Available online at http://scik.org
Commun. Math. Biol. Neurosci. 2026, 2026:68
https://doi.org/10.28919/cmbn/10003
ISSN: 2052-2541

OPTIMIZING MULTIVARIABLE KERNEL REGRESSION WITH GCV AND UBR FOR SYSTOLIC BLOOD PRESSURE MODELING IN TYPE 2 DIABETES MELLITUS

KHAURA ANISA, SUPARTI*, MASITHOH YESSI ROCHAYANI

Department of Statistics, Diponegoro University, Semarang, Indonesia

Abstract: Random blood glucose and age can influence the increase in systolic blood pressure in patients with type 2 diabetes mellitus. The data pattern forms a random pattern, so the modeling uses multivariable kernel nonparametric regression. This study uses the Nadaraya-Watson estimator because the calculation is simpler and uses the Gaussian kernel function with bandwidth optimization methods are Generalized Cross-Validation (GCV) and Unbiased Risk (UBR). The model with GCV optimization obtained an R^2 of 99.9713%, while the model with UBR optimization obtained an R^2 of 99.9726%, so the model produced by UBR optimization was the best. The model with UBR optimization revealed that systolic blood pressure in patients with type 2 diabetes mellitus was strongly influenced by random blood glucose and age. Evaluation of the UBR optimization model performance on the testing data yielded an Symmetric Mean Absolute Percentage Error (SMAPE) value of 18.05096%. This means that the model performance with UBR optimization has a good ability to predict data.

Keywords: multivariable kernel regression; Nadaraya-Watson; generalized cross-validation; unbiased risk.

2020 AMS Subject Classification: 62G08.

---

*Corresponding author
E-mail address: suparti@live.undip.ac.id
Received May 26, 2026

1. INTRODUCTION

Type 2 diabetes mellitus is a chronic disease characterized by increased random blood glucose due to insulin resistance [1]. Type 2 diabetes mellitus can affect blood pressure because elevated random blood glucose levels impair endothelial function, ultimately leading to blood vessel narrowing, vascular stiffness, and increased blood pressure [2]. High blood pressure in type 2 diabetes mellitus patients is also caused by age due to physiological changes in hormones, the heart and blood vessels along with the aging process [3]. High blood pressure is a chronic disease that causes the number one death in the world, and its condition can be detected through checking systolic blood pressure or diastolic blood pressure (Kementerian Kesehatan, 2024). This study used systolic blood pressure for analysis.

The extent to which random blood glucose levels and age influence systolic blood pressure in patients with type 2 diabetes mellitus can be analyzed using nonparametric regression methods. Nonparametric regression is more flexible than parametric regression because it has less stringent assumptions and can adjust to unknown functional forms [5]. The unknown function form can be estimated using smoothing techniques, one of which is kernel regression for random data patterns [6]. The most commonly used kernel regression function estimator in the kernel regression method is the Nadaraya-Watson estimator because its calculations are simpler [7]. The kernel regression function estimator uses a kernel density estimator to obtain accurate estimates calculated using the kernel function and bandwidth. Types of kernel functions include uniform, triangle, Epanechnikov, and Gaussian functions [8]. The Gaussian kernel function is often used because its calculations are simple and it has the advantage that its weights are defined for all real numbers [9]. Bandwidth is a key determinant of model performance, so a method is needed to obtain optimal bandwidth. Possible methods include Generalized Cross-Validation (GCV) and Unbiased Risk (UBR) [10].

Research conducted by Muhajiriansyah and Binuko [11] on type 2 diabetes mellitus patients provided results that HbA1c levels and random blood glucose levels were related to blood pressure in type 2 diabetes mellitus patients which were tested using the Chi-Square test. Research by Naseri et al. [12] also examined type 2 diabetes mellitus patients and concluded that age and systolic blood pressure had a significant positive relationship, while age and diastolic blood pressure did not have a significant relationship, which was tested using a 2-tailed Pearson correlation. Another study conducted by Rasyid et al. [13] in comparing CV and GCV bandwidth optimization in modeling the Human Development Index with kernel regression showed that the best model was obtained with GCV optimization because the R^2 value was higher and the MAPE value was lower than the model with CV optimization. Azhar et al. [14] conducted a study by comparing UBR and GCV optimization in selecting knot points in spline regression to model the Open Unemployment Rate, providing results that the model with GCV optimization was better than UBR because the R^2 value was higher and the MSE value was lower than the model with UBR optimization.

Based on the description that has been presented, this study analyzes the effect of GDS levels and age on systolic blood pressure in patients with type 2 diabetes mellitus at PKU Muhammadiyah Hospital, Temanggung Regency in 2024. The method used is nonparametric regression with a kernel approach using the Gaussian kernel function and the GCV and UBR optimization methods. This study does not use CV optimization like the study of Rasyid et al. (2025) because if CV optimization is applied in this study, it will produce many undefined CV values because there are observation data with zero weighting. Therefore, this study compares GCV and UBR optimization to determine whether the study of Azhar et al. (2024) [14] which provides results that GCV is superior to UBR in spline regression also applies to kernel regression.

2. PRELIMINARIES

Nonparametric regression is a statistical technique applied to identify patterns of relationships between predictor and response variables without knowing the form of the function or regression curve. The unknown function form can be estimated using smoothing techniques, including kernels, local polynomials, splines, and Fourier series. The nonparametric regression model can be formulated in Equation (1) [15].

$$Y_i = m(x_i) + \varepsilon_i, \quad i = 1,2,\dots,n \tag{1}$$

where:

Y_i : response variable at observation i

$m(x_i)$: unknown function representing the predictor variable at observation i

ε_i : error at observation i with mean = 0 and varians = σ^2

Random data patterns can use kernel regression smoothing techniques. Kernel regression requires a kernel density estimator, which is an estimator used to smooth the data distribution by weighting each observation data using the formula in Equation (2) [16].

