VOL. 15, NO. 9, MAY 2020 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

THE PERFORMANCE OF NONPARAMETRIC REGRESSION FOR TREND AND SEASONAL PATTERN IN LONGITUDINAL DATA

M. Fariz Fadillah Mardianto[1,2], Sri Haryatmi Kartiko[3] and Herni Utami[3]
[1]Ph.D. Candidate in Department of Mathematics, Gadjah Mada University, Yogyakarta, Indonesia
[2]Department of Mathematics, Airlangga University, Surabaya, Indonesia
[3]Department of Mathematics, Gadjah Mada University, Yogyakarta, Indonesia
E-Mail: m.fariz.fadillah.m@fst.unair.ac.id

ABSTRACT

Recently, nonparametric regression does not only develop in cross section data but also in longitudinal data. Longitudinal data have repeated measurements in each subject. In the measurements for each subject sometimes there is a trend, seasonal, also combination between trend and seasonal data pattern. In this study, the performance of nonparametric regression estimators for longitudinal data related to model trend seasonal data pattern is compared by using Mean Square Error (MSE), Generalized Cross Validation (GCV) and determination coefficient value as goodness of indicator. The estimators that be used is truncated spline, Nadaraya Watson kernel, and Fourier series with include cosines and sines bases. This paper has contribution to introduce Fourier series, the new estimator for longitudinal data, as an alternative estimator for modeling trend and seasonal data. The result, the Fourier series estimator has the best performance indicators in modeling trend and seasonal data pattern for longitudinal data when compared with the estimator that be developed early in nonparametric regression for longitudinal data, such as spline and kernel. The result is important for data analysis in nonparametric regression for longitudinal data because there is data pattern with trend seasonal in many applications that need suitable estimator.

Keywords: nonparametric regression, trend and seasonal data pattern, longitudinal data, performance indicator.

1. INTRODUCTION

Nonparametric regression is an alternative method in regression analysis. Nonparametric regression can be used to investigate the relationship between predictor variables and response variables where regression curves that be estimated by certain functions will approach the correspond data patterns based on their characteristics. Nonparametric regression has flexibility, it means that the unknown data pattern that presented on a plot can determine the shape of regression curve based on estimators in the nonparametric regression (Asrini and Budiantara, 2014). Other characteristic of nonparametric regression is including quantity measure for smoothing like knot point, bandwidth, and oscillation parameter (Mardianto *et al*., 2019).

There are some estimators in nonparametric regression that often be used like spline, kernel, and Fourier series that have similarity to make smooth and flexible estimator as explained. The basic concept about spline estimator is proposed by Wahba (1990), Budiantara *et al*. (1997), and Eubank (1999) that determine optimal knot for selecting the best model. The spline estimator in nonparametric regression study for cross section data is growing rapidly until in multi response, multi predictor, and heteroscedasticity case that be proposed Lestari *et al*., (2012). The basic concept about kernel estimator is proposed by Hardle (1990) that determine optimal bandwidth for selecting the best model. The kernel estimator in nonparametric regression study for cross section data is developing until in multi predictor case that be proposed by Okumura and Naito (2006) and multi response case that be proposed by Chamidah and Saifudin (2013). The basic concept about Fourier series estimator is proposed by Bilodeau (1992), Dette *et al*. (2006), and Biedermann *et al.,* (2006) that determine optimal oscillation parameter for selecting the best model. Recently, the study about Fourier series estimator in nonparametric regression study for cross section data is growing rapidly until in more than one response case that proposed by Tjahjono *et al*. (2018).

