BAREKENG: Journal of Mathematics and Its Applications
September 2025 Volume 19 Issue 3 Page 1525-1536
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol19iss3pp1525-1536

MODELING GENDER DEVELOPMENT INDEX IN SOUTHEAST SULAWESI PROVINCE USING SEMIPARAMETRIC KERNEL REGRESSION

Andi Tenri Ampa[1], Lilis Laome[2*], Muhammad Ridwan [3], Baharuddin[4], Makkulau[5]

[1,2,3,4,5]*Statistics Study Program, Faculty of Mathematics and Natural Sciences, Universitas Halu Oleo Jln. H.E.A. Mokodompit, Kendari, 93232, Indonesia*

*Corresponding author's e-mail: * **lhi2slaome@gmail.com***

Article History:

Received: 30th August 2024
Revised: 27th January 2025
Accepted: 3rd February 2025
Published: 1st July 2025

Keywords:

Bandwidth;
Gender Development Index;
Nadaraya-Watson Estimator;
Semiparametric Regression

ABSTRACT

The issue of gender equality in Southeast Sulawesi still needs further attention, as indicated by the uneven value of the Gender Development Index (GDI) in each district/city in the region. Therefore, an in-depth analysis is needed to identify factors that affect the GDI. One method that can be used is semiparametric regression with the Nadaraya-Watson estimator, which allows modeling the relationship between variables with more flexibility. This study aims to build a semiparametric regression model to identify factors that contribute to HDI in Southeast Sulawesi Province. The results of the analysis showed that the optimal bandwidth values obtained were h_1= 1.57, h_2=0.49, h_3=2.50 and h_4=4.61. The resulting model has an R^2 and MSE values of 99.8% and 0.14% respectively, indicating that the model has high accuracy in explaining the overall variation in GDI.

How to cite this article:

A. T. Ampa, L. Laome, M. Ridwan, Baharuddin and Makkulau., "MODELING GENDER DEVELOPMENT INDEX IN SOUTHEAST SULAWESI PROVINCE USING SEMIPARAMETRIC KERNEL REGRESSION," *BAREKENG: J. Math. & App.,* vol. 19, no. 3, pp. 1525-1536, September, 2025.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

One of the factors that can be used as the competitiveness of a country is the quality of human resources, both in terms of ability, skills, or productivity. Therefore, in realizing a nation that has competitiveness, it is necessary to make an effort to build the quality of these human resources. Human development efforts are intended for the entire population in a country, regardless of gender differences. In the context of human development, there is a term regarding gender-based development that measures development achievements between women and men [1].

Gender issues are a multidimensional subject of discussion. Because gender covers all sides of health, education, and the economy which is the focus of sustainable development [2]. Gender differences do not matter if they are accompanied by justice and equality. However, the fact is that disparities still occur between men and women in Indonesia today. For example, there is still violence experienced by women according to data from the National Commission on Violence Against Women, where in 2019 there were 1,413 cases and then an increase of 60% to 2,389 cases in 2020 (http://komnasperempuan.go.id). This shows that the gender gap still exists. To overcome this, gender equality still needs to be improved, so that in Indonesia gender equality is still a serious problem. Equal opportunities obtained by men and women in human rights such as participating in decision-making, playing a role in various economic, socio-cultural, political activities and equality in enjoying the results of development are referred to as gender equality. Therefore, there is a need for an indicator to see the level of success of a human development that addresses gender issues, namely the Gender Development Index (HDI) [3].

The HDI of Southeast Sulawesi Province has experienced ups and downs over the past few years [4]. The occurrence of this phase in the GDI of Southeast Sulawesi is related to the factors that influence it, as evidenced by the uneven GDI value and has different values. This means that the Southeast Sulawesi provincial government is still struggling to deal with gender development index issues in its region, not only by the Southeast Sulawesi government, but also including the government of each district/city. Therefore, it is necessary to conduct a more in-depth analysis in determining the factors that have the most significant effect on the GPA in Southeast Sulawesi Province.

Regression analysis is one of the methods to determine the relationship between response variables (Y) and predictor variables (X). There are three types of regression analysis approaches, namely parametric, nonparametric, and semiparametric. The parametric approach is an approach with data variables whose patterns are known. While the nonparametric approach is an approach with data variables with smooth curves whose patterns are unknown, so that the data will find its own curve shape [5].

Semiparametric regression combines the advantages of parametric and nonparametric regression which enables it to handle complex data patterns [6]. Research on semiparametric regression has been conducted, including [7] who used spline estimators in the semiparametric regression model. Its application to HIV data, where the results obtained that the level of HIV patients after therapy (y) is influenced by the initial CD4 level (x) and the time of examination (t). [8] studied the use of specific nonlinear functions and tree-based methods, nonparametric components developed on nonlinear smooth functions and tree methods. Furthermore, [9] developed an algorithm for creating semiparametric regression modeling on real data, such as stock market, real estate and airline data to produce real-time analysis. In semiparametric regression, there are several estimators that can be used, namely the Nadaraya Watson Kernel estimator [10], Local Polynomial Kernel [11], Gasser Muller Kernel [12], and others. Each of these estimators has its own characteristics. The Nadaraya-Watson Kernel Estimator is the most widely used estimator because the Nadaraya-Watson estimator is used to handle data that is not normally distributed, nonlinear, and has irregular patterns [13].

Based on the previous description, the purpose of this article is to model the IPG data using a semiparametric regression model with the Nadaraya-Watson Kernel estimator. The data used includes GDI (Y) and the population aged 5 years and above who are no longer in school (X), the gender ratio (Z_1), the open unemployment rate (Z_2), the labor force participation rate (Z_3), and the population with health complaints (Z_4) as predictors [14]. The variables used are expected to represent all fields including educational, economic and social fields.

