Available online at: http://journal.uny.ac.id/index.php/pythagoras
PYTHAGORAS: Jurnal Matematika dan Pendidikan Matematika, 19 (1), 2024, 64-76

Comparative Analysis of Local Polynomial Regression and ARIMA in Predicting Indonesian Benchmark Coal Price

Arinda Mahadesyawardani, Utsna Rosalin Maulidya, Barnabas Anthony Philbert Marbun, Fachriza Yosa Pratama, Nur Chamidah*
Department of Mathematics, Universitas Airlangga, Surabaya, Indonesia
* Corresponding Author. E-mail: nur-c@fst.unair.ac.id

ARTICLE INFO

Article History:
Received: 16-Jun. 2024
Revised: 7-Nov. 2024
Accepted: 18-Nov. 2024

Keywords:
Nonparametric Regression, local Polynomial, ARIMA, Benchmark Coal Price.

ABSTRACT

As one of the world's biggest coal producers, it is essential for Indonesia to follow the trend of benchmark coal price fluctuations for any future possibilities. This study compared two methods of forecasting benchmark coal prices to evaluate the accuracy of the predictions used a nonparametric regression based on the local polynomial estimator and a parametric ARIMA method. Local polynomial analysis obtained a MAPE of 2.929278% using a CV method based on optimal bandwidth of 5.06 at order 2 with a cosine kernel, which means highly accurate forecasting accuracy. As for the ARIMA analysis, the data does not meet the assumption of normality, but forecasting is still continued with the best model ARIMA (1,2,1) model so that the MAPE is 12.6327%, which means good forecasting accuracy. Therefore in this study, the use of nonparametric regression methods using local polynomial estimators on data with non-normal distribution are more suitable to obtain accurate prediction results.

Sebagai salah satu produsen batubara terbesar di dunia, penting bagi Indonesia dalam mengikuti tren fluktuasi harga batubara acuan untuk mengetahui potensi yang mungkin terjadi di masa depan. Penelitian ini membandingkan dua metode peramalan harga batubara acuan untuk mengevaluasi keakuratan prediksi dengan menggunakan regresi nonparametrik berdasarkan estimator polinomial lokal dan metode parametrik ARIMA. Analisis polinomial lokal memperoleh MAPE sebesar *2.929278%* menggunakan metode CV berdasarkan bandwidth optimal 5.06 pada orde 2 dengan kernel cosinus, yang berarti hasil peramalan ini sangat akurat. Sementara untuk analisis ARIMA, data tidak memenuhi asumsi normalitas namun peramalan tetap dilanjutkan dengan model terbaik yaitu ARIMA (1,2,1) dan diperoleh MAPE sebesar 12.6327% yang berarti akurasi peramalan yang baik. Metode nonparametrik dengan menggunakan estimator polinomial lokal pada data yang berdistribusi tidak normal lebih sesuai untuk mendapatkan hasil prediksi yang akurat.

Scan me

How to Cite: Mahadesyawardani, A., Maulidya, U., Marbun., B., Pratama, F., Chamidah, N. (2024). Comparative analysis of local polynomial regression and ARIMA in predicting indonesian benchmark coal price. *Pythagoras: Jurnal Matematika dan Pendidikan Matematika,* 19(1), 64-76. https://doi.org/10.21831/pythagoras.v19i1.74889

https://doi.org/10.21831/pythagoras.v19i1.74889

INTRODUCTION

Coal represents a primary source of energy in Indonesia as one of the main export commodities in Indonesia with increased production reaching 775.2 million tons in 2023 (Kementerian ESDM, 2024). A significant increase in coal production, driven by rising international prices and increased regional demand, has positioned Indonesia as

the most flexible exporter in 2023, with exports reaching nearly 500 Mt (IEA, 2023). The successful realization has led to increased domestic utilization of coal, which will contribute to national energy security and economic growth. The Benchmark Coal Prices are calculated as the average of the Global Coal Newcastle Index (GCNC), Newcastle Export Index (NEX), Indonesia Coal Index (ICI), and Platt's 5900 in the previous month with equalized quality. This price is utilized in the direct exchange of coal commodities. (Gunarto & Wulansari, 2020).

The fluctuation of Indonesia's benchmark coal price is irrefutable. It reached its highest level in November 2021 at USD 215.01 per ton and then decreased to USD 158.50 per ton in January 2022. This was the consequence of an increase in domestic coal production within China. In the same year, the intensification of geopolitical tensions between Russia and Ukraine led to a surge in global coal commodity prices. This was beneficial for Indonesia, so that the Ministry of Energy and Mineral Resources has established the price at USD 288.4 per ton in April 2022, until finally the dynamics of the benchmark coal price reached the highest value in October 2023 at USD 330.97 per ton (Kementerian ESDM, 2022).

