Kernel Nonparametric Regression for Forecasting Local Original Income

Joji Ardian Pembargi[1], **Mustika Hadijati**[2], **Nurul Fitriyani**[3]

[1,2,3]University of Mataram, Indonesia

ABSTRACT

Regional Original Revenue (ROR) is an income collected based on regional regulations under statutory regulations. ROR aims to give authority to Regional Governments to sponsor the implementation of regional autonomy following regional potential. Every year, the Central Lombok Regency government sets ROR targets to assist the government in formulating regional policies. The targets set by the government are sometimes not following their realization. This study aims to determine a model that can be used in forecasting ROR targets. One way to predict the value of ROR is by using a nonparametric regression approach. This approach is flexible since it is not dependent on a particular model. The use of the nonparametric kernel regression method with the Gaussian kernel function obtained a minimum GCV value of 1,769688931 with an optimum bandwidth value of h_1 of 0,212740452 and h_2 of 0,529682589. Modeling with optimum bandwidth produces a coefficient of determination of 87,55%. The best model is used for forecasting and produces a MAPE value of 5,4%. The analysis results show that what influences the value of ROR is ROR receipts in the previous month and the previous 12 months.

Corresponding Author:

Joji Ardian Pembargi,
Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Mataram
Email: jojiardian15@gmail.com

A.   INTRODUCTION

Regional Original Revenue (ROR) is region income based on regional regulations according to statutory regulations. It consists of regional taxes and retribution, separated regional wealth management results, and other legitimate regional original revenues (Law of The Republic of Indonesia Number 23 of 2014 Concerning Regional Government, 2014). ROR aims to give authority to Regional Governments to fund the regional autonomy implementation following regional potential. In implementing regional autonomy, each region must develop and optimize all regional possibilities extracted from within the area concerned. As one of the regencies in the West Nusa Tenggara Province, Central Lombok Regency is one of the regions that implement autonomy like other regions. A series of incomes from various sectors are expected to contribute to regional development so that regional autonomy can run smoothly (Regional Regulation of Central Lombok Regency Number 6 of 2021 Concerning Budget Regional Revenue and Expenditures, 2021; Regional Regulation of West Nusa Tenggara Province Number 5 of 2018 Concerning Regional Fees, 2018).

Every year, the Government of Central Lombok Regency sets a ROR target. This target is made to assist the government in formulating policies towards the regions, especially in regulating regional budgets. The targets the government has set are sometimes not following their realization. Sometimes the target value exceeds the realized value, and vice versa; sometimes, the target value is below the completion (Regional Regulation of Central Lombok Regency Number 6 of 2021 Concerning Budget Regional Revenue and Expenditures, 2021). Therefore, the Central Lombok Regency government needs an accurate forecast of the amount of ROR for the future in formulating policies in determining future targets. Several methods can be used in forecasting, such as regression, time series, etc. Forecasting uses regression analysis, such as linear regression and multiple regression, while forecasting uses time

series analysis, such as ARIMA, SARIMA, and others (Palma, 2016). One method that is often used is the regression analysis method. Regression analysis is a statistical method used to see and study the relationships between one or more variables. Two regression approaches can be used to see the relationship pattern between the dependent and independent variables, i.e., parametric and nonparametric regression (Fitriyani and Budiantara, 2014; Sauri et al., 2021).

Parametric regression must meet normality, multicollinearity, heteroscedasticity, and autocorrelation assumptions. Another weakness of parametric regression is that not all relationship pattern problems can be estimated because the relationship between the variables does not yet have a particular curve pattern. While the nonparametric regression model is not bound by the assumption that the regression curve must be known in its shape, allowing the data to determine the shape of the curve itself so that it is flexible. In addition, nonparametric regression does not require assumptions that must be met and the pattern of relationships between variables. Nonparametric regression is an appropriate regression model used to model the presence of nonlinearity. Another advantage of the nonparametric regression method is that the calculations and procedures are more straightforward (Fitriyani and Budiantara, 2014; Ghosh, 2018; Hollander et al., 2013; Sauri et al., 2021).