The development of data analysis indicate that regression analysis is not only for cross section data but also longitudinal data. Longitudinal data structures are more complex because longitudinal data contain elements of cross section and time series data. The advantage of using longitudinal data, can to find out the changes that occur in subjects, because the observations are repeated for each subject (Wu and Zhang, 2006). Some estimator in nonparametric regression for longitudinal data have been proposed. Spline estimator for longitudinal data is proposed You and Zhou (1999), Wu and Zhang (2006), and Fernandes *et al*. (2014). Kernel estimator for longitudinal data is proposed by Wu and Chiang (2006), also Shen and Ramos (2016). Fourier series estimator is being proposed by authors as a new approach in nonparametric regression for longitudinal data. However, there have been no studies that comparing the performance of the three estimators in nonparametric regression for longitudinal data on the same data pattern. In this case we use trend and seasonal data pattern in longitudinal data. These data patterns are found in various fields such as meteorology, economics, and health sciences. In time series analysis, Fourier series suitable to be used for trend seasonal data pattern, compared with kernel (Mardianto, *et al.,* 2019). The result from Mardianto *et al*. (2019) are the Fourier series estimator has better value in goodness of indicator than kernel estimator.

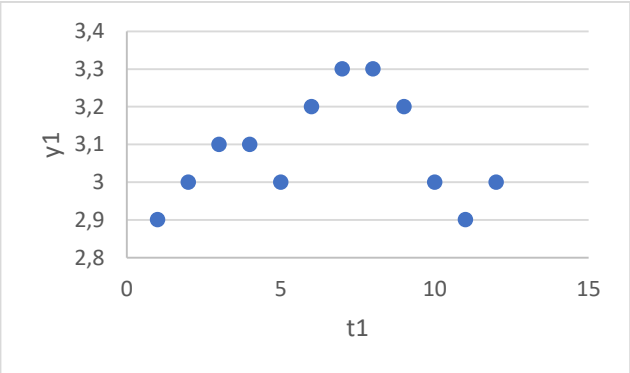
In this study, the purpose of the comparison is to find out what estimator is suitable for modeling trend and seasonal data patterns for longitudinal data based on the performance of indicator. So, the reason to compare is making idea based on research that Fourier series estimator is suitable for trend and seasonal data pattern, especially in longitudinal data case. The performance of indicator that be used is the smallest Mean Square Error (MSE) for optimal quantity measure in smoothing, and the biggest determination coefficient for each selected model in every estimator. Estimator that be used in this study is truncated spline, Nadaraya-Watson kernel, and the Fourier series which contains linear functions and complete trigonometric bases. The data example that be simulated in this study is based on the analogue in Box *et al.* (1976).

2. MATERIALS AND METHODS

2.1 Data and procedure

Simulation data consists of one response and one predictor. The response data used represents monthly wind speed data in 10 cities as subjects, while the predictor data used represents the observation period. In this case study there were 10 cities, each of which was observed for 12 months. Based on the results of the scatter plot between responses and predictors, there are trends and seasonal patterns. Simulation process used analogue of data pattern in Box *et al.* (1976) that be designed in such a way as a longitudinal data.

Figure-1 gives the pattern data that be used. There are four procedures in data analysis. The first, data is analyzed based on spline estimator until determine the optimal knot point. The second, data is analyzed based on kernel estimator until determine the optimal bandwidth. The third, data is analyzed based on Fourier series estimator until determine the oscillation parameter. The last, after we choose the selected model for every estimator, the best estimator is compared to conclude which estimator that has good performance in modelling trend and seasonal data pattern based on the performance indicators. Figure-1 presents the data pattern for first and last subject. Based on Figure-1 can be shown that mostly, the data have trend and seasonal pattern.



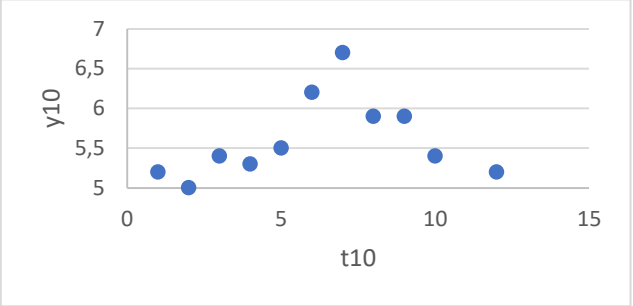


Figure-1. The data pattern for the first and the last subject in this study.

