Available online at http://scik.org

Commun. Math. Biol. Neurosci. 2025, 2025:36

https://doi.org/10.28919/cmbn/8911

ISSN: 2052-2541

THE USE OF LIKELIHOOD-BASED THRESHOLD IN ESTIMATING NONPARAMETRIC REGRESSION MODELS THROUGH THE ADAPTIVE NADARAYA-WATSON ESTIMATOR

NUR ASFIRAH, ANNA ISLAMIYATI*, NIRWAN

Department of Statistics, Faculty of Mathematics and Natural Sciences, Hasanuddin University, Makassar 90245, Indonesia

Abstract: This study develops a nonparametric regression model using the Likelihood-Based Threshold through the Adaptive Nadaraya-Watson Estimator to model rice productivity data in South Sulawesi in 2023. The issue of data variability can be addressed by simultaneously using bandwidth and threshold to improve estimation accuracy, compared to using only bandwidth. This problem is solved by integrating an adaptive threshold, which allows the estimator to adjust to the characteristics of the data. This method considers the distance between data points and the variation, enabling a more responsive estimation of changes in data patterns. This research aims to obtain the best nonparametric regression model to forecast rice productivity data. The best model is determined using the criterion of the minimum Mean Squared Error (MSE). The analysis results show that the optimal values are h=0.92 and δ=0.99, with the smallest MSE value of 0.075, it produces accurate predictions.

Keywords: threshold; bandwidth; likelihood; Nadaraya Watson; rice productivity; land area.

2020 AMS Subject Classification: 62G08.

1. INTRODUCTION

The primary objective of statistical analysis in observational or clinical research is to discover

*Corresponding author

E-mail address: annaislamiyati701@gmail.email

Received September 19, 2024

and determine the characteristics of the relationship between predictor and response variables. Regression analysis, especially parametric regression, can be more effective under certain conditions, but it is limited due to strict assumptions and difficulty detecting nonlinear relationships [1]. On the other hand, non-parametric regression has become an essential tool because it does not rely on predefined assumptions about the functional form of the relationship, making it more flexible and allowing the data itself to find the appropriate relationship pattern [2]. Several estimators in non-parametric regression include spline [3], kernel [4], local polynomial [5], and Fourier series [6]. Each estimator has different parameters to obtain accurate estimation results. In the spline estimator, optimal knot points indicate where data patterns change [7]. The kernel estimator, specifically the Nadaraya-Watson estimator, depends on the bandwidth parameter, which determines the smoothness of the regression model [8]. This bandwidth parameter affects how much area around the observation point is used for estimation, focusing on approximations around the target point with weights varying by distance [9]. The estimation process for the local polynomial estimator depends on the degree of the polynomial chosen and the bandwidth used to determine the local area around each data point [5]. The Fourier series estimator, on the other hand, depends on the number of terms used in the Fourier series. This research uses a kernel approach in non-parametric regression, which requires the selection of appropriate bandwidth and kernel functions to control the smoothness of the estimated curve [10]. Incorrect bandwidth selection can lead to high bias or variance, making it essential to choose the optimal bandwidth, often measured using Mean Squared Error (MSE), to achieve the best estimation [4]. The kernel approach is more suitable for analyzing agricultural data with complex and variable characteristics [11].

The Nadaraya-Watson estimator without *threshold* was used, taking advantage of the bandwidth parameter to account for variations caused by seasonal changes and geographical differences, resulting in more precise agricultural production estimates [8]. The proposed study focuses on the simultaneous use of bandwidth and threshold. Using both parameters in kernel regression aims to improve estimation accuracy by adjusting the estimator to the characteristics of the data [12]. This research will apply non-parametric regression using the kernel approach, specifically the Nadaraya-Watson estimator with a Gaussian function on rice productivity data in South Sulawesi. This method is considered highly suitable for agricultural data analysis due to its flexibility in handling nonlinear data and its ability to capture variations influenced by factors such as seasonal fluctuations and geographical differences. This method shows the potential for generating more responsive estimates compared to using bandwidth alone.