Nonparametric methods can use several smoothing techniques, including kernel estimators, histograms, wavelets, Fourier series, splines, k-NN, and orthogonal series. The kernel estimator is an expansion of the histogram estimator. The kernel estimator is identical to other nonparametric regression estimators, the only difference being that the kernel estimator is more specialized in using bandwidth methods. Bandwidth is a smoothing parameter that controls the smoothness of the estimated curve. A bandwidth that is too small will result in a curve estimate that is too rough, while a bandwidth that is too large will result in a curve estimate that is too smooth. The kernel estimator's advantage is that it can model data that does not have a particular pattern (Chacón and Duong, 2018; Scott, 2015; Hollander et al., 2013).

Previous studies have used nonparametric kernel regression, such as modeling the Jangkok Watershed, Lombok Island's water discharge by autoregressive pre-whitening. The kernel approach also applies in the statistical downscaling model to predict the daily rainfall in Dodokan Watershed, simulate monthly rainfall data of Dodokan Watershed, and expect the temperature in Selaparang Lombok, based on (Hadijati et al., 2016; Hadijati et al., 2017; Hadijati et al., 2021, Hadijati and Irwansyah, 2022). Previous studies tend to model time series data related to climate data by utilizing the kernel approach. Research related to the prediction of ROR using the kernel nonparametric regression method, especially in Central Lombok, is not been widely explored.

Based on the previous background, this study aims to forecast the total receiving regional original revenue in Central Lombok using the nonparametric kernel regression. This study is expected to assist regional governments in formulating policies in determining future targets, especially related to the original regional revenue.

B.   RESEARCH METHOD

The data used is secondary data obtained from the Office of the Central Lombok Regency Revenue Management Agency which is located at Jalan Raden Puguh PrayaPuyung 2, Praya. The variables used are data on the amount of ROR receipts in Central Lombok for the last five years, from January 2016 to December 2020, with 60 data as in-samples used to obtain the nonparametric kernel model and 2021 data as many as 12 data as out-samples. The out-sample data is used to predict the ROR value 12 months later. The prediction results obtained will be compared with actual data to see the accuracy of the predictions. ROR data consists of the results of regional taxes, regional levies, separated wealth management, and other legitimate regional income. ROR data is time series data with monthly time. Data analysis was assisted by R software. This study aims to determine the nonparametric kernel model and predict ROR acceptance in Central Lombok. The method used is nonparametric regression analysis, with research procedures in Figure 1.
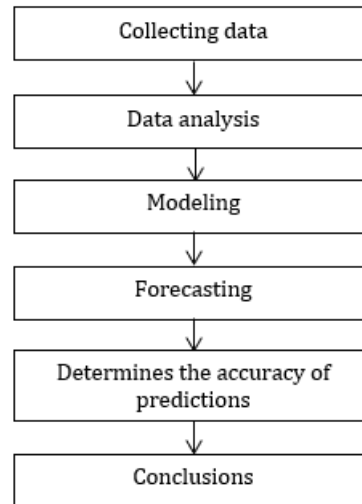
Figure 1. Research procedures

The steps in this research are divided into several stages:

1. Determine the dependent variable (Y) and independent variable (X) by looking at the correlated lag value on the autocorrelation function (ACF) plot (Palma, 2016).

2. Create a scatterplot. A scatterplot is used to see the pattern of data distribution.

3. Perform a multicollinearity test. The occurrence of multicollinearity is a condition that indicates a strong correlation or relationship between independent variables in the regression model. Determining multicollinearity can be known by looking at the Variance Inflation Factors (VIF) value. If the VIF value > 10, then multicollinearity occurs. The multicollinearity test can be done with Equation 1 (Michael et al., 2015):

$$VIF_j = \frac{1}{1 - R_j^2} \quad j = 1, 2, 3, \ldots, m \tag{1}$$

R_j^2 is the coefficient of determination.

4. Determine the optimal bandwidth value that minimizes the value of Generalized Cross Validation (GCV). The formula is in Equation 2 (Budiantara et al., 2015; Fitriyani and Budiantara, 2014)

$$GCV(h) = \frac{n^{-1} \sum_{i=1}^{n} (y_i - \widehat{y})^2}{\left(n^{-1} tr \left(I - X(h) \right) \right)^2} \tag{2}$$

y_i is the actual data, $\widehat{y_i}$ is the estimated data, I is the identity matrix, $X(h)$ is the weighting matrix, n is the number of data and tr is the number of main diagonal elements of weighting matrix.