2. RESEARCH METHODS

2.1 Parametric Regression

Parametric regression analysis is a method for analyzing functions or regression curves of known shape. The following is a multiple linear parametric regression model:

$$y_i = \beta_0 + \sum_{k=1}^{p} \beta_i X_{ik} + \varepsilon_i \; ; \; i = 1,2,\dots,n \; ; \; k = 1,2,\dots,p \tag{1}$$

with y_i is the dependent variable for the i-th observation, X_{ik} is the k-th independent variable at the i-th observation, β_k is the regression coefficient on X_{ik}, β_0 is the regression model constant and ε_i is the i-th residual.

2.2 Nonparametric Regression

The nonparametric approach is a method used to analyze unknown functions or regression curves. Suppose X is the predictor variable and Y is the response variable for n paired observations $\{(x_i, y_i)\}$; $i = 1, 2, \dots, n$, then the linear relationship between the two variables can be known with the nonparametric regression equation model, which is as follows:

$$y_i = m(z_i) + \varepsilon_i \;\; ; \; i = 1,2,\dots,n \tag{2}$$

where y_i is the i-th observation dependent variable, $m(z_i)$ is the unknown regression function and is the i-th residual [5].

2.3 Semiparametric Regression

Semiparametric regression is a combination of parametric and nonparametric regression [15]. The curve estimation is determined based on the behavior of the data, so smoothing techniques are required. In some cases, the response variable may have a linear relationship with one of the predictor variables, but with other predictor variables the pattern of the relationship is unknown. Variables that have a known data pattern, or there is information about the data pattern are classified in the parametric component. Meanwhile, variables with unknown data patterns are classified into nonparametric components [16]. Given a data pair (x_i, z_i, y_i) where $i = 1,2,\dots,n$ the semiparametric regression model can be written as follows:

$$y_i = \mu(x_i, z_i) + \varepsilon_i \tag{3}$$

with the curves assumed to be additive, i.e.:

$$\mu(x_i, z_i) = f(x_i) + m(z_i) \tag{4}$$

Therefore, equation 3 can be written as:

$$y_i = f(x_i) + m(z_i) + \varepsilon_i \tag{5}$$

Where

$$m(z_i) = \frac{\sum_{i=1}^{n} K\left(\frac{z_j - z_{ji}}{h_j}\right)\left(y_i - f(x_i)\right)}{\sum_{i=1}^{n} K\left(\frac{z_j - z_{ji}}{h_j}\right)} \tag{6}$$

Such that

$$\hat{y}_i = \sum_{k=1}^{p} \beta_k X_{ik} + \sum_{k=1}^{q} m_k(z_{ki}) + \varepsilon_i \;\; ; \; i = 1,2, \quad \dots,n \; ; \; k = 1,2,\dots,p \tag{7}$$

where x_i is the parametric dependent variable, z_i is the nonparametric dependent variable, and y_i is the dependent variable of the i-th observation, $\sum_{k=1}^{p} \beta_k X_{ik}$ the parametric component, $\sum_{k=1}^{p} m_k(z_{ki})$ while is the unknown regression function or is the nonparametric component, and ε_i is the random error where $\varepsilon_i \sim N(0,2)$ [17].

2.4 Estimation of Semiparametric Kernel Regression Model

In the semiparametric regression model in **Equation (7)**, m and β are functions and parameters to be estimated from the data. The estimation uses a kernel estimator and parameter estimation uses the least squares method. Suppose if there are n observations with p independent variables x and 1 independent variable z, then **Equation (6)** can be written as follows:

$$y_i - \sum_{k=1}^{p} \beta_k X_{ik} = m(z_i) + \varepsilon_i \;\; ; \; i = 1,2,\dots,n \; ; \; k = 1,2,\dots,p \tag{8}$$

If $y_i^* = y_i - \sum_{k=1}^{p} \beta_k X_{ik}$, then the value of $m(z)$ is equivalent to the expected value of the response variable when the variable $T = t$ is known. It is assumed that the response and predictor variables are random variables. Mathematically written:

$$m(z) = E(y^*|T = t) = \int_{-\infty}^{\infty} y^* m(y|z)dy = \frac{\int_{-\infty}^{\infty} y^* m(y|z)dy^*}{m(z)} \tag{9}$$

Furthermore, by using the Nadaraya-Watson estimator such as kernel nonparametric regression, the following equation is obtained:

$$\hat{m}(z) = \frac{\sum_{i=1}^{n} K\left(\frac{z - z_i}{h}\right) y_i^*}{\sum_{i=1}^{n} K\left(\frac{z - z_i}{h}\right)} = \sum_{i=1}^{n} W_i(z)\left(y_i - \sum_{k=1}^{p} \beta_k x_{ik}\right) \tag{10}$$

Thus obtained $\hat{m}(z)$ to estimate $m(z)$ with the estimation model.

$$y_i = \sum_{k=1}^{p} \beta_k x_{ik} + \sum_{k=1}^{p} m_k(z_{ki}) + \varepsilon_i \tag{11}$$

Equation (11) can also be written as $Y = X\beta + \hat{M}(z) + \varepsilon$. According to **[17]** with the least squares method, the estimate of β is obtained by minimizing the sum of the squared errors, which is as follows:

$$\begin{aligned} \varepsilon^T\varepsilon &= \big(Y - X\beta - W(z)\big)^T\big(Y - X\beta - W(z)\big) \\ &= (Y^T - \beta^T X^T + Y^T W(z)^T)^T(Y - X\beta + W(z)y) \end{aligned}$$

