used as a food ingredient [20]. Chocolate is usually given as gifts or parcels or expressions on days of celebration. Diksa shows that there is the existence of the keyword "Chocolate" in Google Trends searches at certain times and certain seasons [21]. Using Google Trends, [20] also show predictions of bread production from chocolate using the Double Exponential Smoothing method. This research aims to model the search trend for the keyword "Chocolate" because it is exciting and up-to-date since the data patterns formed tend to show repeating patterns. This study selects optimal oscillations using Generalized Cross-Validation (GCV).

## 2. LITERATURE REVIEW
2.1. Nonparametric Regression

Nonparametric regression is used when the form of the relationship between the dependent variable and the independent variable is unknown [7], [22]–[24]. Suppose there are paired data (x_i, y_i), where $i = 1, 2, \ldots, n$, then the nonparametric regression model is generally presented in Equation (1) as follows:

$$y_i = f(x_i) + \varepsilon_i \tag{1}$$

where y_i is the response variable, x_i is the predictor variable, ε_i is the error which is assumed to be identical, independent, and normally distributed with zero mean and variance σ^2, while $f(x_i)$ is a regression function whose pattern shape is unknown [25], [26]. In this study, $f(x_i)$ is a function of the Fourier Series.

2.2. Fourier Series

The Fourier Series is a form of a trigonometric polynomial with such flexibility that it can adjust and estimate the regression curve effectively [27], [28]. The Fourier Series is highly dependent on precisely determining the oscillation parameters [5], [29], [30]. An illustration of the repeating pattern and oscillation parameters is shown in Figure 1.
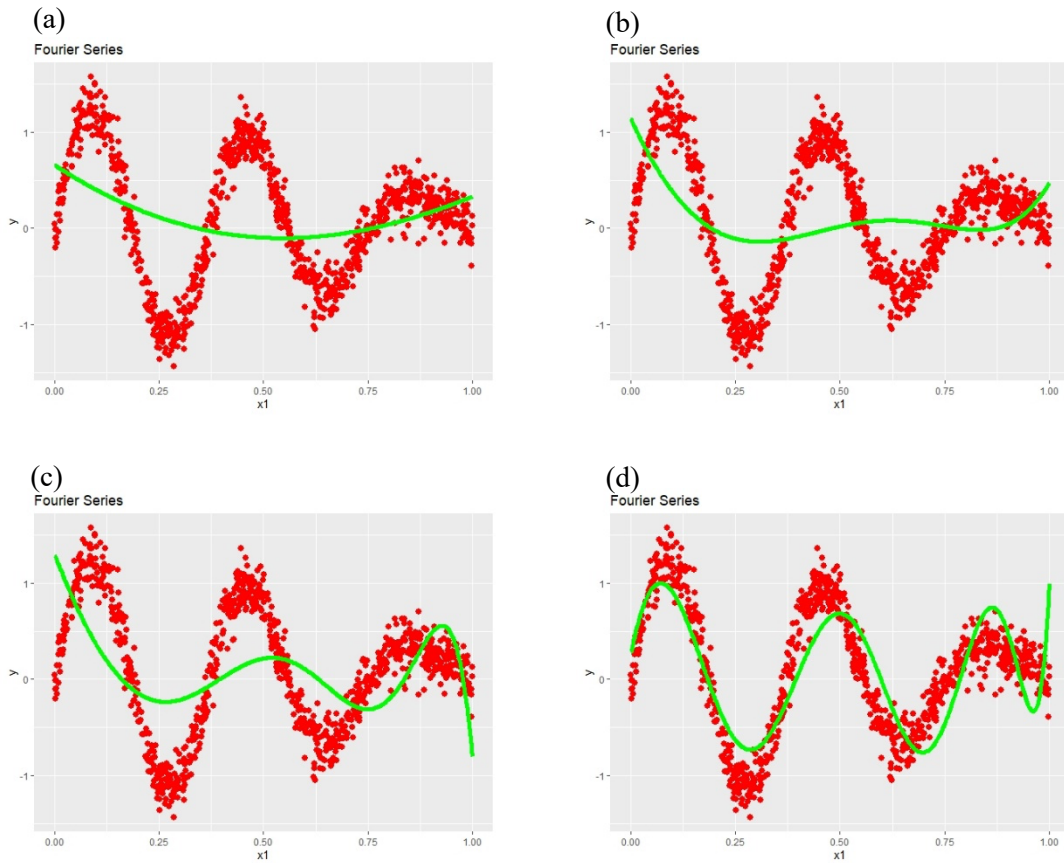
Fig 1. Illustration of Oscillation (a) *K*=1, (b) *K*=3, (c) *K*=5, (d) *K*=7

