International Journal of Science and Engineering Invention (IJSEI)
Volume 11, Issue 04, June 2025
https://doi.org/10.23958/ijsei/vol11-i04/284

Research Article

Optimizing Education-Based HDI Modeling in Indonesia: A Multivariable Kernel Regression Approach with CV and GCV

Muhammad Rafi'Ar Rasyid, Suparti Suparti *, Masithoh Yessi Rochayani

Department of Statistics, Diponegoro University, Indonesia.

*Corresponding author: Suparti Suparti; *suparti@live.undip.ac.id*

Received 07 May 2025; Accepted 05 June 2025; Published 09 June 2025

Abstract

This study aims to model the Human Development Index (HDI) in Indonesia based on education quality indicators using multivariable kernel regression, and to identify the optimal bandwidth selection method through Cross-Validation (CV) and Generalized Cross-Validation (GCV). Employing a quantitative modeling design, the research utilizes secondary data comprising educational and HDI indicators from 34 Indonesian provinces in 2023. The analysis applies multivariable kernel regression with a triangle kernel function, with mean years of schooling and expected years of schooling as predictor variables. Bandwidth optimization is performed using CV and GCV, and model performance is assessed through the coefficient of determination (R^2) and Mean Absolute Percentage Error (MAPE). The results indicate that the GCV method yields a slightly better model, with R^2 of 86.32% and MAPE of 1.94%, compared to the CV method, which as an R^2 of 85.92% and MAPE of 1.96%. While both models show excellent forecasting accuracy, GCV demonstrates superior stability and predictive performance. These findings confirm that multivariable kernel regression, particularly when optimized with GCV, is an effective approach for modeling complex data patterns such as HDI based on educational indicators in Indonesia.

Keywords: ***Cross-Validation, Education, Generalized Cross-Validation, Human Development Index, Kernel Regression, Nonparametric Regression.***

Introduction

Nonparametric regression is a flexible and effective method for modeling data without requiring strict parametric assumptions (Abdy, 2019). One commonly used approach in nonparametric regression is kernel regression, which utilizes the Nadaraya-Watson estimator to measure the relationship between response and predictor variables (Lamusu et al., 2020). This estimator relies on two parameters that called kernel function and bandwidth parameter. The kernel function assigns weights based on the distance between observation X_i and point x, while bandwidth controls the smoothness of the estimated density (Ogden, 1997). There are several types of kernel functions, including Gaussian, triangle, Epanechnikov, biweight, and uniform. Accurate bandwidth selection is crucial for determining the best regression model (Hardle, 1994). Optimization methods such as Cross-Validation (CV) and Generalized Cross-Validation (GCV) are used to achieve optimal bandwidth, preventing overfitting and underfitting to the model (Suparti et al., 2018).

Several studies have confirmed the advantages of kernel nonparametric regression in data analysis. (Sadek & Mohammed, 2024) found that kernel regression achieved a coefficient of determination of 95%, significantly higher than the 23% obtained using parametric regression. Puspitasari et al. (2012) compared parametric and nonparametric kernel regression in stock market data, finding that the lowest Mean Squared Error (MSE) occurred in the nonparametric regression with a triangle kernel. Similarly, Astuti et al. (2018) analyzed kernel regression using various kernel functions and found consistent estimation results. Razak et al. (2019) applied multivariable kernel regression to malnutrition data in Indonesia, obtaining a coefficient of determination of 84.73%. Furthermore, Lamusu et al. (2020) compared CV and GCV optimization methods in kernel regression and concluded that GCV provided a better model evaluation than CV for corn production estimation.

Based on the above findings, this study employs multivariable kernel regression using the triangle kernel function and compares CV and GCV optimization. The study analyzes the relationship between education quality and Human Development Index (HDI) in Indonesia in 2023. HDI is a comparative measure that encompasses life expectancy, education, and living standards globally (Raghuvanshi & Verma, 2024). The education quality indicators used in HDI include the years of schooling and expected years of schooling (Badan Pusat Statistik, 2024). This study focuses on the education dimension due to its crucial role in improving quality of life and contributing to economic and social development.

Material and Methods

2.1. Nonparametric regression

Nonparametric regression is one method used to model data when the form of the regression function form is unknown. In some cases, the observed data analyzed may form a certain pattern such as linear, but do not meet the assumptions of the parametric model. In this condition, modeling using the parametric approach can be less precise and can potentially be misleading (Suparti et al., 2018). Alternatively, nonparametric regression can be used in modeling. The function in nonparametric regression is assumed to be smooth so that it has high flexibility to estimate the regression function (Eubank, 1999). Function estimation is done based on observation data using certain modeling techniques. The nonparametric regression model can be systematically written as follows:

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

where:

Y_i: response variable of observation i
$m(x_i)$: function representing the predictor variable of observation i
ε_i: error term with mean 0 and variance σ^2

There are several techniques for modeling functions that can be used, namely kernels, splines, local polynomials, Fourier series, and wavelets (Hardle, 1994).

2.2. Kernel Density Estimator

The kernel density estimator is a development of the histogram estimator and the naive estimator, where the main goal is to smooth the data distribution by giving weight to each observation data, so that closer observations contribute more to the estimation results (Ogden, 1997). The histogram estimator has a weakness in describing the existing data distribution because its dependence on the initial value x_0 and the binwidth h results in a graph that is too rigid. Similar to the naive estimator, this estimator depends on a rigid function weight because its value will always be the same in a certain weight which causes a less smooth distribution graph (Silverman, 1986). (Rosenblatt, 1956) and (Parzen, 1962) introduced an alternative estimator also called the kernel density estimator (Machkouri, 2011).

Kernel density estimators are more flexible than histogram and naive estimators because they do not rely on fixed bin sizes or rigid function weights and allow for smoother results. Kernel density estimators are defined as follows (Ogden, 1997):

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

The kernel density estimator is influenced by two parameters: the kernel function $K(\cdot)$ and the smoothing parameter or bandwidth h. The kernel function serves to provide weights based on the distance between observation X_i and point x, while the bandwidth controls how smooth or coarse the resulting density estimate is (Ogden, 1997). The kernel function is denoted as:

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

where $K(\cdot)$ is the kernel function and h denote the bandwidth (Hardle, 1994).