Next minimize $\varepsilon^T\varepsilon$ to get the value of $\hat{\beta}$ by differentiating against β

$$\begin{aligned} \frac{\partial \varepsilon^T\varepsilon}{\partial \beta} &= \frac{\partial}{\partial \beta}(y - X\beta + W(z)y)^T(y - X\beta + W(z)y) \\ &= -2X^TY + 2X^TX\hat{\beta} - 2X^TW(z)y \\ \hat{\beta} &= (X^TX)^{-1}X^T(W(z) + I)y \end{aligned} \tag{12}$$

Where

$$X = \begin{bmatrix} 1 & x_1 & z_{11} & \cdots & z_{41} \\ 1 & x_2 & z_{12} & \cdots & z_{42} \\ 1 & x_3 & z_{13} & \cdots & x_{43} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & z_{1n} & \cdots & x_{4n} \end{bmatrix}; y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}; I_{n\times n} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

$$W(z) = \sum_{k-1}^{q} P_k(h_k)$$

where:

$$P_k(h_k)_{n\times n} = \begin{bmatrix} n^{-1}\dfrac{\frac{1}{h_j}K\left(\frac{z_{j1}-z_{j1}}{h_j}\right)}{n^{-1}\sum_{i=1}^{n}\frac{1}{h_j}K\left(\frac{z_{j1}-z_{ji}}{h_j}\right)} & n^{-1}\dfrac{\frac{1}{h_j}K\left(\frac{z_{j2}-z_{j2}}{h_j}\right)}{n^{-1}\sum_{i=1}^{n}\frac{1}{h_j}K\left(\frac{z_{j2}-z_{ji}}{h_j}\right)} & \cdots & n^{-1}\dfrac{\frac{1}{h_j}K\left(\frac{z_{j3}-z_{jn}}{h_j}\right)}{n^{-1}\sum_{i=1}^{n}\frac{1}{h_j}K\left(\frac{z_{j3}-z_{ji}}{h_j}\right)} \\ n^{-1}\dfrac{\frac{1}{h_j}K\left(\frac{z_{j2}-z_{j1}}{h_j}\right)}{n^{-1}\sum_{i=1}^{n}\frac{1}{h_j}K\left(\frac{z_{j2}-z_{ji}}{h_j}\right)} & n^{-1}\dfrac{\frac{1}{h_j}K\left(\frac{z_{j2}-z_{j2}}{h_j}\right)}{n^{-1}\sum_{i=1}^{n}\frac{1}{h_j}K\left(\frac{z_{j2}-z_{ji}}{h_j}\right)} & \cdots & n^{-1}\dfrac{\frac{1}{h_j}K\left(\frac{z_{j3}-z_{jn}}{h_j}\right)}{n^{-1}\sum_{i=1}^{n}\frac{1}{h_j}K\left(\frac{z_{j3}-z_{ji}}{h_j}\right)} \\ \vdots & \vdots & \ddots & \vdots \\ n^{-1}\dfrac{\frac{1}{h_j}K\left(\frac{z_{jn}-z_{j1}}{h_j}\right)}{n^{-1}\sum_{i=1}^{n}\frac{1}{h_j}K\left(\frac{z_{jn}-z_{ji}}{h_j}\right)} & n^{-1}\dfrac{\frac{1}{h_j}K\left(\frac{z_{jn}-z_{j2}}{h_j}\right)}{n^{-1}\sum_{i=1}^{n}\frac{1}{h_j}K\left(\frac{z_{jn}-z_{ji}}{h_j}\right)} & \cdots & -1\dfrac{\frac{1}{h_j}K\left(\frac{z_{jn}-z_{jn}}{h_j}\right)}{n^{-1}\sum_{i=1}^{n}\frac{1}{h_j}K\left(\frac{z_{jn}-z_{ji}}{h_j}\right)} \end{bmatrix}$$

So that the equation model is obtained

$$\hat{y} = X\beta + \widehat{M}(z) + \varepsilon \tag{13}$$

$$\hat{y} = (M^* + N^*)y \tag{14}$$

where:

$$M^* = X(X^TX)^{-1}X^T(2W(z) + I)$$

$$N^* = \sum_{i=1}^{n} W_i(z)\left(y_i - \sum_{k=1}^{p}\beta_k x_{ik}\right)$$

2.5 Determination of Optimal Bandwidth

According to [18], to obtain an optimal curve, it is necessary to smooth the curve using the most optimal bandwidth. The selection of the optimum bandwidth in this study uses Silverman's Rule of Thumb, where the estimate used is the standard deviation and interquartile range divided by 1.34 to overcome non-linear density.

In its application, [19] gives the kernel width (bandwidth) with the following formula:

$$h = 0.9 \times min\left\{S, \frac{IQR}{1.34}\right\} \times n^{-\frac{1}{5}} \tag{15}$$

With

h : bandwidth

S : standard deviation of the variable

IQR : $Q_3 - Q_1$

n : amount of data

2.6 Test for Linearity

One of the assumptions often used in regression is to look at the linearity of the data (Test for linearity Statistic). Thus, linearity testing is required, to show that the variables being tested have a linear relationship with each other. This basic assumption of linearity is only required for linear regression models. If the data between variables is not linear with the data of other variables, then the regression model must use non-linear regression methods [20].