Suppose given n independent observations, namely data pairs (x_i, y_i) where $i = 1, 2, \ldots, n$. The relationship pattern between x_i and y_i s as in Equation (1). Based on Equation (1), the regression curve of $f(x_i)$ can be estimated using the Fourier Series estimator with the cosine component as follows:

$$f(x_i) = \frac{1}{2}\alpha_0 + \gamma x_i + \sum_{k=1}^{K} \alpha_k \cos(k\, x_i) \tag{2}$$

If Equation (2) is substituted in Equation (1), then

$$y_i = \frac{1}{2}\alpha_0 + \gamma x_i + \sum_{k=1}^{K} \alpha_k \cos(k\, x_i) + \varepsilon_i \tag{3}$$

Equation (3) if implemented for $i = 1, 2, \ldots, n$, then we obtain

$$y_1 = \frac{1}{2}\alpha_0 + \gamma x_1 + \sum_{k=1}^{K} \alpha_1 \cos(k\, x_1) + \varepsilon_1$$

$$y_2 = \frac{1}{2}\alpha_0 + \gamma x_2 + \sum_{k=1}^{K} \alpha_1 \cos(k\, x_2) + \varepsilon_2$$

$$\vdots$$

$$y_n = \frac{1}{2}\alpha_0 + \gamma x_n + \sum_{k=1}^{K} \alpha_1 \cos(k\, x_n) + \varepsilon_n$$

in such a way that, based on Equation 3, we can write it in matrix form as in Equation (4) as follows

$$y = X\varphi + \varepsilon \tag{4}$$

Where:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad ; X = \begin{bmatrix} \frac{1}{2} & x_1 & \cos x_1 & \dots & \cos K\, x_1 \\ \frac{1}{2} & x_2 & \cos x_2 & \dots & \cos K\, x_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{2} & x_n & \cos x_n & \dots & \cos K\, x_n \end{bmatrix} \quad ; \varphi = \begin{bmatrix} \alpha_0 \\ \gamma \\ \alpha_1 \\ \vdots \\ \alpha_K \end{bmatrix} \quad ; \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Using the Maximum Likelihood Estimation (MLE) [31], an estimate for the parameter $\widehat{\varphi}$ is obtained, namely:

$$\widehat{\varphi} = (X^T X)^{-1} X^T y. \tag{5}$$

2.3. Generalized Cross-Validation (GCV)

The GCV method has several advantages compared to other methods, such as Cross Validation (CV) and the Unbiased Risk (UBR) method. Theoretically, the GCV method has asymptotically optimal properties [32]. The formula does not contain the variance σ^2, and invariance of the transformation [12]. The GCV function for selecting optimal oscillation parameters can be shown in Equation (6).

$$GCV\ (K) = \frac{MSE\ (K)}{\left(\frac{1}{n} trace\ (I - Q(K)\right)^2} \tag{6}$$

Where:

$$MSE\ (K) = \frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{n}$$

$$Q(K) = X(X^T X)^{-1} X^T$$

3. METHODOLOGY
3.1. Data and Source Data

The data used in this study is secondary data in the form of weekly data series starting from January 2019 to January 2023. The data is collected by the Google search engine via the Google Trends page. The research variables are divided into two, namely, the independent variable and the dependent variable.

