Multiple Time Series Modeling of Autoregressive Distributed Lags with Forward Variable Selection for Prediction

ACHMAD EFENDI[1,*], YUSI TYRONI MURSITYO[2], NINIK WAHJU HIDAJATI[3], NUR ANDAJANI[3], ZURAIDAH ZURAIDAH[4], SAMINGUN HANDOYO[1,5]

[1]Statistics Department,
Brawijaya University,
Malang, 65145, East Java,
INDONESIA

[2]Information System Department,
Brawijaya University,
Malang, 65145, East Java,
INDONESIA

[3]Civil Engineering Department,
Universitas Negeri Surabaya, Surabaya,
60231, East Java,
INDONESIA

[4]Islamic Banking Study Program,
State Islamic Institute of Kediri,
Kediri, 64127, East Java,
INDONESIA

[5]EECS – IGP Department,
National Yang Ming Chiao Tung University,
Hsinchu, 30100,
TAIWAN

**Corresponding Author*

Abstract: - The conventional time series methods tend to explore the modeling process and statistics tests to find the best model. On the other hand, machine learning methods are concerned with finding it based on the highest performance in the testing data. This research proposes a mixture approach in the development of the ARDL (Autoregressive Distributed Lags) model to predict the Cayenne peppers price. Multiple time series data are formed into a matrix of input-output pairs with various lag numbers of 3, 5, and 7. The dataset is normalized with the Min-max and Z score transformations. The ARDL predictor variables of each lag number and dataset combinations are selected using the forward selection method with a majority vote of four criteria namely the Cp (Cp Mallow), AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion), and adjusted R^2. Each ARDL model is evaluated in the testing data with performance metrics of the RMSE (Root Mean Square Error), MAE (Mean Absolute Error), and R^2. Both AIC and adjusted R^2 always form the majority vote in the determining optimal predictor variable of ARDL models in all scenarios. The ARDL predictor variables in each lag number are different but they are the same in the different dataset scenarios. The price of Cayenne pepper yesterday is the predictor variable with the most contribution in all of the 9 ARDL models yielded. The ARDL

lag 3 with the original dataset outperforms in the RMSE and MAE metrics while the ARDL lag 3 with the Z score dataset outperforms in the R^2 metric.

Key-Words: - ARDL model, time series, autoregressive, forward selection, machine learning, normalization method, prediction.

Received: May 24, 2023. Revised: March 8, 2024. Accepted: March 26, 2024. Published: April 19, 2024.

1 Introduction

The price stability of staple food commodities must be maintained by the government because they have an important impact on various sectors including the economic, social, and political aspects of a nation, [1]. However, the Indonesian government still does not have a consistent list of staple food commodities until now. The commodity of chilies is not part of staple food commodities, but it is highly sought after by customers. In particular, Javanese people don't even care about the very expensive price, [2]. Meanwhile, there are various types of chilies including large chilies, curly chilies, and cayenne peppers, [3]. The Cayenne pepper has the spiciest taste and has a very large gap of price fluctuation, [4]. The government should guarantee the availability of cayenne pepper and control the price adequately so that it can help ease the burden on society, [5]. A model that can predict the price of cayenne pepper accurately can be a tool for the government to control the price.

Modeling with a statistics approach has the goal of proving the hypothesis through experiment designing, data collecting, and data analyses, [6]. A regression model captures the relationship between independent and dependent variables and can draw causal inferences from the data, with attention to causal confounding, [7]. From an engineering field's point of view, [8], modeling with machine learning which is concerned with optimizing a score function and finding the best model performance on the testing data is more popular than statistical modeling. Unsupervised learning is defined in which there is no target variable in a dataset such as clustering methods, [9], and ranking methods, [10]. While, supervised learning can be the predictive model to predict class labels called classification models, [11], and to predict numerical values with regression models, [12]. Nevertheless, in real and practical life, there are many data in the ordered sequence as single or multiple datasets called time series data that can be modeled by regression approaches such as with both a conventional time series and machine learning regression.

The availability of big time series data provided by both private and public organizations has supported and motivated many researchers in various fields to develop forecasting models that can give advantages for winning competition in their business, [13]. The ARIMA (Autoregressive Integrated Moving Average) model and its derivative have been implemented successfully in the stock market, [14], public health, [15], agricultural business, [16], and so forth. Unfortunately, the ARIMA model is tricky enough in the identification stage, and it is not clear in the model performance evaluation. It had shown that hybrid models yielded better model performance in forecasting non-stationary univariate time series data where the input lag number is determined by using input-output pairs formed on various lag numbers, [17]. Developing a VAR (Vector Autoregression) model for multiple time series such as in Gričar, [18], has a drawback of inflexibility and inefficiency in application because all variables must have the same lag numbers. On the other side, the ARDL model can accommodate multiple time series modeling with more effective stages and it

does not require the dimension of the response variables to be the same as predictor variables, [19]. The forward variable selection is carried out to obtain the predictor variables with significant contributions to the response variable, [20].

The models to predict agricultural commodities, either the conventional statistic or machine learning approaches, are rarely acquired in literature. In particular, the developed predictive models employ multiple time series data where the data are available in almost every sector of life. The time series modeling such as ARIMA or VAR can explain the interpretation model well but they are relatively poor in the model performance. On the other side, machine learning models tend to acquire performance well but they are difficult to interpret and too complicated model. The implementation of multiple regression models with forward selection can be found in many literatures. However, the ARDL modeling with the forward selection employed in the multiple time series forecasting faces a sophisticated identification of the optimal subset predictor variables. Rarely works published in the domain. Hence, the development of predictive models that are easy to interpret and powerful in performance is still an open problem.

Picking up the advantages of both econometric and machine learning approaches ensures the acquisition of better models. In the case of finding a useful model that can support controlling the Cayenne pepper prices, the model has to be not only easily interpreted but also highly performed. The ARDL is an interpretable model and the machine learning approach offers systematic stages in the modeling process. The identification issue of relevant predictor variables of the ARDL structure is tackled by the formatting of input-output pairs of multiple time series into various lag numbers. Dividing the input-output pairs into the training and testing data where the testing data is employed in the selection model and the training data is employed in parameter estimation will lead to a model with high performance to predict future values.