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-X_i}{h}\right) \tag{2}$$

where:

$K(.)$: kernel function

h : bandwidth

x : point on the estimated predictor variable

X_i : observation data i on the predictor variable

The kernel density estimator is calculated using the kernel function K(.) and bandwidth h. The kernel function is used to regulate the shape of the kernel weights used, while the bandwidth is used to control how smooth the estimated curve is [8]. The requirements for a kernel function in kernel regression are that it must be continuous, real-valued, symmetrical, bounded, and $\int_{-\infty}^{\infty}K(u)du = 1$ [5]. The kernel function is symmetric about the line $u = 0$, so that it can be written $\int_{-\infty}^{\infty}uK(u)du = 0$. The kernel function scaled with bandwidth (h) can be formulated in Equation (3) [16].

$$K_h(u) = \frac{1}{h}K\left(\frac{u}{h}\right), \quad -\infty < u < \infty \text{ dan } h > 0 \tag{3}$$

The value $u = x - X_i$, so the kernel function equation which is scaled by bandwidth (h) and translated as far as X_i on the x axis can be written as Equation (4).

$$K_h(x - X_i) = \frac{1}{h}K\left(\frac{x-X_i}{h}\right) \tag{4}$$

The types of kernel functions used can be seen in Table 1 [8].

Table 1. Types of Kernel Functions

No	Kernel Function Name	Kernel Function Form
1	Uniform	$K(u) = \begin{cases} \frac{1}{2}, & \text{for } \lvert u\rvert \le 1 \\ 0, & \text{for other } u \end{cases}$
2	Triangle	$K(u) = \begin{cases} 1-\lvert u\rvert, & \text{for } \lvert u\rvert \le 1 \\ 0, & \text{for other } u \end{cases}$
3	Epanechnikov	$K(u) = \begin{cases} \frac{3}{4}-(1-u^2), & \text{for } \lvert u\rvert \le 1 \\ 0, & \text{for other } u \end{cases}$
4	Gaussian	$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}, \text{for } \lvert u\rvert < \infty$

This study uses the Gaussian kernel function with its advantage that all data have defined weights with data that is far from the observation data will have a smaller weight [9]. The kernel density estimator can be used to estimate multivariable data, namely data with more than one predictor variable, for example as many as p predictor variables. The kernel density estimator formula in p-dimensional space can be formulated in Equation (5) [8].

$$\hat{f}_{h_1,h_2,\dots,h_p}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\left[\prod_{d=1}^{p}\frac{1}{h_d}K\left(\frac{x_d-X_{id}}{h_d}\right)\right] \tag{5}$$

Kernel regression is a nonparametric regression to estimate the conditional expectation of a random variable that satisfies Equation (1) [5]. The unknown function $m(x_i)$ can be estimated using the Nadaraya-Watson estimator. Estimation of the multivariable kernel regression function with one response variable (Y_i) and p the predictor variables, namely $\mathbf{X_i} = (X_{i1}, X_{i2}, \cdots, X_{ip})$ with $i = 1,2,\cdots,n$ can be seen in Equation (6) [8].

$$\hat{m}(\mathbf{x}) = \frac{\sum_{i=1}^{n}\left[\prod_{d=1}^{p}K\left(\frac{x_d-X_{id}}{h_d}\right)\right]Y_i}{\sum_{i=1}^{n}\left[\prod_{d=1}^{p}K\left(\frac{x_d-X_{id}}{h_d}\right)\right]} \tag{6}$$

where:

$\hat{m}(\mathbf{x})$: multivariable kernel regression function estimation

$K(\cdot)$: kernel function

x_d : observation point of the d predictor variable

X_{id} : i observation point of the d predictor variable

h_d : bandwidth d

Y_i : i observation data on the response variable

n : number of observations

p : number of predictor variables

The bandwidth value significantly influences the estimation results. A bandwidth value that is too small can result in the estimation function experiencing overfitting, a condition where the model follows the training data pattern too closely, resulting in high accuracy on the training data but low accuracy on the testing data. A bandwidth value that is too large can result in the estimation function experiencing underfitting, a condition where the model is too simple in capturing the training data pattern, resulting in low accuracy on the training data, while accuracy on the testing data can be high or low depending on the data pattern [16]. Therefore, an optimization method is needed to obtain the optimal bandwidth, such as the Generalized Cross-Validation (GCV) and Unbiased Risk (UBR) methods. The optimal bandwidth with the GCV method for kernel regression is selected with the smallest GCV value. The GCV formula can be written in Equation (7) [8].

$$\mathrm{GCV}(h) = \frac{1}{n}\frac{n^2\mathrm{MSE}(h)}{\left[n-\sum_{j=1}^{n} H_{jj}\right]^2} \tag{7}$$

where:

$\mathrm{GCV}(h)$: Generalized Cross-Validationvalue with a certain bandwidth (h)

H_{jj} : kernel weight for the j observation point on the predictor variable with the j observation data on the predictor variable

The $\mathrm{MSE}(h)$ value is the Mean Squared Error value with bandwidth (h) which can be formulated in Equation (8).

$$\mathrm{MSE}(h) = \frac{1}{n}\sum_{j=1}^{n}\left(Y_j - \widehat{m}_h(X_j)\right)^2 \tag{8}$$

The weight H_{ji} is an element of the hat matrix $\mathbf{H}(h)$ [5]. The formula for H_{ji} can be written in Equation (9).

$$H_{ji} = \frac{K\left(\frac{X_j - X_i}{h}\right)}{\sum_{k=1}^{n} K\left(\frac{X_j - X_k}{h}\right)} \tag{9}$$

The UBR optimization method in kernel regression is a method used to determine the optimal bandwidth by considering risk and bias. The optimal bandwidth selected is the bandwidth with the smallest UBR value. The UBR formula can be written in Equation (10).

$$\mathrm{UBR}(h) = \frac{1}{n}\|(\mathbf{I}-\mathbf{H}(h))\mathbf{Y}\|^2 + \frac{2\sigma^2}{n}\mathrm{tr}\,(\mathbf{H}(h)) \tag{10}$$

If the error variance (σ^2) is not known, you can estimate the formula in Equation (11).