2.2 Nonparametric regression for longitudinal data

Consider pairs of data with form (y_{ij}, t_{ij}), t_{ij} denotes predictor variable for j^{th} observation in i^{th} subject. Here, $i = 1,2,\dots,n$ denote the number of subjects, $j = 1,2,\dots,n_i$ denote the number of observations for each subject, and p represents the number of predictors. Response variable for j^{th} observation in i^{th} subject is denoted by y_{ij}. The pairs of data that be presented in Table 1, follows nonparametric regression equation for longitudinal data

$$y_{ij} = f(t_{ij}) + \varepsilon_{ij}, \varepsilon_{ij} \sim N(0, \sigma^2) \quad (1)$$

$f(t_{ij})$ represents a regression curve. Random error for j^{th} observation in i^{th} subject is denoted by ε_{ij} that independent, identically normal distributed with mean 0, and variance σ^2.

Table-1. The structure of longitudinal data for this study.

Subject	Response (y_{ij})	Predictor (t_{ij})
1st Subject	y_{11} y_{12} ⋮ y_{1n_1}	t_{11} t_{12} ⋮ t_{1n_1}
2nd Subject	y_{21} y_{22} ⋮ y_{2n_2}	t_{21} t_{22} ⋮ t_{2n_2}
⋮	⋮	⋮
nth Subject	y_{n1} y_{n2} ⋮ y_{nn_n}	t_{n1} t_{n2} ⋮ t_{nn_n}

2.2.1 Spline estimator

Spline is a p-order piecewise polynomial that has continuous segmentation or truncated, so that it is effective to explain local characteristics of a data pattern and has flexibility as an approach to a local function in estimation theory, especially nonparametric regression (Eubank, 1999). The spline equation for paired data with form (y_{ij}, t_{ij}) is given in equation (2) as follows:

$$f(t_{ij}) = \sum_{r=1}^{m} a_{ri} t_{ij}^{r} + \sum_{l=1}^{q} b_{li}(t_{ij} - J_{li})_{+}^{m} \quad (2)$$

l is representation of knot point, q is the number of knot point, J_{li} is the value of knot point for l^{th} knot point in i^{th} subject, r is representation of spline order, m denotes the maximum of spline order. Parameters that their values can be determined based on WLS result denoted by a_{ri} and b_{li}. The nonparametric regression equation based on spline estimator for longitudinal data can be formed with substitutes equation (2) in equation (1) with result as follows:

$$y_{ij} = \sum_{r=1}^{m} a_{ri} t_{ij}^{r} + \sum_{l=1}^{q} b_{li}(t_{ij} - J_{li})_{+}^{m} + \varepsilon_{ij}, \varepsilon_{ij} \sim N(0, \sigma^2) \quad (3)$$

Form equation (3) and based on WLS result, the estimation form for nonparametric regression curve can be obtained as follows:

$$\hat{y}_{ij} = \sum_{r=1}^{m} \hat{a}_{ri} t_{ij}^{r} + \sum_{l=1}^{q} \hat{b}_{li}(t_{ij} - J_{li})_{+}^{m} \quad (4)$$

2.2.2 Kernel estimator

Kernel estimator has flexible form and its mathematical calculations are easily adjusted. The most popular estimator in nonparametric regression based on kernel approach that be used is Nadaraya-Watson kernel estimator. It can be used because has good flexibility in nonparametric regression with kernel approach (Hardle, 1990). The kernel estimator function based on Nadaraya-Watson for paired data with form (y_{ij}, t_{ij}) is given in equation (5) as follows:

$$\hat{f}(t_{ij}) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n_i} \kappa_h(t_{ij} - t) w_i^{jj} y_{ij}}{\sum_{i=1}^{n} \sum_{j=1}^{n_i} \kappa_h(t_{ij} - t) w_i^{jj}} \quad (5)$$

$\kappa_h(t_{ij} - t)$ denotes kernel function for bandwidth h that includes predictor that has corresponding with other. Weight for i^{th} subject can be presented based on w_i^{jj} that related with covariance matrix and WLS optimization. Further theoretical study about kernel estimator can be studied in Hardle (1990), Wu and Chiang (2006), also Shen and Ramos (2016). Because of $\hat{y}_{ij}$ equals with $\hat{f}(t_{ij})$, so the estimation form for nonparametric regression curve can be obtained with change $\hat{f}(t_{ij})$ becomes $\hat{y}_{ij}$.