The research has the main goal of finding the ARDL model with a hybrid approach of econometric and machine learning to predict the Cayenne pepper prices by considering 3 kinds of datasets namely the original time series, the min-max transformation, and the Z-score transformation. The remaining sections are organized into some sections. Section 2 presents the conceptual methods used in the research. The third section presents the summary of multiple time series data and research stages. Meanwhile, section 4 discusses the modeling process, model interpretation, and performance evaluation of the testing data. And finally, the conclusion part is given in the last section.

2 Literature Reviews

The section discusses the conceptual theory and formula related to the stages in the building of the proposed model and also discusses the model performance metrics which are used to evaluate the performance of models yielded on the testing data in all development scenarios.

2.1 The Min-max and Z Score Transformation Methods

Data engineering such as data normalization is a process of how transforming data to not only have the same measurement unit but also all predictor variables to have the same domain value which ensures commensurate measurement of predictor variables, [21]. There are two popular types of data normalization in machine learning namely min-max and Z-score transformations. The min-max method transforms numerical variables into the range of 0 to 1, and the Z score method transforms numerical variables into the range of -4 to 4. The formulas of both methods are given in (1) and (2) as the following, [22] :

$$\mathrm{T(x)} = (x_i - x_{min})/x_{max} \quad (1)$$

$$\mathrm{T(x)} = (x_i - \bar{x})/Sd(X) \quad (2)$$

Where x_{min}, x_{max}, $\bar{x}$, and $Sd(X)$ are respectively the minimum of X, maximum of X, mean of X, and standard deviation of X. The transformation outputs yielded by the min-max method seem more natural because they are all positive values, but the Z score method yields the transformation output which can be used to investigate the outlier data.

2.2 The Autoregressive Distributed Lags (ARDL) Model

Developing a model that involves many variable predictors relating to a response variable will lead to better model performance if all of the variables have a significant contribution to predicting the response variable value, [23]. Some time series data are related to each other so time series modeling by considering some related ones to develop a predictive multivariate time series model is supposed to be a comprehensive approach. An ARDL model is one type of econometrics model referring to a model involving some lags coming from both the dependent and independent variables, [24]. A structure of the ARDL model with order (p, q) which means it has p lags in the independent variable x, and q lags in the dependent variable is stated in (3) as follows, [25]:

$$\begin{aligned} y_t = {} & \phi_0 + \phi_1 x_t + \phi_2 x_{t-1} + \phi_3 x_{t-2} + \cdots \\ & + \phi_{p+1} x_{t-p} + \theta_1 y_{t-1} + \theta_2 y_{t-2} \\ & + \cdots + \theta_q y_{t-q} + \epsilon_t \end{aligned}$$

$$y_t = \sum_{J=0}^{P} \phi_j x_{t-j} + \sum_{J=1}^{q} \theta_j y_{t-j} + \epsilon_t \quad (3)$$

Where x_{t-p} is the observation of variable X on the p lag, and the y_{t-q} is the observation of variable Y on the q lag. The order autoregressive q and order distributed p in (3) are ordered from 1 to q and from 1 to p respectively. From a simple point of view, equation (3) can be considered a linear multiple regression model with the total number of predictor variables of $q + p + 1$. The coefficients of the ARDL model can be estimated by using the ordinary least squares (OLS) method, [26].

An ARDL model can involve more than two variables such as (r, q, p) for the model with two independent variables and one response variable. Practically, not all of the r, q, and p lags did contribute significantly to the response variable so only predictor variables which contributed significantly are maintained in the final model. The model that only involves the predictor variables with significant contributions to the response variable is called the best ARDL which is obtained by conducting a filter variable selection method such as forward selection, [27].

2.3 The Forward Variable Selection Method

A statistical model should satisfy the parsimony principle which is a simpler model involving fewer parameters is favored over more than complex models involving many parameters, [28]. The principle is also applied in machine learning modeling with conducting variable selection algorithms. In practice, there are known 3 types of variable selection algorithms i.e., the heuristic approach which is not determined model structure yet, the filter approach when variable selection is done in the forward selection or backward elimination using the goodness of fit criteria, and the wrapper approach if the best model is obtained through the greedy search method on a specific machine learning algorithm, [29].

Considering the development of an ARDL model by determining the input lag number of k on d time series datasets. It will lead to a model that has the number of (d*k) predictor variables. In nature, any predictor variables did not contribute to predicting the response variable significantly. The best ARDL model is obtained by selecting predictor variables one by one starting with the predictor

variable having the highest contribution to the explaining response variable, [30]. The forward selection method worked firstly by searching a predictor variable having the minimum value of residual sum square (RSS), and then calculating the model goodness of fit criteria such as the Cp statistic, Akaike's Information Criterion (AIC), Bayesian's Information Criterion (BIC), and adjusted R^2 that associated to the minimum RSS value. All of the above criteria are calculated involving the RSS value or its derivation where their formulas are given in (4) to (7) as the following, [31].

$$C_p = (1/m) * (RSS + 2 * p * \hat{\sigma}^2) \quad (4)$$

$$\text{AIC} = (1/(m * \hat{\sigma}^2)) * (RSS + 2 * p * \hat{\sigma}^2) \quad (5)$$

$$\text{BIC} = (1/(m * \hat{\sigma}^2)) * (RSS + \log(m) * p * \hat{\sigma}^2) \quad (6)$$

$$\text{R}^2 adj. = 1 - \left(\frac{\text{RSS}}{m-p-1} \Big/ \frac{\text{TSS}}{m-1} \right) \quad (7)$$

Where $\text{RSS} = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$, $\text{TSS} = \sum_{i=1}^{m} (y_i - \bar{y})^2$, and $\hat{\sigma}^2 = (1/(m-p-1)) * \min(RSS)$ with m is the number of instances, and p is several predictor variables.

