Jurnal Varian
Vol. 9, No. 1, February 2026, pp 29–38
e-ISSN: 2581-2017
DOI: https://doi.org/10.30812/varian.v9i1.4949

Simulation Study for Nonparametric Regression Model with Quartic Kernel Function

Esmar Suprianto, Sifriyani, Andrea Tri Rian Dani
Universitas Mulawarman, Samarinda, Indonesia

Article Info

Article history:

Received : 03-06-2025
Revised : 11-26-2025
Accepted : 12-29-2025

Keywords:

Nonparametric Regression;
Quartic Kernel Function;
Simulation Study.

ABSTRACT

Nonparametric regression is a method for estimating the pattern of the relationship between predictor variables and response variables when the functional form of the regression curve is unknown. One estimator applicable to nonparametric regression is the Kernel estimator. The kernel estimator has a more flexible form, and the calculations are straightforward. The performance of the Kernel estimator is significantly affected by the Kernel function and the smoothing parameter (bandwidth). The method used in this study is the Kernel estimator, applied to a simulation study using a quartic kernel for optimal bandwidth selection via generalized cross-validation (GCV). This study aims to evaluate simulation results across various combinations of sample sizes and variances and to present a prediction plot of the Quartic Kernel function based on the simulation study. The results of this study are based on the Quartic Kernel function; larger sample sizes yield smaller Mean Squared Error (MSE) and GCV values and a larger coefficient of determination. In addition to sample size, variance is also very influential. The larger the variance, the larger the MSE and GCV values, and the smaller the coefficient of determination. The results of this study are prediction plots against the simulation studies used, showing that the Quartic Kernel function is less effective at predicting simulation study results. This is also evident from the accuracy obtained across different sample sizes and data with varying levels of variance, indicating that, in simulation studies using the quartic kernel estimator, predictive performance is poorer.

Corresponding Author:

Sifriyani,
Department of Mathematics and Natural Science,
Universitas Mulawarman,
Email: sifriyani@fmipa.unmul.ac.id

How to Cite:
Suprianto, E., Sifriyani, S., & Dani, A. T. R. (2026). Simulation Study for Nonparametric Regression Model with Quartic Kernel Function. *Jurnal Varian*, *9*(1), 29–38.

A. INTRODUCTION

Along with the rapid development of information technology, the volume, variety, and speed of data generated have increased significantly. This creates new challenges in identifying patterns of relationships between variables. In situations like this, statistical modeling using regression analysis has limitations in getting the relationship patterns hidden in the data. One statistical method for analyzing and modeling the relationship between predictor variables and the response variable is regression analysis. The three approaches in regression analysis are parametric, nonparametric, and semiparametric regression (Nurdin et al., 2018; Padatuan et al., 2021).

Parametric regression is one of the approaches used in regression analysis when the shape of the regression is already well known, nonparametric regression used when the exact shape of the regression curve is an unknown, while semiparametric regression is an approach that combines parametric regression and nonparametric regression, it is used when the shape of the regression curve is partially known and partially unknown (Dani et al., 2021; Fadlirhohim et al., 2024). The nonparametric regression approach does not rely on the assumption of a particular regression curve shape, thereby offering great flexibility (Handayani et al., 2024; Pasarella et al., 2022). In nonparametric regression, several estimators have been developed to support nonparametric modeling, including the truncated-spline estimator, Fourier series, and kernel methods (Dani & Adrianingsih, 2021; Rahmania et al., 2024).

One estimator in nonparametric regression is the kernel estimator, which does not require a specific data pattern. In contrast, the truncated spline is suitable for data patterns with trend change points (knots), and the Fourier series is suitable for periodic or repetitive data patterns. The kernel takes a more flexible form, and the calculations are straightforward (Yuliati & Sihombing, 2020). Another advantage is that it can model data that lack clear patterns and exhibits a relatively fast convergence rate. The performance of the Kernel estimator is significantly affected by the Kernel function and the smoothing parameter (bandwidth). There are various types of kernel functions, namely triangle, uniform, Epanechnikov, quartic, triweight, cosine, and Gauss (Budiantara et al., 2015). In this study, researchers examine the quartic kernel function in greater detail.

The quartic kernel is a kernel function that is more sensitive to local changes in the data. According to Sifriyani et al. (2023), the Kernel estimator is similar to other estimators, but its use of a more specialized bandwidth distinguishes it from other nonparametric regression methods. Bandwidth is a parameter that controls the smoothness of the estimated curve. When the bandwidth is overly large, it produces an oversmoothed curve, with greater estimation bias and smaller variance. Conversely, if the bandwidth is too small, the resulting estimated curve will be coarser or follow the pattern of the data; the bias is smaller, and the variance is greater (Budiantara et al., 2015)—bandwidth selection using the Generalized Cross Validation (GCV) method. GCV is the most widely used and preferred method because it has asymptotically optimal properties (Pratama, 2022; Purnaraga et al., 2020).