The hypothesis is that if the F value of the linearity test shows the F-count $>$ F-table value or a significance value smaller than the alpha level (0.05), then the relationship between the two sample groups is linear. Conversely, if the F-count value $<$ F-table, it will be insignificant because the significance value of F is greater than the alpha level (0.05) [20]. The formula for finding F count is as follows:

$$F = \frac{\dfrac{R^2}{(k-1)}}{\dfrac{1-R^2}{(n-k)}} \tag{16}$$

With

F : F-value

R^2 : Coefficient of Determination

k : Number of Variables

n : Number of Observations

2.7 Evaluation of Model Goodness

One of the objectives of regression analysis is to get the best model that can explain the relationship between predictor variables and response variables based on certain criteria. One of the criteria used in selecting the best model is by using the coefficient of determination R^2. In general, the greater the R^2 value, the better the model obtained. According to [21], [22], the coefficient of determination is defined as follows:

$$R^2 = 1 - \frac{JKG}{JKT} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{17}$$

With $\hat{y}_i$ is the estimated value of y_i in the semiparametric regression model with optimal bandwidth and $\bar{y}$ is the average of y values.

Mean Absolute Percentage Error (MAPE) is used to measure the accuracy of the estimation. The smaller the percentage error, the greater the forecasting accuracy. According to [15] to determine the MAPE value can use the following formula:

$$MAPE = \frac{1}{n}\left(\sum_{i=1}^{n}\frac{|\hat{y}_i - y_i|}{y_i}\right) \times 100\% \tag{18}$$

Based on the previous description, this research will examine semiparametric kernel regression in modeling the factors affecting the Gender Development Index in Southeast Sulawesi Province with the Nadaraya-Watson estimator method on the Gaussian Kernel function with a measure of model goodness using MAPE and R^2.

2.8 Source of Data

The data used in this study were obtained from the 2022 publication of the Central Bureau of Statistics (BPS). One of the reasons for selecting 2022 data is that, unlike the previous year, offline activities had resumed, potentially influencing the observed trends. The observation units in this study consist of 17 districts/cities in Southeast Sulawesi Province. The variables analyzed in this study include

Table 1. Research Variables

Variable		Variable Name	Units
Dependent	y	Gender Development Index (GDI)	Index
Independent	x	Population aged 5 years and over who are no longer in school	Percent
	z_1	Sex Ratio	Index
	z_2	Open Unemployment Rate	Percent
	z_3	Labor Force Participation Rate	Percent
	z_4	Population with Health Concerns	Percent

2.9 Research Steps

In this study using R-Studio software. The following steps were taken:

1. Collecting data.
2. Creating descriptive statistics.
3. Identifying data by creating a scatterplot
4. Conduct a linearity test for each independent variable against the dependent variable. From the results of the linearity test and scatterplot can determine the parametric component variables and nonparametric components.

5. Determine the parameter β using the Weight Least Square method
6. Determining the bandwidth (h) using Silverman's Rule of Thumb method.
7. Determining the multivariable kernel semiparametric regression model.
8. Determining the coefficient of determination (R^2)
9. Determining the Mean Absolute Percentage Error (MAPE) value.
10. Interpreting the results and drawing conclusions

3. RESULTS AND DISCUSSION

3.1 Descriptive Statistical Analysis of Data

In this study, the data used was obtained from the official website of the Central Bureau of Statistics. Data presentation with descriptive statistical methods displays the maximum, minimum, average, standard deviation, quartile, and interquartile values. The data obtained are districts/cities in Southeast Sulawesi. The following is a table of descriptive statistics for each research variable.

Table 2. Descriptive Statistics

Variable	Mean	StDev	Min	Q1	Median	Q3	Max	IQR
y	88.92	6.27	73.43	85.94	89.05	93.17	99.29	7.24
x	70.28	1.66	67.76	68.83	70.40	71.13	73.93	2.30
z_1	102.34	3.09	96.16	99.98	101.96	104.79	106.97	4.81
z_2	3.05	1.13	1.47	2.16	2.86	3.85	5.39	1.69
z_3	70.77	5.85	61.14	66.61	69.46	75.28	82.12	8.67
z_4	30.84	10.99	17.23	23.51	30.95	35.63	60.62	12.12

Table 2 characterizes the variable (y) which is the Gender Development Index (GDI) in Southeast Sulawesi Province and the variables which are the Population Aged 5 Years and Over Who Are No Longer in School (x), Sex Ratio (z_1), Open Unemployment Rate (z_2), Labor Force Participation Rate (z_3) and Population with Health Complaints (z_4).

Based on the results of **Table 2**, it can be seen that the variable (y), namely the Gender Development Index in Southeast Sulawesi Province, has an average value of 88.92%. The variance of the Gender Development Index data shows a fairly large value of 39.31, this means that the Gender Development Index data has a fairly high variation or the data is diverse. The lowest Gender Development Index of 73.43% is in South Buton Regency. The highest Gender Development Index of 99.29% is in East Kolaka Regency.

The variable (x) Population aged 5 years and over who are no longer in school in Southeast Sulawesi Province has an average value of 70.28%. The variance of the data of Population Aged 5 Years and Over who are no longer in school shows a value of 2.75, this means that the data of Population Aged 5 Years and Over who are no longer in school has sufficient variation or the data is diverse. The lowest percentage of the population aged 5 years and above who are no longer in school, 67.76%, is in South Buton District. The highest percentage of the population aged 5 years and above who are no longer in school, 73.93%, is in Kendari City.