The forward selection algorithm, in summary, is given in the following stages, [32]:

a. Building a simple linear model between the response variable and each of the predictor variables, calculating the RSS of each model, picking up the predictor variable having the minimum value of RSS as the first selected variable to enter into the linear model, and computing the values of Cp, AIC, BIC, and adjusted R^2 using the minimum RSS.
b. Building a multiple linear model with two predictor variables where one predictor variable is the first selected variable and the second variable is each of the remaining predictor variables, calculating the RSS of each model with two predictor variables, picking up the predictor variable having the minimum value of RSS as the second selected variable to enter into the linear model, and computing the values of Cp, AIC, BIC, and adjusted R^2 using the minimum RSS.
c. Building a multiple linear model with 3 predictor variables where 2 predictor variables are the selected variables in the previous stage and the third variable is each of the remaining predictor variables, calculating the RSS of each model with 3 predictor variables, pick up the predictor variable having the minimum value of RSS as the third selected variable to enter into the linear model, and computing the values of Cp, AIC, BIC, and adjusted R^2 using the minimum RSS.
d. The process of building a linear model, selecting the minimum RSS, and calculating of 4 criteria of model goodness of fit is repeated until all of the predictor variables have entered into the linear model.
e. The best ARDL model is determined by the majority vote of the model goodness of fit criteria.

2.4 The Model's Performance Metrics

The gap between the actual and prediction values has an important role in the measuring of the model's performance. There are many terms to call it including error, residual, and bias. A regression model performance is evaluated based on the accuracy measures including mean absolute error (MAE), root mean square error (RMSE), and determination coefficient (R^2) which are calculated directly from the gap values, [33]. MAE is calculated by considering the same weight to all of absolute errors to produce a positive value metric. On other side, the RMSE metric is calculated by giving a large weight to the large error. Both the MAE and RMSE have the same measure unit as the original response variable.

The correlation between the actual and prediction values squared will yield a metric called the

coefficient of determination (R^2) which in this case measures how coincidence between the prediction value generated by the model and the associated actual value. The R^2 value lies in the range of 0 to 1 where the R^2 = 1 means that the regression model can predict the actual value perfectly with 100% accuracy, [34]. When it is presented in a scatter plot of the actual versus prediction values, both values coincide and overlap perfectly with each to other. Here are the formulas to calculate the MAE, RMSE, and R^2, [35]:$MAE = \sum_{i=1}^{n} |y_i - \hat{y}_i|/n$
(8)

$$RMSE = \sqrt{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2/n} \quad (9)$$

$$R^2 = \left[cov(y_i, \hat{y}_i)/\left(\sqrt{Var(y_i)} * \sqrt{var(\hat{y}_i)}\right)\right]^2 \quad (10)$$

Where the $\hat{y}$ is the predicted value, and the y is the corresponding actual value.

3 Materials and Methods

The multiple time series studied in this study are consumer prices of 3 commodities namely the large chilies (L), Curly chilies (C), and Cayenne peppers (P) in Malang City of Indonesia. The data can be obtained through https://siskaperbapo.jatimprov.go.id/ provided by the East Java provincial industry office, which records basic commodity prices per kg in 37 cities' wholesale markets. The observation values were recorded from March/1/2020 to April/24/2023. These 3 commodities are supposed to have a correlation with each other that it was shown in Figure 1.

The red plot shows the time series as the response variable which is supposed to be influenced by 2 other time series. The original prices dataset will be transformed using min-max transformation to scale observation values in the range of 0 to 1, and also be transformed using Z score transformation to scale observation values in the range of -4 to 4.

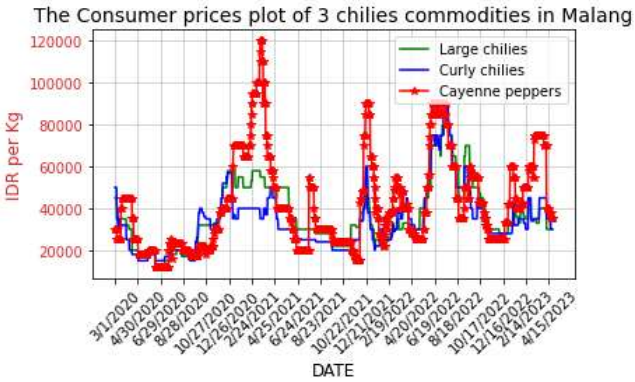

Fig. 1: The patterns of multiple time series plots of 3 chili commodities in Malang Indonesia

The development of ARDL models is expected to acquire a tool to predict the current value of the time series as the response variable with predictor variables formed by various previous values of the multiple time series involved in the dataset. As a case study, the multiple time series consisting of the prices of three commodities namely the Cayenne pepper, Large chilies, and Curly chilies with the current values of the Cayenne peppers as the response variable employed in the dataset for developing and evaluating the proposed ARDL model. The stages of developing the ARDL model with a machine learning approach are presented in Figure 2, it can be described in summary as follows:

a. Formatting the input and output pairs matrix based on several determined lags
b. Dividing the input and output pairs matrix into the training and testing sets.
c. Transforming both the min-max and Z score to the training set means getting 9 different types of training sets which are the combination of 3 training set types and 3 input lag numbers.
d. Conducting forward selection by the fitting of linear regression models om each training set
e. Choosing the ARDL predictor variables based on the majority voting of 4 goodness of fit criteria.
f. Training the ARDL model with the associated training set.

g. Evaluating the model's performance using the testing set with the associated ARDL predictor variables.

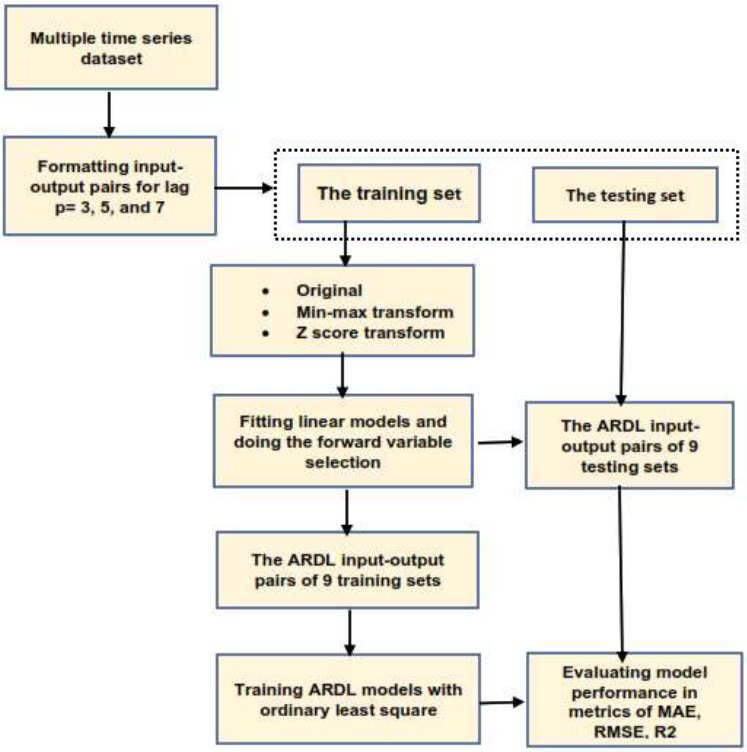

Fig. 2: The research schema on developing ARDL models using a machine-learning approach