Kernel function K is a continuous, real-valued, bounded, and $\int_{-\infty}^{\infty} K(u)du = 1$. In addition, the kernel function is an even function, meaning it is symmetric to the origin, thus $\int_{-\infty}^{\infty} uK(u)du = 0$ (Suparti et al., 2018). Some types of kernel functions that are commonly used to estimate are presented in Table I (Ogden, 1997):

The triangle kernel used in this study, provides a linearly decreasing weight, with higher weights on data close to the symmetric center and decreasing gradually for data further from the symmetric center. This kernel provides stable and easy-to-interpret estimates.

The kernel density estimator in Equation (2) can also be written by substituting the kernel function as in Equation (3) so that it can be denoted as Equation (4).

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x - X_i}{h}\right) \tag{4}$$

where K is the kernel function, and h is the smoothing parameter or bandwith. According to Suparti et al., (2018), the kernel function K is a continuous, real-valued, bounded, symmetric, and the function satisfies $\int_{-\infty}^{\infty} K(u)du = 1$.

The kernel density estimator can also be extended to estimate the density of multivariable data. For instance, if there are d predictor variables, the kernel density estimation in d-dimensional space can be obtained using a multivariable kernel function $\kappa(u_1, u_2, \dots, u_d)$. This function can be simplified by multiplying the kernel functions for each dimension d, assuming that $x_1, x_2, \dots, x_n$ are independent. The multivariable kernel density estimator is given by:

$$\hat{\boldsymbol{f}}_{\boldsymbol{h}}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}\left\{\prod_{j=1}^{d}\frac{1}{h_j}K\left(\frac{x_j - X_{ij}}{h_j}\right)\right\} \tag{5}$$

Table I. Types of Kernel Functions

Name	Kernel Function
Uniform	$K(x) = \begin{cases} \frac{1}{2}, \text{if } \lvert x\rvert \le 1 \\ 0, \text{otherwise} \end{cases}$
Triangle	$K(x) = \begin{cases} 1 - \lvert x\rvert, \text{if } \lvert x\rvert \le 1 \\ 0, \text{ otherwise} \end{cases}$
Epanechnikov	$K(x) = \begin{cases} \frac{3}{4}(1 - x^2)^2, \text{if } \lvert x\rvert \le 1 \\ 0, \text{otherwise} \end{cases}$
Gaussian	$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, -\infty < \text{x} < \infty$
Biweight	$K(x) = \begin{cases} \frac{15}{16}(1 - x^2)^2, \text{if } \lvert x\rvert \le 1 \\ 0, \text{otherwise} \end{cases}$

2.3. Kernel Regression

Kernel regression is a nonparametric regression technique used to predict the regression function values that satisfy Equation (1). One method of estimating the regression function $m(x_i)$ is by using the Nadaraya-Watson estimator (Hardle, 1994). If the data contains more than one predictor variable, for example, d variables, represented as $\boldsymbol{X_i} = (X_{i1}, X_{i2}, \dots, X_{id})$, the regression model can be written as follows (García, 2023):

$$\hat{m}(\boldsymbol{x}) = \frac{\sum_{i=1}^{n}\prod_{j=1}^{d}\frac{1}{h_j}K\left(\frac{x_j - X_{ij}}{h_j}\right)Y_i}{\sum_{i=1}^{n}\prod_{j=1}^{d}\frac{1}{h_j}K\left(\frac{x_j - X_{ij}}{h_j}\right)} \tag{6}$$

where:

$\hat{m}(\boldsymbol{x})$: estimated regression function
Y_i : observed response variable
$K(.)$: kernel function
x_j : observation point for predictor variable j
X_{ij} : observed value for predictor variable j in observation i

h_j : bandwith vlue with vector $\boldsymbol{h} = (h_1, h_2, \dots, h_d)$
n: number of observations
d : number of predictor variables

2.4. Optimal Bandwidth Selection

In kernel regression analysis, choosing the right optimal bandwidth (h) plays an important role in controlling the smoothness of the estimated curve. If the tested bandwidth value is too small, the resulting curve will be too rough or have a jagged structure. Conversely, if the tested bandwidth value is too large, the resulting curve will be too smooth or oversmooth and produce high bias because too much smoothing is done and the variance is low (Hardle, 1994). The optimal bandwidth value will not produce high or low variance and bias. Based on this, it is necessary to choose the right optimal bandwidth so that it produces the best estimated value.

In the linear estimation approach, for each value of the model complexity parameter h, there is a matrix $H(h)$ of size $n \times n$ which is symmetric and positive semidefinite (Takezawa, 2006). The value of the function weight $W_i(x)$ is the same as the matrix $H(h)$, so that $W_i(x) = H(h)$ with the elements of the matrix $H(h)$ are H_{ij} (Astuti et al., 2018). H_{ij} can be denoted as follows:

$$H_{ij} = \frac{K\left(\frac{x - X_i}{h}\right)}{\sum_{k=1}^{n} K\left(\frac{x - X_k}{h}\right)} \qquad (7)$$

where $K(\cdot)$ is the kernel function and h is bandwidth. The methods used to determine the optimal bandwidth are Cross-Validation (CV) and Generalized Cross-Validation (GCV).

2.5. Cross-Validation (CV)

The optimal bandwidth is selected based on the minimum CV value (Carmack et al., 2011). The systematic form of the CV optimization method for multivariable data is given as follows:

$$CV(\boldsymbol{h}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_j - \hat{m}_h(\boldsymbol{X_j})}{1 - H_{jj}(\boldsymbol{h})}\right)^2 \qquad (8)$$

where:

$CV(\boldsymbol{h})$: Cross-Validation value containing the bandwith vector $\boldsymbol{h} = (h_1, h_2, \dots, h_d)$
n : number of observations
$H_{jj}(\boldsymbol{h})$: diagonal element of the smoothing matrix $\boldsymbol{H}$
$\hat{m}_h(\boldsymbol{X_j})$: estimated value at X_j

2.6. Generalized Cross-Validation (GCV)

Generalized Cross-Validation (GCV) is another optimization method that can be used to determine the optimal model. GCV minimizes the GCV function and is derived from CV by replacing H_{jj} with its mean over all observations $H_{jj} = \frac{\sum_{j=1}^{n} H_{jj}}{n}$. The GCV function for multivariable data is given as follows:

$$GCV(\boldsymbol{h}) = \frac{n^2 MSE(\boldsymbol{h})}{\left(n - \sum_{j=1}^{n} H_{jj}\right)^2} \qquad (9)$$

where:

$GCV(\boldsymbol{h})$: *Generalized Cross-Validation* value with bandwith vector $\boldsymbol{h} = (h_1, h_2, \dots, h_d)$
n : number of observations
$MSE(\boldsymbol{h})$: Mean Squared Error at bandwith $h = (h_1, h_2, \dots, h_d)$
$\sum_{j=1}^{n} H_{jj}$: total weight of the smoothing matrix with bandwith h for row j and column j

2.7. Coefficient of Determination

A regression model is considered optimal if the predictor variables can explain the response variable. The goodness of fit of a model can be measured using the coefficient of determination (Gujarati, 1972). The coefficient of determination is formulated as follows:

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \qquad (10)$$

where

R^2 : coefficient of determination
SSE: sum of squares error
SST: sum of squares total
Y_i : the actual value of the response variable at observation i
$\hat{Y}_i$: the estimated value of the i-th observation with $\hat{Y}_i = \hat{m}(X_i), i = 1,2, \dots, n$
$\bar{Y}$: average actual value of response variable

The coefficient of determination has a value between 0 and 1. A coefficient value approaching 1 will produce a better model and vice versa if the coefficient of determination value approaches 0 then the model created will be less good. The criteria for model quality based on R^2 are categorized as follows: an R^2 value greater than 67% indicates a strong model, whereas an R^2 value between 19% and 33% suggests a weak model (Chin, 1998).

2.8. Mean Absolute Percentage Error (MAPE)

In addition to R^2, another method used to evaluate model performance is the MAPE. MAPE is a measure of accuracy used to calculate the average absolute percentage error. This value can be calculated using the following formula:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left(\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right|\right) \times 100\% \qquad (11)$$

The lower the MAPE value, the more accurate the forecasting model used, conversely, the higher the MAPE value, the poorer the forecasting model's ability (Maricar, 2019). According to Chang et al., (2007), the evaluation criteria for forecasting performance are as follows: a MAPE value below 10% indicates excellent forecasting ability, while a MAPE value of 50% or more suggests poor forecasting performance with high error rates and low reliability.

2.9. Data and Research Stages

This study utilizes secondary data on the Human Development Index (HDI) as the response variable and education quality data—including the mean years of schooling and expected years of schooling as predictor variables for Indonesia's provinces in 2023. The dataset consists of 34 observations. The data was randomly divided into training (80%) and testing (20%) sets.

The study employs RStudio. The data analysis steps include:

1. Splitting the dataset into training and testing sets
2. Analyzing the scatterplot relationships between predictor and response variables
3. Conducting a multicollinearity test on predictor variables in the training data
4. Determining bandwidth values for evaluation
5. Selecting the appropriate kernel function
6. Applying optimization methods (CV and GCV)
7. Identifying the best model from the optimization methods
8. Evaluating the model's goodness-of-fit using R^2
9. Estimating test data using the best model obtained
10. Comparing the best models from CV and GCV

Results and Discussion

3.1. Data Description

The research data contains education quality data in the form of average length of schooling (X_1) and expected length of schooling (X_2) as predictor variables and HDI (Y) s as response variable. The research data was randomly divided into two parts, training data and testing data. The training data comprise of 80% of the total data, which is 27 data based on provincial data from Indonesia in 2023. The testing data comprise 20% of the total data, or 7 data. The training data are used to build regression model, while the testing data are used to evaluate the model's performance. Descriptive statistics of education quality data, including average length of schooling (X_1) and expected length of schooling (X_2) against HDI (Y) can be shown in Table II.

Based on Table II, the HDI value has an average of 72.25 with the lowest HDI value in Papua Province, which is 62.25 and the highest HDI value in DKI Jakarta Province, which is 82.46. In addition, the predictor variable for the average length of schooling found that the variable has an average of 8.89 with the lowest average length of schooling in Papua Province, which is 7.17 and the highest average length of schooling in DKI Jakarta Province, which is 11.45. Expected length of schooling has an average of 13.31 with the lowest value in Papua Province, which is 11.15 and the highest value in DI Yogyakarta Province, which is 15.66.

The form of the relationship between the predictor variables of average length of schooling and expected length of schooling with the response variable HDI can be seen in 3D in Fig. 1 and in 2D in Fig. 2 and Fig. 3. Based on the 3D data pattern in Fig. 1, it is found that the relationship between the variables of average length of schooling and expected length of schooling with the HDI variable appears to form a random pattern.

However, there is a linear pattern that is formed and there is data that has quite different values so that there are several outlier data that can be seen in Fig. 1. Based on Fig. 2, the relationship between the average length of schooling and the HDI form a linear pattern, while the random pattern is formed by the relationship between the expected length of schooling and the HDI in Fig. 3.

To ensure there is no correlation between predictor variables, a multicollinearity test is conducted using the VIF value. Both predictor variables have a VIF value of 1.28, which is less than 10. Therefore, there is no multicollinearity between the variables of average length of schooling and expected length of schooling.

Table II. Data Description

Statistic	X_1	X_2	Y
Observation count	27	27	27
Minimum value	7.15	11.15	62.25
Maximum value	11.45	15.66	82.46
Median	8.81	13.22	7.77
Mean	8.89	13.31	72.25