5. Model the ROR data using nonparametric kernel regression. In general, kernel functions is defined in Equation 3 (Deshpande et al., 2017):

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

K is the kernel function and h is the bandwidth. Kernel functions must meet several conditions, i.e.:

  (a) $K_h(x) > 0, \forall x$
  (b) $\int_{-\infty}^{\infty} K(x) \, dx = 1$
  (c) $\int_{-\infty}^{\infty} x^2 K(x) \, dx = \sigma^2 > 0$
  (d) $\int_{-\infty}^{\infty} x K(x) \, dx = 0$

A kernel estimator is a nonparametric approach to estimating curves. The kernel estimator is very sensitive to the choice of bandwidth as a curve smoothness controller. Nadaraya and Watson define a kernel regression estimator called the Nadaraya-Watson estimator in Equation 4 (Ghosh, 2018; Pratiwi et al., 2020):

$$m(x_i) = \frac{\sum_{i=1}^{n} \left(\prod_{j=1}^{m} K \left(\frac{x_j - x_{ij}}{h_j} \right) \right) y_i}{\sum_{i=1}^{n} \left(\prod_{j=1}^{m} K \left(\frac{x_j - x_{ij}}{h_j} \right) \right)} \tag{4}$$

$K\left(\dfrac{x_j - x_{ij}}{h_j}\right)$ is a kernel function. The kernel function used is the Gaussian kernel in Equation 5:

$$K\left(\frac{x_j - x_{ij}}{h_j}\right) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x_j - x_{ij}}{h_j}\right)^2\right) \tag{5}$$

6. Calculate the goodness of the model with the coefficient of determination in Equation 6 (Cheng et al., 2014; Zhang, 2017):

$$R^2 = \frac{\sum_{i=1}^n (\widehat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \tag{6}$$

$\widehat{y}_i$ represent the forecasting values of y_i, and $\bar{y}$ indicates the average value of y_i.

7. Check the random residual assumptions, i.e., to be identical and independent ($\varepsilon \sim IID(0, \sigma^2)$).
8. Forecast with the model obtained will be tested for forecasting accuracy with the Mean Absolute Percentage Error (MAPE) criteria in Equation 7 (Palma, 2016):

$$MAPE = \frac{\sum_{i=1}^n \left|\dfrac{y_i - \widehat{y}_i}{y_i}\right|}{n} \times 100\% \tag{7}$$

$\widehat{y}_i$ represent the forecasting values of y_i

C. RESULTS AND DISCUSSION

The data used in this research is secondary data, namely monthly ROR data in Central Lombok from January 2016 to December 2020. The total data from January 2016 to December 2020 is 60 as in sample data and 12 from January 2021 to December 2021 as data out samples. The in-sample data will be used to form the best model, and the out-sample data will be used to measure the accuracy of ROR predictions in Central Lombok for the next 12 months based on the model that has been obtained.

Plotting the ACF on ROR data was used to determine the independent variable. It determined the autocorrelation between the time lag. The following is an ACF plot of ROR data:
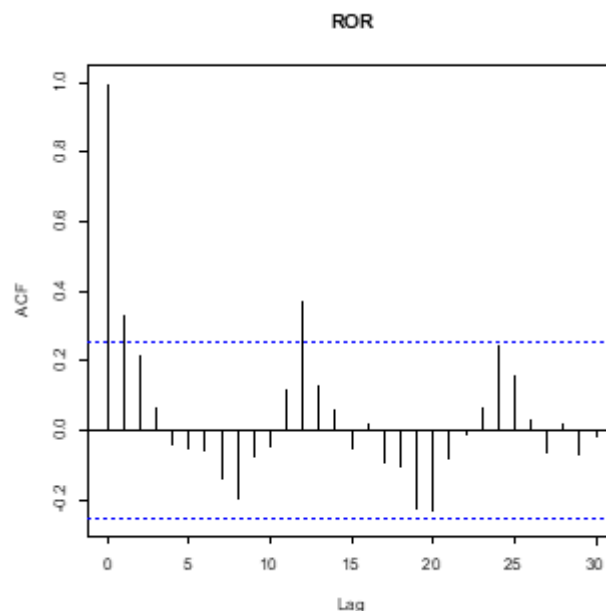

Figure 2. Data autocorrelation function plot

It can be noticed in Figure 2 that the ACF plot cut off at lag one and lag 12. Therefore, in this study, lag one and lag 12 were used as independent variables that were indicated to affect the ROR value at time t, with $t = 1, 2, 3, \ldots, 60$. Furthermore, ROR lag 1 data or data to x_{t-1} is referred to as x_1, ROR lag data 12 or data to x_{t-12} is referred to as x_2 and data x_t is referred to as y. Figure 2 shows the scatter plots between the variables and is used to identify the data distribution pattern.