2. PRELIMINARIES

1. Adaptive Nadaraya Watson Estimator

Nadaraya and Watson in 1994 defined the kernel regression estimator hence it is called the Nadaraya-Watson estimator. The general form of the Nadaraya-Watson estimator function is as follows.

$$m(x) = \frac{\sum_{i=1}^{n}(\frac{K_h(x-x_i)}{h})y_i}{\sum_{j=1}^{n}(\frac{K_h(x-x_i)}{h})}$$

The Nadaraya-Watson kernel is one type of kernel function that applies a weighting concept to provide further influence on data points that are closer to the evaluation point $\hat{x}$. These weights are calculated using the kernel function, which decreases rapidly as the distance from its center increases [13]. Adding a threshold to kernel estimation is necessary when the data has a considerable variation or significant noise, such as unevenly distributed, dense, or sparse data clusters. The purpose of the threshold function is to help avoid overfitting or underfitting by ensuring that the selected bandwidth provides optimal results and does not merely adjust to noise in the data. If the general form of the Nadaraya-Watson estimator function is used without a threshold, there is a risk of overfitting in dense data clusters because nearby points will overly influence the estimation. Conversely, sparse data clusters may experience underfitting as the estimation becomes too smooth and fails to capture essential patterns [4]. Therefore, it is crucial to set the threshold parameter adaptively. The formula for the Adaptive Nadaraya-Watson Estimator with a Gaussian function is presented as follows:

$$m(x) = \frac{\sum_{i=1}^{n} exp\left(-\frac{(x-x_i)^2}{2h^2\delta^2}\right)y_i}{\sum_{i=1}^{n} exp\left(-\frac{(x-x_i)^2}{2h^2\delta^2}\right)}$$

In the adaptive estimator, the local bandwidth $h_i$is adjusted based on the threshold δ to further improve estimation accuracy [4].

2. Likelihood-based threshold

A likelihood-based threshold is a threshold selection method based on a statistical model's likelihood or log-likelihood values. In the context of kernel regression, specifically the Nadaraya-Watson Estimator, selecting the optimal bandwidth is crucial to maintaining a balance between bias and variance. If the bandwidth is too small, the model will suffer from overfitting, while a too large bandwidth leads to underfitting. Using a likelihood-based threshold, the threshold value for

bandwidth selection is adjusted according to the likelihood values produced by the model. Bandwidth selection is done by maximizing the log-likelihood, which means choosing the bandwidth with the highest likelihood value. It is then evaluated using the Mean Squared Error (MSE), indicating that the model with the smallest MSE is the best [12].

The primary advantage of the likelihood-based threshold approach is its adaptive nature, as the threshold is determined based on the characteristics of the data being analyzed. This offers greater flexibility compared to universal or fixed thresholds, which use static values and do not account for variations in the data. This method is often used in model selection, where the threshold is adjusted to produce more accurate estimates [1].

3. Bandwidth

Bandwidth is a parameter denoted by h used in kernel regression that aims to smooth the shape of the curve obtained from the estimation. Optimal bandwidth selection helps control the balance between the smoothness of the function and fit to the data. According to Kerpicci et al. (2020), a bandwidth value that is too large will make the curve very smooth but is prone to losing data variability, while a bandwidth value that is too small will form a fluctuating and rough curve [14]. Therefore, bandwidth selection must be done carefully to obtain an optimal estimator. Fadillah (2022) suggests several methods commonly used to select bandwidth aside from the bandwidth rule of thumb formulae, such as unbiased cross-validation, biased cross-validation, complete cross, and trial and error [9]. The method used is trial and error, where the bandwidth value h is chosen randomly, with $h \in (0,1)$, while the threshold δ is obtained based on that bandwidth value and depends on the data used.

