Implementation of the ARIMA(p,d,q) method to forecasting CPI Data using forecast package in R Software

Ansari Saleh Ahmar\textsuperscript{1*}, Achmad Daengs GS\textsuperscript{2}, Tri Listyorini\textsuperscript{3}, Castaka Agus Sugianto\textsuperscript{4}, Y Yuningsih\textsuperscript{5}, Robbi Rahim\textsuperscript{6}, & Nuning Kurniasih\textsuperscript{7}

\textsuperscript{1}Department of Statistics, Universitas Negeri Makassar, Indonesia
\textsuperscript{2}Universitas 45 Surabaya, Surabaya, Indonesia
\textsuperscript{3}Teknik Informatika, Universitas Muria Kudus
\textsuperscript{4}Teknik Informatika - Politeknik TEDC Bandung
\textsuperscript{5}Department of Management, Faculty Economic and Business, Universitas Pembangunan Nasional “Veteran”, Jawa Timur, Indonesia
\textsuperscript{6}School of Computer and Communication Engineering, Universiti Malaysia Perlis, Perlis, Malaysia
\textsuperscript{7}Faculty of Communication Sciences, Universitas Padjadjaran, Bandung, Jawa Barat 45363, Indonesia

ansarisaleh@unm.ac.id

Abstract. The Consumer Price Index is an index which calculates the average change in prices over a period, of a set of goods and services consumed by the population/households within a specific time. In this paper will be discussed about the forecasting of consumer price index of Indonesia using forecast package with R Software. The forecasting process this data using algorithms popularized by Rob J. Hyndman and Yeasmin Khandakar in 2008. By using this method, it is obtained a suitable ARIMA model to forecast CPI data Indonesia. The model most suitable time series is ARIMA (1,0,0).

1. Introduction
Consumer Price Index (CPI) is an index that is used to describe the rate of inflation or deflation of goods. Based on information from the Badan Pusat Statistik (BPS-Statistics Indonesia), the consumer price index (CPI) is an index that measures the average change in the intertemporal price of a package of goods and services consumed by the population/households in urban areas by a given period. The calculation of the CPI is intended to determine changes in the price of a fixed group of goods/services that are consumed by the public. CPI change from time to time describe the rate of increase (inflation) or the rate of decrease (deflation) of goods/services household needs every day.

CPI change from time to time be predictable. Forecasting is an activity to predict what will happen in the future \cite{1}. CPI data associated with time, so it can be said that the CPI data is time-series data. To forecast a time series then we can use forecasting method, e.g., ARIMA Box-Jenkins method \cite{2} and $\alpha$-Sutte Indicator \cite{3, 4}. The purpose of this study was to determine forecasting CPI data using the ARIMA Box-Jenkins method.
2. Literature

2.1. Consumer Price Index (CPI)
Consumer Price Index (CPI) is an index, which calculates the average change in prices over a period, of a set of goods and services consumed by the population/households within a specific time [5].

Furthermore, BPS explains that the CPI of Indonesia is calculated with modified Laspeyre's formula. In calculating the average price of commodities, the measure used is the arithmetic mean, but for some products such as rice, cooking oil, gasoline, and so used the geometry means [5]. Since January 2014, CPI have presented using the base year 2012=100 and includes 82 cities which cover 33 provincial capitals and 49 major cities throughout Indonesia. And CPI previously used the base year 2007=100 and only covered 66 cities of Indonesia.

CPI is calculated using a modified Laspeyres formula as follows [6]:

$$\text{IHK}_n = \frac{\sum_{i=1}^{k} \frac{P_{ni}}{P_{(n-1)i}} P_{(n-1)i} Q_{oi}}{\sum_{i=1}^{k} P_{oi} Q_{oi}} \times 100$$

where:
- $\text{IHK}_n$ = CP at $n^{th}$ (n: January, February, ..., December)
- $\frac{P_{ni}}{P_{(n-1)i}} \times 100$ = Relative price of commodity $i$ at month-to-month $n^{th}$ (n-1)
- $P_{ni}$ = Price of commodity $i$ at month $n$
- $P_{(n-1)i}$ = Price of commodity $i$ at $(n-1)^{th}$
- $P_{(n-1)i} Q_{oi}$ = Value of commodity consumption at the month $i$ to $(n-1)^{th}$
- $J$ = Value of commodity consumption in the base year
- $k$ = Number of commodities which are included in the sub-group / groups / total expenditure

2.2. ARIMA
If the observation time series can be predicted with certainty and does not require further investigation, it is called deterministic time series and if observations can only show the probabilistic structure of state will come a time series, the time series is called stochastic.