Based on Figure 3, indicates the relationship pattern between the dependent variable y and each independent variable x_1 and x_2. It can be seen that the pattern of relationships produced by both variables between x_1 and y as well as variables between x_2 and y can be said to be spread or not have specific patterns such as linear, quadratic, and others. Therefore, the appropriate approach for the relationship pattern is a nonparametric regression oncoming. In this study, a nonparametric kernel regression approach was used.
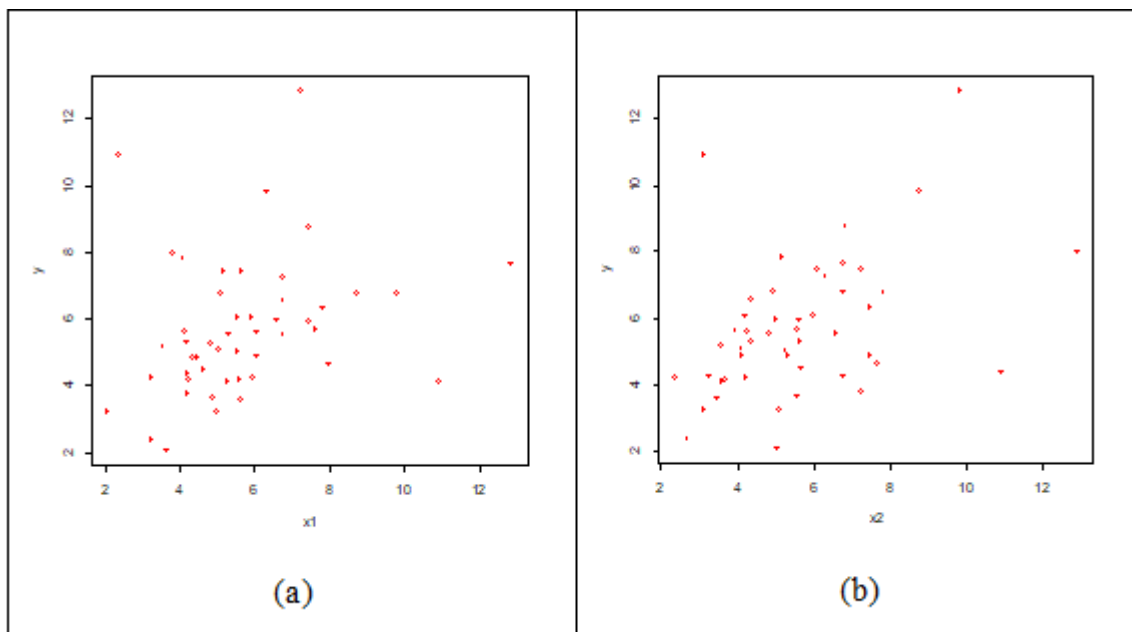

Figure 3. Data autocorrelation function plot

Moreover, the VIF values were used to determine whether or not multicollinearity occurs. If the VIF value is more than 10, multicollinearity exists, so there is a strong relationship within the independent variables. Based on the analysis, it is found that both variables have a VIF value less than 10, i.e., the VIF value of 1,016, which indicates that there is no multicollinearity.