The high demand for coal has an impact on the fluctuation of the benchmark coal price. In preparing advanced strategies for all possibilities in the future, a prediction of the fluctuating pattern of benchmark coal price data can be made. A study by Prahesti et al. (2023) has predicted the benchmark coal price using Autoregressive Integrated Moving Average (ARIMA) and obtained a Mean Absolute Percentage Error (MAPE) of 23.14%. ARIMA is a univariate parametric time series method commonly used to make predictions based on a synthesis of historical data patterns (Hendrawan, 2013). Furthermore, a study by Hidayanti et al. (2022) also predicted the benchmark coal price using quadratic parametric regression and obtained a MAPE of 9.031%. In practice, quadratic regression method is limited to identify parabolic trend, while the reference coal price data shows a fluctuating non-stationary pattern. Time series analysis using both parametric methods will require an assumption for get valid results (Ardianti et al., 2020). Non-stationary fluctuations in the reference coal price data can be more effectively handled with ARIMA through a differentiation process and with a non-parametric approach. In nonparametric regression, there are no requirements for stationarity or model error assumptions in making decisions about the goodness of the model so that it is more flexible, so it can use quantitative measures such as R^2 and the Mean Absolute Percentage Error (MAPE) (Suparti & Santoso, 2024). In this study, researchers analyze the accuracy of benchmark coal price (HBA) predictions using nonparametric regression analysis with a local polynomial approach to assess the accuracy benchmark coal price predictions. Local polynomial regression is an estimator of the kernel regression function which is formed based on a polynomial order, where the weight size is determined by the bandwidth parameter that regulates the estimation at local points (Prahutama et al., 2018). Local polynomials have several advantages, including the ability to reduce asymptotic bias and produce accurate estimates (Welsh & Yee, 2006). Moreover, this approach was selected due to its ability to overcome data with fluctuating distributions, and then compared it with the ARIMA method which can also handle fluctuating data to identify the minimum MAPE value. The purpose of this analysis is to identify the most accurate method for future predictions.

METHOD

The data used in this study is a type of secondary data, that is monthly data on the Indonesian Benchmark Coal Price (HBA) in units of USD/ton that obtained from the official website of the Ministry of Energy and Mineral Resources. The response variable in this study is the reference coal price, while the monthly time index is the predictor variable. A total of 87 data points were identified for analysis, 75 representing the in-sample data (January 2017 to March 2023) and 12 representing the out-sample data (April 2023 to March 2024). The analysis procedures in this study are as follows:

1. Conduct descriptive statistical analysis to provide general information of the data, including the mean, standard deviation, maximum and minimum values.
2. Estimating the parameter values of the regression model using a non-parametric regression based on a local polynomial estimator with the following steps:
    a. Determine the optimal bandwidth for orders 1, 2, and 3. Bandwidth selection for time series analysis is conducted using the cross-validation (CV) method, then comparing Gaussian and Cosine kernels on in-sample data based on the best minimum order and CV.

b. Using the optimal bandwidth that has been selected in the previous step to estimate the value of the regression model parameters on the entire data set, including in-sample, out-of-sample, and all sample data, then evaluate it with the MAPE value.

3. Modeling with the Autoregressive Integrated Moving Average (ARIMA) approach with the following steps:
   a. Test the stationarity of the data in variance using Box-Cox transformation and for stationarity in mean using the Augmented Dickey-Fuller (ADF) test.
   b. Identify existing ARIMA models using ACF and PACF plots.
   c. Tests for parameter significance and diagnostic tests for the ARIMA model. If there are multiple significant models, the model with the smallest AIC value will be chosen.
   d. Diagnostic tests such as white noise using the Ljung-Box test, homoscedasticity using the Langrange Multiplier Engle test, and normality tests using Jarque-Bera test.
   e. Make forecasts based on the best model and evaluate the prediction results using the MAPE value then write the ARIMA model equation.
4. Comparing the accuracy of prediction results using the local polynomial estimator and ARIMA method based on Mean Absolute Percentage Error (MAPE).

Nonparametric Regression

In nonparametric regression models, there is no assumption on the shape of the regression function. Instead, it is only assumed to be smooth and contained in the Sobolev space. Nonparametric regression is a method approach wherein the shape of curve is unknown (Chamidah & Lestari, 2022). The univariate nonparametric regression model for n observations is written in the following equation (1).

$$y_i = g(x_i) + \varepsilon_i\,,\ i = 1,2,\dots,n \tag{1}$$

with,

y : response variable

x : predictor variable

$g(x_i)$: differentiable and continuous regression function

ε : random error.

Local Polynomial Estimator

The local polynomial estimator is a statistical technique employed in nonparametric regression model for estimating $g(x)$. The function $g(x)$ is approximated by a Taylor expansion around the point $x0$ which is written in equation (2) below (Chamidah & Lestari, 2022).

$$g(x) \approx \sum_{k=0}^{p} \frac{g^{(k)}(x_0)}{k!}(x - x_0)^k = \sum_{k=0}^{p} \beta_k^*(x_0)(x - x_0)^k \tag{2}$$

with,

$g(k)(x0)$: value of the k-th derivative of $g(x)$ to x at the point $x = x0$, for $x \in (x0 - h, x0 + h)$.

According to the nonparametric regression model that expressed in equation (1), equation (2) can be represented in matrix notation as follows

$$\boldsymbol{y}^* = \boldsymbol{x}_{x_0}^*\boldsymbol{\beta}^*(x_0) + \boldsymbol{\varepsilon}^* \tag{3}$$

With,

$$\boldsymbol{X}_{\boldsymbol{x_0}}^* = \begin{pmatrix} 1 & (\mathrm{x}_1 - \mathrm{x}_0) & \cdots & (\mathrm{x}_1 - \mathrm{x}_0)^p \\ \vdots & & \ddots & \vdots \\ 1 & (\mathrm{x}_n - \mathrm{x}_0) & \cdots & (\mathrm{x}_n - \mathrm{x}_0)^p \end{pmatrix}, \boldsymbol{y}^* = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \boldsymbol{\varepsilon}^* = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \tag{4}$$

The estimation of $\boldsymbol{\beta}^*(x_0)$ in equation (3) is based on the local polynomial approach, which involves taking n samples of paired data $\{x_i, y_i\}, i = 1,2,\dots,n$. So that n equations can be formed as written in equation 5 below.