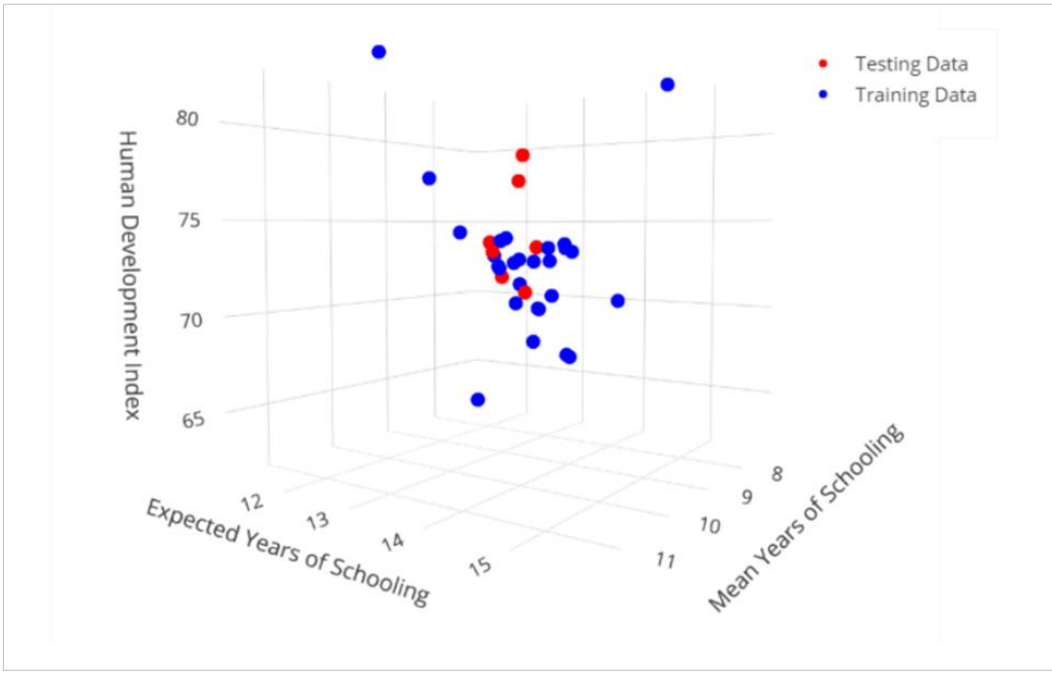

Fig. 1: 3D Plot of Predictor Variables against Response Variables

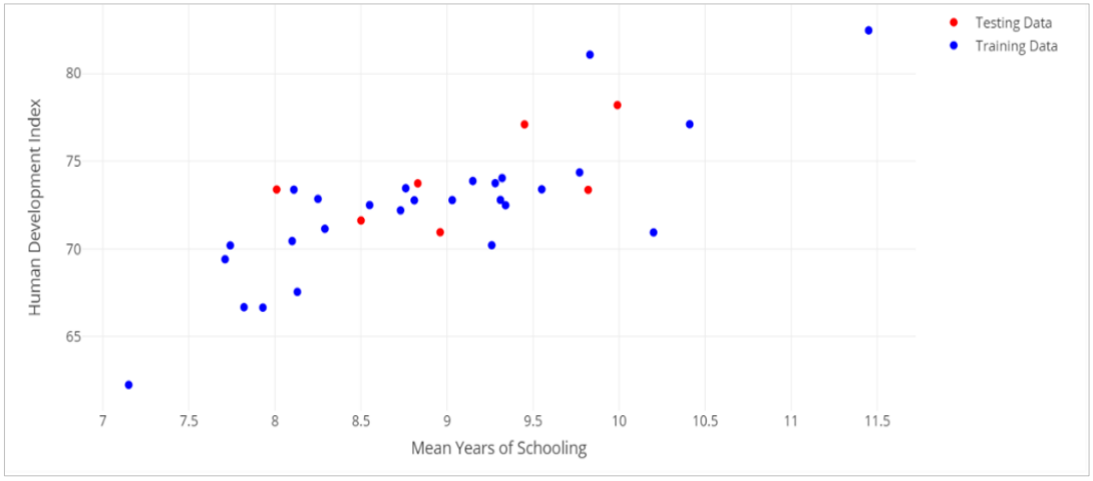

Fig. 2. 2D plot between Mean years of Schooling and HDI

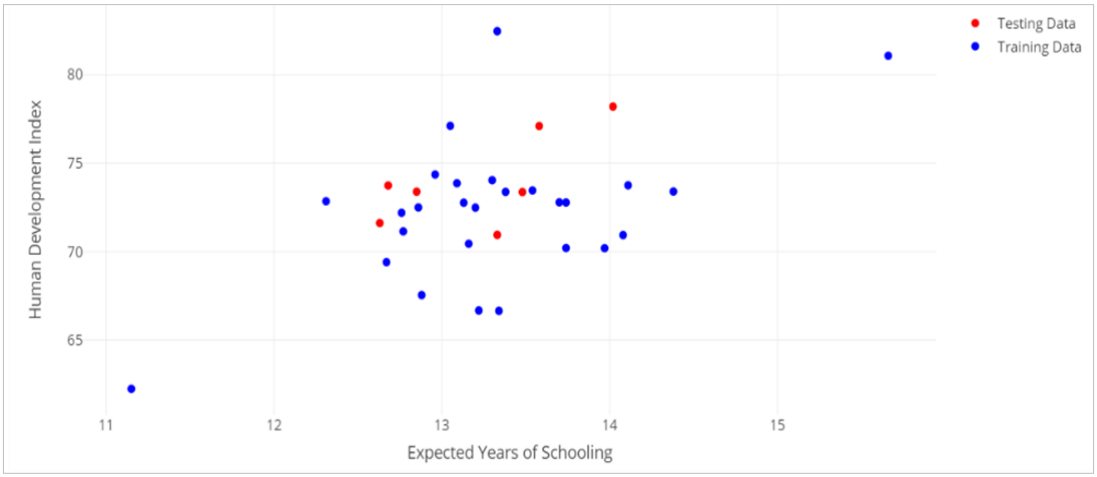

Fig. 3. 2D plot between Expected Years of Schooling and HDI

3.2. Kernel Regression Modeling

Based on Fig. 1, the data has a pattern that tends to be random and in Fig. 2, a linear pattern is seen formed by the relationship between the average length of schooling and the HDI. With the random pattern seen in the data plot, multivariable kernel nonparametric regression modeling can be used to overcome complex patterns more accurately.

Multivariable kernel modeling begins by determining the bandwidth limits to be tested on previously adjusted training data. The bandwidth limits in question include the lower limit of the bandwidth value, the upper limit of the bandwidth value, and the increase between the bandwidth values of each predictor variable as in Table III. Based on Table III, the minimum bandwidth value tested is 1, the maximum bandwidth value is 2, and the bandwidth increase is 0.1 on both predictor variables. The number of bandwidth combinations from the two predictor variables is 121. The number of bandwidths is formed from each element tested, namely the combination of each element of the value h_1 and $h_2 = 11 \times 11$.