The variable (z_1) Sex Ratio in Southeast Sulawesi Province has an average value of 102.34. The variance of the Sex Ratio data shows a value of 9.54, this means that the Sex Ratio data has considerable variation or the data is diverse. The lowest Sex Ratio of 96.16 is in Muna Regency. The highest Sex Ratio data in Southeast Sulawesi of 106.97 is in Konawe Regency.

The variable (z_2) Open Unemployment Rate in Southeast Sulawesi Province has an average value of 3.04. The variance of the Open Unemployment Rate data shows a value of 1.27, which means that the Open Unemployment Rate data has a small variation or the data is not diverse. The lowest Open Unemployment Rate of 1.47 is in Bombana Regency. The highest Open Unemployment Rate data in Southeast Sulawesi is 5.39 in Bau-bau City.

The variable (z_3) Labor Force Participation Rate in Southeast Sulawesi Province has an average value of 70.77%. The variance of the Labor Force Participation Rate data shows a value of 34.22, this means that the Labor Force Participation Rate data has considerable variation or the data is diverse. The lowest Labor

Force Participation Rate of 61.14% is in Bau-bau City. The highest Labor Force Participation Rate of 82.12% is in West Muna Regency.

The dependent variable (z_4) Population with Health Complaints in Southeast Sulawesi Province has an average value of 30.84%. The variance of the Labor Force Participation Rate data shows a value of 120.78, this means that the data on Population with Health Complaints has a very large variation or the data is very diverse. The lowest percentage of population with health complaints is 17.23% in Wakatobi district. The highest percentage of people who have health complaints is 60.62% in North Kolaka Regency.

3.2 Scatterplot

To see the relationship between the dependent variable and the independent variable and the pattern of data distribution, a scatterplot or scatter diagram can be used to determine linear and nonlinear data patterns [22].

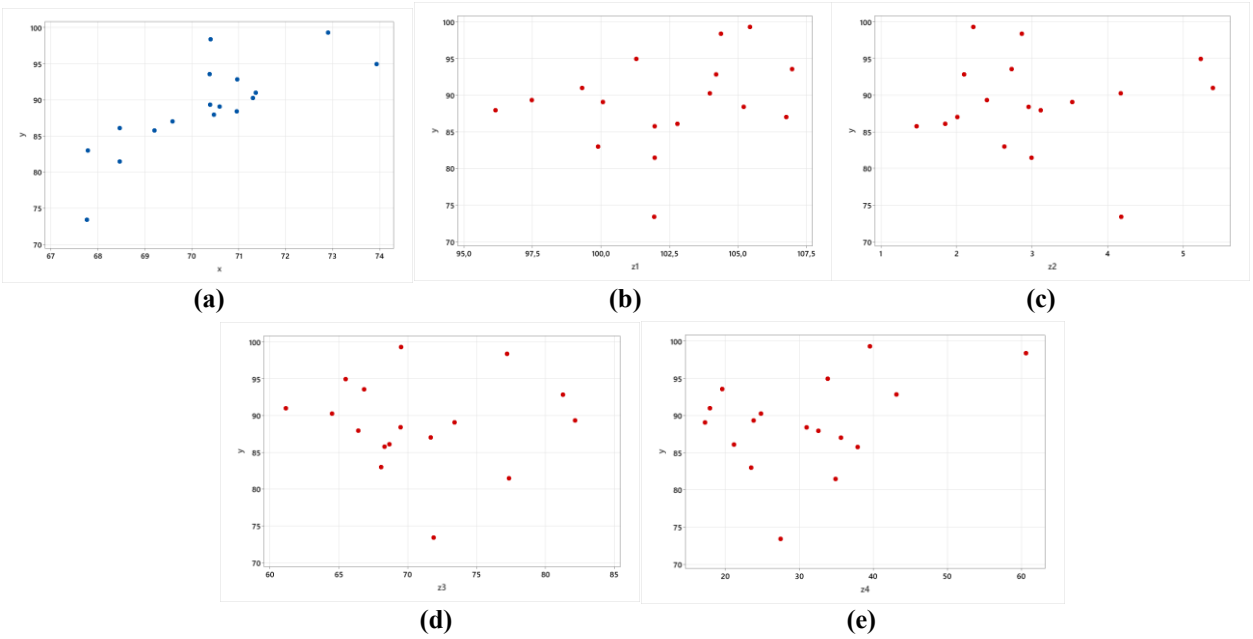

Figure 1. Scatterplot of Independent Variables against Dependent Variables, (a) x and y, (b) z_1 and y, (c) z_2 and y, (d) z_3 and y, (e) z_4 and y

Based on Figure 1, it can be seen that the shape of the relationship pattern between the response variable Gender Development Index (y) and the variable Population Aged 5 Years and Over who are no longer in school (x) tends to be linear, so it is tried to be modeled using a linear parametric function.

Figure 1. also show that the shape of the relationship pattern between the response variable Gender Development Index (y) and the predictor variables Sex Ratio (z_1), Open Unemployment Rate (z_2), Labor Force Participation Rate (z_3) and Population with Health Complaints (z_4), do not have a relationship pattern or tend not to follow a certain pattern, so they are modeled by nonparametric kernel.

3.3 Linearity Test

The procedure for testing the linearity relationship between the independent variable and the dependent variable is carried out using simple regression, namely one independent variable each on the dependent variable by looking at the F test value (simultaneous test) [19].

Table 3. Linearity Test

Dependent Variable: y				**data patterns**
Independent Variable	***F*-count**	***F*-table**	**Sig**	
x	31.7896	4.493998	0.000	linear
z_1	1.9600	4.493998	0.189	nonlinear
z_2	4.2803	4.493998	0.062	nonlinear
z_3	0.0083	4.493998	0.929	nonlinear
z_4	0.6168	4.493998	0.448	nonlinear

According to [19] the following is the linearity test hypothesis:

H_0 : The relationship between the independent and dependent variables is nonlinear.