Furthermore, it is compulsory to determine the optimum bandwidth value used in the modeling. A bandwidth that is too small will produce a too-rough curve; on the other hand, a too-large bandwidth will create a too-smooth curve that does not fit the data pattern. Therefore, it is necessary to choose the optimal bandwidth. One method to get the optimal bandwidth is using the GCV criteria, a criterion usually used in nonparametric regression approaches (Fitriyani and Budiantara, 2014; Ghosh, 2018). Some of the GCV values obtained are as follows:

Table 1. Bandwidth and GCV values

h_1	h_2	**CGVV**
0,212740452	0,529682589	1,769688931
0,223299858	0,543917101	1,796199603
0,225825242	0,624568984	1,858577982
0,247327220	0,573076774	1,883614084
0,150406123	0,573632985	1,945966741

Based on the GCV criteria in Equation 2, the minimum GCV value is 1,769688931, with the optimal bandwidth $h_1 = 0,212740452$ and $h_2 = 0,529682589$. The kernel regression curve estimation model with the estimator used is Nadaraya-Watson multivariate form with $j = 1, 2$, where j is the number of independent variables. The data used are $i = 1, 2, 3, \ldots, 48$. From Equation 4, it is obtained:

$$\widehat{y} = n^{-1} \sum_{i=1}^{48} \frac{\prod_{j=1}^{2} \frac{1}{h_j} K\left(\frac{x_j - x_{ij}}{h_j}\right)}{n^{-1} \sum_{i=1}^{48} \left(\prod_{j=1}^{2} \frac{1}{h_j} K\left(\frac{x_j - x_{ij}}{h_j}\right)\right)} y_i$$

$$= n^{-1} \sum_{i=1}^{48} W_{hi}(x) y_i$$

$$= n^{-1} W_{h1}(x) y_1 + n^{-1} W_{h2}(x) y_2 + \ldots + n^{-1} W_{h48}(x) y_{48}$$

The $W_{hi}(x_j)$ function is a multivariate kernel density function, namely:

$$W_{hi}(x_j) = \frac{\prod_{j=1}^{2} \frac{1}{h_j} K\left(\frac{x_j - x_{ij}}{h_j}\right)}{n^{-1} \sum_{i=1}^{n} \left(\prod_{j=1}^{2} \frac{1}{h_j} K\left(\frac{x_j - x_{ij}}{h_j}\right)\right)}$$

The kernel function used is the Gaussian kernel in Equation 5. Therefore, the form of the function $W_{h_i}(x_j)$ becomes,

$$W_{hi}(x_j) = \frac{\prod_{j=1}^{2} \frac{1}{h_j} \exp\left(-\frac{1}{2} \frac{x_j - x_{ij}}{h_j}\right)}{n^{-1} \sum_{i=1}^{n} \prod_{j=1}^{2} \frac{1}{h_j} \exp\left(-\frac{1}{2} \frac{x_j - x_{ij}}{h_j}\right)}$$

The summation form of $\widehat{y}_i$ is:

$$\widehat{y}_i = n^{-1} W_{h1}(x) y_1 + n^{-1} W_{h2}(x) y_2 + \ldots + n^{-1} W_{h48}(x) y_{48}$$

Since the model applies from the $i = 1$ to $i = 48$, so the model can also be written in matrix form.

$$\begin{bmatrix} \widehat{y}_1 \\ \widehat{y}_2 \\ \vdots \\ \widehat{y}_{48} \end{bmatrix} = \begin{bmatrix} n^{-1} W_{h1}(x_1) y_1 + n^{-1} W_{h2}(x_1) y_2 + \ldots + n^{-1} W_{h48}(x_1) y_{48} \\ n^{-1} W_{h1}(x_2) y_1 + n^{-1} W_{h2}(x_2) y_2 + \ldots + n^{-1} W_{h48}(x_2) y_{48} \\ \vdots \\ n^{-1} W_{h1}(x_{48}) y_1 + n^{-1} W_{h2}(x_{48}) y_2 + \ldots + n^{-1} W_{h48}(x_{48}) y_{48} \end{bmatrix}$$

$$= \begin{bmatrix} n^{-1} W_{h1}(x_1) & n^{-1} W_{h2}(x_1) & \ldots & n^{-1} W_{h48}(x_1) \\ n^{-1} W_{h1}(x_2) & n^{-1} W_{h2}(x_2) & \ldots & n^{-1} W_{h48}(x_2) \\ \vdots & \vdots & \ddots 7\vdots & \\ n^{-1} W_{h1}(x_{48}) & n^{-1} W_{h2}(x_{48}) & \ldots & n^{-1} W_{h48}(x_{48}) \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{48} \end{bmatrix}$$

The estimation curve of the nonparametric kernel model is written in the matrix form in Equation 8.