$$\hat{\sigma}^2 = \frac{\|(\mathbf{I}-\mathbf{H}(h))\mathbf{Y}\|^2}{n-\mathrm{tr}\,(\mathbf{H}(h))} \tag{11}$$

where:

$\mathbf{H}(h)$: $n \times n$ size hat matrix

$\mathrm{tr}\,(\mathbf{H}(h))$: sum of all diagonal elements $\mathbf{H}(h)$, $\mathrm{tr}\,(\mathbf{H}(h)) = \sum_{j=1}^{n} H_{jj}$

$\mathbf{Y}$: vector of response variable observation data, $\mathbf{Y} = (Y_1, Y_2, \cdots, Y_n)$

$\hat{\sigma}^2$: estimation of error variance

The model with optimal bandwidth of GCV and UBR optimization methods can be tested for its goodness of fit using the coefficient of determination R^2. The coefficient of determination is a method used to determine how much the predictor variable is able to explain the response variable [17]. The value of the coefficient of determination is in the range $0 \le R^2 \le 1$. The closer the R^2 value is to 1, the stronger the model's ability. The formula for the coefficient of determination can be seen in Equation (12).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i-\bar{Y})^2} \tag{12}$$

where:

R^2 : coefficient of determination

Y_i : actual data of the i observation on the response variable

$\hat{Y}_i$: the predicted data of the i observation on the response variable with $\hat{Y}_i = \hat{m}(\mathbf{X_i})$

$\bar{Y}$: the average of the actual data of the response variable

The criteria for the coefficient of determination according to Ndruru et al. [18] are that the model is said to have very strong ability if the R^2 value is in the range of $0.8 \le R^2 < 1.0$, strong model ability is in the range of $0.6 \le R^2 < 0.8$, moderately strong model ability is in the range of $0.4 \leq \mathrm{R}^2 < 0.6$, low model ability is in the range of $0.2 \leq \mathrm{R}^2 < 0.4$, while very low model ability is in the range of $0 \leq \mathrm{R}^2 < 0.2$.

The Symmetric Mean Absolute Percentage Error (SMAPE) value can be used to evaluate model performance. The SMAPE method evaluates how large the error of the predicted results is compared to the actual data. The SMAPE calculation formula is in accordance with Equation (13).

$$\mathrm{SMAPE} = \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{|\mathrm{Y_i} - \hat{\mathrm{Y}}_\mathrm{i}|}{\frac{1}{2}(|\mathrm{Y_i}| + |\hat{\mathrm{Y}}_\mathrm{i}|)} \times 100\% \tag{13}$$

A smaller SMAPE value indicates a more accurate model performance in forecasting. The SMAPE criteria according to Hu et al. [19] are that the model has excellent forecasting ability with a SMAPE value ≤ 10%, good forecasting ability has a range of 10% < SMAPE ≤ 20%, fair forecasting ability is in the range of 20% < SMAPE ≤ 50%, and poor forecasting ability has a SMAPE value > 50%.

The research data comes from PKU Muhammadiyah Temanggung Hospital in the form of secondary data collected through the medical records unit. The data used are data from patients with type 2 diabetes mellitus who were examined in 2024 and have data on random blood glucose levels, age, and systolic blood pressure for study. The research data used were 135 data which were randomly divided into training data and testing data with a ratio of 85:15. The variables in this study consisted of one response variable, systolic blood pressure (Y) measured in mmHg, and two predictor variables, random blood glucose (X_1) measured in mg/dL and age (X_2) expressed in years. Stages of multivariable kernel nonparametric regression analysis using R-studio software with the following steps:

1. Collect the data and identify the variables required for the study.
2. Perform descriptive statistical analysis to provide an overview of the data.
3. Split the dataset into training and testing sets using an 85:15 ratio.
4. Visualize the training data using two-dimensional and three-dimensional scatterplots to examine the data distribution.
5. Select the kernel function to be used in the analysis.
6. Evaluate a range of bandwidth values through trial-and-error to identify candidate

bandwidths for the training data.

7. Estimate the response variable $(\widehat{m}(\mathbf{x}))$ for each candidate bandwidth using Equation (6).
8. Compute the GCV and UBR values for each estimated model using Equations (7) and (10), respectively.
9. Select the optimal model based on the minimum GCV and UBR values.
10. Assess the goodness of fit of the obtained model by calculating the coefficient of determination (R^2) using Equation (12).
11. Use the selected model to estimate the response variable $(\widehat{m}(\mathbf{x}))$ for the testing data using Equation (6).
12. Evaluate the predictive performance of the selected model by calculating the SMAPE value for the testing data using Equation (13).

3. Main Results

The research data was randomly divided into training data and testing data. The training data was used for regression modeling, while the testing data was used to evaluate the performance of the best model obtained from the training data modeling. The training data used a proportion of 85% of the total data, consisting of 114 data on type 2 diabetes mellitus patients. The testing data used a proportion of 15% of the total data, consisting of 21 data on type 2 diabetes mellitus patients. Descriptive statistics of the entire research data can be seen in Table 2.

Table 2. Descriptive Statistics

Statistics	X_1 (mg/dL)	X_2 (Year)	Y (mmHg)
Minimum Value	93	24	106
Maximum Value	638	88	239
Mean	283.7	59.86	149.9
Median	266	60	146

Based on Table 2, the number of observations used was 135 data from type 2 diabetes mellitus patients at PKU Muhammadiyah Hospital Temanggung in 2024 who had systolic blood pressure

(Y) ranging from 106 mmHg to 239 mmHg with an average of 149.9 mmHg and a median of 146 mmHg. The average systolic blood pressure value was ≥140 mmHg indicates that most patients with type 2 diabetes mellitus have high blood pressure. Random blood glucose levels (X_1) in patients with type 2 diabetes mellitus ranged from 93 mg/dL to 638 mg/dL, with an average of 283.7 mg/dL and a median of 266 mg/dL. This average value indicates that most patients had blood glucose levels above the normal limit, which is ≥200 mg/dL. The age variable (X_2) ranged from 24 to 88 years, with an average of 59.86 years and a median of 60 years. This age indicates that the patients were in the adult to elderly age group.