$$\begin{aligned} y_1 &= \beta_0^*(x_0) + \beta_1^*(x_0)(\mathrm{x}_1 - \mathrm{x}_0) + \beta_2^*(x_0)(\mathrm{x}_1 - \mathrm{x}_0)^2 + \cdots + \beta_p^*(x_0)(\mathrm{x}_1 - \mathrm{x}_0)^p + \varepsilon_1 \\ y_2 &= \beta_0^*(x_0) + \beta_1^*(x_0)(\mathrm{x}_2 - \mathrm{x}_0) + \beta_2^*(x_0)(\mathrm{x}_2 - \mathrm{x}_0)^2 + \cdots + \beta_p^*(x_0)(\mathrm{x}_2 - \mathrm{x}_0)^p + \varepsilon_2 \\ &\vdots \\ y_n &= \beta_0^*(x_0) + \beta_1^*(x_0)(\mathrm{x}_n - \mathrm{x}_0) + \beta_2^*(x_0)(\mathrm{x}_n - \mathrm{x}_0)^2 + \cdots + \beta_p^*(x_0)(\mathrm{x}_n - \mathrm{x}_0)^p + \varepsilon_n \end{aligned} \tag{5}$$

The estimator $\widehat{\boldsymbol{\beta}}^*(x_0)$ is obtained through the application of a local polynomial approximation, utilizing a kernel weight function $K_{h^*}(x_i - x_0)$. The shape of the local polynomial weights is determined by the kernel function K, with the size of the weights determined by the bandwidth (h^*). The estimator $\widehat{\boldsymbol{\beta}}^*(x_0)$ is obtained through a process of minimizing the WLS function. So that, the local polynomial estimator for the regression function $g(x)$ is given by equation (6).

$$\hat{g}(x) = \boldsymbol{X}^*_{\boldsymbol{x_0}} [\boldsymbol{X}^*_{\boldsymbol{x_0}}{}^{\boldsymbol{T}} K_{h^*}(x_0) \boldsymbol{X}^*_{\boldsymbol{x_0}}]^{-1} \boldsymbol{X}^*_{\boldsymbol{x_0}}{}^{\boldsymbol{T}} K_{h^*}(x_0) \boldsymbol{y}^* \quad (6)$$

Kernel Density Estimator

One type of regression commonly used is kernel regression, which is a nonparametric method. The advantage of the kernel is that it can achieve a relatively fast convergence rate. The kernel estimator has several functions, including kernel uniform, triangle, epanechnikov, gaussian, quartic, and Cosinus (Kurniasih et al., 2013). In solving problems with fluctuating data, this research used Gaussian and Cosine kernel estimators.

Gaussian kernel function is commonly used in nonparametric regression analysis because it has the advantage of being easier to use that is more efficient in some cases, and the weight function of the kernel is defined for all real number (Shantika Martha, 2019). Gaussian kernel density function has the form written in equation (7).

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{1}{2}(-x^2)\right) \quad (7)$$

with $-\infty < x < \infty$.

Cosine kernel function is used in various applications, such as time series forecasting, nonparametric regression model analysis. The cosine kernel is efficient for estimation and can handle data with a fluctuating distribution, allowing it to approximate data patterns effectively (Salim et al., 2022). Cosine kernel density function has the form written in equation (8).

$$K(x) = \frac{\pi}{4} \cos\left(\frac{\pi}{2} x\right) \quad (8)$$

with $|x| \leq 1$ and 0 for others.

Cross Validation Method

One of the methods that can be used to select the optimal smoothing parameters in nonparametric regression is the cross validation (CV) method that defined in equation (9) below (Chamidah & Lestari, 2022).

$$CV(h) = \frac{1}{n} \sum_{i=1}^{n} \left[y_i - \hat{g}_{h,-i}(x_i)\right]^2 \quad (9)$$

with,

y_i : response variable at the i-th observation

$\hat{g}_{h,-i}(x_i)$: estimated value of the regression function at x_i without considering the i-th observation

The optimal bandwidth value is determined by identifying the bandwidth that result in the minimum CV value.

Autoregressive Integrated Moving Average (ARIMA)

The most common method for forecasting is ARIMA Box-Jenkins, which is used to process univariate time series. In order to be processed using the ARIMA Box-Jenkins method, a time series data set must meet the criteria for stationarity (Makridakis et al., 1999). A time series is considered stationary if the average and variance are constant, the data has no trend and no seasonal element. To handle non-stationary time series data, an appropriate d th differencing process is used. The ARIMA(p,d,q) model can be written as follows (Wei, 2006).

$$\phi_p(B)(1-B)^d Z_t = \theta_0 + \theta_q(\mathrm{B}) a_t \quad (10)$$

With,

$\phi_p(B) = (1 - \phi_1 B - \cdots - \phi_1 B^p)$ is an AR operator

$\theta_q(\mathrm{B}) = (1 + \theta_1 B + \cdots + \theta_q B^q)$ is a MA operator.

The Augmented Dickey-Fuller (ADF) test is commonly used to evaluate the stationarity of data in the mean by examining the existence or absence of unit roots within the model (Pitaloka et al., 2019). The hypothesis testing is

$H_0 : \delta = 0$ (there is a unit root or data is not stationary)

$H_1 : \delta < 0$ (there is no unit root or data is stationary)

With a significance level of α, the test criteria H_0 is rejected if τ-statistic < $\tau_{(\alpha;n-1)}$ or p-value < α.