2.2.3 Fourier series estimator

The Fourier series function is one of several estimators in nonparametric regression that interest to be studied. One of the advantages of the nonparametric regression using the Fourier series estimator, can overcome data pattern that be approached based on trigonometry pattern. Data patterns that correspond to the Fourier series approach are repeated data patterns. The process of repeating data patterns occurs in the value of the dependent variable for different independent variable (Mardianto *et al*., 2019). Fourier series in nonparametric regression has high flexibility in modelling the relationship between predictor and response that oscillated (Tjahjono *et al*., 2018). The Fourier series equation for paired data with form (y_{ij}, t_{ij}) is given in equation (6) as follows:

$$f(t_{ij}) = \frac{\alpha_{0i}}{2} + \gamma_i t_{ij} + \sum_{k=1}^{K} (\alpha_{ki} \cos k t_{ij} + \beta_{ki} \sin k t_{ij}) \quad (6)$$

k is representation of oscillation parameter, K is the number of oscillation parameter. Parameters that their values can be determined based on WLS result denoted by α_{0i}, γ_i, α_{ki} and β_{ki}. The nonparametric regression equation based on Fourier series estimator for longitudinal data can be formed with substitutes equation (6) in equation (1) with result as follows:

$$\hat{y}_{ij} = \frac{\hat{\alpha}_{0i}}{2} + \hat{\gamma}_i t_{ij} + \sum_{k=1}^{K} (\hat{\alpha}_{ki} \cos k t_{ij} + \hat{\beta}_{ki} \sin k t_{ij}) \quad (7)$$

2.3 The performance indicator

The performance indicators are associated with goodness of fit in the selected estimator based on the quantity measured for smoothing. In this case, we use three kinds of performance indicator commonly used in nonparametric regression study - Generalized Cross Validation (GCV), Mean Square Error (MSE), and Coefficient of Determination (R^2). Theoretically, GCV has optimal asymptotic properties, which are not shown by other methods (Wahba, 1990). For determining an optimal quantity measure for smoothing can be seen based on the smallest GCV value, which is that of MSE in this case. Also, the last performance indicator, the main thing is to select a model with the highest value of R^2 (Takezawa, 2006).

3. RESULTS AND DISCUSSIONS

The longitudinal data that the pattern of relationships between predictor and response for the first and last subject is presented in Figure-1 is modelled by spline, kernel, and Fourier series estimator in order to determine the performance of nonparametric regression for trend and seasonal data pattern in longitudinal data based on the performance indicators such as GCV, MSE, R^2, and the concept of parsimony model. The weighting that be used in this case is variance weighting because accommodate variability of errors.

3.1 Result based on spline estimator

The spline estimator is used to model longitudinal data consisting of trend and seasonal pattern. The GCV values for 1, 2 and 3 ordos are shown in Table-2, with the number of knot points inputted as 1, 2, and 3 respectively. Based on the result, the selected model is the quadratic or two ordos spline, with minimum GCV value equals to 0,1226, the number of knots is one and this minimum value was reached at a knot value of 11,7. The selected model has MSE value equals to 0,12 and R^2 equals to 0,8617.

Table-2. GCV value result for spline estimator.

Number knot	Minimum GCV		
	Linear spline	Quadratic spline	Cubic spline
1	0,1430	**0,1226**	0,1313
2	0,1429	0,1259	0,1325
3	0,1427	0,1285	0,1358

3.2 Result based on kernel estimator

The Kernel estimator is used to model longitudinal data consisting of trend and seasonal pattern. Table-3 shows the GCV for the value of bandwidth inputted. Considering Table-3, the smallest GCV is reached when the value of bandwidth is 0,4. The selected model has minimum GCV value equals to 0,0319025 and its MSE value equals to 0,3827 while R^2 is equal to 0,7685.