Several studies have examined nonparametric kernel regression on IHSG data, including Karimuse et al. (2023), which reported results for the Gaussian kernel with an optimal bandwidth of 0.332 and a minimum GCV of 0.246. Previous research on related simulations, namely Dani et al. (2022), examined a simulation study using a Fourier series estimator. This study used a sample size of 1000 and a variance of 0.15, with a trigonometric function regression curve. The results showed that the estimated curve follows the pattern of the actual data, provided that the number of oscillations used is appropriate and optimal. Dani et al. (2021) used a sample size of 500, a variance of 0.1, a trigonometric function regression curve, and a uniformly distributed predictor variable and normally distributed errors. The results showed that nonparametric truncated spline regression has excellent capabilities for handling data with changing behavior. Ratnasari et al. (2021) compared the accuracy of mixed spline and kernel estimators. This study used sample sizes of 25, 50, 100, and 200 with variances of 0.05, 0.5, and 1, respectively. The results showed that GCV provided superior accuracy and performance compared to CV and UBR. Based on these studies, researchers are interested in employing kernel estimators in simulation studies because no prior studies have used them alone. The only studies available employ Fourier series estimators, such as the simulation study by Dani et al. (2022) and the study by Ratnasari et al. (2021), which used a combined estimator of truncated spline and Gaussian kernel.

The difference between this research and prior research is that it employs simulated data, whereas most studies use real data. The next difference lies in the method employed: existing nonparametric regression research is limited to truncated-spline and Fourier-series estimators; therefore, this research introduces kernel estimators as a novelty. Therefore, researchers are interested in using simulation studies with kernel estimators. This study aims to evaluate simulation results across various combinations of sample sizes and variances, and to plot the predictions of the quartic kernel function against the simulation results. Given this background, the researcher intends to conduct a study titled "Simulation Study: Nonparametric Regression Model with Quartic Kernel Function". The contribution of this study is to provide readers with an initial overview of the use of kernel estimators, particularly quartic, across various data conditions and combinations.

B. RESEARCH METHOD

The study employed an experimental design in which researchers manipulated the research subject through simulation and observed multiple characteristics. This research employs a causal design, which aims to investigate causal relationships. The research variables consisted of one response variable and one predictor variable. The predictor variable is generated from a trigonometric function using a Uniform distribution $(0, 1)$, and the error follows a normal distribution $(0, \sigma^2)$. Furthermore, the response variable is obtained by adding the error term to the regression function.

The steps used to analyze the data in this study are as follows:

1. Generate data with sample sizes ($n = 50,\ 100,\ 200,$ and 250) and variances ($\sigma^2 = 0.01;\ 0.05;\ 0.5$ and 1), where the response variables are generated with errors following a normal distribution $(0, \sigma^2)$ and the predictor variables following a Uniform distribution $(0, 1)$.
2. Determining optimal bandwidth using GCV. Possible bandwidth values are obtained from the predictor variable, with a lower limit of 0 and an upper limit of the maximum value, subtracted from the minimum value in the data. For example, the desired

DOI: https://doi.org/10.30812/varian.v9i1.4949

bandwidth is 50. Still, during the process, the lower and upper limits are removed, yielding 48 possible values, and the optimal bandwidth is selected as the value that minimizes GCV.

3. Modeling nonparametric Kernel regression with the Quartic Kernel function as in Equation (1).

$$\widehat{y_i} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\frac{15}{16}\left(1-\left(\frac{x-x_i}{\alpha}\right)^2\right)^2}{\frac{1}{n}\sum_{i=1}^{n}\frac{15}{16}\left(1-\left(\frac{x-x_i}{\alpha}\right)^2\right)^2}\right] y_i \quad (1)$$

with $\widehat{y_i}$ is the estimate of the response variable in the i-th observation, n is the number of observations, x is a predictor variable that applies to each $x = x_1, x = x_2, \ldots, x = x_n$ and x_i is the predictor variable in the i-th observation.

4. Calculate the MSE in Equation (2) and also coefficient of determination values for the various combination of sample size as well as the variance being tested.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (2)$$

where n is the number of observations, y_i is the response variable in the i-th observation data, $i = 1, 2, \ldots, n$, $\widehat{y_i}$ is the estimate of the response variable in the i-th observation.