The test statistic for the Dickey-Fuller test is expressed by Equation (11), with $\hat{\delta}$ representing the estimator of δ and SE($\hat{\delta}$) representing the standard error of δ.

$$\tau_{statistic} = \frac{\hat{\delta} - \delta}{SE(\hat{\delta})} \tag{11}$$

Non-stationary time series data in terms of the mean can be addressed by applying differencing. Stationarity testing for variance can be conducted by examining the value of λ in the Box-Cox transformation. A λ value that is not equal to one indicates that the data is not stationary in variance, and thus the data must be transformed using the Box-Cox method (Suparti & Santoso, 2024). The Box-Cox transformation is expressed by Equation (12).

$$Z_t^* = \begin{cases} \dfrac{Z_t^{\lambda-1}}{\lambda}, & \lambda \neq 0 \\ ln\, Z_t\,, & \lambda = 0 \end{cases} \tag{12}$$

Data that are already stationary can be used to build ARIMA models with the following steps (Ardi et al., 2017).

a. Identify the model by determining the order p and q from the ACF and PACF plots.
b. Estimation of the model parameters, with criteria for testing the model as significant if the p-value $< \alpha$.
c. Model diagnostic test, where the residuals obtained must fulfill the hypothesis assumptions of white noise or independence using Ljung-Box, normal distribution using the Jarque-Bera Test, and homoscedasticity or the model has a constant variance using Lagrange Multiplier Test. The model passes the diagnostic test if the p-value $> \alpha$.

At the forecasting step, the best candidate ARIMA model is selected that has the smallest Akaike Information Criterion (AIC) value (Wei, 2006) and by considering the concept of parsimony that defined by selecting the minimum number of model parameters that can adequately explain the model.

Mean Absolute Percentage Error (MAPE)

MAPE is a statistical measure used to evaluate the accuracy in a forecasting methodology. Here is the formula to calculate MAPE

$$\text{MAPE} = \frac{1}{N}\sum_{t=1}^{N}\left|\frac{Y_t - \hat{Y}_t}{Y_t}\right| \times 100\% \tag{13}$$

with,

N : number of observations
Y_t : actual value at time t
$\hat{Y}_t$: forecasted value at time t

Table 1 below explains the meaning of the MAPE value (Chang et al., 2007).

Table 1. MAPE values

MAPE	Forecasting Accuracy
<10%	Highly accurate

10-20%	Good
20-50%	Reasonable
>50%	Inaccurate

RESULT AND DISCUSSION

Descriptive Analysis

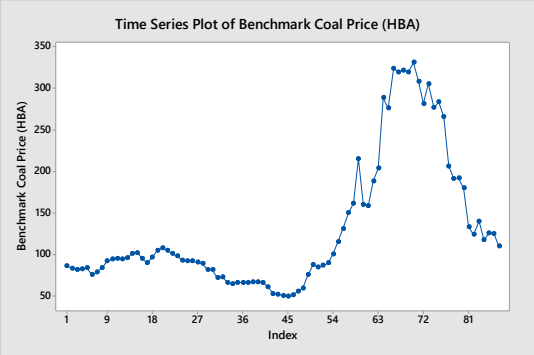

Figure 1. Time series plot of benchmark coal price

The results of the descriptive statistics in Table 2 show that the mean of benchmark coal price from January 2017 to March 2024 is 131.06 USD/ton, with a standard deviation of 80.72. This indicates that each data in the period tends to be homogeneous or spread around its average value, characterized by a standard deviation value that is lower than the mean. The lowest HBA of 49.42 USD/ton was observed in September 2020, a consequence of China and India implementing import restrictions to reduce their import needs. A significant increase was observed when the HBA value reached 330.97 USD/ton in October 2022. The intensification of geopolitical tensions between Russia and Ukraine at the time led to a surge in global coal commodity prices. Fluctuations in reference coal price data are shown in Figure 1.

Table 2. Statistics descriptive

Variable	Mean	StDev	Minimum	Maximum
HBA	131.06	80.72	49.42	330.97

Optimal Bandwidth Using Local Polynomial Estimator

The selection of optimal bandwidth was selected on 75 in-sample data of benchmark coal prices for monthly time periods in January 2017 to March 2023 using the Cross Validation (CV) method with Gaussian and Cosine kernels. In this study, parameter estimation in local polynomial regression uses orders 0, 1, and 2 until the optimal bandwidth value is obtained by considering the minimum CV value. The results of selecting the optimal bandwidth for the Gaussian and Cosine kernels based on the minimum CV are shown in Table 3 below.

Table 3. Bandwidth test result comparison

Kernel Function	Orde	Optimal Bandwidth	CV Minimum	MSE	MAPE (%)

Gaussian	0	1.18	143.0941	40.14696	2.883984
	1	1.29	146.9723	44.77373	3.096486
	2	1.98	142.3610	66.89165	2.890855
Cosinus	0	3.00	138.2877	46.93627	3.239961
	1	3.18	145.9416	50.67470	3.464946
	2	5.06	136.1164	66.31777	2.929278

Based on Table 3, the optimal bandwidth is located at order 2 using the Cosine kernel function, which is 5.06 with a minimum CV value of 136.1164. This optimal bandwidth at order 2 will be used on in-sample data to estimate the model. The goodness of fit of the in-sample model can be measured using R^2 or the coefficient of determination (Suparti & Santoso, 2024). In this case, the R^2 value with the optimal bandwidth is obtained at 99.05207%, indicating that the model is classified as strong. A plot of the optimal bandwidth towards the CV value is shown in Figure 2 below.