4 Results and Discussion

To make simpler in naming the predictor variables used in the study, some terms are used namely the 'L-1', 'L-2', 'L-3', 'L-4', 'L-5', 'L-6', and 'L-7' of the predictor variable names for the large chilies on the lag of p = 1, 2, 3, 4, 5, 6, and 7 respectively. It is also used for naming the datasets of Curly chilies (C), and Cayenne peppers(P) by replacing the 'L' to be the 'C' or 'P' respectively. The model that be will developed has the Pt response variable and it is considered by setting the input lag number of p = 3, 5, and 7.

In detail, let the input lag k =5. The response variable Pt has to be influenced by 15 predictor variables which are 'P-1' to 'P-5', 'L-1' to 'L-5', and 'C-1' to 'C-5'. A similar condition is run for k = 3 and k = 5 where the response variable Pt has to be influenced respectively by 9 and 21 predictor variables. The union of the response variable and associated predictor variables creates the input-output data pairs in a matrix data structure. There are three matrices of input-output data pairs. One of the characteristics of the machine learning approach is the existence of the testing dataset which is picked up 100 last rows of each input-output matrix.

4.1 The Forward Selection and Selected Predictor Variables of ARDL Models

The selection of predictor variables was carried out separately for each input-output pairs matrix. The predictor variables were selected using the forward selection method based on the criteria of Cp, AIC, BIC, and adjusted R^2 statistics. The best ARDL predictor variables are determined by using majority votes of 4 goodness of fit criteria where they were

selected in each input-output pairs matrix of the training data on the original dataset and both normalized datasets with the min-max and Z score transformation.

Table 1. The forward selection stages in the input-output pairs of lag 5 on the Z-score dataset

Step	**Selected variable**	**Cp**	**AIC**	**BIC**	**adj. R^2**
1	['P-1']	0.01832	1.02013	1.02487	0.98206
2	['P-1', 'P-4']	0.01820	1.01375	1.02323	0.98219
3	['P-1', 'P-4', 'L-1']	0.01811	1.00888	**1.02310**	0.98229
4	['P-1', 'P-4', 'L-1', 'L-5']	0.01808	1.00687	1.02583	0.98234
5	['P-1', 'P-4', 'L-1', 'L-5', 'C-1']	0.01806	1.00577	1.02946	0.98238
6	***['P-1', 'P-4', 'L-1', 'L-5', 'C-1', 'C-5']***	**0.01795**	**0.99963**	1.02806	**0.98251**
7	['P-1', 'P-4', 'L-1', 'L-5', 'C-1', 'C-5', 'C-4']	0.01797	1.00080	1.03397	0.98250
8	['P-1', 'P-4', 'L-1', 'L-5', 'C-1', 'C-5', 'C-4', 'P-3']	0.01799	1.00195	1.03986	0.98250
9	['P-1', 'P-4', 'L-1', 'L-5', 'C-1', 'C-5', 'C-4', 'P-3', 'L-2']	0.01801	1.00325	1.04590	0.98249
10	['P-1', 'P-4', 'L-1', 'L-5', 'C-1', 'C-5', 'C-4', 'P-3', 'L-2', 'C-2']	0.01803	1.00442	1.05180	0.98249
11	['P-1', 'P-4', 'L-1', 'L-5', 'C-1', 'C-5', 'C-4', 'P-3', 'L-2', 'C-2', 'P-5']	0.01806	1.00609	1.05821	0.98248
12	['P-1', 'P-4', 'L-1', 'L-5', 'C-1', 'C-5', 'C-4', 'P-3', 'L-2', 'C-2', 'P-5', 'P-2']	0.01809	1.00777	1.06463	0.98246
13	['P-1', 'P-4', 'L-1', 'L-5', 'C-1', 'C-5', 'C-4', 'P-3', 'L-2', 'C-2', 'P-5', 'P-2', 'L-3']	0.01813	1.00960	1.07120	0.98245
14	['P-1', 'P-4', 'L-1', 'L-5', 'C-1', 'C-5', 'C-4', 'P-3', 'L-2', 'C-2', 'P-5', 'P-2', 'L-3', 'L-4']	0.01816	1.01148	1.07782	0.98243
15	['P-1', 'P-4', 'L-1', 'L-5', 'C-1', 'C-5', 'C-4', 'P-3', 'L-2', 'C-2', 'P-5', 'P-2', 'L-3', 'L-4', 'C-3']	0.01820	1.01340	1.08448	0.98241

The process of the variable selection was presented in both the forward selection matrix and the goodness of fit criteria curve. As an example, Table 1 shows a matrix describing the variable selection process in the Z score dataset on the input lag 5, while, Figure 3 presents a plot of 4 criteria values in Table 1. The plot will help to find the optimal criteria value of Cp, BIC, and AIC, and adjust R^2 easily. The ARDL predictor variables are determined using the goodness of fit criterion with the majority vote.

The forward selection method is started by the fitting of a simple linear model between the response variable (Pt) and each of the predictor variables which are 15 variables namely the 'L-5' to 'P-1'. Based on each simple linear model, the calculation of each residual sum square (RSS) and finding the minimum one to select the predictor variable that enters the linear model. The best RSS is used to calculate the associated criteria of the Cp, AIC, BIC, and adjusted R^2 values which are the goodness of fit criteria for selecting the best predictor variables of the ARDL model.