Data plots using the training data for modeling can be seen in three dimensions in Figure 1 and in two dimensions in Figure 2. Based on Figure 1 and Figure 2, it can be seen that the relationship between random blood glucose and age on systolic blood pressure forms a random pattern so it is suitable to use multivariable kernel nonparametric regression. Multivariable kernel regression is influenced by the kernel function and bandwidth. This study uses a Gaussian kernel function with a bandwidth tested on the random blood glucose variable (X_1) namely a lower limit of 1 and an upper limit of 5 with an increase of 0.1 while the bandwidth tested on the age variable (X_2) namely a lower limit of 0.1, an upper limit of 3, and an increase of 0.1. The combination of bandwidths tested will form a model of 1230. The selected model is the model that has the optimal bandwidth using the GCV and UBR methods.

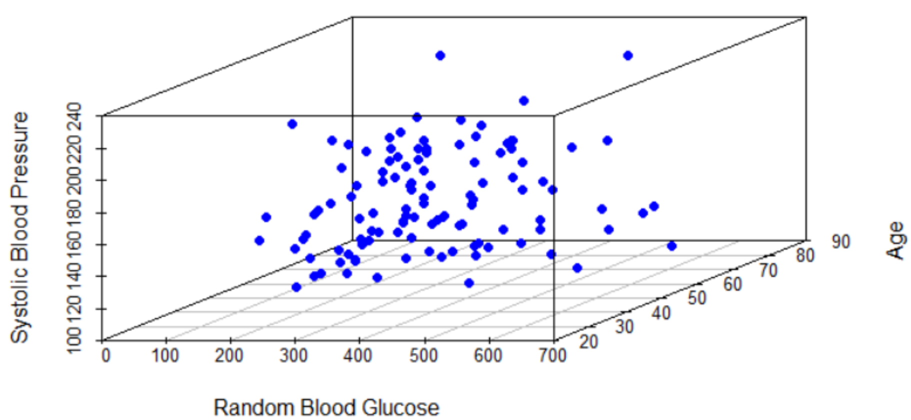
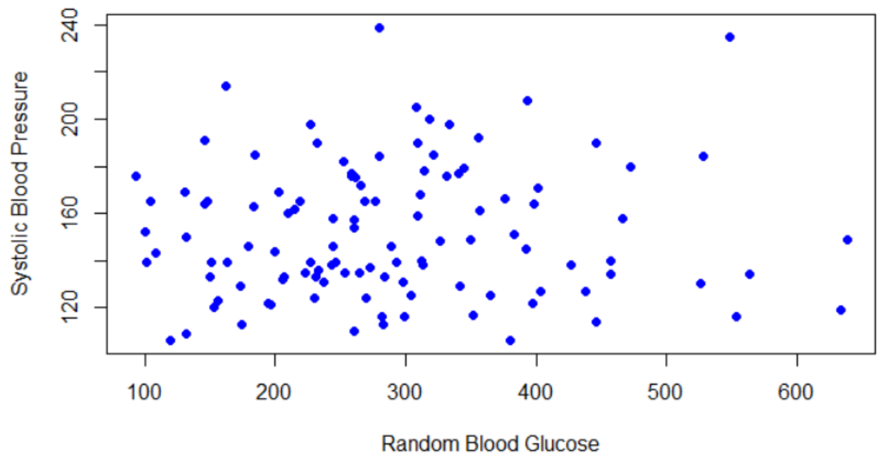
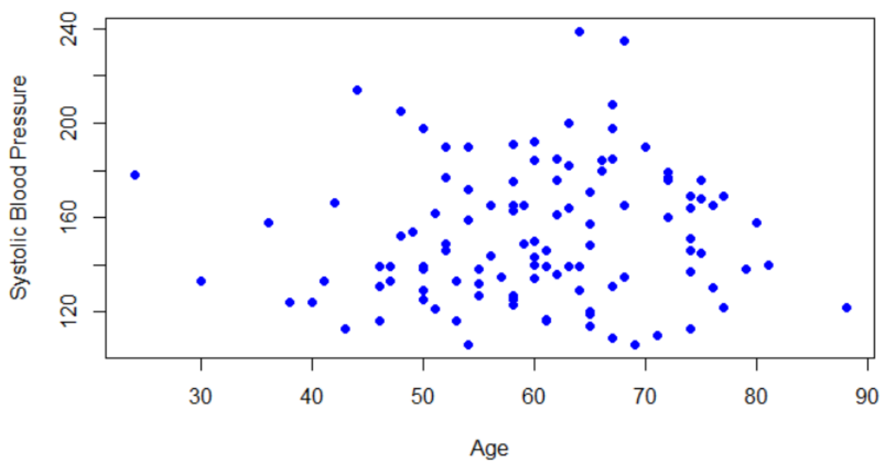


FIGURE 1. 3D Data Plot of Predictor Variables against Response Variables



(a)



(b)

FIGURE 2. (a) 2D Plot of Random Blood Glucose and Systolic Blood Pressure; (b) 2D Plot of Age and Systolic Blood Pressure