Table-3. GCV value result for kernel estimator.

Bandwidth	GCV
0,1	0,0423467
0,2	0,0319045
0,3	0,0319041
0,4	0,0319025
0,5	0,0319616
0,6	0,0323154
0,7	0,0331965
0,8	0,0346497
0,9	0,0365874
1,0	0,0388975
1,1	0,0414944
1,2	0,0443230
1,3	0,0473454
1,4	0,0505301
1,5	0,0538463

3.3 Result based on Fourier series estimator

The Fourier series estimator is used to model longitudinal data consisting of trend and seasonal pattern. Table 4 shows the GCV value for oscillation parameter inputted. Based on Table 4, the smallest GCV can be reached when the oscillation parameter equals to 3. The selected model has minimum GCV value equals to 0,0699511. The selected model has MSE value equals to 0,0074 and the value of R^2 equals to 0,9669.

Table-4. GCV value result for Fourier series estimator.

Bandwidth	GCV
1	0,16342640
2	0,07089285
3	0,06995110
4	0,07324524
5	0,1945920
6	0,2018173
7	0,2066809
8	0,2091073
9	0,213128
10	0,236036
11	0,264483
12	0,395900

3.4 Comparison based on the performance indicator

Based on the performance indicator, the results of spline, kernel, and Fourier series estimator in modelling trend and seasonal longitudinal data pattern are compared and Figure-2 shows the comparison plots based on the values of MSE, GCV, and R^2.

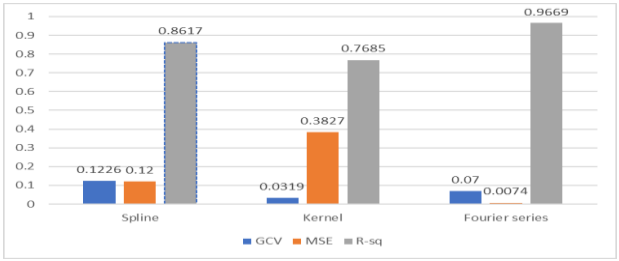


Figure-2. Comparison based on performance indicator.

Figure-2 is used as a reference for determining which estimator has the best performance in modelling trend and seasonal data pattern in nonparametric regression. Considering the outcome, the Fourier series estimator has the smallest MSE value which is equal to 0,0074 and the highest R^2 equals to 0,9669, and the spline estimator has a smaller MSE value equals to 0,1200, compared with kernel. Also, the spline has a bigger R^2 value of 0,8617 compared with kernel estimator. Hence, there is a significant difference between the MSE with spline and kernel estimators. Also, there is a significant difference between the value of R^2 with spline and kernel. These results show that the estimator with the best performance in modelling trend and seasonal data pattern in nonparametric regression is the Fourier series.

4. CONCLUSIONS

The Fourier series estimator has the best performance for modelling trend, seasonal and longitudinal data pattern in nonparametric regression. It has the smallest MSE value compared with spline and kernel as well as the highest R^2 value. This result shows the Fourier series form in nonparametric regression consist of trigonometric components which accommodate

seasonal patterns, and linear functions which accommodate trend patterns. In addition, the Fourier series estimator is the most parsimony model in nonparametric regression for longitudinal data.

ACKNOWLEDGEMENTS

The authors appreciate the support of the Gadjah Mada University and Airlangga University, as well as Lembaga Pengelola Dana Pendidikan (LPDP) for the success of this study. We are also grateful to The Ministry of Finance in Indonesia for the doctoral scholarship given to the authors.

REFERENCES

Asrini L. J. and Budiantara I. N. 2014. Fourier series semiparametric regression models (case study: the production of lowland rice irrigation in Central Java). ARPN Journal of Engineering and Applied Sciences. 9(9): 1501-1506.

Biedermann S., Dette H. and Hoffmann P. 2009. Constrained optimal discrimination designs for Fourier regression models. Ann Inst Stat Math Journal. 61: 143-157.