5. Repeating steps 1 through 4 for 10 repetitions.
6. Calculate the average MSE in Equation (3) by adding up all the MSE values obtained from each repetition, divided by the number of repetitions, and also the coefficient of determination for each repetition and combination of sample size, as well as the variance being tested.

$$Average(MSE) = \frac{MSE_1 + MSE_2 + \cdots + MSE_r}{r} \quad (3)$$

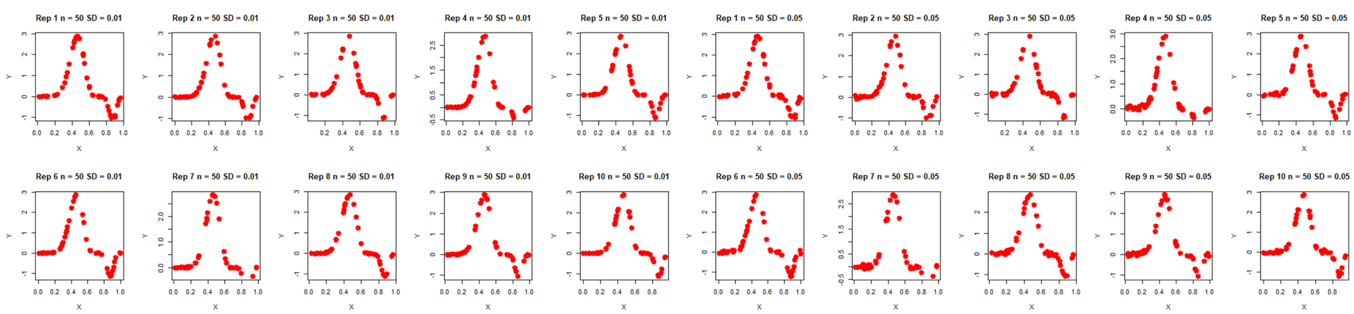
where r is the number of repetitions.

7. Obtain the prediction results of the quartic Kernel function with the prediction plot using the MSE value and the coefficient of determination.
8. Making conclusions.

C. RESULT AND DISCUSSION

1. Scatter Plot of Simulated Data