Table 1. Research Variables

Notation	Variables
y	Trending searches for the keyword "chocolate"
x_1	Time, $t = 1, 2, \dots, n$
x_2	y_{t-1}, the $(t-1)$-th lag

The independent variables used in this study are time, $t = 1,2, \dots, n$ and first lag. Using the first lag (y_{t-1}) based on the identification results on the PACF graph where there is a cut off at lag 1.

3.2. Steps of Analysis

The stages of modeling using the Fourier series approach are detailed as follows:

1. Exploring data by creating time series graphs.
2. Make a scatter diagram on each of the dependent and independent variables.
3. Modeling uses nonparametric regression with the Fourier series approach. The oscillation parameters tested are limited in this study, namely $K = 1,2,3,4,5$.
4. Selection of the optimal oscillation based on the minimum GCV value.
5. Visualization of actual data and predictions from the best models.

4. RESULTS AND DISCUSSION

In this section, we will describe the analysis results and discussion of modeling the search trend for the keyword "Chocolate" using the Fourier series approach.

4.1. Data Exploration

Exploration of time series data can be used to see patterns of data changes. In general, the exploration of time series data is displayed using the time series graph presented in Figure 2.
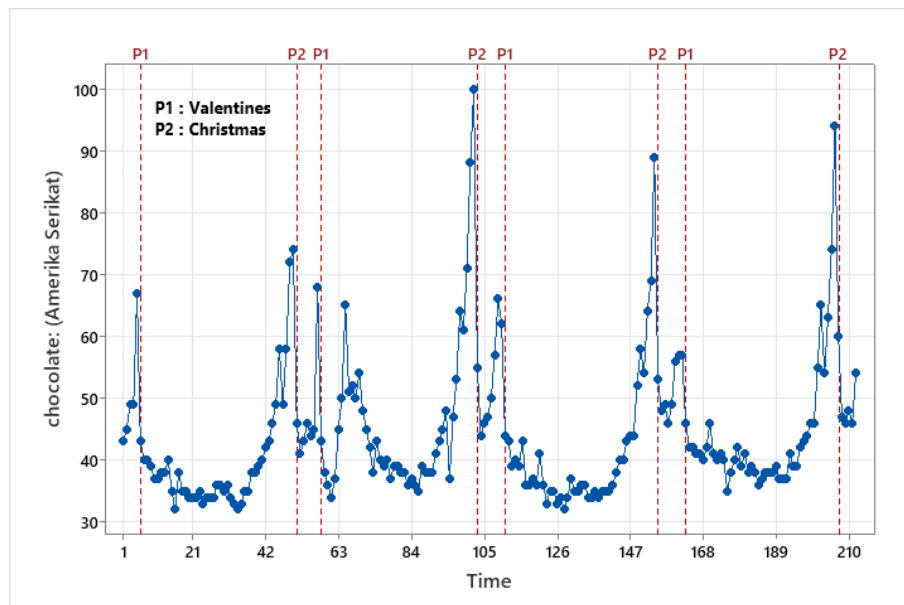
Fig 2. Time series graph of search trend data for the keyword "Chocolate"

Noted:

P1: Indexed denote valentine day

P2: Indexed denote Christmas & New Year

The time series graph displayed starts from January 2019 to January 2023 for each week. Based on Figure 2, it can be seen that the search trend for the keyword "Chocolate" shows a recurring pattern. This repeating pattern is fathomed to be caused by Valentine's Day in February and Christmas New Year at the end of each year.