In addition, kernel regression modeling also requires kernel functions and optimization methods in determining the best model of each combination of bandwidth tested. In this study, the kernel triangle function is used in modeling so that the model shape is in accordance with the following equation

$$\hat{m}(\boldsymbol{x}) = \frac{\sum_{i=1}^{n} \prod_{j=1}^{d} \frac{1}{h_j} K\left(\frac{x_j - X_{ij}}{h_j}\right) Y_i}{\sum_{i=1}^{n} \prod_{j=1}^{d} \frac{1}{h_j} K\left(\frac{x_j - X_{ij}}{h_j}\right)}$$

$$= \frac{\sum_{i=1}^{n} \frac{1}{h_1 h_2} K\left(\frac{x_1 - X_{i1}}{h_1}\right) K\left(\frac{x_2 - X_{i2}}{h_2}\right) Y_i}{\sum_{i=1}^{n} \frac{1}{h_1 h_2} K\left(\frac{x_1 - X_{i1}}{h_1}\right) K\left(\frac{x_2 - X_{i2}}{h_2}\right)}$$

$$= \frac{\sum_{i=1}^{n} \left\{1 - \left|\frac{x_1 - X_{i1}}{h_1}\right|\right\} I\left(\left|\frac{x_1 - X_{i1}}{h_1}\right| \leq 1\right) \left\{1 - \left|\frac{x_1 - X_{i2}}{h_2}\right|\right\} I\left(\left|\frac{x_2 - X_{i2}}{h_2}\right| \leq 1\right) Y_i}{\sum_{i=1}^{n} \left\{1 - \left|\frac{x_1 - X_{i1}}{h_1}\right|\right\} I\left(\left|\frac{x_1 - X_{i1}}{h_1}\right| \leq 1\right) \left\{1 - \left|\frac{x_1 - X_{i2}}{h_2}\right|\right\} I\left(\left|\frac{x_2 - X_{i2}}{h_2}\right| \leq 1\right)}$$

3.3. Optimization using Cross-Validation (CV)

After calculating the entire combination of tested bandwidths as in Table III, various models are formed from each combination of bandwidths. Therefore, an optimization method is needed, one of which is CV, to determine the best model from the various models generated through the calculation of each existing bandwidth. The steps in determining the best model using the CV optimization method are as follows:

1. Determining Optimal Bandwidth

The determination of the optimal bandwidth is done through the kernel regression modeling process by imputing bandwidth as in Table III and using the kernel triangle function by selecting the CV optimization method in the GUI application. From the analysis results, 10 combinations were obtained that produced the smallest CV value from each combination of bandwidth tested and can be seen in Table IV. Based on the table, the optimal bandwidth values achieved using the CV optimization method include the optimal bandwidth value on the X_1 variable which is 1.1 and the optimal bandwidth value on the X_2 variable which is 1.6 with a CV value of 8.7675. The combined form of the results of the CV optimization is presented in Fig. 4.

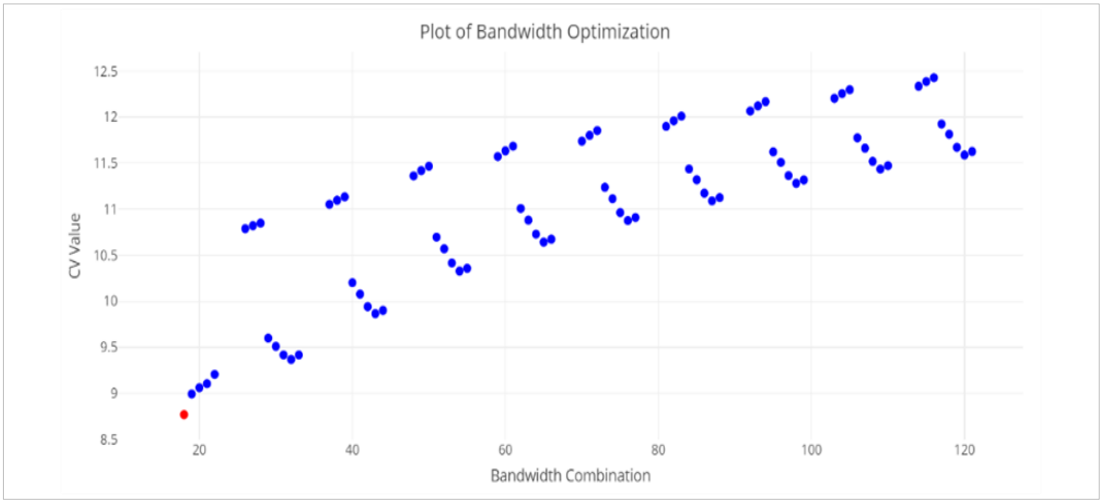

Fig. 4. Plot of bandwidth combination using CV

Fig. 4 shows a plot of the tested bandwidth combinations using the CV method. From the test results, the minimum CV value is obtained at the 18th combination which is marked with a red dot on the graph. Based on the tested bandwidth combinations, the combinations have a range of values from 1 to 2 with an increase of 0.1 as shown in Table 3. However, Fig. 4 only shows starting from the 18th combination due to constraints on the previous combinations. Combinations lower than 18 are not shown in the graph because there is a value of 1 in the $H_{jj}(h)$. This causes the denominator in the CV calculation to be infinite, making it impossible to calculate. Therefore, the graph only shows valid bandwidth combinations. The kernel regression model of optimal bandwidth can be written as follows:

$$\hat{m}(\boldsymbol{x}) = \frac{\sum_{i=1}^{27}\left\{1-\left|\frac{x_1-X_{i1}}{1,1}\right|\right\}I\left(\left|\frac{x_1-X_{i1}}{1,1}\right|\leq 1\right)\left\{1-\left|\frac{x_1-X_{i2}}{1,6}\right|\right\}I\left(\left|\frac{x_2-X_{i2}}{1,6}\right|\leq 1\right)Y_i}{\sum_{i=1}^{27}\left\{1-\left|\frac{x_1-X_{i1}}{1,1}\right|\right\}I\left(\left|\frac{x_1-X_{i1}}{1,1}\right|\leq 1\right)\left\{1-\left|\frac{x_1-X_{i2}}{1,6}\right|\right\}I\left(\left|\frac{x_2-X_{i2}}{1,6}\right|\leq 1\right)}$$

2. Model Feasibility Test

After obtaining the best model with optimal bandwidth from the CV optimization method, the model is used to calculate the estimated value of each actual value of its response. Furthermore, the model will be assessed for its feasibility using the coefficient of determination based on its estimation results. The plot of the model formed from the actual value with the calculated estimated value can be seen as follows:

Based on the output in Fig. 5, the estimated data results can be said to be close to the actual data well. The coefficient of determination obtained is 85.92%. This means that the predictor variables of average length of schooling and expected length of schooling have an influence of 85.92% on the HDI response variable, while the remaining 14.08% is influenced by other variables not tested in the study.