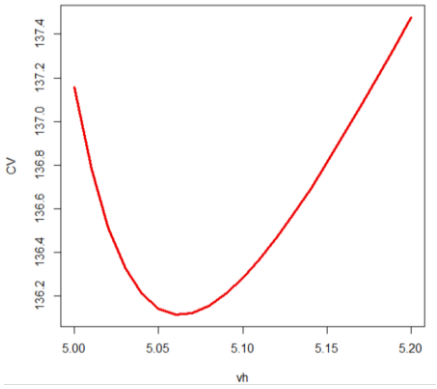

Figure 2. Plot of benchmark coal price data with optimal bandwidth

Parameter Model Estimation

The results of the estimated values of the model parameters $(\hat{\beta}_0,\ \hat{\beta}_1,\ \hat{\beta}_2)$ on the in-sample data are shown in Table 4 below.

Table 4. Parameter estimation of in-sample data

x	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
1	86.04789	-3.09512440	1.18167418
2	84.25379	-1.11696093	-0.06834917
3	83.06080	-1.17415948	-0.03807204
4	81.25085	-0.98551312	0.43005397
5	79.72908	-0.08927099	0.98929643
⋮	⋮	⋮	⋮
75	281.45478	-2.85195240	3.48422055

Based on equation 5, the local polynomial model for in-sample data of benchmark coal price are presented in equation 14.

$$\begin{aligned} y_1 &= 86.04789 + (-3.09512440)(\mathrm{x}_1 - \mathrm{x}_0) + 1.18167418(\mathrm{x}_1 - \mathrm{x}_0)^2 + \varepsilon_1 \\ y_2 &= 84.25379 + (-1.11696093)(\mathrm{x}_2 - \mathrm{x}_0) + (-0.06834917)(\mathrm{x}_2 - \mathrm{x}_0)^2 + \varepsilon_2 \\ &\vdots \\ y_{75} &= 281.45478 + (-2.85195240)(\mathrm{x}_{75} - \mathrm{x}_0) + 3.48422055(\mathrm{x}_{75} - \mathrm{x}_0)^2 + \varepsilon_{75} \end{aligned} \quad (14)$$

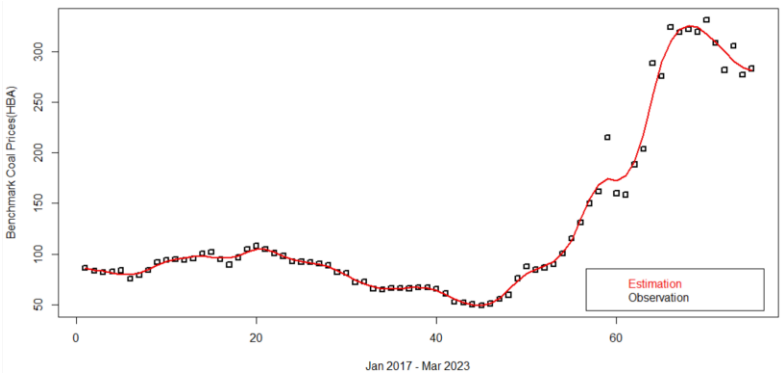

Figure 3. Comparison plot of estimated and observed results of in-sample data

The estimation results in Figure 3 using the CV method with optimal bandwidth show an optimal curve that is neither too smooth nor too rough. This indicates that the estimation results are optimal, thereby rendering them statistically representative and suitable for prediction purposes. The MAPE test results on modeling in-sample data with an optimal bandwidth of 5.06 obtained a MAPE of 2.929278%.

While the results of the estimated values of the model parameters $(\hat{\beta}_0,\ \hat{\beta}_1,\ \hat{\beta}_2)$ on the out-sample data are shown in Table 5 below.

Table 5. Parameter estimation of out-sample data

x	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
76	259.9202	-50.726069	15.9392891
77	223.0153	-27.926971	5.5783321
78	201.4235	-21.898272	2.2682386
79	180.2933	-19.026492	2.6825221
80	162.9474	-16.072860	2.5552968
⋮	⋮	⋮	⋮
87	112.4648	-8.0133	-1.8478

The local polynomial model for out-sample data of benchmark coal price is presented in equation 15.

$$\begin{aligned} y_{76} &= 259.9202 + (-50.726069)(\mathrm{x}_{76} - \mathrm{x}_0) + 15.9392891(\mathrm{x}_{76} - \mathrm{x}_0)^2 + \varepsilon_{76} \\ y_{77} &= 223.0153 + (-27.926971)(\mathrm{x}_{77} - \mathrm{x}_0) + 5.5783321(\mathrm{x}_{77} - \mathrm{x}_0)^2 + \varepsilon_{77} \\ &\vdots \\ y_{87} &= 112.4648 + (-8.0133)(\mathrm{x}_{87} - \mathrm{x}_0) + (-1.8478)(\mathrm{x}_{87} - \mathrm{x}_0)^2 + \varepsilon_{87} \end{aligned} \quad (15)$$

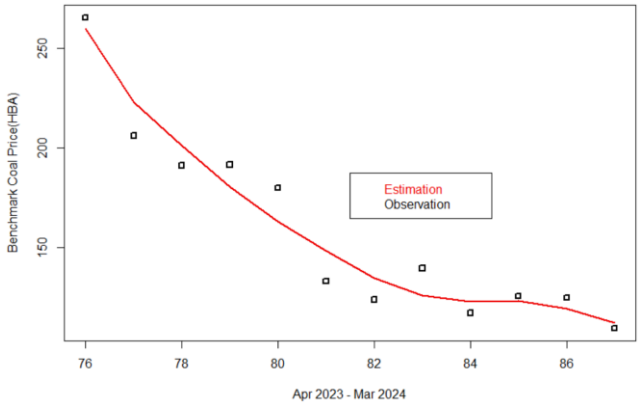

Figure 4. Comparison plot of estimated and observed results of out-sample data

Based on out-sample data modelling, the MAPE is 6.173094%. Overall, parameter estimation on all-sample data shows a MAPE of 3.486737%. This indicates that the model has highly accurate predictive capabilities regarding the Indonesian benchmark coal price.