$$\widehat{y}_i = X(h) y_i \tag{8}$$

with,

$$\widehat{y}_i = \begin{bmatrix} \widehat{y}_1 \\ \widehat{y}_2 \\ \vdots \\ \widehat{y}_{48} \end{bmatrix}, y_i = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{48} \end{bmatrix}, X(h) = \begin{bmatrix} n^{-1} W_{h1}(x_1) & n^{-1} W_{h2}(x_1) & \ldots & n^{-1} W_{h48}(x_1) \\ n^{-1} W_{h1}(x_2) & n^{-1} W_{h2}(x_2) & \ldots & n^{-1} W_{h48}(x_2) \\ \vdots & \vdots & \ddots 7\vdots & \\ n^{-1} W_{h1}(x_{48}) & n^{-1} W_{h2}(x_{48}) & \ldots & n^{-1} W_{h48}(x_{48}) \end{bmatrix}$$

Furthermore, a coefficient of determination (R^2), a criterion for the model's goodness, was used to see the estimate's accuracy. The value of R^2 is an indicator of how good the resulting model can describe the variability of the data. Based on Equation 6, the coefficient of determination is 0,8755. The coefficient of determination for the Nadaraya-Watson kernel estimator indicates that the independent variable used influences the dependent variable by 87,55%.

The model obtained with the optimum bandwidth can be used for forecasting in the future. In this study, the out-sample data used is 2021, with $i = 1, 2, 3, \ldots, 12$. Figure 4 indicates the comparison between the actual and estimated data.
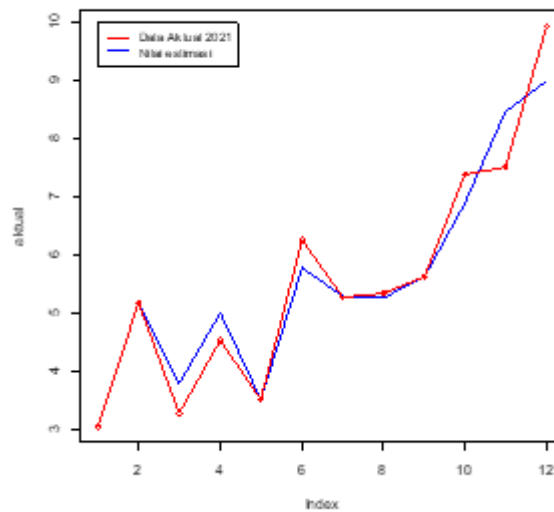
Figure 4. The actual and estimated data

Based on Figure 4, comparing the estimated data's value to the actual data shows results close to the actual data. Based on the forecasting results, the forecast accuracy is measured by a MAPE value of 5,4%. The results indicate that the prediction results obtained are very good because the MAPE value is less than 10%, so it can be said that the prediction results are very accurate. From the analysis results, it is found that what influences the value of ROR is ROR receipts in the previous month and 12 months before so that the relevant government can consider the analysis results. Compared to the research conducted by Latipah (Latipah et al., 2019), who predicted ROR using the Grey-Markov model (1,1), the comparison of prediction accuracy is still better than the results obtained by the kernel nonparametric method. Similar research was also conducted by Fitriyani et al. (2023), who performed modeling and comparison of estimated kernel nonparametric regression curves.

D. CONCLUSION AND SUGGESTION

Based on the analysis performed, it can be concluded that the kernel nonparametric regression model obtained with optimum bandwidth h_1 of 0,212740452 and h_2 of 0,529682589, with a minimum GCV value is 1,769688931. Based on the modeling that has been done using nonparametric kernel regression in the ROR case in Central Lombok, it was found that the forecasting results had a very accurate prediction. Based on the analysis performed, it can be concluded that the kernel nonparametric regression model obtained with optimum bandwidth h_1 of 0,212740452 and h_2 of 0,529682589, with a minimum GCV value is 1,769688931. Based on the modeling that has been done using nonparametric kernel regression in the ROR case in Central Lombok, it was found that the forecasting results had a very accurate prediction. The results of this study suggest the Regional Revenue Management Agency for Central Lombok Regency considers the state of the ROR data for the previous month and 12 months in determining ROR targets. In addition, future research is expected to utilize optimization algorithms in determining optimal bandwidth values in kernel approaches, such as genetic algorithms, Particle Swarm Optimization (PSO), and Simulated Annealing (SA).