In the modeling of time series analysis, assuming that the data are stationary. Time series is said to be stationary if there is no change in the average trends and changes in the variance. Stationary of time series is relatively no increase or decrease in value extremes in the data or data reside on the fluctuations around the average value constant.

Stationary can be viewed by using time series diagram is a scatter diagram between the variables with time $t$. If the diagram of time series fluctuates around a line parallel to the axis of time ($t$), then it is said to be stationary in average. If the condition is stationary on average not met the necessary process of differentiating [7].

The process of differencing the first order represents the difference between the data all the data into $(t-1)$, i.e $\Delta Z_t = Z_t - Z_{t-1}$, forms for differencing second order is

$$\Delta^2 Z_t = \Delta Z_t - \Delta Z_{t-1}$$

$$= (Z_t - Z_{t-1}) - (Z_{t-1} - Z_{t-2})$$

$$= Z_t - 2Z_{t-1} + Z_{t-2}$$
If the condition is stationary in variance is not met, Box & Cox [8] introduced the rank transformation that is \( Z^{(r)}_t = \frac{Z_t^{(r)} - 1}{\lambda} \), (\( \lambda \) is a parameter), known as Box-Cox transformation. Here are some provisions to stabilize the variance is:

1. The transformation may be done only for series \( Z_t \) positive.
2. Transformation is done before differencing and modeling of time series.
3. Value of \( \lambda \) selected based on the sum of squares error (SSE) of the series transformed. Value of SSE smallest deliver results most constant variance.
4. Transformation is not only stabilized the variance but also be able to normalize the distribution.

Autoregressive Integrated Moving Average (ARIMA) models [2], [4], [9] has been studied by George Box and Gwilym Jenkins in 1976 [10], and their names are then often synonymous with ARIMA process which applied to the analysis of time series. In general ARIMA model is written with the notation ARIMA(p, d, q), where p denotes the order of the process autoregressive (AR), d is express differencing, and q is the specific order of the moving average (MA). Model autoregressive (AR) was first introduced by Yule in 1926 [11] and later developed by Walker in 1931 [12], while Slutsky first used the model Moving Average (MA) in 1937 [13]. And in 1938, then Wold produce the basic theory of combination process of ARMA and combinations is then often used [14].

Box and Jenkins have effectively managed to reach agreement on the relevant information required to understand and use ARIMA models for time series one variable [15]. A process (\( Z_t \)) can be said to follow a mixed model autoregressive-moving average or ARMA (p, q) if it satisfies:

\[
\phi_p(B)Z_t = \theta_q(B)a_t
\]

with \( \phi_p(B) = (1 - \phi_1B - \phi_2B^2 - \ldots - \phi_pB^p) \) (for AR (p))

and \( \theta_q(B) = (1 - \theta_1B - \theta_2B^2 - \ldots - \theta_qB^q) \) (for MA (q))

If there is a differencing then becomes ARIMA models as follows:

\[
\phi_p(B)(1-B)^dZ_t = \theta_q(B)a_t
\]

with \( \phi_p(B) = (1 - \phi_1B - \phi_2B^2 - \ldots - \phi_pB^p) \) (for AR (p)), \( (1-B)^d \) (for non-seasonal differencing) and \( \theta_q(B) = (1 - \theta_1B - \theta_2B^2 - \ldots - \theta_qB^q) \) (for MA (q)) [15]

In doing forecasting of time Series data using ARIMA (p, d, q) methods, there are called steps or phases. The phases in forecasting that [14].

1. Identification of model
   Identification of the model is made to see the significance of autocorrelation and stationary of data, so whether or not the transformation or differencing process. From this stage, the model is obtained while the model testing process will be conducted in accordance whether or not the data.

2. Assessments and testing model
   After the process of identifying of the model, then the next step is the estimation and model testing. At this stage is divided into two parts, namely parameter estimation and diagnostic models.
   a) Parameter estimation
      Having obtained one or several models while the next step is to seek estimates for the parameters in the model.
   b) Diagnostic Model
      Diagnostic checking is passed to check whether the model is estimated quite suitable or adequate with existing data. Diagnostic checking based on the residual analysis. The
The underlying assumption of ARIMA model is that residuals are independent normally distributed random variables with mean zero constant variance.