4.2. Scatter Diagram

In conducting regression modeling, it is necessary to know in advance the pattern of the relationship between the independent variables and the dependent variable. If the relationship pattern between the independent and dependent variables is unknown, then a nonparametric regression approach can be used. Analysis of data patterns between the dependent variable and each independent variable is shown in Figure 3.
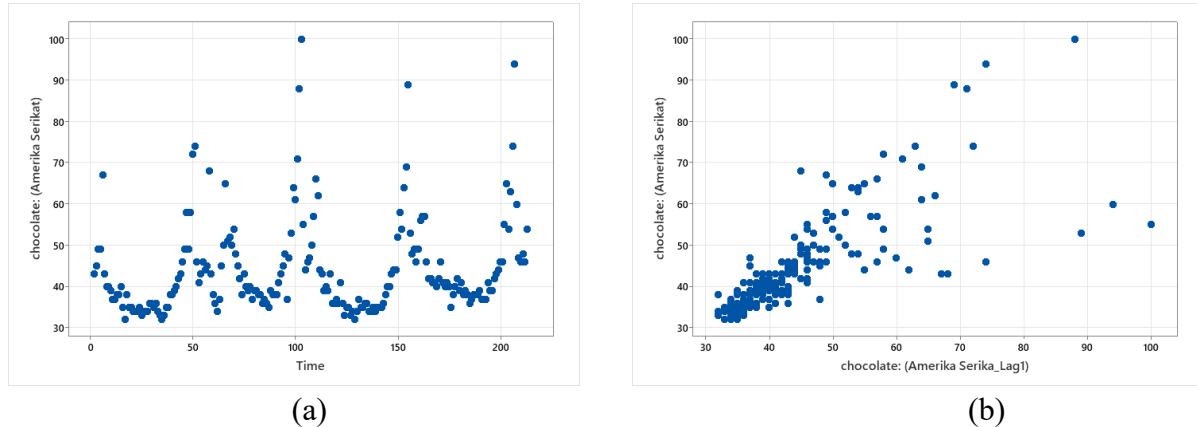

(a)   (b)

Fig 3. Scatter Diagram (a) chocolate (b) chocolate and Lag1

4.3. Modeling with Fourier series

In the nonparametric Fourier series regression modeling, the first step that needs to be considered is to determine the number of oscillations used. The number of oscillations will be limited to 1 to 5 oscillations. The results of the Fourier series nonparametric regression modeling are shown in Table 2.

**Table 2**. Minimum Oscillation and GCV

Number of Oscillations	GCV (K)	R^2
1	145947.10	59.48
2	64782.07	59.53
3	36393.70	59.59
4	22806.23	60.43
5	**15647.74**	**60.91**

Based on Table 2, the minimum GCV value is obtained when the number of oscillations used is five with GCV value of 15647.74, with a coefficient of determination of 60.91%. As a note, the number of oscillations that are attempted can actually be more than 5, but this will directly impact the number of parameters that need to be estimated so that the resulting model will not be parsimony.

The best Fourier series nonparametric regression model with the number of oscillations used is five can be written:

$$\hat{y}_i = 10.30 - 0.008x_{1i} + 0.021\cos x_{1i} + 0.354\cos 2x_{1i} + 0.161\cos 3x_{1i} + 0.431\cos 4x_{1i} + 0.554\cos 5x_{1i} +$$
$$0.745x_{2i} + 1.038\cos x_{2i} - 0.050\cos 2x_{2i} + 0.326\cos 3x_{2i} + 1.564\cos 4x_{2i} + 0.999\cos 5x_{2i}$$