3. MAIN RESULTS

1. Adaptive Nadaraya-Watson Kernel Model Estimation

The predicted value $\hat{\boldsymbol{y}}$ is calculated as the ratio of the sum of the exponential weights of y_i to the sum of the exponential weights. These weights are determined by the Gaussian kernel function, which considers the distance between x and x_i, as well as the bandwidth parameter h and the threshold δ. Using the Gaussian kernel in Kernel Density Estimation (KDE) helps smooth the data and estimate the probability distribution of the observed data. By using the bandwidth h and the threshold δ, the model's sensitivity to the distance between data points can be adjusted. There is a threshold parameter δ that will be estimated using the Kernel Density Estimation (KDE) and Maximum Likelihood Estimation (MLE) methods. The Gaussian kernel is used in Kernel Density

Estimation (KDE) to smooth the data.

The Gaussian kernel function is written as follows:

$$K(u) = \frac{1}{\sqrt{2\pi}} exp\left(-\frac{u^2}{2}\right)$$

with bandwidth h and threshold δ be:

$$K_{h.\delta^2} = \frac{1}{h\delta\sqrt{2\pi}} exp\left(-\frac{u^2}{2h^2\delta^2}\right)$$

The Likelihood function estimates model parameters with the Maximum Likelihood Estimation (MLE) approach. The likelihood function $L(h)$ is used to estimate the bandwidth parameter h and the threshold δ. This likelihood function is calculated based on the error between the observed value y and the estimated value $\hat{f}(x)$. The likelihood function $L(h)$ is written as follows:

$$L(h) = \prod_{i=1}^{n} \frac{1}{h\delta\sqrt{2\pi}} exp\left(-\frac{u^2}{2h^2\delta^2}\right)$$

with $u = y - \hat{f}(x)$ being the residual or error between the observed value y and estimate value $\hat{f}(x)$.

$$L(h) = \prod_{i=1}^{n} \frac{1}{h\delta\sqrt{2\pi}} exp\left(-\frac{\left(y - \hat{f}(x)\right)^2}{2h^2\delta^2}\right)$$

To simplify the calculation, the likelihood function is transformed into the natural logarithmic form:

$$ln(e) = ln(2\pi)^{-n/2} + ln(h)^{-n} + ln(\delta)^{-n} + ln\, exp\left(\sum_{i=1}^{n} -\frac{(y - \hat{f}(x))^2}{2h^2\delta^2}\right)$$

Through the estimation process using the MLE method, the threshold parameter δ is obtained as follows:

$$\delta = \sqrt{\frac{1}{nh^2}\sum_{i=1}^{n}\left[y_i - \frac{\sum_{i=1}^{n} exp\left(-\frac{(x - x_i)^2}{2}\right) y_i}{\sum_{i=1}^{n} exp\left(-\frac{(x - x_i)^2}{2}\right)}\right]^2}$$

Next, after the threshold parameter has been estimated, the values obtained from that formula are presented in the adaptive Nadaraya-Watson estimation model as follows:

$$\hat{y} = \hat{m}(x) = \frac{\sum_{i=1}^{n} exp\left(-\frac{(x-x_i)^2}{2h^2\delta^2}\right) y_i}{\sum_{i=1}^{n} exp\left(-\frac{(x-x_i)^2}{2h^2\delta^2}\right)}$$

2. Application of the actual example

Given the data on rice productivity in South Sulawesi Province for the year 2023, $n = 1,2,3,\ldots,24$ with x representing the land area in units of hectares, y represents rice productivity in quintals per hectare. The likelihood-based threshold estimation on the Nadaraya-Watson estimator is called the Adaptive Nadaraya-Watson estimation. The selection of adaptive bandwidth depends on the threshold value. In the rice productivity data, Table 1 shows the obtained threshold values. Using the values of h and δ from Table 1, the next step is to calculate the smallest Mean Squared Error (MSE) to obtain the optimum h and δ. In this discussion, the Adaptive Nadaraya-Watson estimation model with the Gaussian kernel function and the addition of a threshold is used to obtain the estimated value $\hat{\boldsymbol{y}}$ or $\hat{\boldsymbol{m}}(\boldsymbol{x})$.