ARIMA Analysis

ARIMA analysis using 75 in-sample data. The stationarity of the data was tested with a significance level of 5%. The stationarity analysis in variance using the Box-Cox transformation shows a rounded value (λ) of -0.5, so the data will be transformed into the form $\lambda^{-0.5}$ to make the data stationary for forecasting. While the results of testing for stationarity in the mean using the ADF test are presented in Table 6.

Table 6. ADF test results

	t-statistics	Prob	Decision
HBA	-0.370730	0.9084	Accept H_0
D(HBA)	-5.117139	0.0000	Reject H_0
D(HBA,2)	-11.43442	0.0001	Reject H_0

Based on Table 6, the critical value of the test in 5% level is -2.895924. The original data shows that it is not stationary with a probability 0.9084 > 0.05 or the t-statistic value is less than the Dickey Fuller critical value. The data is continued at the differencing stage, and the probability value is less than 0.05 at the first and second differencing, so the there is no unit root or the benchmark coal price data is stationary. The stationary data can be continued to identify the ARIMA model based on the ACF and PACF plots, as shown in the figure below.

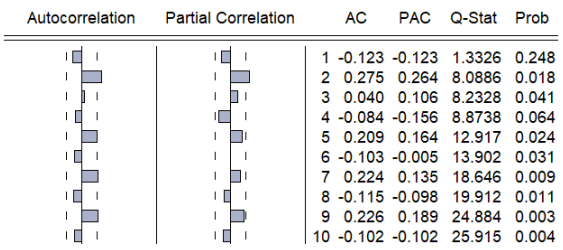

Figure 5. ACF and PACF plots at d=1

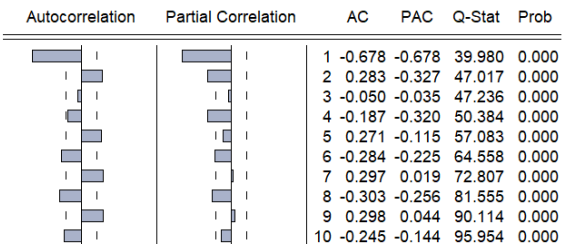

Figure 6. ACF and PACF plots at d=2

The ACF and PACF plots of the first differencing results in Figure 5 indicate that there is one significant lag out, so that the tentative models that can be identified are ARI(1,1), MA(1,1), and ARIMA(1,1,1). Figure 6 presents the plot of the second differencing results, with the ACF plot indicating a cut-off at lag 1 and the PACF plot indicating a

cut-off at lag 2, so that the tentative models that can be identified are ARI(1,2), ARI(2,2), MA(2,1), ARIMA(1,2,1), and ARIMA(2,2,1).

After determining the ARIMA model, diagnostic tests can be conducted to determine the optimal model for estimation in forecasting. The results of the parameter significance test and diagnostic test show that none of the three models in the first differencing were identified as significant at the 5% significance level. Therefore, the model with the second differencing was used for further test.

Table 7. Parameter significance test

Model	Parameter	Coefficient	Prob	AIC
ARI(1,2)	$\hat{\Phi}_1$	-0.679284	0.0000	8.797549
ARI(2,2)	$\hat{\Phi}_2$	0.286397	0.0001	9.329350
IMA(2,1)	$\hat{\theta}_1$	-1.000000	0.9985	8.628219
ARIMA(1,2,1)	$\hat{\Phi}_1$	-0.271823	0.0056	8.629623
	$\hat{\theta}_1$	-0.808003	0.0000	
ARIMA(2,2,1)	$\hat{\Phi}_2$	0.0263351	0.0057	8.587727
	$\hat{\theta}_1$	-1.000000	0.9988	

Based on Table 7, the results of the model parameter significance test indicate that at the 5% significance level, only three of the five models fulfill the criteria for testing the significant parameters of the model, these are ARI(1,2), ARI(2,2), and ARIMA(1,2,1), where ARIMA(1,2,1) has the smallest AIC value among the three significant models, which is 8.629623. So that ARIMA(1,2,1) can be used as the best preliminary model choice to be continued in the diagnostic test.

Diagnostic testing of the white noise assumption is by using the Ljung box to test whether there is a correlation in the residuals between the lags of each model. The ACF and PACF plots of the ARIMA(1,2,1) model residuals are shown in Figure 7.

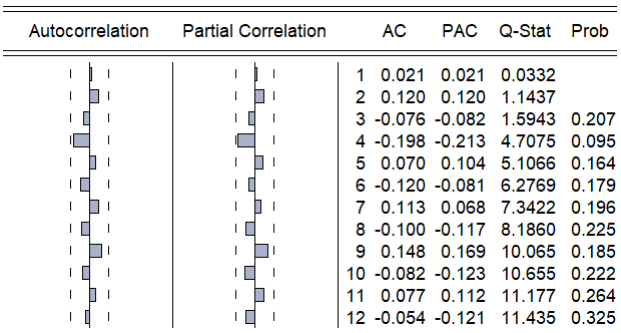

Figure 7. ACF and PACF plots of ARIMA(1,2,1) model residuals

Figure 7 shows that the probability value of the Ljung box test is greater than 0.05 at all lags, which means that there is no correlation in the residuals between lags (independent residuals) in the ARIMA (1,2,1) model.