A graphical illustration of a comparison of actual data with predictions from the best Fourier series nonparametric regression model with five oscillations is shown in Figure 4.
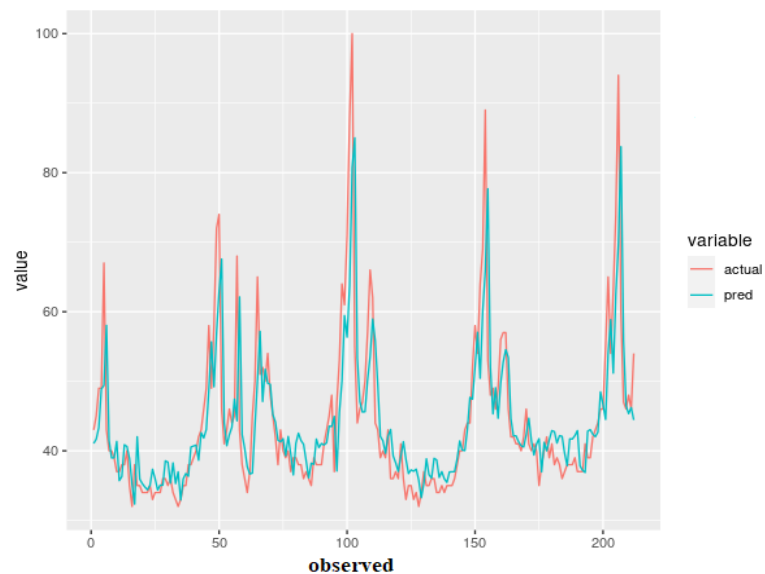
Fig 4. Visualization of actual vs predicted data

Based on Figure 4, it can be seen that the comparison graph of actual and predicted data from the Fourier series nonparametric regression model tends to follow the actual data, although there are some parts that show that there are still quite high gaps.

5. CONCLUSION

Based on the analysis results, the best Fourier series nonparametric regression model is obtained with the number of oscillations of 5, which is indicated by the minimum GCV value and coefficient of determination of 60.91%. By looking at the comparison graph of actual and predicted data, it can be seen that the predicted results from the Fourier series nonparametric regression model tend to follow the actual data.

The suggestion that can be submitted for further research is to use other Fourier components, for example, adding *sin* and *cos* components combined. In addition, this study also could be improved further by incorporating higher lag of *Y* as independent variable with initial inspection of lag plots. Moreover, the addition of dummy variable for special events of outlier could be considered.

ACKNOWLEDGMENT

REFERENCES

[1] Suparti, R. Santoso, A. Prahutama, A. R. Devi, and Sudargo, "Modeling longitudinal data based on Fourier regression," in *Journal of Physics: Conference Series*, Institute of Physics Publishing, Jun. 2019. doi: 10.1088/1742-6596/1217/1/012105.

[2] N. P. A. M. Mariati, I. N. Budiantara, and V. Ratnasari, "Combination Estimation of Smoothing Spline and Fourier Series in Nonparametric Regression," *Journal of Mathematics*, vol. 2020, 2020, doi: 10.1155/2020/4712531.

[3] A. T. R. Dani and N. Y. Adrianingsih, "Pemodelan Regresi Nonparametrik dengan Estimator Spline Truncated vs Deret Fourier," *Jambura Journal of Mathematics*, vol. 3, no. 1, pp. 26–36, 2021, doi: 10.34312/jjom.v3i1.7713.

[4] Suparti, R. Santoso, A. Prahutama, and A. R. Devi, "Indonesia's Inflation Analysis Using Hybrid Fourier - Wavelet Multiscale Autoregressive Method," in *Journal of Physics: Conference Series*, Institute of Physics Publishing, Sep. 2019. doi: 10.1088/1742-6596/1306/1/012041.

[5] M. A. D. Octavanny, I. N. Budiantara, H. Kuswanto, and D. P. Rahmawati, "Modeling of Children Ever Born in Indonesia Using Fourier Series Nonparametric Regression," in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Feb. 2021. doi: 10.1088/1742-6596/1752/1/012019.

[6] B. Lestari and I. Nyoman Budiantara, "Spline estimator and its asymptotic properties in multiresponse nonparametric regression model."

[7] A. Wahab, I. N. Budiantara, and K. Fitriasari, "The Estimation of Residual Variance in Nonparametric Regression," *Jurnal Matematika, Statistika dan Komputasi*, vol. 17, no. 3, pp. 438–446, 2021, doi: 10.20956/j.v17i3.13192.