H_1 : The relationship between the independent and dependent variables is linear.

If F-count $<$ F-table or the significance level of the data linearity test is greater than the alpha level (0.05), then H_0 is accepted or the relationship between variables is not linear. Conversely, if F-count $>$ F-table or the significance level of the data linearity test is smaller than the alpha level (0.05), then H_1 is accepted or the relationship between variables is linear.

Based on the results of the F significance in Table 3, it is concluded that variable x has an F-count $>$ F-table value or a significant value of F ($0.000 < 0.05$) so that it can be said that H_0 is rejected, meaning that variables y and x have a linear relationship. Meanwhile, the variables z_1, z_2, z_3, and z_4, which have an F-count value $<$ F-table significant F (p-value > 0.05) so that it can be said that it fails to reject H_0, meaning that the variables z_1, z_2, z_3, and z_4 have a non-linear relationship to variable y.

3.4 Determination of Semiparametric Kernel β Parameters

In determining the β Semiparametric Kernel Parameter, you can use the least square estimation method which is one of the techniques for estimating the β parameter using the ordinary least square estimation method or obtained by minimizing the sum of the squared errors [17].

Based on the equation $Y = X\beta + \widehat{M}(z) + \varepsilon$. With the least squares method, the estimation of β is obtained as follows:

$$\hat{\beta} = (X^T X)^{-1} X^T (\Omega_k + I) y$$

obtained β Semiparametric Kernel parameters are as follows:

$$\hat{y}_i = 89.49 + 4.57 x_i \tag{19}$$

3.5 Nadaraya Watson Estimator with Gaussian Kernel Function

To produce a kernel regression model with a Gaussian kernel function using the Nadaraya Watson estimator, namely:

$$\widehat{m}(z_{ji}) = \sum_{j=1}^{4} \frac{\sum_{i=1}^{n} K\left(\frac{z_j - z_{ji}}{h_j}\right) y_i^*}{\sum_{i=1}^{n} K\left(\frac{z_j - z_{ji}}{h_j}\right)} = \sum_{i=1}^{n} W_i(z) \left(y_i - \sum_{k=1}^{p} \beta_k x_{ik} \right)$$

Since a Gaussian kernel function is used, then

$$\widehat{m}(z_{ji}) = \sum_{j=1}^{4} \frac{\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{z_j - z_{ji}}{h_j}\right)^2\right) y_i^*}{\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{z_j - z_{ji}}{h_j}\right)^2\right)}$$

So that the following model is obtained:

$$\widehat{m}(z_{ji}) = \sum_{j=1}^{4} \frac{\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{z_1 - z_{1i}}{h_1}\right)^2\right) y_i^*}{\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{z_1 - z_{1i}}{h_1}\right)^2\right)} + \frac{\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{z_2 - z_{2i}}{h_2}\right)^2\right) y_i^*}{\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{z_2 - z_{2i}}{h_2}\right)^2\right)}$$
$$+ \frac{\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{z_3 - z_{3i}}{h_3}\right)^2\right) y_i^*}{\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{z_3 - z_{3i}}{h_3}\right)^2\right)} + \frac{\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{z_4 - z_{4i}}{h_4}\right)^2\right) y_i^*}{\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{z_4 - z_{4i}}{h_4}\right)^2\right)} \tag{20}$$

3.6 Determination of Optimal Bandwidth

The most important thing in obtaining regression curve estimation results in the kernel approach is to obtain the appropriate bandwidth. As for this research, the optimal bandwidth selection uses the Silverman rule so that the following results are obtained:

Table 4. Optimal Bandwidth Value

h_1	h_2	h_3	h_4
1.57	0.49	2.50	4.61

Based on Table 4, it is known that by using the calculation of the bandwidth formula from Silverman, the optimal bandwidth (h) value is obtained, namely $h_1 = 1.57, h_2 = 0.49, h_3 = 2.50$ and $h_4 = 4.61$.

3.7 Modeling Gender Development Index in Southeast Sulawesi Province using Semiparametric Kernel Regression

Estimation of each parameter that has been obtained, both parametric components obtained using the ordinary least square method and nonparametric components obtained using the Silverman rule of thumb bandwidth and substituting the values of x, z_1, z_2, z_3, z_4 and y in Equation (9), where the parameters obtained are β_0 = 89.49, and β_1 = 4.57 and the optimal bandwidth obtained is $h_1 = 1.57, h_2 = 0.49, h_3 = 2.50$ and $h_4 = 4.61$. then the semiparametric regression Nadaraya Watson estimator Gaussian kernel function is obtained as follows:

$$y_i = f(x_i) + m(z_i) + \varepsilon_i \text{ or } Y = X\beta + M(z) + \varepsilon$$

The equation of the multivariable kernel semiparametric model is as follows:

$$\hat{y}_i = 89.49 + 4.57x_i + \left(\frac{\sum_{i=1}^{17} K\left(\frac{z_1 - z_{1i}}{1.57}\right)\left(y_i - (89.49 + 4.57x_i)\right)}{\sum_{i=1}^{17} K\left(\frac{z_1 - z_{1i}}{1.57}\right)} \right) + \left(\frac{\sum_{i=1}^{17} K\left(\frac{z_2 - z_{2i}}{0.49}\right)\left(y_i - (89.49 + 4.57x_i)\right)}{\sum_{i=1}^{17} K\left(\frac{z_2 - z_{2i}}{0.49}\right)} \right) + \left(\frac{\sum_{i=1}^{17} K\left(\frac{z_3 - z_{3i}}{2.50}\right)\left(y_i - (89.49 + 4.57x_i)\right)}{\sum_{i=1}^{17} K\left(\frac{z_3 - z_{3i}}{2.50}\right)} \right) + \left(\frac{\sum_{i=1}^{17} K\left(\frac{z_4 - z_{4i}}{0.61}\right)\left(y_i - (89.49 + 4.57x_i)\right)}{\sum_{i=1}^{17} K\left(\frac{z_4 - z_{4i}}{4.61}\right)} \right)$$

3.8 Accuracy Evaluation Model

The accuracy of the Gender Development Index estimation can be seen by using the Coefficient of Determination and the MAPE value.