Simulations were conducted on the different combinations used. In this study, the sample sizes to be tested are 50, 100, 200, and 250. The variation of the variance σ^2 to be tried is 0.01, 0.05, 0.5, and 1. In generating data in this study, 10 repetitions of data generation were carried out, and for predictor variables, a uniform distribution $(0, 1)$ was followed. Then for response variables generated with trigonometric functions $\frac{\sin(2\pi x^2)^5}{\sin(\pi x^3)}$ that follows uniform distribution (0,1) and are added with the errors that follow a normal distribution $(0, \sigma^2)$. The scatter plot for a sample size of 50, 100, 200 and 250 with a variance of 0.01 and 0.05 is show in Figure 1.



Continued to next page.

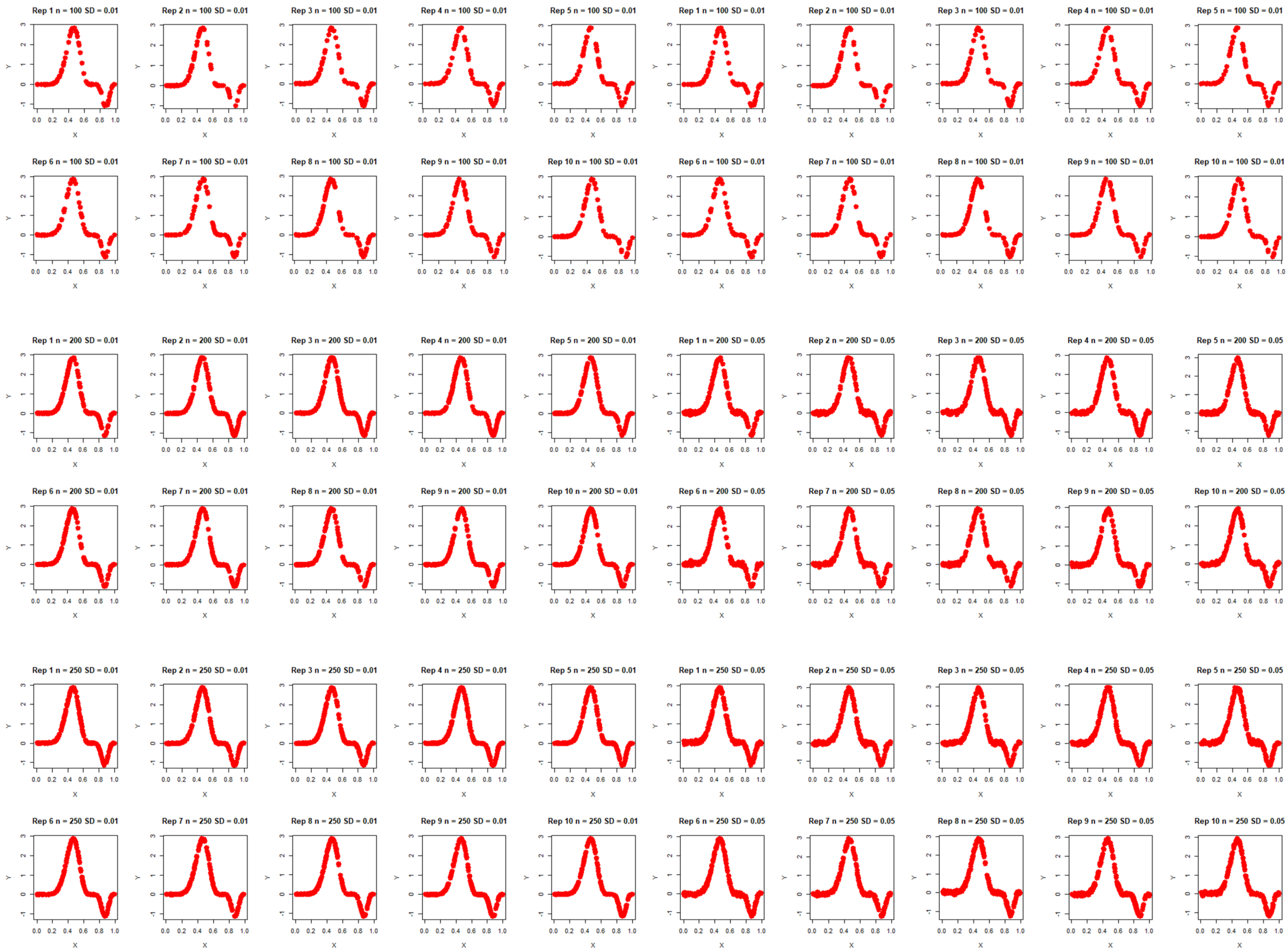
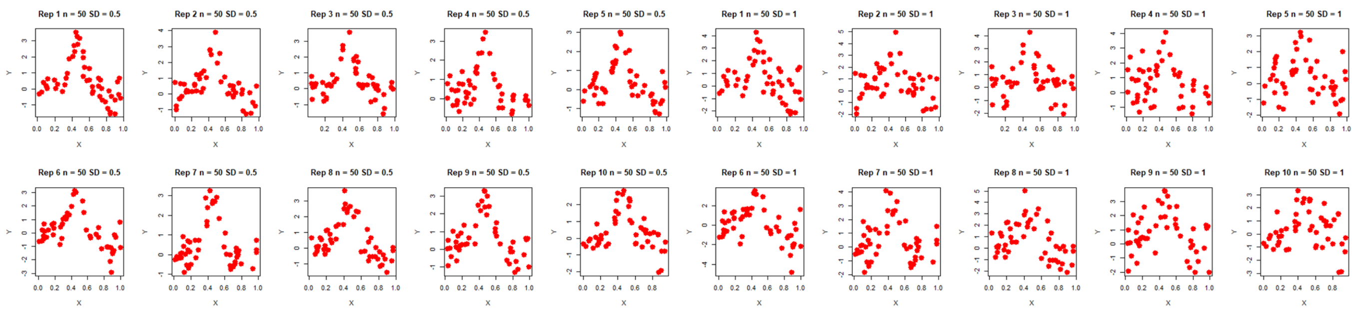