3. Model Performance Evaluation

The best model that has been formed using the CV method will be evaluated for its performance based on the MAPE measure. The MAPE value obtained is 1.96%. This indicates that the constructed model demonstrates excellent forecasting performance.

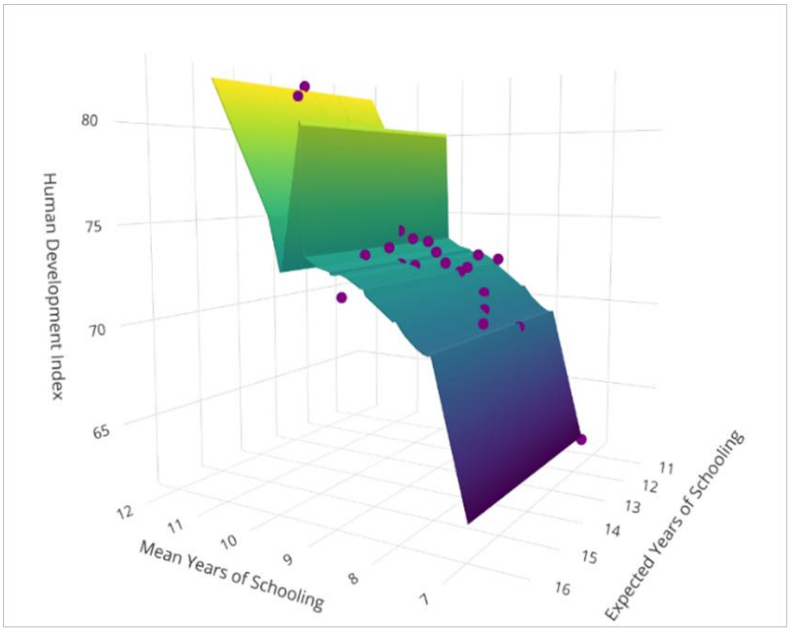

Fig. 5: Estimation Curve and Actual Data with CV Optimization

3.4. Optimization Using Generalized Cross-Validation (GCV)

The application of GCV is expected to provide a more stable and flexible model for the data. Similar to the CV method, the optimization process with GCV involves testing various combinations of bandwidth, but with a calculation formula that gives different weights and avoids validity issues in some kernel functions. The steps to determine the best model using the GCV method are described as follows:

1. Determining Optimal Bandwidth

The determination of the optimal bandwidth is done through the kernel regression modeling process with the determination of bandwidth as in Table III and the kernel triangle function so that the smallest GCV value is obtained. Table V shows the combination that produces the smallest GCV value. Based on Table V, the optimal bandwidth value obtained from the GCV optimization method is 1.1 on the X_1 variable and 1.5 on the X_2 variable with a minimum GCV value of 4.1835.

Fig. 6 shows a graph of the tested bandwidth combinations using the GCV method. Unlike the previous CV method, this graph contains all tested bandwidth combinations as in Table III. From the test results, it was obtained that the minimum GCV value was in the 17th combination which was marked with a red dot on the graph. The kernel regression model formed can be written as follows:

$$\hat{m}(\boldsymbol{x}) = \frac{\sum_{i=1}^{27}\left\{1-\left|\frac{x_1-X_{i1}}{1,1}\right|\right\}I\left(\left|\frac{x_1-X_{i1}}{1,1}\right|\leq 1\right)\left\{1-\left|\frac{x_1-X_{i2}}{1,5}\right|\right\}I\left(\left|\frac{x_2-X_{i2}}{1,5}\right|\leq 1\right)Y_i}{\sum_{i=1}^{27}\left\{1-\left|\frac{x_1-X_{i1}}{1,1}\right|\right\}I\left(\left|\frac{x_1-X_{i1}}{1,1}\right|\leq 1\right)\left\{1-\left|\frac{x_1-X_{i2}}{1,5}\right|\right\}I\left(\left|\frac{x_2-X_{i2}}{1,5}\right|\leq 1\right)}$$

2. Model Feasibility Test

Through the previously formed model, the estimated value can be calculated using the actual data of the response variable. The estimated value will be assessed for its feasibility using the coefficient of determination by calculating the estimated value from the actual value of the response variable. The graph between the actual data and the estimated data is shown in Fig. 7.

The model formed has a coefficient of determination of 86.32% (very good). This means that the predictor variables of average length of schooling and expected length of schooling have an influence of 86.32% on the IPM response variable, while the remaining 13.68% is influenced by other variables not tested in the study.

3. Model Performance Evaluation

Next, the best model that has been formed using the GCV method will be assessed for its performance evaluation based on the MAPE measure. The MAPE value obtained is 1.94%. This means that the model created has very good performance in forecasting.