The forward selection method is started by the fitting of a simple linear model between the response variable (Pt) and each of the predictor variables which are 15 variables namely the 'L-5' to 'P-1'. Based on each simple linear model, the

calculation of each residual sum square (RSS) and finding the minimum one to select the predictor variable that enters the linear model. The best RSS is used to calculate the associated criteria of the Cp, AIC, BIC, and adjusted R^2 values which are the goodness of fit criteria for selecting the best predictor variables of the ARDL model.

Let's consider the first step in Table 1, the step implies that the 'P-1' variable has the minimum RSS value and it is picked up as the first predictor variable of the linear model. The succeeding columns present the associated Cp, AIC, BIC, and adjusted R^2 values. The second step is obtained by fitting a linear model with 2 predictor variables where the first variable is 'P-1' which is one obtained in the previous step and the second predictor variable is one of the remaining variables. The linear model with 2 predictor variables of ['P-1', 'P-4'] has the minimum RSS value which means the 'P-4' variable entered into the linear model with 2 predictor variables then 4 associated goodness of fit criteria are calculated. The sequence processes of the fitting linear model with adding one variable of remaining variables, calculating their RSS, picking up the variable with the minimum RSS, and calculating the associated goodness of fit criteria are carried out until the remaining variable is empty.

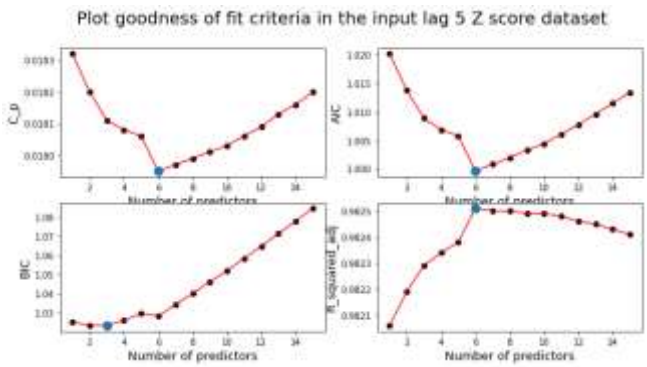

Fig. 3: Plots of 4 goodness of fit criteria of the ARDL with input-output pairs of lag 5 on the Z score dataset

Based on Figure 3, the marked point with the majority vote was found in the criteria of the AIC and adjusted R^2 curves where the number of predictor variables is 6 which means it refers to predictor variables at step 6 in Table 1. The list variable in the step 6 is the selected predictor variable of the best ARDL model in the input-output pairs of lag 5 on the Z score dataset. There are as many as 8 other matrices similar to Table 1, and 8 other figures similar to Figure 3. Nevertheless, the selected predictor variables in the same input lag number in all datasets yielded the same predictor variables although the values of criteria are different. The selected predictor variables of the ARDL models in all of the input lags and datasets are presented in Table 2.

Table 2. The selected predictor variables of ARDL models for 3 input lags and 3 datasets

Input	The selected predictor variables
Lag 3	['P-1', 'L-1', 'P-3', 'L-2', 'C-1', 'C-2']
Lag 5	['P-1', 'P-4', 'L-1', 'L-5', 'C-1', 'C-5']
Lag 7	['P-1', 'P-7', 'L-1', 'C-1', 'C-4', 'L-5', 'P-5', 'C-5']

Table 2 shows the selected predictor variables yielded by the forward selection method where there are differences in the variable's name for the different input lags. Table 2 implies the influencing order of the predictor variables on the response variable. The variable 'P-1' is the most influencing the response variable 'Pt' in all of the input lag scenarios. However, the influencing order of predictor variables in the second place, and higher order place are not definite.

Table 3. The model's coefficients and performances of all scenarios of input lags and datasets

Dataset	**The ARDL coefficients at the input lag 3** ['P-1', 'L-1', 'P-3', 'L-2', 'C-1', 'C-2']	**RSME**	**MAE**	**R^2**
Original	[1.0249 0.2103 -0.0473 -0.1721 -0.1512 0.1329]	3949.079	1462.045	0.9260
Min-max	[1.0249 0.1422 -0.0473 -0.1164 -0.105 0.0923]	3980.853	1716.289	0.9263
Z score	[1.0249 0.1417 -0.0473 -0.1159 -0.0968 0.085]	3976.087	1689.691	0.9263
Dataset	**The ARDL coefficients at the input lag 5** ['P-1', 'P-4', 'L-1', 'L-5', 'C-1', 'C-5']	**RSME**	**MAE**	**R^2**
Original	[1.0226 -0.0434 0.1465 -0.1175 -0.0989 0.0867]	4002.883	1511.172	0.9239
Min-max	[1.0226 -0.0434 0.099 -0.0794 -0.0687 0.0602]	4015.202	1730.366	0.9249
Z score	[1.0226 -0.0434 0.0987 -0.0792 -0.0633 0.0555]	4012.223	1713.412	0.9249
Dataset	**The ARDL coefficients at the input lag 7** ['P-1', 'P-7', 'L-1', 'C-1', 'C-4', 'L-5', 'P-5', 'C-5']	**RSME**	**MAE**	**R^2**
Original	[1.0103 -0.0778 0.1337 -0.107 0.0301 -0.1023 0.0424 0.0666]	4005.283	1615.107	0.9242
Min-max	[1.0103 -0.0778 0.0904 -0.0743 0.0209 -0.0692 0.0424 0.0462]	4044.207	1871.268	0.9251
Z score	[1.0103 -0.0778 0.0901 -0.0685 0.0193 -0.0689 0.0424 0.0426]	4040.766	1857.270	0.9251

4.2 The Coefficients of ARDL Models and Their Interpretation

The coefficients of each ARDL model are estimated by using the OLS method on the input-output pairs of the training dataset that all of not selected predictor variables are dropped from the dataset. The dropping of not selected predictor variables is also conducted on the testing datasets which are used for the evaluation of the model's performances. The model's coefficients yielded for all scenarios of input lags and datasets are presented in Table 3.