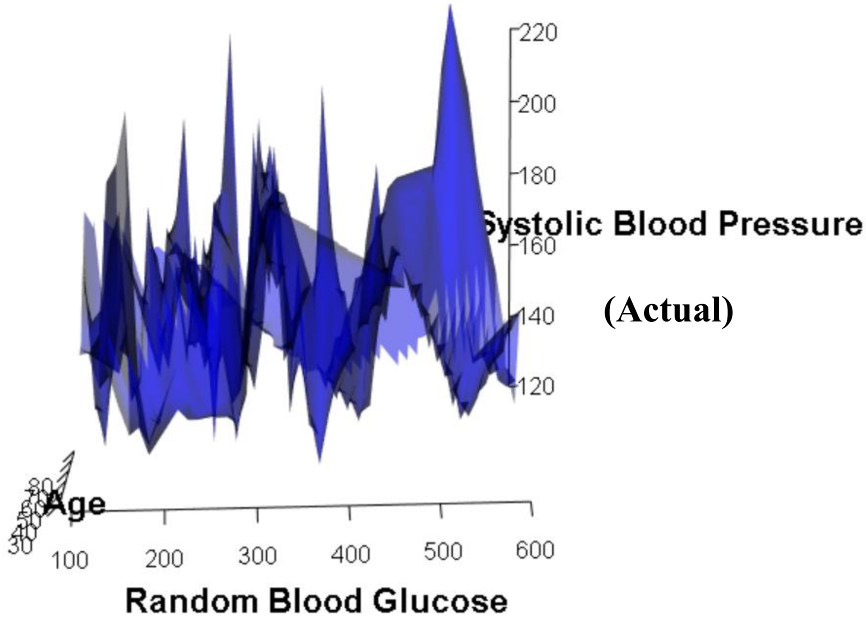
The bandwidth combinations that produce the 10 smallest GCV values can be seen in Table 3. Based on Table 3, the smallest GCV value is 283.2651 which is located in the 333^{rd} combination with a bandwidth value for the random blood glucose predictor variable of 2.1 and a bandwidth for the age predictor variable of 0.3. The model of multivariable kernel regression with the optimal bandwidth of the GCV method and Gaussian function can be written in Equation (14).

$$\hat{m}(\mathbf{x}) = \frac{\sum_{i=1}^{114}\left\{\frac{1}{\sqrt{2\pi}}e^{\frac{-\left(\frac{x_1-X_{i1}}{2.1}\right)^2}{2}}\right\}\left\{\frac{1}{\sqrt{2\pi}}e^{\frac{-\left(\frac{x_2-X_{i2}}{0.3}\right)^2}{2}}\right\}Y_i}{\sum_{i=1}^{114}\left\{\frac{1}{\sqrt{2\pi}}e^{\frac{-\left(\frac{x_1-X_{i1}}{2.1}\right)^2}{2}}\right\}\left\{\frac{1}{\sqrt{2\pi}}e^{\frac{-\left(\frac{x_2-X_{i2}}{0.3}\right)^2}{2}}\right\}} \tag{44}$$

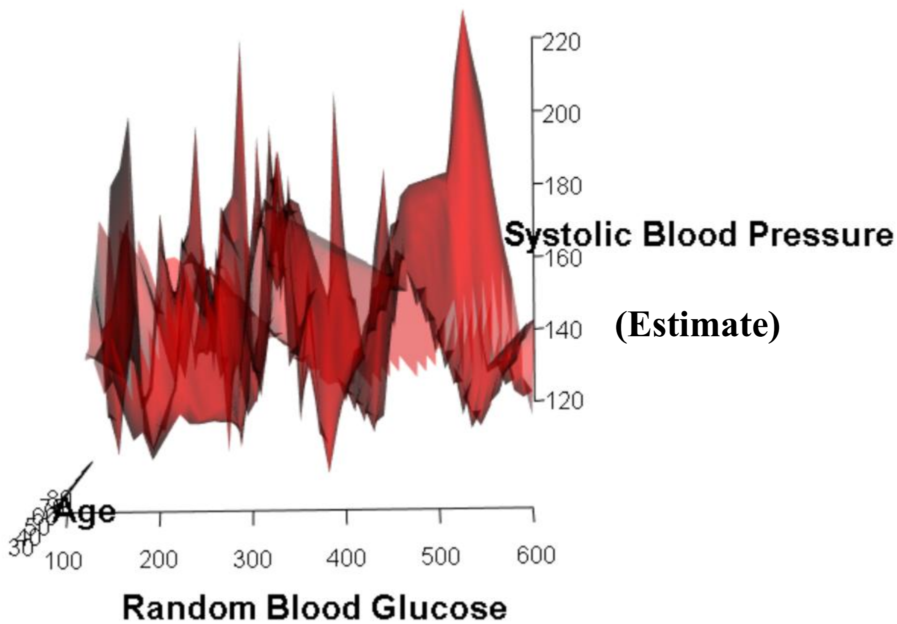
Table 3. Combination of Bandwidth and GCV Value

Combination of	h_1	h_2	GCV Value
333	2.1	0.3	283.2651
303	2.0	0.3	284.0018
332	2.1	0.2	285.5187
331	2.1	0.1	285.5219
302	2.0	0.2	286.1773
301	2.0	0.1	286.1804
273	1.9	0.3	287.3938
363	2.2	0.3	287.4995
272	1.9	0.2	289.5137
271	1.9	0.1	289.5168

The goodness of fit of the obtained model can be assessed using the coefficient of determination (R^2). The R^2 for the model with GCV optimization is 0.999713. This means that systolic blood pressure is influenced by random blood glucose levels and age by 0.999713 or 99.9713% while the remaining 0.0287% is influenced by other factors not tested in this study. Based on the coefficient of determination criteria, the model formed from GCV optimization is very strong. The curve of the actual value with the estimated value of the response variable can be seen visually in Figure 3 and Figure 4.



(a)



(b)

FIGURE 3. (a) Actual Value Plot of GCV Optimization Training Data; (b) Estimated Value Plot of GCV Optimization Training Data

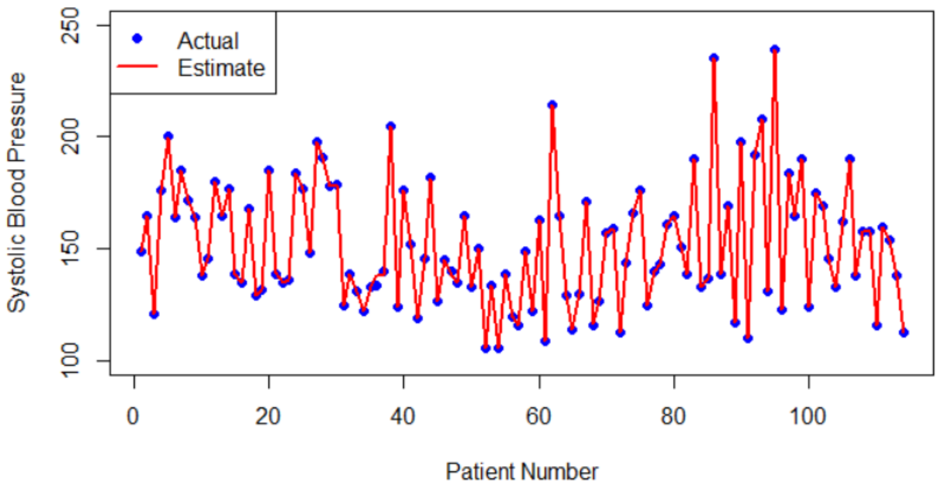


FIGURE 4. 2D for Actual Value and Estimated Value of GCV Optimization Training Data