Table 5. Evaluation Model

R^2 (%)	MAPE (%)
99.80	0.14

Based on Table 5, the MAPE value is 0.14%. From this value, it can be concluded that the estimated kernel semiparametric regression model has a very accurate forecasting ability. While the coefficient of determination (R^2) value is 99.80%, this value indicates that all independent variables have a very good ability to explain the variance of the response variable by 99.80% and it is very unlikely to be explained by other variables not included in this study. From the resulting model, the results of the comparison of actual data and estimated data are as follows:

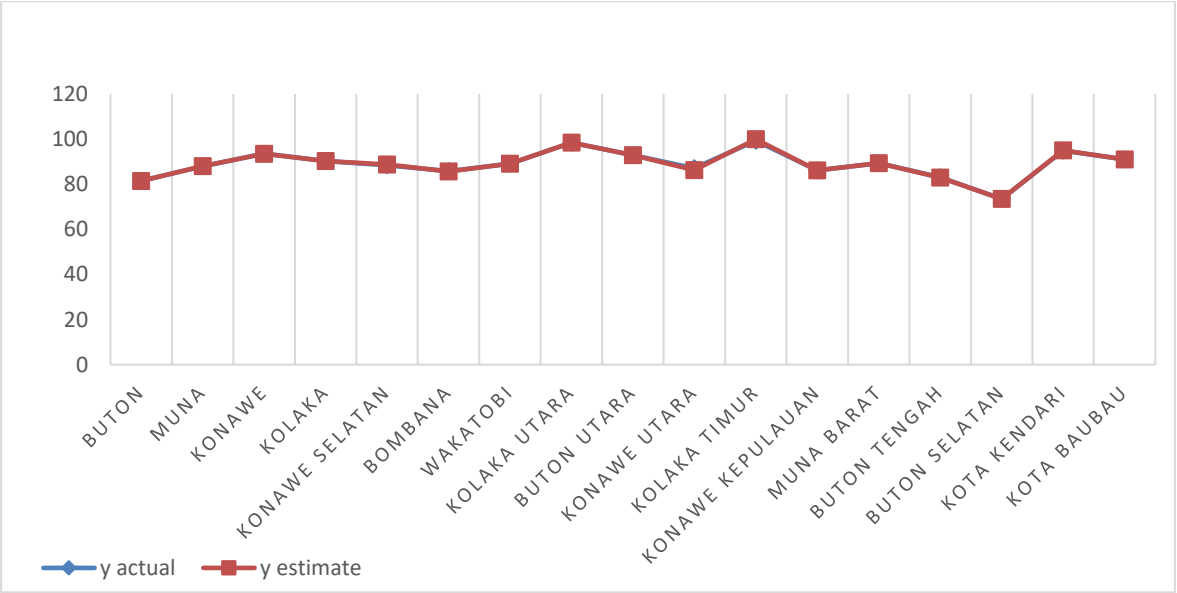

Figure 2. Comparison of Actual Data and Estimated Data

Based on the graph, it can be said that the estimated y tends to be close to the actual y, meaning that the prediction data is able to cover the actual data. It can be stated that the modeling obtained is suitable for predicting the gender development index. The region with the highest predicted value is in East Kolaka district, while the region with the lowest predicted value is in South Buton district.

4. CONCLUSIONS

Based on the results of the Gender Development Index (GDI) modeling analysis using Gaussian kernel semiparametric regression with the Nadaraya-Watson estimator, a well-fitting model was obtained, achieving R^2 and MAPE values of 99.80% and 0.14, respectively. These R^2 and MSE values indicate that almost all variations in the GDI data can be explained by the predictor variables used. Furthermore, from a policy perspective, careful consideration should be given to the selection and impact of these predictor variables to effectively address issues related to GDI.

REFERENCES

[1] A. Indrasetianingsih, F. Fitriani, and P. J. Kusuma, "KLASIFIKASI INDEKS PEMBANGUNAN GENDER DI INDONESIA TAHUN 2020 MENGGUNAKAN SUPERVISED MACHINE LEARNING ALGORITHMS," *Inferensi*, vol. 4, no. 2, p. 129, 2021, doi: https://doi.org/10.12962/j27213862.v4i2.10940.

[2] I. E. Lestari, S. N. Sarfiah, and G. Jalunggono, "ANALISIS FAKTOR-FAKTOR YANG MEMPENGARUHI INDEKS PEMBANGUNAN GENDER DI PROVINSI JAWA TENGAH TAHUN 2010-2019," *Dinamic*, vol. 1, pp. 182–194, 2021.