Figure 1. Scatter plot of data distribution for $n = 50,\ 100,\ 200$, and 250 with a variance of 0.01 and 0.05

Based on Figure 1, the data appear to form a regular, consistent, and visually identifiable pattern. The level of data dispersion is also low, with most data points following the specified trigonometric function, and the data are relatively homogeneous compared with other variance conditions. This is because the data does not deviate too far from the data center. We can also observe the data patterns that form; they consistently emerge after 10 repetitions. Furthermore, another scatter diagram is given, the same predictor and response variables with 10 repetitions for sample sizes n = 50, 100, 200, and 250 with variances of 0.5 and 1 are shown in Figure 2.



Continued to next page.

DOI: https://doi.org/10.30812/varian.v9i1.4949

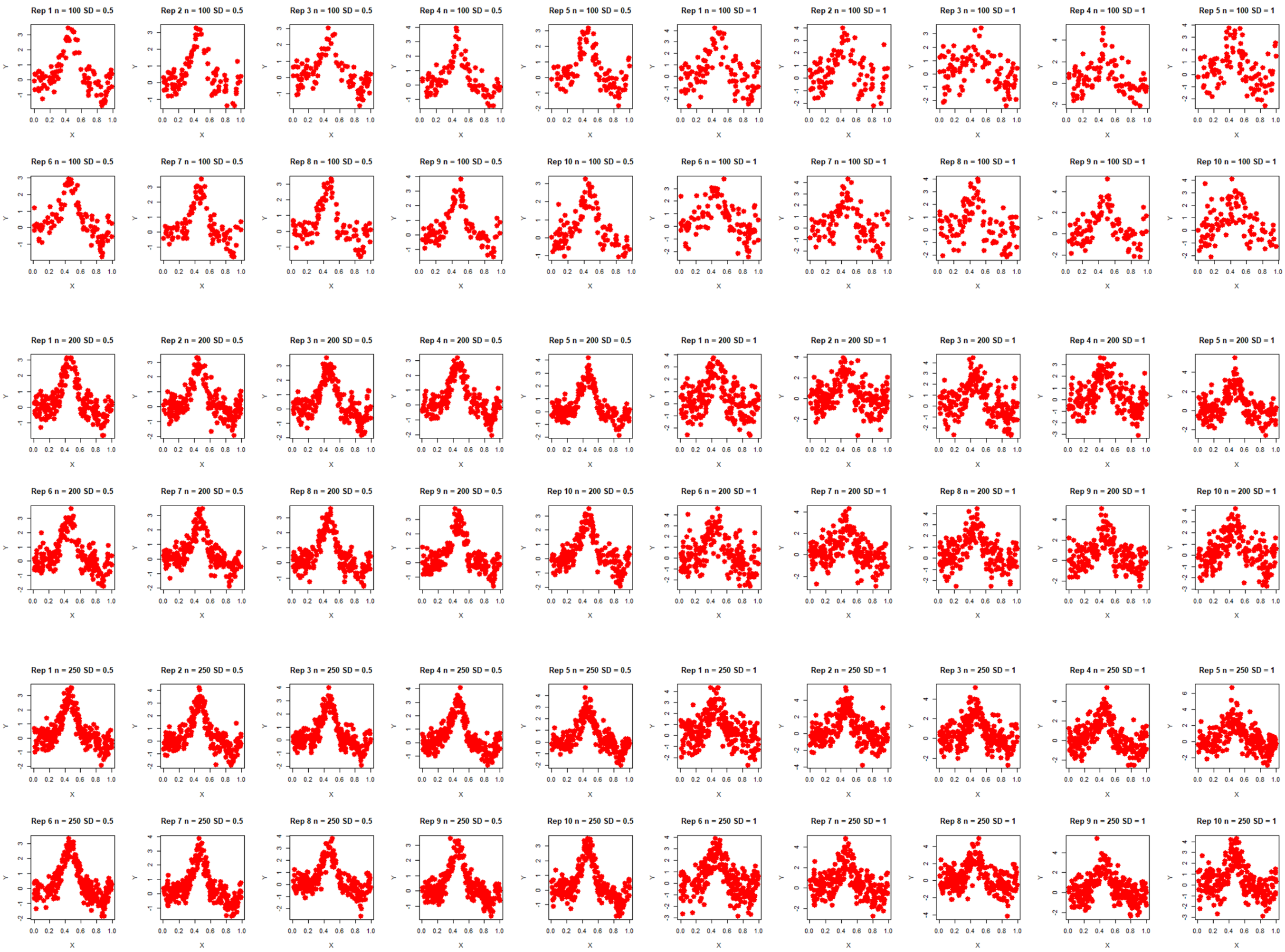


Figure 2. Scatter plot of data distribution for $n = 50,\ 100,\ 200$ and 250 with a variance of 0.5 and 1

Based on Figure 2, it can be seen that in these two combinations, the data pattern begins to show a wider spread than the previous variance. The extent of the data spread, which causes some data points to deviate further from the data pattern based on the trigonometric function used and indicates the occurrence of heterogeneity in the data. The higher the value of the variance tried, the more likely the data will spread away from its mean value and the more heterogeneous the simulated data will be. This greater spread of data makes the patterns formed less regular and more difficult to identify visually. This shows that with a wider spread of data, the relationship between data points becomes more complex, and the pattern formed is no longer as clear as in the smaller variance condition.

Overall, it can be seen that the variation of the tested variance causes the relationship pattern between these predictor variables and the response to be more dispersed. This indicates that the greater the variance, the higher the level of heterogeneity in the data.

2. Simulation Results

Simulation aims to evaluate the ability of a method with various scenarios or data combinations. The simulation results in this study are in the form of the average bandwidth value (α), Mean Square Error (MSE), coefficient of determination (R^2) and Generalized Cross-Validation (GCV) of each repetition or repetition carried out based on the Kernel function and sample size and variance variation tried.

Table 1. Simulation Results of Quartic Kernel Function with Various Combinations of Sample Size and Variance

Sample Size (n)	Variance (σ^2)	Average			
		α	MSE	R^2	GCV
50	0,01	0,431	0,469	56,29%	0,494
	0,05	0,431	0,470	56,23%	0,495
	0,5	0,431	0,674	47,14%	0,710
	1	0,431	1,323	31,16%	1,393
100	0,01	0,422	0,463	58,10%	0,474
	0,05	0,424	0,465	58,01%	0,476
	0,5	0,434	0,701	48,29%	0,720
	1	0,448	1,421	32,47%	1,460
200	0,01	0,433	0,459	55,41%	0,465
	0,05	0,435	0,463	55,03%	0,469
	0,5	0,446	0,727	42,10%	0,737
	1	0,464	1,482	24,96%	1,503
250	0,01	0,428	0,463	56,43%	0,468
	0,05	0,428	0,466	56,33%	0,471
	0,5	0,426	0,709	46,30%	0,716
	1	0,426	1,439	30,25%	1,453

Based on Table 1, it can be seen that the average value of each accuracy measure is used for each repetition. For example, with a variance of 0.01 and a variable sample size, MSE and GCV decrease as sample size increases, whereas the coefficient of determination increases. For example, for a sample size of 100 and a variance of 0.01, using GCV in the optimal bandwidth selection process yields an average GCV of 0.474, an MSE of 0.463, and a coefficient of determination of 58.10%. As another example, if the sample of 250 has a variance of 0.01, using GCV in the optimal bandwidth selection process yields an average GCV of 0.468, an MSE of 0.463, and a coefficient of determination of 56.43%.