Table 8. Heteroskedasticity test

F-statistic	2.791349	Prob. F(1,70)	0.0992
Obs*R-squared	2.761003	Prob. Chi-Square(1)	0.0966

Assumption of residual homoscedasticity is tested using the Lagrange Multiplier test (LM test). The probability value of Obs * R-squared = 0.0966 > 0.05. This indicates that the ARIMA (1,2,1) model does not have an ARCH effect or a homogeneous residual model.

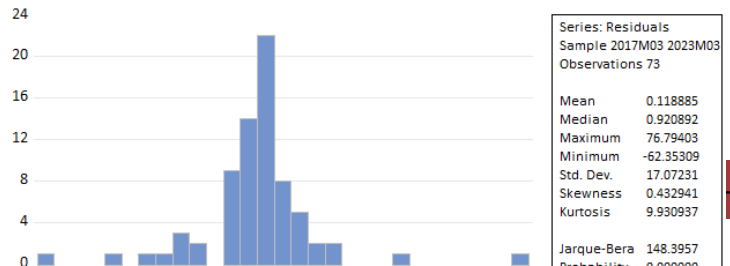

Figure 8. Normality test

Residual normality test result with the Jarque-Bera test as shown in Figure 8, indicate that the p-value of the ARIMA(1,2,1) model is 0.000 < 0.05, which indicates that the residuals are not normally distributed. The researchers also tested two other ARIMA models such as ARIMA(1,2) and ARIMA(2,2), the results indicated that all these tentative models do not meet the assumption of residual normality. Therefore, the decision to proceed with the ARIMA(1,2,1) model was based on its lowest AIC criteria compared to the other ARIMA models despite none of them fulfilling normality. The subsequent analysis will compare the MAPE values with the local polynomial regression method, which does not require any assumptions. This is why, in forecasting Indonesian benchmark coal price data, it is recommended to use nonparametric methods, which can be used on data that has a normal distribution or not. In this study, researchers continue to forecast to see the MAPE using ARIMA(1,2,1) model. Therefore using the equation (16), the equation of the ARIMA(1,2,1) model is written in equation (17).

$$\phi_p(B)(1-B)^d Z_t = \theta_0 + \theta_q(\mathrm{B})a_t \tag{16}$$

$$(1-B)^d(1-\phi_1 B-\cdots-\phi_1 B^p)Z_t = (1+\theta_1 B+\cdots+\theta_q B^q)a_t$$

$$(1-B)^2(1+0.271823B)Z_t = (1-0.808003B)a_t$$

$$(1-1.728177B+0.456354B^2+0.271823B^3)Z_t = (1-0.808003B)a_t$$

$$\frac{1}{\sqrt{Z_t}} = \frac{1}{\sqrt{1.728177Z_{t-1}}} - \frac{1}{\sqrt{0.456354Z_{t-2}}} - \frac{1}{\sqrt{0.271823Z_{t-3}}} + a_t - 0.808003a_{t-1} \tag{17}$$

The following table presents the results of forecasting the benchmark coal price on the out-sample data for the next 12 periods.

Table 9. ARIMA (1,2,1) model forecast results

Actual	Forecast	APE
265.26	278.5298221	5.002572
206.16	265.4029411	2873639
191.26	203.4953327	6.397225
191.6	178.5731304	6.798993
179.9	189.1608528	5.147778
133.13	182.9680472	37.43562
123.96	131.3392935	5.952963
139.8	113.32341	18.93891
117.38	138.419476	17.92424
125.85	123.3801045	1.962571
124.95	121.4915376	2.767877
109.77	125.7166125	14.5273
	MAPE (%)	12.6327

As shown in Table 9, the forecast results using the ARIMA(1,2,1) model on 12 out-sample data gives a MAPE of 12.6327%, which means that the model's forecasting ability is good.

Table 10. Prediction results of ARIMA and local polynomial regression

Method	Parameter	MAPE (%)	Accuracy
Local Polynomial Regression	Bandwith 5.06 ; order 2	2.929278	Highly accurate
ARIMA	ARIMA (1,2,1)	12.6327	Good

Table 10 compares the prediction accuracy of ARIMA and Local Polynomial Regression models for Indonesia's benchmark coal price. Local Polynomial Regression achieved a much lower MAPE of 2.93% labeled as "Highly accurate," while ARIMA had a higher MAPE of 12.63% labeled as “Good”. Thus, Local Polynomial Regression outperformed ARIMA in prediction accuracy.

CONCLUSION

In forecasting reference coal price fluctuations, a nonparametric regression method with a local polynomial approach is used, with an optimal bandwidth is 5.06 and a minimum CV value of order 2 is 136.1164 obtained using the Cosine kernel. The MAPE value is 2.929278%, indicating that the forecasting ability is highly accurate. While parametric testing using ARIMA with the optimal model is ARIMA (1,2,1) but this model did not meet the assumption of normality. However, this model was still used for forecasting and obtained a MAPE value of 12.6327%, which indicates a good forecasting. Therefore, forecasting using a local polynomial approach is more suitable for analysing the accuracy of predicting Indonesian benchmark coal prices.

ACKNOWLEDGEMENT

Researcher would like to express gratitude to the Department of Statistics, Faculty of Science and Technology, Airlangga University for their support of this research, and thank you to the Ministry of Energy and Mineral Resources for providing the Indonesian Benchmark Coal Price public data for successful research.