REFERENCES

Budiantara, I. N., Ratnasari, V., Ratna, M., and Zain, I. (2015). The combination of spline and kernel estimator for nonparametric regression and its properties. *Applied Mathematical Sciences*, 9(122):6083–6094.

Chacón, J. and Duong, T. (2018). Multivariate kernel smoothing and its applications, vol. 160 of. *Monographs on Statistics and Applied Probability*.

Cheng, C.-L., Garg, G., et al. (2014). Coefficient of determination for multiple measurement error models. *Journal of Multivariate Analysis*, 126:137–152.

Deshpande, J. V., Naik-Nimbalkar, U., and Dewan, I. (2017). *Nonparametric statistics: Theory and methods*. World Scientific.

Fitriyani, N. and Budiantara, I. (2014). Metode cross validation dan generalized cross validation dalam regresi nonparametrik spline (studi kasus data fertilitas di jawa timur). In *Prosiding Seminar Nasional Pendidikan Sains. ISBN*, pages 978–602.

Fitriyani, N., Pembargi, J. A., Setiawana, E., et al. (2023). Modeling the hotel tax revenue in central lombok using nonparametric regression. *Jurnal Matematika, Statistika dan Komputasi*, 19(3):498–505.

Ghosh, S. (2018). *Kernel smoothing: Principles, methods and applications*. John Wiley & Sons.

Hadijati, M., Ali, I., Fitriyani, N., et al. (2021). Prediction of daily rainfall in dodokan watershed based on statistical downscaling model: An effort to manage watershed ecosystems mustika hadijati1*, irwansyah 1, nurul fitriyani 1. *Eastern Journal of Agricultural and Biological Sciences*, 1(2):1–7.

Hadijati, M. and Irwansyah, N. F. (2022). Autoregressive prewhitening on the nonparametric regression model of water discharge in the jangkok watershed, lombok island. *Global Journal of Pure and Applied Mathematics*, 18(1):307–318.

Hadijati, M., Komalasari, D., and Fitriyani, N. (2016). Statistical downscaling model using nonparametric regression to predict temperature in selaparang lombok. *Scientific Committe*, page 129.

Hadijati, M., Komalasari, D., and Fitriyani, N. (2017). Simulation of monthly rainfall data of dodokan watersheed using nonparametric statistical downscaling model. In *The 2nd International Conference on Science and Technology 2017*, pages 213–219.

Hollander, M., Wolfe, D. A., and Chicken, E. (2013). *Nonparametric statistical methods*. John Wiley & Sons.

Latipah, L., Wahyuningsih, S., and Syaripuddin, S. (2019). Peramalan pendapatan asli daerah provinsi kalimantan timur menggunakan model grey-markov (1, 1). *Jambura Journal of Mathematics*, 1(2):89–103.

Law of The Republic of Indonesia Number 23 of 2014 Concerning Regional Government (2014). Law of the republic of indonesia number 23 of 2014 concerning regional government.

Michael, O., Hussaini, G., and Agboola, S. (2015). Variance inflation factor: as a condition for the inclusion of suppressor variable (s) in regression anarysis. *Sci Res Publ*, 5:754–67.

Palma, W. (2016). *Time series analysis*. John Wiley & Sons.

Pratiwi, D., Mursy, L. A. A., Rizaldi, M., and Fitriyani, N. (2020). Regresi nonparametrik kernel gaussian pada pemodelan angka kelahiran kasar di provinsi nusa tenggara barat. *EIGEN MATHEMATICS JOURNAL*, pages 100–105.

Regional Regulation of Central Lombok Regency Number 6 of 2021 Concerning Budget Regional Revenue and Expenditures (2021). Regional regulation of central lombok regency number 6 of 2021 concerning budget regional revenue and expenditures.

Regional Regulation of West Nusa Tenggara Province Number 5 of 2018 Concerning Regional Fees (2018). Regional regulation of west nusa tenggara province number 5 of 2018 concerning regional fees.

Sauri, M. S., Hadijati, M., and Fitriyani, N. (2021). Spline and kernel mixed nonparametric regression for malnourished children model in west nusa tenggara. *Jurnal Varian*, 4(2):99–108.

Scott, D. W. (2015). *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.

Zhang, D. (2017). A coefficient of determination for generalized linear models. *The American Statistician*, 71(4):310–316.