The blue area in Figure 3 (a) shows the actual value, while the red area in Figure 3 (b) shows the estimated value. Based on Figures 3 (a) and (b), the estimated value resembles the actual value, indicating that the GCV-optimized model is very robust in explaining the response variable based on the predictor variables. A two-dimensional plot of the actual value versus the estimated value can also be seen in Figure 4, which shows that the red line (estimated value) approaches the blue point (actual value), indicating that the GCV-optimized model is very robust in explaining the response variable based on the predictor variables.

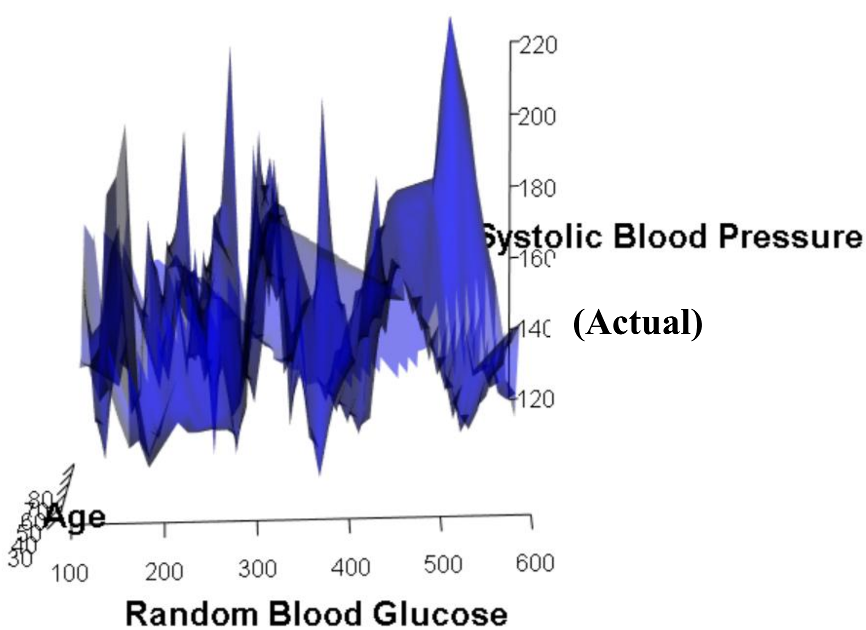
Table 4. Combination of Bandwidth and UBR Values

Combination of-	h_1	h_2	UBR Value
272	1.9	0.3	15.44431
303	2.0	0.3	15.46537
272	1.9	0.2	15.47515
271	1.9	0.1	15.47523
302	2.0	0.2	15.49767
301	2.0	0.1	15.49775
243	1.8	0.3	15.50916
242	1.8	0.2	15.53886
241	1.8	0.1	15.53894
213	1.7	0.3	15.59405

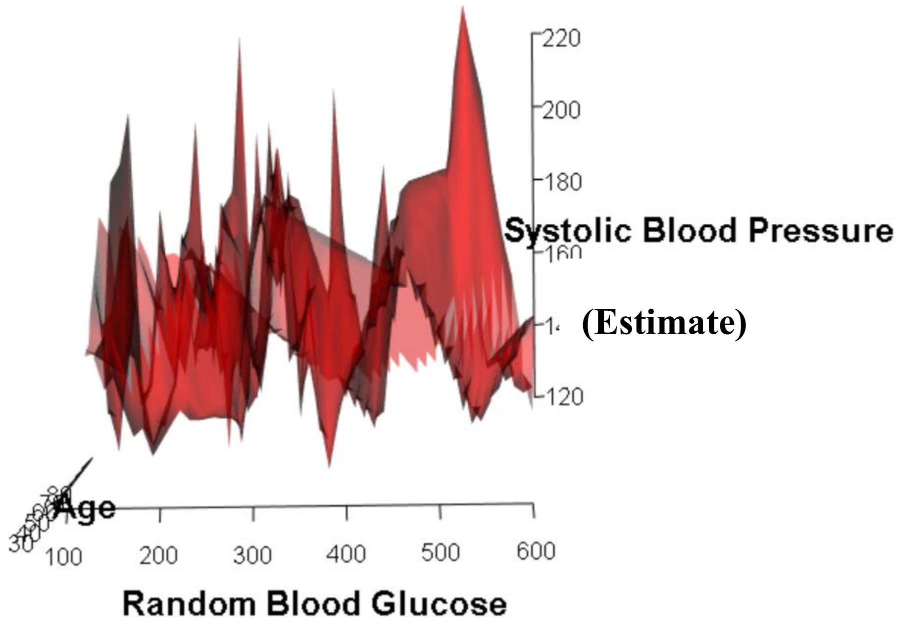
The bandwidth combinations that produce the 10 smallest UBR values are shown in Table 4. Based on Table 4, the smallest UBR value is 15.44431 which is located in the 272^{nd} combination with the bandwidth value for the GDS predictor variable being 1.9 and the bandwidth for the age predictor variable being 0.3. The model of multivariable kernel regression with the optimal bandwidth of the UBR method and using the Gaussian function can be written in Equation (15).