REFERENCES

Ardi, T., Santoso, R., & Prahutama, A. (2017). Implementasi Subset Autoregressive Menggunakan Paket Fitar. *Jurnal Gaussian, 6(4), 510–519.* https://doi.org/10.14710/j.gauss.6.4.510-519

Ardianti, C. W., Santoso, R., & Sudarno, S. (2020). Analisis Arima Dan Wavelet Untuk Peramalan Harga Cabai Merah Besar Di Jawa Tengah. *Jurnal Gaussian, 9(3), 247–262.* https://doi.org/10.14710/j.gauss.v9i3.28906

Chamidah, N., & Lestari, B. (2022). *Analisis Regresi Nonparametrik dengan Perangkat Lunak R*. Surabaya: Airlangga University Press.

Chang, P. C., Y.W, W., & C.H, L. (2007). The Development of a Weighted Evolving Fuzzy Neural Network for PCB Sales Forecasting. *Journal Expert System with Applications, 32(1), 86–96.* https://doi.org/10.1016/j.eswa.2005.11.021

Gunarto, M., & Wulansari, R. (2020). Analisis pergerakan harga saham berdasarkan sarga scuan dan volume penjualan : Ssudi pada PT Bukit Asam Tbk. *Jurnal Manajemen Dan Bisnis Sriwijaya, 18(4), 255-272.* https://doi.org/10.29259/jmbs.v18i4.13021

Hendrawan, B. (2013). *Penerapan Model ARIMA Dalam Memprediksi IHSG*. Politeknik Negeri Batam.

Hidayanti, R. P., Mustikasari, & Hariani. (2022). Prediksi Harga Batu Bara Menggunakan Regresi. *Journal of Embedded System Security and Intelligent System, 3(1), 1-10.* https://doi.org/10.26858/jessi.v3i1.33142

IEA. (2023). *Coal 2023*. International Energy Agency. https://www.iea.org/reports/coal-2023/executive-summary

Kementerian ESDM. (2022). *Laporan Kinerja 2022*. Direktorat Jenderal Mineral dan Batubara. https://www.minerba.esdm.go.id/upload/file_menu/20230329164006.pdf

Kementerian ESDM. (2024). *Harga Acuan*. Direktorat Jenderal Mineral dan Batubara. https://www.minerba.esdm.go.id/harga_acuan

Kementerian ESDM. (in press). *Produksi Batubara Domestik Tembus Target, Ketahanan Energi Nasional Terjaga.* Kementerian ESDM RI. https://www.esdm.go.id/id/media-center/arsip-berita/produksi-batubara-domestik-tembus-target-ketahanan-energi-nasional-terjaga

Kurniasih, D., Mariani, S., & Sugiman. (2013). Efisiensi Relatif Estimator Fungsi Kernel Gaussian Terhadap Estimator Polinomial Dalam Peramalan Usd Terhadap Jpy. *UNNES Journal of Mathematics*, *2*, *79–84.*

Makridakis, S., WheelWright, S., & McGree, V, E. (1999). *Metode dan Aplikasi Peramalan* (2nd ed.). Jakarta: Erlangga.

Pitaloka, R. A., Sugito, S., & Rahmawati, R. (2019). Perbandingan Metode Arima Box-Jenkins Dengan ARIMA Ensemble Pada Peramalan Nilai Impor Provinsi Jawa Tengah. *Jurnal Gaussian, 8(2), 194–207.* https://doi.org/10.14710/j.gauss.v8i2.26648

Prahutama, A., Suparti, D. I., & Utami, T. W. (2018). Pemodelan Bivariate Polinomial Lokal Pada Jumlah Kematian Ibu Dan Bayi Di Jawa Tengah. *Prosiding Seminar Nasional Venue Artikulasi-Riset, Inovasi, Resonansi-Teori, Dan Aplikasi Statistika (VARIANSI)*, *2018(Vol 1), 209–220.* http://ojs.unm.ac.id/variansistatistika/article/view/7208

Salim, M. I., Adnan Sauddin, & M. Ichsan Nawawi. (2022). Model Regresi Nonparametrik Deret Fourier Pada Kasus Tingkat Pengangguran Terbuka Di Sulawesi Selatan. *Jurnal MSA (Matematika Dan Statistika Serta Aplikasinya)*, *10*(2), *48–56.* https://doi.org/10.24252/msa.v10i2.30993

Suci Pujiani Prahesti, Itasia Dina Sulvianti, & Yenni Angraini. (2023). Peramalan Harga Batu Bara Acuan Menggunakan Metode Autoregressive Integrated Moving Average Dan Fungsi Transfer. *Xplore: Journal of Statistics, 12(1), 1–11.* https://doi.org/10.29244/xplore.v12i1.1100

Shantika Martha, N. A. N. N. D. (2019). Estimasi Model Regresi Nonparametrik Kernel Menggunakan Estimator Nadaraya-Watson. *Bimaster : Buletin Ilmiah Matematika, Statistika Dan Terapannya*, *8*(4), *633–638.* https://doi.org/10.26418/bbimst.v8i4.35870

Suparti, S., & Santoso, R. (2024). Analisis Data Time Series Menggunakan Model Kernel: Pemodelan Data Harga Saham MDKA. *Indonesian Journal of Applied Statistics, 6(1), 22.* https://doi.org/10.13057/ijas.v6i1.79385

Wei, W. W. S. (2006). *Time Series Analysis:Univariate and Multivariate Methods*. New York: Pearson.

Welsh, A. H., & Yee, T. Y. (2006). Local Regression for Vector Responses. *Journal of Statistical Planning and Inference, 136(9), 3007–3031.* https://doi.org/10.1016/j.jspi.2004.01.024