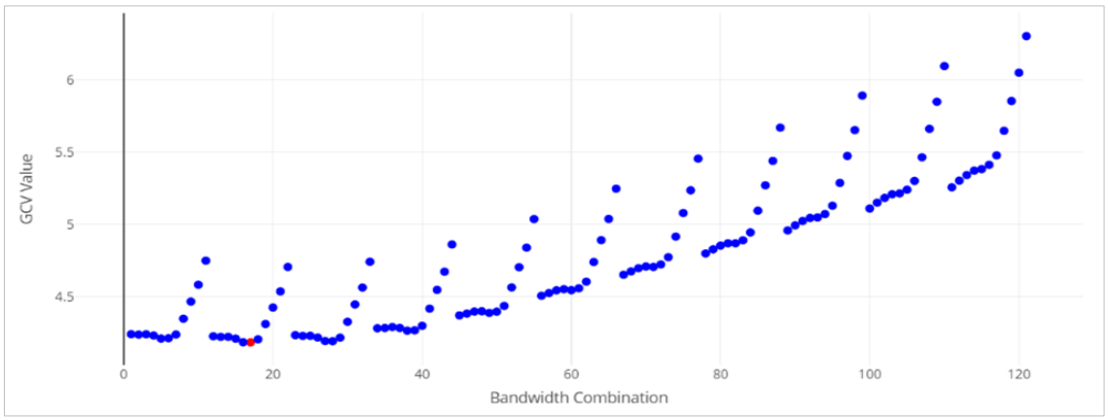

Fig. 6. Plot of Bandwidth and GCV

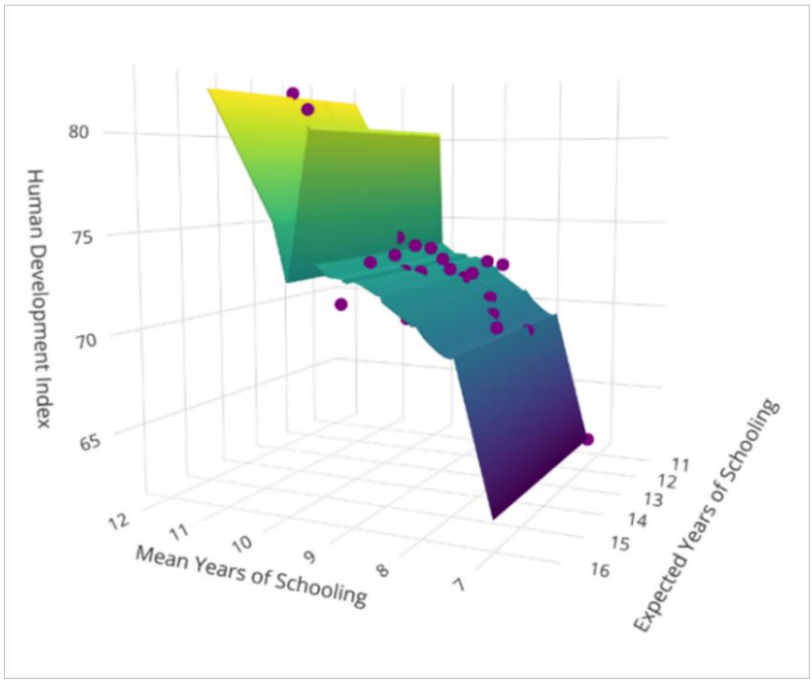

Fig. 7. Estimation Curve and Actual Data with GCV Optimization

Table III. Tested bandwidth

Variable	Minimum bandwidth value	Maximum bandwidth value	Bandwidth increase
X_1	1	2	0,1
X_2	1	2	0,1

Table IV: Ten Bandwidth Combinations with the Smallest CV Values

Combination	h_1	h_2	CV
18	1.1	1.6	8.7675
19	1.1	1.7	8.9918
20	1.1	1.8	9.0599
21	1.1	1.9	9.1037
22	1.2	2	9.2048
32	1.1	1.9	9.3668
31	1.1	1.8	9.4154
33	1.2	2	9.4155
30	1.1	1.7	9.5093
29	1.1	1.6	9.5988

CV: Cross-Validation

Table V: Ten Bandwidth Combinations with the Smallest GCV Values

Combination	h_1	h_2	GCV
17	1.1	1.5	4.1835
16	1.1	1.4	4.1844
28	1.2	1.5	4.1910
27	1.2	1.4	4.1923
18	1.1	1.6	4.2047
15	1.1	1.3	4.2092
5	1	1.4	4.2098
6	1	1.5	4.2122
29	1.2	1.6	4.2157
26	1.2	1.3	4.2167

GCV: Generalized Cross-Validation

Determination of the Best Model Based on CV and GCV

Fig. 8 shows the plot of actual data and estimated data of the response variable using the CV and GCV methods. Figure 8 shows that the actual data curve and the estimated results using different optimization methods, CV and GCV, have almost the same pattern and overlap. However, between the two methods, the GCV method is closer to the actual data. In determining the best method in estimating the HDI value, it can be done by comparing the coefficient of determination and MAPE produced between the two. The calculation results of the coefficient of determination and MAPE values from the CV and GCV methods are shown in Table VI.

Table VI shows the determination coefficient value of GCV which is 86.32% greater than the CV value which is 85.92%, so the GCV method is better than the CV method even though in reality the determination coefficient value of CV is more than 67%. Therefore, a better method to use to estimate the HDI value in Indonesia in 2023 is the GCV method.

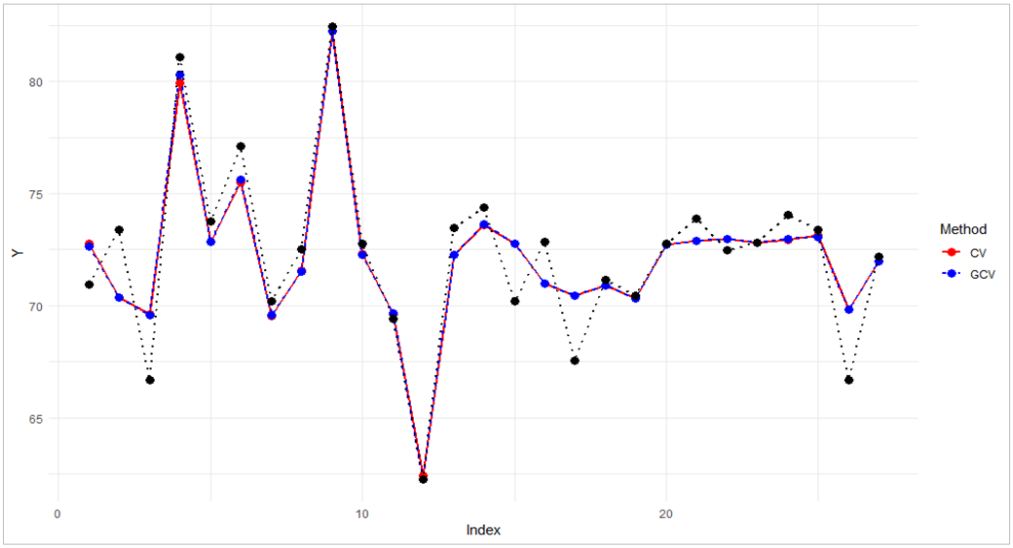

Fig. 8. Plot of Actual Data Estimation using CV and GCV Methods

Table VI. Comparison of Model Evaluations

Method	R^2	MAPE
CV	85.92%	1.96%
GCV	86.32%	1.94%

CV: Cross-Validation; GCV: Generalized Cross-Validation

Conclusion

Based on the analysis conducted, it can be concluded that the Generalized Cross-Validation (GCV) method outperforms the Cross-Validation (CV) method in estimating HDI based on education quality in Indonesia for 2023. The GCV method offers greater flexibility than CV because it addresses the limitations of the CV formula, which may produce invalid results in certain cases.