In addition to sample size, variance significantly affects the simulation results. The greater the variance, the more dispersed the observations are from the mean. For example, for a sample size of 50, increasing the variance tends to increase the MSE and GCV values and decrease the coefficient of determination. The increase in MSE and GCV values indicates that the prediction accuracy of the model is getting worse with increasing variance, while the decrease in the coefficient of determination indicates that the model formed is not able to explain the variability in the data, so the model has difficulty in capturing the original pattern in the data. For example, for a variance of 0.05 and a sample size of 50, using GCV in the optimal bandwidth selection process, the average GCV value generated is 0.495 with an MSE value of 0.470 and a coefficient of determination of 56.23%. As another example, if the variance is 1 and the sample size is 50, using GCV in the optimal bandwidth selection process, the average GCV value generated is 1.393 with an MSE value of 1.323 and a coefficient of determination of 31.16%. The impact of this variance is clear: the variance represents the deviation of the data from the mean, so that larger values of variance tend to disperse the data further from the mean.

3. Data Simulation Prediction Plot

The prediction results and accuracy obtained by varying the sample size and variance are examined. The nonparametric Kernel regression model with a quartic function is written in Equation (1). then, the prediction results for $n = 50$ and $\sigma^2 = 0.01$ are shown in Equation (4):

$$
\begin{aligned}
\hat{y}_1 &= \frac{1}{50}\left[\frac{\frac{15}{16}\left(1-\left(\frac{0.975-0.975}{0.438}\right)^2\right)^2+\cdots+\frac{15}{16}\left(1-\left(\frac{0.975-0.668}{0.438}\right)^2\right)^2}{\frac{1}{50}\left(\frac{15}{16}\left(1-\left(\frac{0.975-0.975}{0.438}\right)^2\right)^2+\cdots+\frac{15}{16}\left(1-\left(\frac{0.975-0.668}{0.438}\right)^2\right)^2\right)}\right]0.009 = 0.073 \\
\hat{y}_{50} &= \frac{1}{50}\left[\frac{\frac{15}{16}\left(1-\left(\frac{0.668-0.975}{0.438}\right)^2\right)^2+\cdots+\frac{15}{16}\left(1-\left(\frac{0.668-0.668}{0.438}\right)^2\right)^2}{\frac{1}{50}\left(\frac{15}{16}\left(1-\left(\frac{0.668-0.975}{0.438}\right)^2\right)^2+\cdots+\frac{15}{16}\left(1-\left(\frac{0.668-0.668}{0.438}\right)^2\right)^2\right)}\right]0.002 = 0.292
\end{aligned} \quad (4)
$$

The same calculation as in Equation (4) was carried out for each combination of data sample sizes $n = 50,\ 100,\ 200$ and 250 with variances of $0.01;\ 0.05;\ 0.5;$ and 1. Then the prediction plot for each combination of data is obtained in Figure 3.