$$\hat{m}(\boldsymbol{x}) = \frac{\sum_{i=1}^{114}\left\{\frac{1}{\sqrt{2\pi}}e^{\frac{-\left(\frac{x_1-X_{i1}}{1.9}\right)^2}{2}}\right\}\left\{\frac{1}{\sqrt{2\pi}}e^{\frac{-\left(\frac{x_2-X_{i2}}{0.3}\right)^2}{2}}\right\}Y_i}{\sum_{i=1}^{114}\left\{\frac{1}{\sqrt{2\pi}}e^{\frac{-\left(\frac{x_1-X_{i1}}{1.9}\right)^2}{2}}\right\}\left\{\frac{1}{\sqrt{2\pi}}e^{\frac{-\left(\frac{x_2-X_{i2}}{0.3}\right)^2}{2}}\right\}} \tag{15}$$

The obtained model can be tested for goodness of fit by calculating the coefficient of determination (R^2). The R^2 value for the model with UBR optimization is 0.999726. This means that systolic blood pressure is influenced by GDS levels and age by 0.999726 or 99.9726% while the remaining 0.0274% is influenced by other factors not tested in this study. Based on the coefficient of determination criteria, the model obtained from UBR optimization is very strong. The curve of the actual value with the estimated value of the response variable can be seen visually in Figure 5 and Figure 6.

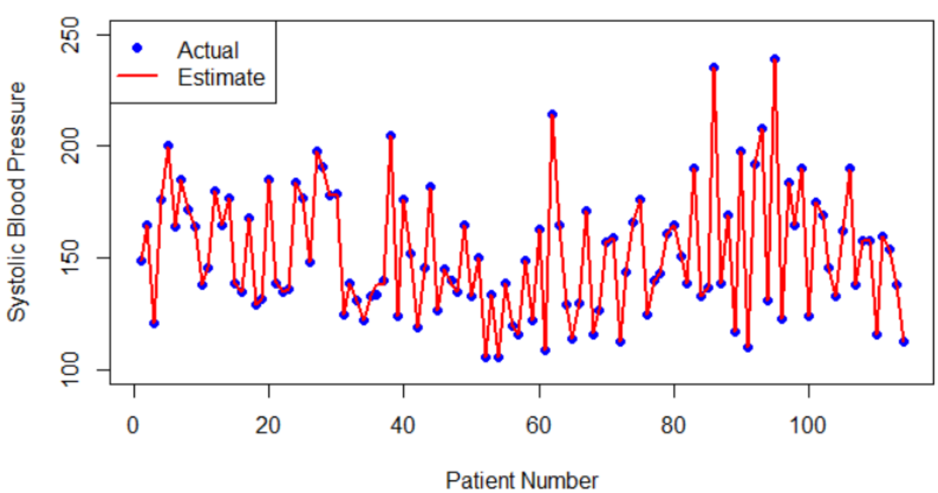


(a)



(b)

FIGURE 5. (a) Actual Value Plot of UBR Optimization Training Data; (b) Estimated Value Plot of UBR Optimization Training Data



(c)

FIGURE 6. 2D for Actual Value and Estimated Value of UBR Optimization Training Data

The blue area in Figure 5 (a) shows the actual value, while the red area in Figure 5 (b) shows the estimated value. Based on Figures 5 (a) and (b), the estimated value resembles the actual value, indicating that the UBR-optimized model is very strong in explaining the response variable based on the predictor variables. A two-dimensional plot of the actual value versus the estimated value can also be seen in Figure 6, which shows that the red line (estimated value) approaches the blue point (actual data), indicating that the UBR optimized model is very strong in explaining the response variable based on the predictor variables.

The best model, formed from GCV and UBR optimization, was selected with the highest R^2 value. The R^2 for the GCV optimization model was 99.9713%, while the UBR optimization model was 99.9726%. The R^2 value for the UBR optimization model was slightly higher than that for the GCV optimization model, indicating that the UBR optimized model was the best model, according to Equation (15).

The performance of the best model was evaluated by measuring the SMAPE value on the testing data to test the model's accuracy in estimating new data. The resulting SMAPE value is 18.05096%, indicating that the best model performance with UBR optimization is able to predict well. The plot between the actual and estimated values in the testing data can be seen visually in Figure 7.

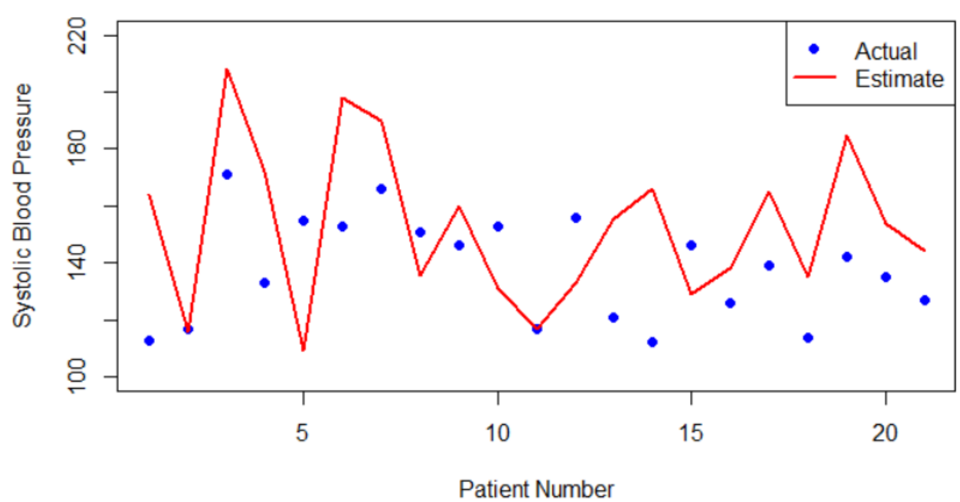


FIGURE 7. Plot of Actual and Estimated Values from UBR Optimization Test Data

The blue dots in Figure 7 represent the actual values, while the red lines represent the estimated values. Based on Figure 7, the estimated values are close to the actual data values, indicating that the UBR-optimized model's forecasting ability is quite good.