[3] I. Elisa, "FAKTOR-FAKTOR YANG MEMPENGARUHI INDEKS PEMBANGUNAN GENDER (IPG) PROVINSI SUMATERA BARAT MENGGUNAKAN ANALISIS REGRESI DATA PANEL," *J. Math. UNP*, vol. 7, no. 2, p. 8, 2022, doi: https://doi.org/10.24036/unpjomath.v7i2.12666.

[4] W. Cahyadi and C. C. Cen, "THE EFFECT OF INCOME DISTRIBUTION, HUMAN DEVELOPMENT INDEX, AND ECONOMIC GROWTH ON POVERTY," *Int. J. Econ. Bus. Appl.*, vol. 1, no. 2, pp. 187–194, 2020, doi: https://doi.org/10.9790/487X-2207011520.

[5] Randall, E. (1999). NONPARAMETRIC REGRESSION AND SPLINE SMOOTHING, SECOND EDITION. United States of America, New York, Basel: Mrcell Dekker.

[6] Ruppert, D., Wand, M.P., & Carroll, R.J. (2003). SEMIPARAMETRIC REGRESSION. New York: Cambridge University Press.

[7] Laome, L. 2010. MODEL REGRESI SEMIPARAMETRIK SPLINE UNTUK DATA LONGITUDINAL PADA KASUS KADAR CD4 PENDERITA HIV, *Paradigma* 13(2), 186-194.

[8] M. Berger and M. Schmid, "SEMIPARAMETRIC REGRESSION FOR DISCRETE TIME-TO-EVENT DATA," *Stat. Modelling*, vol. 18, no. 3–4, pp. 322–345, 2018, doi: https://doi.org/10.1177/1471082X17748084.

[9] J. Luts, T. Broderick, and M. P. Wand, "REAL-TIME SEMIPARAMETRIC REGRESSION," *J. Comput. Graph. Stat.*, vol. 23, no. 3, pp. 589–615, 2014, doi: https://doi.org/10.1080/10618600.2013.810150.

[10] D. Conn and G. Li, "AN ORACLE PROPERTY OF THE NADARAYA–WATSON KERNEL ESTIMATOR FOR HIGH-

DIMENSIONAL NONPARAMETRIC REGRESSION," *Scand. J. Stat.*, vol. 46, no. 3, pp. 735–764, 2019, doi: https://doi.org/10.1111/sjos.12370.
[11] B. Funke and M. Hirukawa, "BIAS CORRECTION FOR LOCAL LINEAR REGRESSION ESTIMATION USING ASYMMETRIC KERNELS VIA THE SKEWING METHOD," *Econom. Stat.*, vol. 20, no. xxxx, pp. 109–130, 2021, doi: https://doi.org/10.1016/j.ecosta.2020.01.004.
[12] H. N, S. S F, N. K, and S. E, "THE NONPARAMETRIC KERNEL METHOD USING NADARAYA-WATSON, PRIESTLEY-CHAO AND GASSER-MULLER ESTIMATORS FOR THE ESTIMATION OF THE RAINFALL DATA IN LAMPUNG," *Int. J. Math. Trends Technol.*, vol. 68, no. 8, pp. 12–20, 2022, doi: https://doi.org/10.14445/22315373/ijmtt-v68i8p502.
[13] Nadaraya, E.A. (1964). ON ESTIMATING REGRESSION. THEORY OF PROBABILITY AND ITS APPLICATIONS, Vol.9(1), 141-142.
[14] BPS. 2022. INDEKS PEMBANGUNAN GENDER. *https://sultra.bps.go.id,* [diakses Mei 2024].
[15] W. Wibowo, S. Haryatmi, and I. N. Budiantara, "KAJIAN METODE ESTIMASI PARAMETER DALAM REGRESI SEMIPARAMETRIK SPLINE," *Berk. MIPA*, vol. 23, no. 1, pp. 102–110, 2013.
[16] T. Hastie and R. Tibsgirani, "ADDITIVE MODELS GENERALIZED," *Stat. Sci.*, vol. 1, no. 3, pp. 297–310, 1986.
[17] A. T. Ampa, I. Nyoman Budiantara, and I. Zain, "KERNEL MULTIVARIABLE SEMIPARAMETRIC REGRESSION MODEL IN ESTIMATING THE LEVEL OF OPEN UNEMPTION IN EAST JAVA PROVINCE," *J. Phys. Conf. Ser.*, vol. 1899, no. 1, 2021, doi: https://doi.org/10.1088/1742-6596/1899/1/012127.
[18] B. Both and Z. Szánthó, "EXPERIMENTAL AND NUMERICAL INVESTIGATION OF AN OFFSET JET USING TANGENTIAL AIR DISTRIBUTION SYSTEM," *Period. Polytech. Mech. Eng.*, vol. 60, no. 3, pp. 129–136, 2016, doi: https://doi.org/10.3311/PPme.8017.
[19] S. F. Muazarah, "REGRESI NONPARAMETRIK KERNEL DENGAN PENDEKATAN FUNGSI GAUSSIAN UNTUK MEMODELKAN INFLASI DI INDONESIA," 2023.
[20] A. T. Ampa, I. N. Budiantara, and I. Zain, "MODELING THE LEVEL OF DRINKING WATER CLARITY IN SURABAYA CITY DRINKING WATER REGIONAL COMPANY USING COMBINED ESTIMATION OF MULTIVARIABLE FOURIER SERIES AND KERNEL," *Sustain.*, vol. 14, p. 13663, 2022, doi: https://doi.org/10.3390/su142013663.
[21] D. N. Gujarati, *BASIC ECONOMITRICS.* 2003.
[22] I. Gani and S. Amalia, "ALAT ANALISIS DATA," *CV Andi Offset*, p. 306, 2014.