The ARDL model's coefficients are given in the second column of Table 3. The coefficients follow the predictor variables in the above row. As an illustration, the ARDL model with input lag 3 and the original dataset can be written as the following:

$$P_t = 1.0249P_{t-1} + 0.2103L_{t-1} - 0.0473P_{t-3} - 0.1721L_{t-2} - 0.1512C_{t-1} + 0.1329C_{t-2}$$

The current price of the Cayenne peppers(P) is influenced by its price yesterday and the previous third day, the large chilies price (L) yesterday and last yesterday, the price of Curly chilies (C) yesterday and last yesterday. The model's coefficients state the important levels of each associated predictor variable on the current price of the Cayenne peppers. The interpretation of each coefficient can be stated as the following: the current Cayenne peppers price will increase IDR 10249 and decrease IDR 473 when it's yesterday and previous third-day price increases IDR 10000, will increase IDR 2103 and decrease IDR 1721 when the yesterday and last yesterday price of the large chilies increases IDR 10000 and will decrease IDR 1512 and increase IDR 1329 when the yesterday and last yesterday price of the Curly chilies increases IDR 10000.

The ARDL model for the prediction of the next price of Cayenne peppers can be obtained by setting t = t+1, so the ARDL prediction model can be stated as follows:

$$P_{t+1} = 1.0249P_t + 0.2103L_t - 0.0473P_{t-2} - 0.1721L_{t-1} - 0.1512C_t + 0.1329C_{t-1}$$

Based on Table 3, there will be obtained as many as 9 ARDL prediction models which are evaluated in their performance by using the

associated testing dataset with the selected predictor variables.

4.3 Performance of ARDL Models and Discussion

The model's performances on the testing data are presented in 3 metrics namely the RMSE, MAE, and R^2 which are given respectively in columns 3 to 5 in Table 3. The best RMSE is IDR 3949 for the ARDL lag 3 with the original dataset. The second and third small RMSE are IDR 3976 and IDR 3981 respectively for the ARDL lag 3 with the Z score and min-max datasets. The other models have the RMSE values higher than IDR 4002 which is between IDR 4002 and IDR 4044. The RMSE values of the ARDL lag 3 in all of the datasets are smaller than other models with lag 5 and lag 7. The models with the original dataset also yield smaller RMSE than the other models with the Min-max and Z score datasets.

The first, second, and third best MAE's are respectively IDR1462, IDR 1511, and IDR 1615 for the ARDL lag 3, lag 5, and lag 7 with the original dataset. The other models have the MAE values higher than IDR 1689 which is between IDR 1689 and IDR 1871. The MAE values of the ARDL with the original dataset in all of the lags are smaller than the MAE values of other models in all of the lags with the Min-max and Z score datasets.

The best R^2 value is 92.63% for the ARDL lag 5 and lag 7 with the original dataset. If the R^2 values are rounded into 3 decimals there are only 3 R^2 values i.e., 92.4%, 92.5%, and 92.6%. The models with lag 5 and lag 7 are better the R^2 values than the models with lag 3. The R^2 value of ARDL models with the original dataset is higher than the models with other datasets.

Both metrics of the RMSE and MAE are closely related to the plot between the actual and prediction values presented in Figure 4. Both the actual and prediction values in Figure 4 seem to coincide with each other but there are some sharp fluctuations in the testing dataset especially on April 9 to April 10, 2023. When the sharp fluctuation occurred the gap between the actual and prediction values was very large. The RMSE metric gives a large weight to the large gap because it is calculated based on the gap squared, [35]. On the other hand, the MAE metric gives the same weight to all of the gaps including the large gap, [35]. The difference in treatment to the gap of both metrics causes the MAE values to be smaller than the RMSE values in all of the lag number and dataset scenarios. Even the MAE values are around 3 times smaller than the RMSE values.

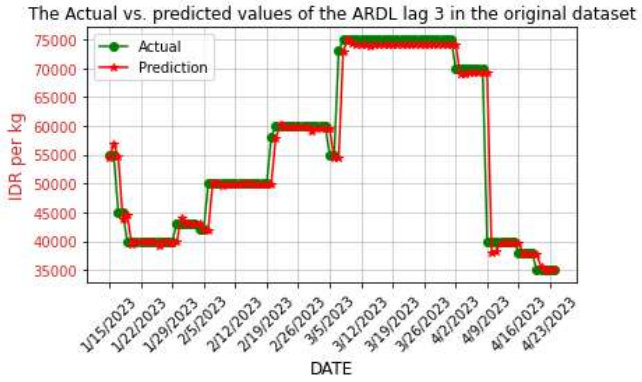

Fig. 4: Plot the actual versus prediction values of the ARDL lag 3 with original data in the testing dataset

The large distance between the actual and prediction value often occurs when the Cayenne pepper prices fluctuate extremely. On some sequence days, the price remains unchanged because the sellers are not supplied each day. The price is uncontrolled by customers but it is the domain of the supplier and seller determining the price. The unchanged prices in a few days are easily predicted but extreme price switching implicates the ARDL model to have a poor prediction. The ARDL model cannot give a fast response to extreme price switching. The advanced models such as the hybridization models published in [17] and [26] or the complicated machine learning models such as those published in [27] and [30] should be tried for tackling in modeling the datasets with present the extreme price switching.

The R^2 metrics describe how the actual and prediction values are close or overlap with each other. The R^2 value is 100% which means the actual and prediction values overlap perfectly, [35]. The scatter plot given in Figure 5 is a tool for describing the

relationship between the actual and predicted values. The R^2 value of the scatter plot is 92.63%. When all of the actual values can be predicted perfectly by the model, the scatter plot will form a solid line which means a perfect correlation between the actual and prediction values. In Figure 5, some points are long distances from the solid line which means a large gap between the actual and prediction value. It also supports the fact that there are many large gaps shown by some points that are far away from the solid line. Presenting some constant prices and extreme switching prices influences the R^2 values. They indirectly cause large variability in the datasets subsequently leading to reduce the acquired R^2 value. The input-output pairs data with various lag numbers of 3, 5, and 7 cannot change the variability in the dataset that affects the produced R^2 values with slight differences, and also the data transformation does not affect variability in the dataset.