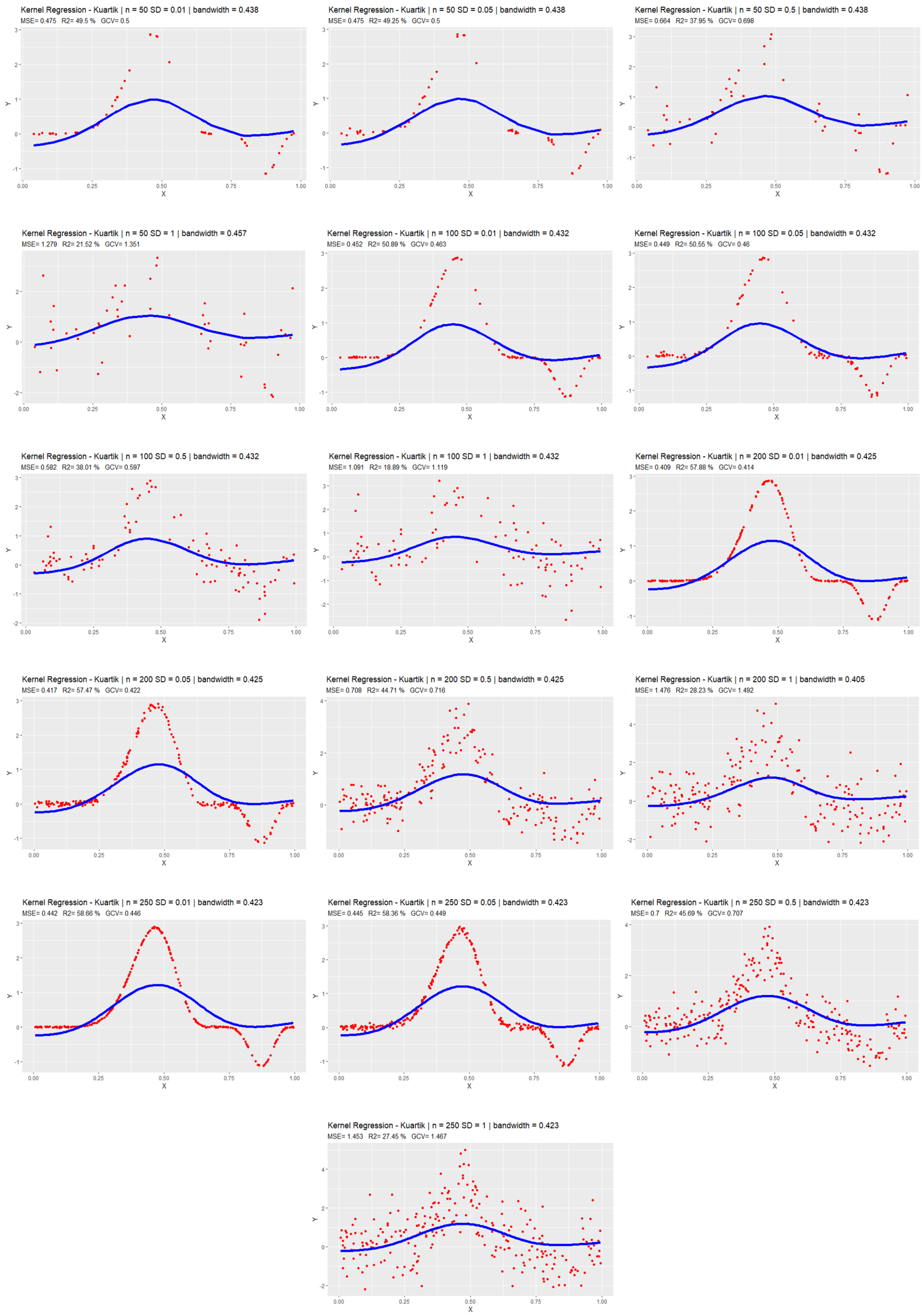


Figure 3. Scatter plot of prediction data for $n = 50,\ 100,\ 200,$ and 250 with a variance of 0.5 and 1

Based on Figure 3, the analysis of the prediction plot and the model accuracy results indicates that the Quartic Kernel function performs poorly when applied to nonparametric regression models with varying sample sizes and levels of variance. The Quartic Kernel function provides poor estimates. This is evident in accuracy measures such as MSE, GCV, and the resulting coefficient of determination. This indicates that the choice of the quartic kernel is inappropriate for certain data combinations. The novelty of this research lies in the use of simulated data as the research dataset and a kernel estimator for prediction. The results of this study align with previous studies that used kernel estimators, but combined them with other estimators, such as Ratnasari et al. (2021), who combined truncated splines and a Gaussian kernel to determine the best parameter selection method using simulation studies, while this study uses a quartic kernel.

D. CONCLUSION AND SUGGESTION

Based on the results of simulation studies with variations in sample size and variance in the estimation of nonparametric regression models with quartic Kernel functions, it is found that with the sample size made variable, it shows that as the sample size is increased, there is a tendency to decrease the MSE and GCV values and increase the coefficient of determination. In addition to sample size, variance is also very influential on the resulting simulation results. The greater the variance, the more scattered the observations are from the mean; as the variance increases, MSE and GCV tend to rise, and the coefficient of determination tends to decrease. Based on the results of the prediction plot analysis and the results of the model accuracy measure, it can be concluded that the quartic Kernel function shows poor results in the application of nonparametric regression models on data with different sample sizes and on data with varying levels of variance, the quartic Kernel function provides poor estimation results. This is evident in accuracy measures such as MSE, GCV, and the resulting coefficient of determination. Suggestions for further research include simulating data using alternative kernel functions, such as the Gaussian kernel. Furthermore, bandwidth selection can be performed using Cross-Validation (CV) or Unbiased Risk (UBR) methods. Nonparametric kernel regression can also be applied to real data, and given this study's limitations, it is advisable to combine multiple kernel functions for prediction rather than relying solely on a single function.

ACKNOWLEDGEMENT

We would like to express our gratitude to Universitas Mulawarman and other related parties for making this research possible.

DECLARATIONS

AUTHOR CONTRIBUTION

According to the authors, this manuscript was prepared through contributions from both parties. The first author carried out the entire research workflow, including conceptualizing the study, processing and analyzing the data, running the software, and writing the full manuscript. The second and the third author contributed by providing academic supervision, guidance, and critical review throughout the research and writing process.

FUNDING STATEMENT

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for profit sectors.

COMPETING INTEREST

The authors declare that they have no competing financial or personal interests that could influence the work reported in this article.