The following is the model for Adaptive Nadaraya-Watson estimation:

$$\hat{y} = \hat{m}(x) = \frac{\sum_{i=1}^{n} exp\left(-\frac{(x-x_i)^2}{2h^2\delta^2}\right) y_i}{\sum_{i=1}^{n} exp\left(-\frac{(x-x_i)^2}{2h^2\delta^2}\right)}$$

Table 1. Threshold Value

h	δ
0.10	9.15
0.70	1.31
0.90	1.02
0.91	1.01
0.92	0.99
0.93	0.98
0.95	0.96
0.97	0.94

Previously, the value of δ was obtained from each predetermined value of h. The optimal bandwidth and threshold are evaluated using MSE. In detail, the MSE values obtained are as follows:

Table 2. MSE Value

h	δ	MSE
0.10	9.15	0.07532281
0.70	1.31	0.07532281
0.90	1.02	0.07532581
0.91	1.01	0.07532893
0.92	**0.99**	**0.07531646**
0.93	0.98	0.07531737
0.95	0.96	0.07531828
0.97	0.94	0.07531798

Based on the Table 2, the smallest MSE value is obtained with $h = 0.92$ and $\delta = 0.99$ with an MSE value of 0.075. Based on the previous equation, the estimated model for rice productivity data in South Sulawesi is as follows:

$$\hat{y} = \hat{m}(x) = \frac{\sum_{i=1}^{24} exp\left(-\frac{(x-x_i)^2}{2(0.92)^2(0.99)^2}\right) y_i}{\sum_{i=1}^{24} exp\left(-\frac{(x-x_i)^2}{2(0.92)^2(0.99)^2}\right)}$$

or

$$\hat{y} = \frac{exp\left[-\frac{1}{2(0.92)^2(0.99)^2}[(x_1-x_1)^2 y_1 + (x_1-x_2)^2 y_2 + \cdots + (x_1-x_{24})^2 y_{24}]\right]}{exp\left[-\frac{1}{2(0.92)^2(0.99)^2}[(x_1-x_1)^2 + (x_1-x_2)^2 + \cdots + (x_1-x_{24})^2]\right]}$$

Next, estimate the value of rice productivity in region-1 (Selayar) using standardization data (x', y'), so that $\hat{y}_1^*$ is obtained based on the estimation equation model as follows:

$$\hat{y}_1^* = \frac{exp\left[\left(\frac{-(x_1-x_1)^2}{2h^2\delta^2}\right) y_1 + \left(\frac{-(x_1-x_2)^2}{2h^2\delta^2}\right) y_2 + \cdots + \left(\frac{-(x_1-x_{24})^2}{2h^2\delta^2}\right) y_{24}\right]}{exp\left[\left(\frac{-(x_1-x_1)^2}{2h^2\delta^2}\right) + \left(\frac{-(x_1-x_2)^2}{2h^2\delta^2}\right) + \cdots + \left(\frac{-(x_1-x_{24})^2}{2h^2\delta^2}\right)\right]}$$

$$\hat{y}_1^* = m(x_1) = -0.204$$

The $\hat{y}^*$ data that has been obtained is returned to the original data scale by inverting with the following formula:

$$\hat{y} = (\hat{y}^*.\sigma_y) + \bar{y}$$

For example, $\hat{y}^* = -0.204$ in the Selayar area, the process of calculating the return to the original scale is as follows:

$$\hat{y} = \big((-0.204)(5.153)\big) + 49.087$$

$$\hat{y} = -1.052 + 49.087$$

$$\hat{y} = 48.033$$

In the same way, the other $\hat{y}$ values are obtained in the following Table 3. It is observed that the most enormous $\hat{y}$ (predicted value) is 51.32 at the 14th observation (the 14th region in South Sulawesi is Sidrap). This indicates that the model predicts an increase in rice productivity by 51.32 units from the average value. The positive value signifies an increase in productivity compared to the average condition. The $\hat{y}$ is 48.033 at the 1st observation (the 1st region in South Sulawesi is Selayar). This suggests that for this observation, the model predicts a decrease in productivity of 48.033 units compared to the average. Here is the plot of estimated rice productivity data in South Sulawesi, as shown in Figure 1.