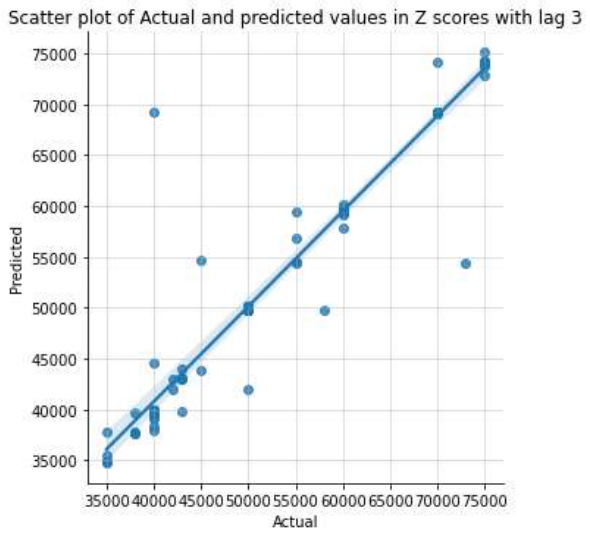

Fig. 5: The scatter plot between the actual and prediction values of the ARDL lag3 with Z score data in the testing dataset

The ARDL lag 3 models outperform the ARDL lag 5 and lag 7 models in all of the performance metrics where the result confirms and supports the basic principle of model identification in the conventional time series modeling that told the lag number of time series models lies in a range of 1 to 3 and almost impossible occurs with the higher lag number, [25]. Nevertheless, A machine learning modeling approach is more flexible than the conventional time series because some stages including model identification and residual diagnostic tests are not conducted and are considered useless stages.

The ARDL models with the original dataset outperform the models with other datasets in the RMSE and MAE metrics. While the R^2 metric of the ARDL models in both the Min-max and Z score datasets is higher than the models with the original dataset although the differences are not significant (around 0.1%). It seems useless to transform a dataset into a normalized dataset because the datasets have a commensurate measurement in the IDR and the domain values do not have a large enough gap. However, the normalized data must be carried out when each time series dataset in multiple time series has various measurement units, [22].

The research has acquired the ARDL models which are not satisfactory in their performance but easily interpreted in modeling multiple time series. The coefficients of the ARDL models have a meaning similar to the popular model of multiple regression. The obtained models are developed systematically which is different from generally developing time series models such as those published in [13], [14] and [15] which are tricky in the process of developing models.

5 Conclusion

The predictor variables selected with the majority vote of the CP, AIC, BIC, and adjusted R^2 have yielded 6 predictors for input lag 3 and input lag 5, but they obtained 8 predictors for input lag 7. Although the values of the 4 criteria are different in 3 types of datasets, the condition does not influence the acquired ARDL predictor variables at the same lag. The ARDL lag 3 with the original dataset is the best one in both performance metrics of the RMSE and MAE where the performance values are IDR 3949 and IDR1462 for metrics of the RMSE and MAE respectively. While the ARDL lag 3 with Z score and Min-max datasets has the highest R^2 metric which is 92.63% although the all of 9 ARDL

models have the R^2 metric in the range of 92.40% to 92.63%. The sequence with the same prices of the Cayenne peppers in a few days and the large switching prices occurrence has caused the R^2 metric of ARDL models to have almost same values. The recommendation of the next works should employ the producer prices of three chili commodities and also add some related time series datasets to develop the ARDL model and hybrid ARDL with popular machine learning models.

Acknowledgment:
We are grateful for the funding from Universitas Brawijaya, with the grant of Hibah Doktor Lektor Kepala, No. 3110.4/UN10.F09/PN/2022.

References:

[1] L. A. Qodri, D. Wulandari, and H. Sumarsono. Food stability analysis in East Java, *International Journal of Scientific and Technology Research*, vol. 9, no. 2, 3712-3716, 2020.

[2] S. M. Khasanah, M. Maksum, and E. Suwondo. Trend Analysis of Red Chili Price-Formation Models, *agriTECH*, vol. 40, no. 1, 57-63, 2020.

[3] R. C. Sutomo, S. Subandiyah, A. Wibowo, and A. Widiastuti. Description and Pathogenicity of Colletotrichum Species Causing Chili Anthracnose in Yogyakarta, Indonesia, *Agrivita*, vol. 44, no. 2, 312-321, 2022.

[4] Megawati, M. I. Sulaiman, and S. Zakaria. Effect of Planting Season on the Residue of Organophosphate in Chili (Capsicum annuum L.), *Indian J Agric Res*, vol. 56, no. 5, 614-620, 2022.

[5] R. N. Ihsan, S. Saadah, and G. S. Wulandari. Prediction of Basic Material Prices on Major Holidays Using Multi-Layer Perceptron, *Jurnal Media Informatika Budidarma*, vol. 6, no. 1, 443-452, 2022.

[6] B. Sisman, J. Yamagishi, S. King, and H. Li. An overview of voice conversion and its challenges: From statistical modeling to deep learning, *IEEE/ACM Trans Audio Speech Lang Process,* vol. 29, 132-157, 2021.

[7] J. Bulbulia, U. Schjoedt, J.H. Shaver, R. Sosis and W.J. Wildman. Causal inference in regression: advice to authors, *Religion, Brain & Behavior*, 11:4, 353-360, 2021.

[8] P. Chatterjee, M. Yazdani, F. Fernández-Navarro, and J. Pérez-Rodríguez. *Machine Learning Algorithms and Applications in Engineering*. 2023.

[9] J. Pérez-Ortega, S.S. Roblero-Aguilar, N.N. Almanza-Ortega, J.F. Solís, C. Zavala-Díaz, Y. Hernández, and V. Landero-Nájera, Hybrid Fuzzy C-Means Clustering Algorithm Oriented to Big Data Realms. *Axioms*. 11(8):377, 2022.

[10] H. Alharthi, N. Sultana, A. Al-amoudi, and A. Basudan. An Analytic Hierarchy Process–based Method to Rank the Critical Success Factors of Implementing a Pharmacy Barcode System, *Perspect Health Inf Manag*. Winter: 12, 2015.

[11] S. Handoyo, Y. P. Chen, G. Irianto, and A. Widodo. The varying threshold values of logistic regression and linear discriminant for classifying fraudulent firm, *Mathematics, and Statistics,* vol. 9, no. 2, 135-143, 2021.