[8] A. A. Khalil, I. Budiantara, and I. Zain, "Comparison of linear and quadratic bi-response semiparametric regression models using spline truncated," *J Phys Conf Ser*, vol. 1511, no. 1, 2020, doi: 10.1088/1742-6596/1511/1/012046.

[9] Sifriyani, I. N. Budiantara, S. H. Kartiko, and Gunardi, "A new method of hypothesis test for truncated spline nonparametric regression influenced by spatial heterogeneity and application," *Abstract and Applied Analysis*, vol. 2018, 2018, doi: 10.1155/2018/9769150.

[10] Wahba, *Spline Models for Observational Data*, 2nd ed. Pennsylvania: SIAM, 1990.

[11] A. T. R. Dani and L. Ni'matuzzahroh, "Penerapan Keluarga Model Spline Truncated Polinomial pada Regresi Nonparametrik," *Inferensi*, vol. 5, no. 1, p. 37, 2022, doi: 10.12962/j27213862.v5i1.12537.

[12] V. Ratnasari, I. N. Budiantara, and A. T. R. Dani, "Nonparametric Regression Mixed Estimators of Truncated Spline and Gaussian Kernel based on Cross-Validation (CV), Generalized Cross-Validation (GCV), and Unbiased Risk (UBR) Methods," *Int J Adv Sci Eng Inf Technol*, vol. 11, no. 6, pp. 2400–2406, 2021.

[13] M. Bilodeau, "Fourier smoother and additive models," *Canadian Journal of Statistics*, vol. 20, no. 3, pp. 257–269, 1992, doi: 10.2307/3315313.

[14] N. P. A. M. Mariati, I. N. Budiantara, and V. Ratnasari, "The application of mixed smoothing spline and fourier series model in nonparametric regression," *Symmetry (Basel)*, vol. 13, no. 11, Nov. 2021, doi: 10.3390/sym13112094.

[15] N. P. A. M. Mariati, I. N. Budiantara, and V. Ratnasari, "Modeling Poverty Percentages in the Papua Islands using Fourier Series in Nonparametric Regression Multivariable," in *Journal of Physics: Conference Series*, Institute of Physics Publishing, Dec. 2019. doi: 10.1088/1742-6596/1397/1/012071.

[16] A. T. Ampa, I. N. Budiantara, and I. Zain, "Selection of Optimal Smoothing Parameters in Mixed Estimator of Kernel and Fourier Series in Semiparametric Regression," *J Phys Conf Ser*, vol. 2123, no. 1, 2021, doi: 10.1088/1742-6596/2123/1/012035.

[17] A. T. R. Dani, V. Ratnasari, and I. N. Budiantara, "Optimal Knots Point and Bandwidth Selection in Modeling Mixed Estimator Nonparametric Regression," *IOP Conf Ser Mater Sci Eng*, vol. 1115, no. 1, p. 012020, Mar. 2021, doi: 10.1088/1757-899x/1115/1/012020.

[18] H. Husain, I. N. Budiantara, and I. Zain, "Mixed estimator of spline truncated, Fourier series, and kernel in biresponse semiparametric regression model," *IOP Conf Ser Earth Environ Sci*, vol. 880, no. 1, 2021, doi: 10.1088/1755-1315/880/1/012046.

[19] A. T. R. Dani, A. F. Dewi, and L. Ni'matuzzahroh, "Studi Simulasi dan Aplikasi: Estimator Deret Fourier pada Pemodelan Regresi Nonparametrik," *Prosiding Seminar Nasional Matematika, Statistika, dan Aplikasinya*, 2022, pp. 279–288.

[20] R. Asrianto and Y. Anggraini, "Prediksi Produksi Roti Coklat Berdasarkan Data Trends Google Menggunakan Metode Double Exponential Smoothing (Studi Kasus : Rotte Bakery)," *Software Engineering and Information Systems (SEIS*, vol. 1, no. 1, pp. 43–57, 2021.