Table 3. $\hat{y}$ Value

No	Region	$\hat{y}$	y_i	No	Region	$\hat{y}$	y_i
1	Selayar	48.033	40.71	13	Wajo	49.900	45.43
2	Bulukumba	49.341	50.60	14	Sidrap	51.324	54.31
3	Bantaeng	48.238	48.90	15	Pinrang	50.890	58.38
4	Jeneponto	48.705	44.57	16	Enrekang	48.166	46.20
5	Takalar	48.787	45.64	17	Luwu	49.709	56.69
6	Gowa	49.582	47.71	18	Tana Toraja	48.399	45.76
7	Sinjai	48.614	48.20	19	Luwu Utara	49.121	52.55
8	Maros	49.149	48.19	20	Luwu Timur	49.222	57.77
9	Pangkajene	48.742	42.99	21	Toraja Utara	48.533	43.40
10	Barru	48.579	50.21	22	Makassar	48.055	52.15
11	Bone	48.652	49.76	23	Parepare	48.019	40.08
12	Soppeng	49.603	52.98	24	Palopo	48.066	54.93

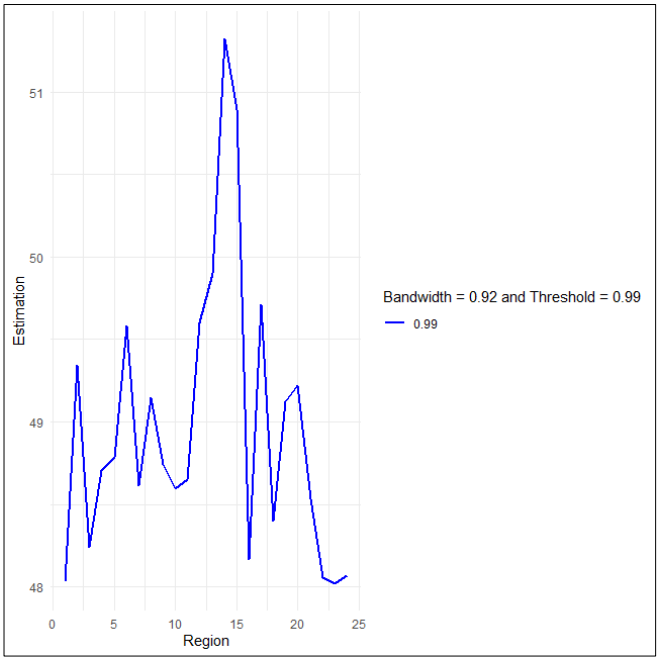

Figure 1. Estimated Productivity Data Plot of South Sulawesi

4. CONCLUSION

The estimated y, with bandwidth h, equals 0.92 and threshold delta equals 0.99, provides useful insights into understanding the relationship between land area and rice productivity. The estimates generated by the Nadaraya-Watson model indicate variations in predicted rice productivity across different observations. Positive values, $\hat{y} > 0$, indicate an increase in productivity compared to the average, while negative values, $\hat{y} < 0$, suggest a relative decrease in productivity. The model can capture variations in productivity across different observations with varying prediction values. In agriculture, it is often observed that productivity per unit area tends to decrease as the land area increases. This does not mean that larger land areas are unproductive but that the productivity increase rate is slower than smaller plots, which are typically managed more intensively. This estimation shows that larger land areas still provide stable productivity results, although there is a marginal decline in productivity growth per unit area. The model disregards minor fluctuations and focuses on significant trends, helping farmers understand how to maximize productivity through more efficient land management. These results help optimize agricultural strategies on larger land areas without losing potential yield.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interest.