Bilodeau M. 1992. Fourier smoother and additive models, The Canadian Journal of Statistics. 3: 257-269.

Box G. E. P., Jenkins G. M., and Reinsel G. C. 1976. Time Series Analysis, Forecasting and Control. John Wiley and Sons, Inc., New York.

Budiantara I. N., Subanar and Soejoeti Z. 1997. Weighted spline estimator. Bulletin of the International Statistical Institute. 51: 333-344.

Budiantara I. N., Ratnasari V., Zain I., Ratna M. and Mardianto M.F.F. 2015. Modeling of HDI and PQLI in East Java (Indonesia) using bi-response semiparametric regression with Fourier series approach. ATABS Journal. 5(4): 21-28.

Chamidah N. and Saifudin T. 2013. Estimation of Children Growth Curve Based on Kernel Smoothing in Multi-Response Nonparametric Regression. Journal of Applied Mathematical Science. 7(37): 1839-1847.

Dette H., Melas V. B. and Shpilev P. 2016. T -optimal discriminating designs for Fourier regression models. Journal of Computational Statistics and Data Analysis. 26: 1-11.

Eubank R. L. 1999. Spline Smoothing and Nonparametric Regression 2nd Edition, Marcel Dekker, New York.

Fernandes A. A. R., Budiantara I. N., Otok B. W. and Suhartono. 2014. Spline estimator for bi-responses nonparametric regression model for longitudinal data. Applied Mathematical Sciences. 8(114): 5653-5665.

Hardle W. 1990. Applied Nonparametric Regression. Cambridge University Press, New York.

Lestari B., Budiantara I. N., Sunaryo S. and Mashuri M. 2012. Spline Smoothing for Multi-Response Nonparametric Regression Model in Case of Heteroscedasticity of Variance. Journal of Mathematics and Statistics. 8(3): 377-384.

Mardianto M. F. F., Kartiko S. H., and Utami H. 2019. Forecasting Trend-Seasonal Data Using Nonparametric Regression with Kernel and Fourier Series Approach Proceedings of the Third International Conference on Computing. Mathematics and Statistics (iCMS2017) Springer Series, Langkawi.

Mardianto M. F. F., Tjahjono E. and Rifada M. 2019. Statistical modelling for prediction of rice production in Indonesia using semiparametric regression based on three forms of Fourier series estimator. ARPN Journal of Engineering and Applied Sciences. 14(15) : 2763-2770.

Mardianto M. F. F., Kartiko S. H. and Utami H. 2019. Prediction the Number of Students in Indonesia who Study in Tutoring Agency and Their Motivations based on Fourier Series Estimator and Structural Equation Modelling. International Journal of Innovation, Creativity and Change. 5(3): 708-731.

Okumura H. and Naito K. 2006. Nonparametric kernel regression for multinomial data, Journal of Multivariate Analysis. 97: 2009-2022.

Shen D. and Ramos F. 2016. Kernel Embeddings of Longitudinal Data. Proceedings of 29th Australasian Joint Conference on Artificial Intelligence, Hobart.

Takezawa K. 2006. Introduction to Nonparametric Regression. John Wiley & Sons, Inc., New Jearsey.

Tjahjono E., Mardianto M. F. F., Chamidah N. 2018. Prediction of electricity consumption using Fourier series estimator in bi-response nonparametric regression model. Far East Journal of Mathematical Sciences. 103 (8): 1251-1263.

Wahba, G. 1990. Spline Model for Observational Data, SIAM XII, Philadelphia.

Wu C. O. and Chiang C. T. 2006. Kernel smoothing on varying coefficient model with longitudinal dependent variable. Statistica Sinica. 10: 433-456.

Wu H. and Zhang J. T. 2006. Nonparametric Regression Methods for Longitudinal Data Analysis. John Wiley and Sons, Inc., New Jersey.

You J. and Zhou X. 2009. Partially linear models and polynomial spline approximations for the analysis of unbalanced panel data. Journal of Statistical Planning and Inference. 139: 679-695.