The GCV method produces optimal bandwidth values of and with a coefficient of determination (R^2) of 86.32%. This result indicates that the predictor variables, which are the mean years of schooling and the expected years of schooling, significantly explain the response variable, which is HDI. These findings confirm the robustness of the model. Furthermore, the MAPE of 1.94% indicates that the model achieves excellent forecasting accuracy.

Declarations

Conflict of interest

The authors declare that they have no conflict of interest.

Funding/ financial support

The work is self-funded and No fund is received from any agency.

Acknowledgments

None

References

[1] Abdy, M. (2019). Tinjauan Singkat Tentang Regresi Parametrik dan Regresi non Parametrik. Saintifik, 5(1), 58–62. https://doi.org/10.31605/saintifik.v5i1.199

[2] Lamusu, F., Machmud, T., & Resmawan, R. (2020). Estimator Nadaraya-Watson dengan Pendekatan Cross Validation dan Generalized Cross Validation untuk Mengestimasi Produksi Jagung. Indonesian Journal of Applied Statistics, 3(2), 93. https://doi.org/10.13057/ijas.v3i2.42125

[3] Ogden, R. T. (1997). Essential Wavelets for Statistical Applications and Data Analysis (Vol. 16, Issue 1).

[4] Hardle, W. (1994). Applied Nonparametric Regression. In Cambridge University Press (Vol. 156, Issue 1). https://doi.org/10.2307/2982873

[5] Suparti, S., Santoso, R., Prahutama, A., & Devi, A. R. (2018). Regresi Nonparametrik. Wade Group.

[6] Sadek, A. M., & Mohammed, L. A. (2024). Evaluation of the Performance of Kernel Non-parametric Regression and Ordinary Least Squares Regression. International Journal on Informatics Visualization, 8(3), 1352–1360. https://doi.org/10.62527/joiv.8.3.2430

[7] Puspitasari, I., Suparti, & Wilandari, Y. (2012). Analisis Indeks Harga Saham Gabungan (IHSG) dengan Menggunakan Model Regresi Kernel. Jurnal Gaussian, 1(1), 93–102.

[8] Astuti, D. A. D., Srinadi, I. G. A. M., & Susilawati, M. (2018). Pendekatan Regresi Nonparametrik Dengan Menggunakan Estimator Kernel Pada Data Kurs Rupiah Terhadap Dolar Amerika Serikat. E-Jurnal Matematika, 7(4), 305. https://doi.org/10.24843/mtk.2018.v07.i04.p218

[9] Razak, R. A., Nur, I. M., & Arum, P. R. (2019). Penerapan Cross Validation (CV) dalam Pemilihan Bandwidth Optimal pada Pemodelan Regresi Nonparametrik Kernel (Studi Kasus: Gizi Buruk pada Balita Di Indonesia). Prosiding Mahasiswa Seminar Nasional Unimus, 2, 364–372.

[10] Raghuvanshi, G., & Verma, D. P. (2024). Human Development Index: A Critical Review. International Journal for Multidisciplinary Research, 6(2), 1–15. https://doi.org/10.36948/ijfmr.2024.v06i02.18146

[11] Badan Pusat Statistik. (2024). Indeks Pembangunan Manusia 2023 (Vol. 18). Badan Pusat Statistik.

[12] Eubank, R. (1999). Nonparametric Regression and Spline Smoothing (Vol. 96, Issue 338).

[13] Silverman, B. W. (1986). Density Estimation for Statistics and Data Analysis. In Statistics and Applied Probability (Vol. 60, Issue 3). https://doi.org/10.3311/PPme.8017

[14] Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. Proceedings of the National Academy of Sciences, 42(1), 43–47. https://doi.org/10.1073/pnas.42.1.43

[15] Parzen, E. (1962). On Estimation of a Probability Density Function and Mode. The Annals of Mathematical Statistics, 33(3), 1065–1076. https://doi.org/10.1214/aoms/1177704472

[16] Machkouri, M. El. (2011). Asymptotic normality of the Parzen-Rosenblatt density estimator for strongly mixing random fields. Statistical Inference for Stochastic Processes, 14(1), 73–84. https://doi.org/10.1007/s11203-011-9052-4

[17] García, E. (2023). Notes for Predictive Modeling. bookdown.org/egarpor/pm-uc3m.

[18] Takezawa, K. (2006). Introduction To Nonparametric Regression. John Wiley & Sons, Inc.

[19] Carmack, P. S., Spence, J. S., & Schucany, W. R. (2011). Generalized Correlated Cross-Validation (GCCV). In Generalized Correlated Cross-Validation (pp. 1–28).

[20] Gujarati, D. N. (1972). Basic Econometrics. In The Economic Journal (Vol. 82, Issue 326). Gary Burke. https://doi.org/10.2307/2230043

[21] Chin, W. W. (1998). The Partial Least Squares Approach to Structural Equation Modeling. In G. A. Marcoulides (Ed.), Modern Methods for Business Research (Issue January 1998, pp. 295-336.).

[22] Maricar, M. A. (2019). Analisa Perbandingan Nilai Akurasi Moving Average dan Exponential Smoothing untuk Sistem Peramalan Pendapatan pada Perusahaan XYZ. Jurnal Sistem Dan Informatika, 13(2), 36-45.

[23] Chang, P.-C., Wang, Y.-W., & Liu, C.-H. (2007). The development of a weighted evolving fuzzy neural network for PCB sales forecasting. ScienceDirect, 32(1), 86–96. https://doi.org/10.1016/j.eswa.2005.11.021