[12] J. Pek, O. Wong, and C.M. Wong. How to address non-normality: a taxonomy of approaches, reviewed, and ilustrated, *Fronties in Psychology*, Vol. 9, 2014, 2018

[13] S. S. Bakshi, R. K. Jaiswal, and R. Jaiswal. Efficiency Check Using Cointegration and Machine Learning Approach: Crude Oil Futures Markets, *Procedia Computer Science*, Elsevier B.V., 304–311, 2021.

[14] C. Dong, J. Liu, Y. Lu, and L. Zhang. Stock Value Prediction Based on Merging SARIMA

Model and Monte Carlo Model, *ACM International Conference Proceeding Series*, 510-514, 2022.

[15] L. Luo, L. Luo, X. Zhang, and X. He. Hospital daily outpatient visits forecasting using a combinatorial model based on ARIMA and SES models, *BMC Health Serv Res*, vol. 17, no. 1, 1-13, 2017.

[16] T. K. Quartey-Papafio, S. A. Javed, and S. Liu. Forecasting cocoa production of six major producers through ARIMA and grey models, *Grey Systems*, vol. 11, no. 3, 434-462, 2021.

[17] H. Kusdarwati and S. Handoyo. System for prediction of non-stationary time series based on the wavelet radial basis function neural network model, *International Journal of Electrical and Computer Engineering*, vol. 8, no. 4, 2327-2337, 2018.

[18] S. Gričar. Implementation of Vector Auto-Regression Models in Tourism: State of the Art Analysis and Further Development, *Tourism and Hospitality Management*, vol. 28, no. 3, 707-709, 2022.

[19] K. Natsiopoulos and N. G. Tzeremes. ARDL bounds test for cointegration: Replicating the Pesaran et al. (2001) results for the UK earnings equation using R, *Journal of Applied Econometrics*, vol. 37, no. 5, 1079-1090, 2022.

[20] Y. Xie, Y. Li, Z. Xia, and R. Yan. An Improved Forward Regression Variable Selection Algorithm for High-Dimensional Linear Regression Models, *IEEE Access*, vol. 8, 129032-129042, 2020.

[21] J. B. Rounds, R. Dawis, and L. H. Lofquist. Measurement of person-environment fit and prediction of satisfaction in the theory of work adjustment, *J Vocat Behav*, vol. 31, no. 3, 297-318, 1987.

[22] S. Kumar, S. Gupta, S. Arora, and S. Kumar. A comparative simulation of normalization methods for machine learning-based intrusion detection systems using KDD Cup'99 dataset, *Journal of Intelligent and Fuzzy Systems*, vol. 42, no. 3, 1749-1766, 2022.

[23] A. Zeraibi, D. Balsalobre-Lorente, and M. Murshed. The influences of renewable electricity generation, technological innovation, financial development, and economic growth on ecological footprints in ASEAN-5 countries, *Environmental Science and Pollution Research*, vol. 28, no. 37, .51003-51021, 2021.

[24] Kamran Ali, Muhammad Siddique, Muhammad Amir Chaudhry, and Haider Tariq. Financial development and economic growth: An application of ARDL model on developed and developing countries, *Journal of Public Value and Administrative Insight*, vol. 5, no. 1, 170-186, 2022.

[25] K. Z. Javangwe and O. Takawira. Exchange rate movement and stock market performance: An application of the ARDL model. *Cogent Economics and Finance*, vol. 10, no. 1, 2075520, 2022.

[26] S. Handoyo and Marji. The fuzzy inference system with least square optimization for time series forecasting, *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 11, no. 3, 1015-1026, 2018.

[27] S. Mohapatra and N. Chaudhary. Statistical Analysis and Evaluation of Feature Selection Techniques and implementing Machine Learning Algorithms to Predict the Crop Yield using Accuracy Metrics, *Engineered Science*, vol. 21, 787, 2023.

[28] Y. Ma, D. Tsao, and H. Y. Shum. On the principles of Parsimony and Self-consistency for the emergence of intelligence, *Frontiers of Information Technology and Electronic Engineering*, vol. 23, no. 9, 1298-1323, 2022.

[29] J. Linja, J. Hämäläinen, P. Nieminen, and T. Kärkkäinen. Feature selection for distance-based regression: An umbrella

review and a one-shot wrapper, *Neurocomputing*, vol. 518, 462-479, 2023.

[30] E. Hancer, B. Xue, M. Zhang, D. Karaboga, and B. Akay. Pareto front feature selection based on artificial bee colony optimization, *Inf Sci (NY)*, vol. 422, 2018.

[31] R. Rossi, A. Murari, P. Gaudio, and M. Gelfusa. Upgrading model selection criteria with goodness of fit tests for practical applications, *Entropy*, vol. 22, no. 4, 447, 2020.

[32] H. Zhou, K. M. Yu, Y. C. Chen, and H. P. Hsu. A Hybrid Feature Selection Method RFSTL for Manufacturing Quality Prediction Based on a High Dimensional Imbalanced Dataset, *IEEE Access*, vol. 9, 29719-29735, 2021.

[33] V. Plevris, G. Solorzano, N. P. Bakas, and M. E. A. Ben Seghier. Investigation Of Performance Metrics In Regression Analysis And Machine Learning-Based Prediction Models. *World Congress in Computational Mechanics and ECCOMAS Congress*, 2022.

[34] J. de Souza Zanirato Maia, A. P. A. Bueno, and J. R. Sato. Assessing the educational performance of different Brazilian school cycles using data science methods, *PLoS One*, vol. 16, no. 3, e0248525, 2021.

[35] D. Chicco, M. J. Warrens, and G. Jurman. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE, and RMSE in regression analysis evaluation, *Peer J Comput Sci*, vol. 7, e623, 2021.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

- Achmad Efendi did conceptualization and formulation of ideas as well as research aims, did formal statistical analysis, writing discussion of the analysis results, and finally conclusion.
- Yusi T. Mursityo carried out the data curation, data management, and writing result and discussion.
- Ninik W. Hidajati, Nur Andajani, and Zuraidah Zuraidah provided the study materials, supplying data, discussion and revision.
- Samingun Handoyo did formal statistical analysis, intrepretasion, and conclusion.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

We received the funding from Universitas Brawijaya, with the grant of Hibah Doktor Lektor Kepala (Doctoral research funding), No. 3110.4/UN10.F09/PN/2022.

Conflict of Interest

The authors have no conflicts of interest to declare.