4. CONCLUSION

Based on the results and discussion described above, the multivariable kernel regression method with Generalized Cross-Validation (GCV) and Unbiased Risk (UBR) bandwidth optimization can model the effect of random blood glucose levels and age on systolic blood pressure in patients with type 2 diabetes mellitus at PKU Muhammadiyah Hospital, Temanggung, in 2024. The model generated by GCV optimization with bandwidths $h_1 = 2.1$ and $h_2 = 0.3$ had a coefficient of determination of 99.9713%. The model generated by UBR optimization with bandwidths $h_1 = 1.9$ and $h_2 = 0.3$ had a coefficient of determination of 99.9726%. The coefficient of determination of the model generated by GCV and UBR optimization was greater than 80%, indicating that systolic blood pressure was strongly explained by its predictor variables, random blood sugar levels and age.

The kernel regression model with UBR optimization had a slightly higher coefficient of determination than the model with GCV optimization. Therefore, the best model selected was the UBR optimized model. The performance of the best model was evaluated using the SMAPE test on the testing data, resulting in an SMAPE value of 18.05096%. This SMAPE value is less than 20%, indicating that the UBR-optimized model has good estimation capabilities.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests.

REFERENCES

[1] A. Singh, S. Shadangi, P.K. Gupta, S. Rana, Type 2 Diabetes Mellitus: A Comprehensive Review of Pathophysiology, Comorbidities, and Emerging Therapies, Compr. Physiol. 15 (2025), e70003. https://doi.org/10.1002/cph4.70003.

[2] G. Jia, J.R. Sowers, Hypertension in Diabetes: An Update of Basic Mechanisms and Clinical Disease, Hypertension 78 (2021), 1197-1205. https://doi.org/10.1161/HYPERTENSIONAHA.121.17981.

[3] B.F.N.E. Tindangen, F.F.L.G. Langi, N.H. Kapantow, Faktor-Faktor yang Berhubungan dengan Kejadian

Hipertensi pada Guru Sekolah Dasar di Kecamatan Tombariri Timur, Kesmas 9 (2020), 189-196.

[4] Kementerian Kesehatan Republik Indonesia, Bahaya Hipertensi, Upaya Pencegahan dan Pengendalian Hipertensi, Direktorat Pencegahan dan Pengendalian Penyakit Tidak Menular, Jakarta. https://kemkes.go.id/id/bahaya-hipertensi-upaya-pencegahan-dan-pengendalian-hipertensi.

[5] Suparti, R. Santoso, A. Prahutama, A.R. Devi, Regresi Nonparametrik, Wade Group, Ponorogo, 2018.

[6] N. Chamidah, B. Lestari, Analisis Regresi Nonparametrik dengan Perangkat Lunak R, Airlangga University Press, Surabaya, 2022.

[7] N. Wulandari, R.A. Putri, A. Nazhirah, S.A. Novia, Inflation Forecasting for Riau Province: A Comparison of Parametric ARIMA with Nonparametric Nadaraya-Watson Kernel Regression and B-Spline Methods, J. Mat. Stat. Komput. 21 (2025), 763-774. https://doi.org/10.20956/j.v21i3.43647.

[8] W. Härdle, Smoothing Techniques with Implementation in S, Springer Series in Statistics, Springer, New York, 1991. https://doi.org/10.1007/978-1-4612-4432-5.

[9] A. Rahmah, T.W. Utami, A. Fadlurohman, Regresi Nonparametrik Kernel Triangle dan Gaussian dengan Estimator Priestley-Chao untuk Pemodelan Kasus AIDS di Pulau Jawa, Pros. Semin. Nas. Sains Data 5 (2025), 302-313.

[10] A. Prahutama, Suparti, D. Ispriyanti, T.W. Utami, Pemodelan Bivariate Polinomial Lokal pada Jumlah Kematian Ibu dan Bayi di Jawa Tengah, in: Prosiding Seminar Nasional VARIANSI, 2018, pp. 209-220.

[11] Muhajiriansyah, R.S.D. Binuko, Hubungan antara Kadar HbA1c dan Gula Darah Sewaktu (GDS) dengan Kejadian Hipertensi pada Pasien Diabetes Melitus Tipe II di Rumah Sakit Darmayu Ponorogo, Health Inf. J. Penelit. 15 (2023), e1206.

[12] M.W. Naseri, H.A. Esmat, M.D. Bahee, Prevalence of Hypertension in Type-2 Diabetes Mellitus, Ann. Med. Surg. 78 (2022), 103758. https://doi.org/10.1016/j.amsu.2022.103758.

[13] M.R.A. Rasyid, Suparti, M.Y. Rochayani, Optimizing Education-Based HDI Modeling in Indonesia: A Multivariable Kernel Regression Approach with CV and GCV, Int. J. Sci. Eng. Invent. 11 (2025), 52-61. https://doi.org/10.23958/ijsei/vol11-i04/284.

[14] Z.A. Azhar, S.S. Handajani, I. Slamet, Pemodelan Faktor-Faktor yang Mempengaruhi Tingkat Pengangguran Terbuka (TPT) di Provinsi Jawa Tengah Menggunakan Regresi Spline Truncated Multivariabel, J. Sutasoma 2

(2024), 66-73. https://doi.org/10.58878/sutasoma.v2i2.264.

[15] M. Schafgans, V. Zinde-Walsh, Multivariate Kernel Regression in Vector and Product Metric Spaces, J. Econom. 253 (2026), 106168. https://doi.org/10.1016/j.jeconom.2025.106168.

[16] W. Härdle, Applied Nonparametric Regression, Econometric Society Monographs, Cambridge University Press, Cambridge, 1990.

[17] D.N. Gujarati, Basic Econometrics, 4th ed., McGraw-Hill, New York, 2004.

[18] R.E. Ndruru, M. Situmorang, G. Tarigan, Analisa Faktor-Faktor yang Mempengaruhi Hasil Produksi Padi di Deli Serdang, Saintia Mat. 2 (2014), 71-83.

[19] H. Hu, C. Zhao, J. Li, Y. Huang, Stock Prediction Model Based on Mixed Fractional Brownian Motion and Improved Fractional-Order Particle Swarm Optimization Algorithm, Fractal Fract. 6 (2022), 560. https://doi.org/10.3390/fractalfract6100560.