REFERENCES

Budiantara, I. N., Ratnasari, V., Ratna, M., & Zain, I. (2015). The combination of spline and kernel estimator for nonparametric regression and its properties. *Applied Mathematical Sciences*, *9*, 6083–6094. https://doi.org/10.12988/ams.2015.58517

Dani, A. T. R., & Adrianingsih, N. Y. (2021). Pemodelan Regresi Nonparametrik dengan Estimator Spline Truncated vs Deret Fourier. *Jambura Journal of Mathematics*, *3*(1), 26–36. https://doi.org/10.34312/jjom.v3i1.7713

Dani, A. T. R., Adrianingsih, N. Y., Ainurrochmah, A., & Sriningsih, R. (2021). Flexibility of nonparametric regression spline truncated on data without a specific pattern. *Jurnal Litbang Edusaintech*, *2*(1), 37–43. https://doi.org/10.51402/jle.v2i1.30

Dani, A. T. R., Dewi, A. F., & Ni'matuzzahroh, L. (2022). Studi Simulasi dan Aplikasi: Estimator Deret Fourier pada Pemodelan Regresi Nonparametrik.

Fadlirhohim, R. D., Sifriyani, S., & Dani, A. T. R. (2024). Modeling Stunting Prevalence in Indonesia Using Spline Truncated Semiparametric Regression. *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, *18*(3), 2015–2028. https://doi.org/10.30598/barekengvol18iss3pp2015-2028

Handayani, T., Sifriyani, S., & Rian Dani, A. T. (2024). Stunting prevalence modeling using nonparametric regression of quadratic splines. *Jurnal Varian*, *7*(2), 149–160. https://doi.org/10.30812/varian.v7i2.2916

Karimuse, W. Y., Nohe, D. A., & Siringoringo, M. (2023). Pendekatan regresi nonparametrik kernel pada data IHSG periode januari 2020 – desember 2021. *STATISTIKA Journal of Theoretical Statistics and Its Applications*, *23*(1), 1–7. https://doi.org/10.29313/statistika.v23i1.1628

Nurdin, I., Sugiman, S., & Sunarmi, S. (2018). Penerapan kombinasi metode ridge regression (RR) dan metode generalized least square (GLS) untuk mengatasi masalah multikolinearitas dan autokorelasi. *Indonesian Journal of Mathematics and Natural Sciences*, *41*(1), 58–68. https://doi.org/10.15294/ijmns.v41i1.16384

Padatuan, A. B., Sifriyani, S., & Prangga, S. (2021). Pemodelan angka harapan hidup dan angka kematian bayi di kalimantan dengan regresi nonparametrik spline birespon. *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, *15*(2), 283–296. https://doi.org/10.30598/barekengvol15iss2pp283-296

Pasarella, M. D., Sifriyani, S., & Amijaya, F. D. T. (2022). Nonparametrik regression model estimation with the fourier series the fourier series approach and its application to the accumulative covid-19 data in indonesia. *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, *16*(4), 1167–1174. https://doi.org/10.30598/barekengvol16iss4pp1167-1174

Pratama, M. H. (2022). Regresi nonparametrik multivariabel dengan pendekatan spline truncated pada kasus tuberculosis. *STATISTIKA Journal of Theoretical Statistics and Its Applications*, *22*(1), 87–93. https://doi.org/10.29313/statistika.v22i1.506

Purnaraga, T., Sifriyani, S., & Prangga, S. (2020). Regresi nonparamaetrik spline pada data laju pertumbuhan ekonomi di kalimantan. *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, *14*(3), 343–356. https://doi.org/10.30598/barekengvol14iss3pp343-356

Rahmania, R., Sifriyani, S., & Dani, A. T. R. (2024). Modeling open unemployment rate in kalimantan island using nonparametric regression with fourier series estimator. *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, *18*(1), 0245–0254. https://doi.org/10.30598/barekengvol18iss1pp0245-0254

Ratnasari, V., Budiantara, I. N., & Dani, A. T. R. (2021). Nonparametric regression mixed estimators of truncated spline and gaussian kernel based on cross-validation (CV), generalized cross-validation (GCV), and unbiased risk (UBR) methods. *International Journal on Advanced Science, Engineering and Information Technology*, *11*(6), 2400. https://doi.org/10.18517/ijaseit.11.6.14464

Sifriyani, S., Dani, A. T. R., Fauziyah, M., Hayati, M. N., Wahyuningsih, S., & Prangga, S. (2023). Spline and kernel mixed estimators in multivariable nonparametric regression for dengue hemorrhagic fever model. *Communications in Mathematical Biology and Neuroscience*, *2023*, 1–15. https://doi.org/10.28919/cmbn/7790

Yuliati, I. F., & Sihombing, P. (2020). Pemodelan fertilitas di indonesia tahun 2017 menggunakan pendekatan regresi nonparametrik kernel dan spline. *Jurnal Statistika dan Aplikasinya*, *4*(1), 48–60. https://doi.org/10.21009/JSA.04105

[This page is intentionally left blank.]