[21] I. G. B. N. Diksa, "Forecasting the Existence of Chocolate with Variation and Seasonal Calendar Effects Using the Classic Time Series Approach," *Jurnal Matematika, Statistika dan Komputasi*, vol. 18, no. 2, pp. 237–250, Jan. 2022, doi: 10.20956/j.v18i2.18542.

[22] K. Nisa' and I. N. Budiantara, "Modeling East Java Indonesia Life Expectancy Using Semiparametric Regression Mixed Spline Truncated and Fourier Series," *MEDIA STATISTIKA*, vol. 13, no. 2, pp. 149–160, Dec. 2020, doi: 10.14710/medstat.13.2.149-160.

[23] I. N. Budiantara, "Penelitian Bidang Regresi Spline Menuju Terwujudnya Penelitian Statistika Yang Mandiri Dan Berkarakter," *Prosiding Seminar Nasional FMIPA Undiksha*, pp. 9–28, 2011.

[24] E. D. Igustin and I. N. Budiantara, "Pemodelan Faktor-Faktor yang Mempengaruhi Total Fertility Rate di Indonesia Menggunakan Regresi Nonparametrik Spline Truncated," *Jurnal Sains dan Seni ITS*, vol. 9, no. 2, pp. 178–185, 2021, doi: 10.12962/j23373520.v9i2.56791.

[25] S. Sifriyani, A. T. R. Dani, M. Fauziyah, M. N. Hayati, S. Wahyuningsih, and S. Prangga, "Spline and Kernel Mixed Estimators in Multivariable Nonparametric Regression For Dengue Hemorrhagic Fever Model," *Commun. Math. Biol. Neurosci.*, vol. 2023, pp. 1–15, 2023.

[26] Sifriyani, S. H. Kartiko, I. N. Budiantara, and Gunardi, "Development of nonparametric geographically weighted regression using truncated spline approach," *Songklanakarin Journal of Science and Technology*, vol. 40, no. 4, pp. 909–920, 2018, doi: 10.14456/sjst-psu.2018.98.

[27] I. Wayan Sudiarsa, "Simulations Study Combined Estimator Fourier Series and Spline Truncated in Multivariable Nonparametric Regression," in *IOP Conference Series: Materials Science and Engineering*, Institute of Physics Publishing, Jul. 2019. doi: 10.1088/1757-899X/546/5/052074.

[28] L. Laome, I. N. Budiantara, and V. Ratnasari, "Estimation Curve of Mixed Spline Truncated and Fourier Series Estimator for Geographically Weighted Nonparametric Regression," *Mathematics*, vol. 11, pp. 1–13, 2023.

[29] I. Wayan Sudiarsa, I. Nyoman Budiantara, S. Suhartono, and S. W. Purnami, "Combined estimator fourier series and spline truncated in multivariable nonparametric regression," *Applied Mathematical Sciences*, vol. 9, no. 97–100, pp. 4997–5010, 2015, doi: 10.12988/ams.2015.55394.

[30] R. Pane, I. N. Budiantara, I. Zain, and B. W. Otok, "Parametric and nonparametric estimators in fourier series semiparametric regression and their characteristics," *Applied Mathematical Sciences*, vol. 8, no. 101–104, pp. 5053–5064, 2014, doi: 10.12988/ams.2014.46472.

[31] N. Y. Adrianingsih and A. T. R. Dani, "Estimasi Model Regresi Semiparametrik Spline Truncated Menggunakan Metode Maximum Likelihood Estimation (MLE)," *Jambura Journal of Probability and Statistics*, vol. 2, no. 2, pp. 56–63, 2021, doi: 10.34312/jjps.v2i2.10255.

[32] G. Wahba and Y. Wang, "Spline Function," *Encyclopedia of Statistical Sciences*, pp. 1–27, 2014, doi: 10.4135/9781446247501.n3679.