2.3. Forecast Package For R
R Software is a statistical software that is open source for all platforms, ie, Windows, Linux and OS X (Mac). The software can be downloaded at http://cran.r-project.org. To run an analysis of statistical data, it is necessary that the named package. Package on R varies depending on user needs regarding analysis. For example, for the analysis of forecasting to use TSA package. Forecasting package in R is a lot, including the TSA package, FitARMA package, forecast package, mtsdi package, and TSPred package. Hyndman & Khandakar explains that forecast package is the methods and tools used to display and analyze the time series forecasting univariate including exponential smoothing model with state space and automatic ARIMA modelling [16].

3. Method
The data used in this research is secondary data. This secondary data obtained from the BPS-Statistics Indonesia website (www.bps.go.id). CPI data used is the CPI from January 2005 to December 2015. This data will be carried out forecasting by using forecast package on R.

4. Result and Discussion
The first step taken in the modeling of time series at the stage of identification is checked stationery of data by plotting the data are as shown below.

4.1. Model Identification and Assessment and Testing Model
By using the following command in R:
> library(forecast)
> arimaft <- auto.arima(data, stationary=TRUE, trace=TRUE)

Forecasting results obtained as follows:
ARIMA(2,0,2) with non-zero mean : 843.0854
ARIMA(0,0,0) with non-zero mean : 1064.619
ARIMA(1,0,0) with non-zero mean : 835.8355
ARIMA(0,0,1) with non-zero mean : 957.6918
ARIMA(0,0,0) with zero mean : 1650.995

Figure 1. Time Series Data Plot

4
ARIMA(2,0,0) with non-zero mean : 838.9077
ARIMA(1,0,1) with non-zero mean : 837.9421
ARIMA(2,0,1) with non-zero mean : 841.0928
ARIMA(1,0,0) with zero mean : Inf

Best model: ARIMA(1,0,0) with non-zero mean

From the above results showed that the best model is ARIMA (1,0,0). To see the results of forecasting using ARIMA (1,0,0) use the following command in the Software R:

```R
> fit <- Arima(data, order=c(1,0,0))
> summary(fit)
> forecast(fit)
> plot(forecast(fit))
```

Obtained the following results:

Series: data
ARIMA(1,0,0) with non-zero mean

Coefficients:

| ar1   | intercept |
|-------|-----------|
| 0.9067 | 128.2355  |

s.e. 0.0349 4.9941

sigma^2 estimated as 32.44:  log likelihood= - 414.64
AIC=835.27  AICc=835.46  BIC=843.9

Training set error measures:

| ME    | RMSE   | MAE   | MPE   | MAPE  | MASE  | ACF1 |
|-------|--------|-------|-------|-------|-------|------|
| 0.1122859 | 5.695209 | 2.060275 | -0.1118524 | 1.624827 | 1.400521 | 0.01311398 |

Box-Pierce test
data: fit$residuals
X-squared = 0.0225, df = 1, p-value = 0.8807

4.2. Diagnostic Model

From the above model, it can be seen that the intercept did not meet because of se value greater than α=0.05. Value Box-Pierce test meet for p-value > α = 0.05. If we saw the value of the training set error measures, the values of ME, RMSE, MAE, MPE, MAPE, MASE, and ACF1 is very small. So the time series model of the Consumer Price Index (CPI) Indonesia.

ARIMA (1,0,0), with AR coefficient is 0.9067 or modeling in mathematics form as follows:

\[
(1-\phi_1 B)Z_t = a_t
\]

\[
Z_t - \phi_1 Z_{t-1} = a_t
\]

\[
Z_t = \phi_1 Z_{t-1} - a_t
\]

\[
Z_t = 0.9067 \cdot Z_{t-1} - a_t
\]
For forecasting used commands: \textit{forecast (fit)}, and the results:

\begin{tabular}{|c|c|c|c|c|}
\hline
Point Forecast & Lo 80 & Hi 80 & Lo 95 & Hi 95 \\
\hline
132 & 123.4793 & 116.1806 & 130.7780 & 112.3169 & 134.6417 \\
133 & 123.9230 & 114.0707 & 133.7780 & 108.8553 & 138.9907 \\
134 & 124.3253 & 112.7895 & 135.8610 & 106.6829 & 141.9677 \\
135 & 124.6900 & 111.9356 & 137.4445 & 105.1838 & 144.1962 \\
136 & 125.0208 & 111.3455 & 138.6960 & 104.1063 & 145.9352 \\
137 & 125.3207 & 110.9324 & 139.7089 & 103.3158 & 147.3255 \\
\hline
Point Forecast & Lo 80 & Hi 80 & Lo 95 & Hi 95 \\
\hline
138 & 125.5926 & 110.6437 & 140.5415 & 102.7302 & 148.4549 \\
139 & 125.8391 & 110.4445 & 141.2337 & 102.9251 & 149.3831 \\
140 & 126.0626 & 110.3111 & 141.8142 & 101.9727 & 150.1526 \\
141 & 126.2653 & 110.2262 & 142.3044 & 101.7357 & 150.7950 \\
\hline
\end{tabular}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{ARIMA(1,0,0).png}
\caption{Plot Forecasting ARIMA (0,1,0)}
\end{figure}

5. Conclusion
Forecasting the results of the discussion above, it can be concluded that (1) time series model is best for the Consumer Price Index (CPI) Indonesia is ARIMA (1,0,0) or \(Z_t = 0.9067 \cdot Z_{t-1} - \alpha_t\), and (2) By using forecast Software R package on the identification of the ARIMA model (p, d, q) best very quickly obtained.

References
[1] Ahmar A S Rahman A Arifin A N M and Ahmar A A, 2017 Predicting movement of stock of “Y” using sutte indicator \textit{Cogent Econ. Financ.} 5, 1.
[2] Rahman A and Ahmar A S, 2017 Forecasting of primary energy consumption data in the United States: A comparison between ARIMA and Holter-Winters models in \textit{AIP Conference Proceedings} 1885.
[3] Ahmar A S Rahman A and Mulbar U, 2018 α-Sutte Indicator: a new method for time series forecasting IOP Conf. Ser. Mater. Sci. Eng. 366, 1 p. 012018.

[4] Ahmar A S, 2018 A Comparison of α-Sutte Indicator and ARIMA Methods in Renewable Energy Forecasting in Indonesia Int. J. Eng. Technol. 7, 1.6 p. 20–22.

[5] Badan Pusat Statistik, 2015, Indeks Harga Konsumen dan Inflasi Bulanan Indonesia, 2005-2016. [Online]. Available: https://www.bps.go.id/linkTabelStatis/view/id/907. [Accessed: 01-Dec-2016].

[6] BPS Kebumen, 2015 Indeks Harga Konsumen dan Inflasi Kota Kebumen 2014 Kebumen, Indonesia: Badan Pusat Statistik Kabupaten Kebumen.

[7] Wei W W S, 2006 Time Series Analysis : Univariate and Multivariate Methods 2nd ed. New York, NY: Pearson Addison Wesley.

[8] Box G E P and Cox D R, 1964 An analysis of transformations J. R. Stat. Soc. Ser. B p. 211–252.

[9] Ahmar A S et al., 2018 Modeling Data Containing Outliers using ARIMA Additive Outlier (ARIMA-AO) J. Phys. Conf. Ser. 954.

[10] Box G E P and Jenkins G M, 1976 Time series analysis, control, and forecasting San Fr. CA Holden Day 3226, 3228 p. 10.

[11] Yule G U, 1926 Why do we sometimes get nonsense-correlations between Time-Series?--a study in sampling and the nature of time-series J. R. Stat. Soc. 89, 1 p. 1–63.

[12] Walker G, 1931 On periodicity in series of related terms Proc. R. Soc. London. Ser. A, Contain. Pap. a Math. Phys. Character 131, 818 p. 518–532.

[13] Slutsky E, 1937 The summation of random causes as the source of cyclic processes Econom. J. Econom. Soc. p. 105–146.

[14] Wold H, 1938 A Study in the Analisys of Stationery: Time Series Almqvist & Wiksells Boktryckeri.

[15] Makridakis S Wheelwright S C and Hyndman R J, 2008 Forecasting methods and applications John Wiley & Sons.

[16] Hyndman R J and Khandakar Y, 2008 Automatic time series forecasting : the forecast package for R J. Stat. Softw. 27, 3 p. 1–22.