REFERENCES

[1] T.H. Ali, Modification of the Adaptive Nadaraya-Watson Kernel Method for Nonparametric Regression (Simulation Study), Commun. Stat. - Simul. Comput. 51 (2022), 391–403. https://doi.org/10.1080/03610918.2019.1652319.

[2] A. Islamiyati, Spline Longitudinal Multi-Response Model for the Detection of Lifestyle-Based Changes in Blood Glucose of Diabetic Patients, Curr. Diabetes Rev. 18 (2022), e171121197990. https://doi.org/10.2174/1573399818666211117113856.

[3] A. Islamiyati, Raupong, A. Kalondeng, U. Sari, Estimating the Confidence Interval of the Regression Coefficient of the Blood Sugar Model through a Multivariable Linear Spline with Known Variance, Stat. Transit. New Ser. 23 (2022), 201–212. https://doi.org/10.2478/stattrans-2022-0012.

[4] T.H. Ali, H.A.A.M. Hayawi, D.S.I. Botani, Estimation of the Bandwidth Parameter in Nadaraya-Watson Kernel Non-Parametric Regression Based on Universal Threshold Level, Commun. Stat. – Simul. Comput. 52 (2023), 1476–1489. https://doi.org/10.1080/03610918.2021.1884719.

[5] N. Chamidah, I.N. Budiantara, S. Sunaryo, et al. Designing of Child Growth Chart Based on Multi-Response Local Polynomial Modeling, J. Math. Stat. 8 (2012), 342-347. https://doi.org/10.3844/jmssp.2012.342.347.

[6] M.F.F. Mardianto, E. Tjahjono, M. Rifada, Statistical Modeling for Prediction of Rice Production in Indonesia Using Semiparametric Regression Based on Three Forms of Fourier Series Estimator, ARPN J. Eng. Appl. Sci. 14 (2019), 2763-2770.

[7] A. Islamiyati, Anisa, M. Zakir, et al. The Use of the Binary Spline Logistic Regression Model on the Nutritional Status Data of Children, Commun. Math. Biol. Neurosci., 2023 (2023), 37. https://doi.org/10.28919/cmbn/7935.

[8] N. Anisa, N.N. Debataraja, S. Martha, Estimation of Nonparametric Kernel Regression Model Using the Nadaraya-Watson Estimator, Bimaster: Bulet. Ilm. Mat. Stat. Terap. 8 (2019), 231-305.

[9] N. Fadillah, P.A. Dariah, A. Anggraeni, et al. Comparison of Gaussian and Epancehnikov Kernels, Tadulako Soc. Sci. Humaniora J. 3 (2022), 13–22. https://doi.org/10.22487/sochum.v3i1.15745.

[10] S. Ghosh, Kernel Smoothing: Principles, Methods and Applications: Principles, Methods and Applications, John Wiley & Sons, 2017. https://doi.org/10.1002/9781118890370.

[11] A.K. Pani, Non-Linear Process Monitoring Using Kernel Principal Component Analysis: A Review of the Basic and Modified Techniques with Industrial Applications, Brazil. J. Chem. Eng. 39 (2022), 327–344. https://doi.org/10.1007/s43153-021-00125-2.

[12] K. Song, Y. Song, H. Wang, Threshold Reweighted Nadaraya–Watson Estimation of Jump-Diffusion Models, Probab. Uncertain. Quant. Risk 7 (2022), 31. https://doi.org/10.3934/puqr.2022003.

[13] R.L. Eubank, A. Kshirsagar, W.R. Schucany, Nonparametric Regression and Spline Smoothing, CRC Press, (2020).

[14] M. Kerpicci, H. Ozkan, S.S. Kozat, Online Anomaly Detection With Bandwidth Optimized Hierarchical Kernel Density Estimators, IEEE Trans. Neural Netw. Learn. Syst. 32 (2021), 4253–4266. https://doi.org/10.1109/TNNLS.2020.3017675.