EJMO

DOI: 10.14744/ejmo.2020.28273
EJMO 2020;4(2):160–165

Research Article

Modeling and Forecasting for the number of cases of the COVID-19 pandemic with the Curve Estimation Models, the Box-Jenkins and Exponential Smoothing Methods

Harun Yonar,[1] Aynur Yonar,[2] Mustafa Agah Tekindal,[1] Melike Tekindal[3]

[1]Deparment of Biostatistics, Selçuk University, Faculty of Veterinary Medicine, Konya, Turkey
[2]Deparment of Statistics, Selçuk University, Faculty of Science, Konya, Turkey
[3]Department of Social Work, Izmir Katip Çelebi University, Faculty of Health Sciences, Izmir, Turkey

Abstract

Objectives: This study aims to provide statistical information summarizing the general structure about the effects and process of infection in all countries of the world in the light of the data obtained and to model the daily change of infection criteria.

Methods: The number of COVID 19 epidemic cases of Turkey and the selected G8 countries, Germany, United Kingdom, France, Italy, Russia, Canada, Japan between 1/22/2020 and 3/22/2020 has been estimated and forecasted in this study by using some curve estimation models, Box-Jenkins (ARIMA) and Brown/Holt linear exponential smoothing methods.

Results: Japan (Holt Model), Germany (ARIMA (1,4.0)) and France (ARIMA (0,1,3)) provide statistically significant but clinically unqualified results in this data set. UK (Holt Model), Canada (Holt Model), Italy (Holt Model), Turkey (ARIMA (1,4,0)) and Russia?? the results are more reliable. This is specified for the particular model used in this case Turkey.

Conclusion: Certainly, more accurate evaluations can be made with more data in future studies. Nevertheless, since this study provides information about the levels at which the number of cases may extend in case that the current situation is not intervened, it can guide countries to take the necessary measures and to intervene it earlier.

Keywords: Box-Jenkins, COVID-19 SARS-CoV2, exponential smoothing methods

Cite This Article: *Yonar H, Yonar A, Tekindal MA, Tekindal M. Modeling and Forecasting for the number of cases of the COVID-19 pandemic with the Curve Estimation Models, the Box-Jenkins and Exponential Smoothing Methods. EJMO 2020;4(2):160–165.*

COVID-19 is a novel coronavirus that has resulted in an outbreak of viral pneumonia around the world. Although the virus was first seen in Wuhan (China), it spread to the entire world in a short time because of being contagious. The virus can cause the death of people of all ages, particularly those with chronic illnesses or the elderly.[1] As the COVID-19 coronavirus spreads worldwide, it reveals not only health risks, but also challenges economies of the countries. The COVID-19 pandemic has affected communities and economies around the world in an unprecedented way.

The countries have been late in taking a series of measures to stop the epidemic, and their current healthcare capacities are insufficient to treat patients. Although countries had followed different strategies to prevent the epidem-

Address for correspondence: Harun Yonar, MD. Deparment of Biostatistics, Selçuk University, Faculty of Veterinary Medicine, Konya, Turkey
Phone: +90 545 311 20 80 **E-mail:** hyonar@selcuk.edu.tr
Submitted Date: March 18, 2020 **Accepted Date:** April 13, 2020 **Available Online Date:** April 15, 2020

ic, they have started to take various measures to prevent spreading as soon as possible by following each other's strategies.

This study will guide how the capacity of the existing healthcare services (personnel, equipment, etc.) should be increased in the coming days in the face of the expected number of cases. Considering the expected numbers, painful efforts to increase health service capacities are very important. The rapid spread of the epidemic reveals the necessity to do needs to be done immediately and by taking the right steps.

This paper is designed to give societies, as well as governments, an idea of how fast this pandemic is progressing and to inform them about necessary precautions. For this purpose, the number of COVID-19 epidemic cases of Turkey and of the selected G8 countries, Germany, the United Kingdom, France, Italy, Russia, Canada, Japan between 1/22/2020 and 3/22/2020 has been estimated and forecasted by using some curve estimation models, Box-Jenkins (ARIMA) and Brown/Holt linear exponential smoothing methods in this study. The starting date of the epidemic varies according to the countries, and therefore models are evaluated and established separately for each country. The estimates show how the course of the epidemic will be in the coming days, taking into account the increase at the rates of existing cases.

The rest of the paper is organized as follows. In Part 2, the data are introduced and some parametric and curve estimation models used in this study, the Box-Jenkins and Brown/Holt linear exponential smoothing methods which are the linear exponential smoothing methods are explained. Application results are given in Part 3. Finally, the results are given in Part 4.

The main motivation of the study is to model the COVID-19 virus, which increases geometrically in a short time according to the health and social policies of the countries. As a result of these modelings, different time-dependent policies can be developed by obtaining estimated figures from the same or similarly increased virus. A short cross-sectional study (22/1/2020-22/3/2020), the G-8 countries (except America) and aimed to model the example of Turkey with a variety of statistical methods. The modeling will be guiding both in health and social aspects. Of course, there will be slight differences in the modeling after the specified episode time. However, if the study with the specified models has the opportunity to be published early, it will be a very serious guide for policymakers.

Methods

Data Set: The data sets in this study include the number of positive cases at the COVID-19 outbreak between 22/1/2020 and 22/3/2020 in the selected G-8 countries: Germany, the United Kingdom, France, Italy, Russia, Canada, Japan, and Turkey.[1]

In this study, the data was modeled with some curve estimation models to estimate the number of positive COVID-19 cases. Then, using the linear exponential smoothing methods the Box-Jenkins and Brown and Holt, COVID-19 positive conditions were forecasted. The analyses are conducted via IBM Corp. Released 2017. IBM SPSS Statistics for Windows, Version 25.0. Armonk, NY: IBM Corp and RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

Some curve estimation models:[2-4]

Linear: $\hat{y} = b_0 + b_1 x$ (1)

Logarithmic: $\hat{y} = b_0 + (b_1 Ln\ (x))$ (2)

Inverse: $\hat{y} = b_0 + (\frac{b_1}{x})$ (3)

Quadratic: $\hat{y} = b_0 + b_1 x + b_{11} x^2$ (4)

Cubic: $\hat{y} = b_0 + b_1 x + b_{11} x^2 + b_{111} x^3$ (5)

Time-series: It is a series of data derived from the observations made in periodic time intervals. This series enables us to develop a proper model and make prospective estimations using statistical methods.[5] However, stationary series are required to estimate the prospective values of any series using the past values. Since the non-stationary series contain up-and-down values showing high levels of variance, the margin of error in the possible estimates is quite high.[6] Stationarity may be defined as "a probabilistic process whose average and variance do not vary over time and covariance between two periods is based on distance only between two periods, not the period for which this covariance is calculated"[7, 8] The most common methods used to search for stationarity are ACF (Autoregressive Correlation Function) and PACF (Partial Autoregressive Correlation Function) graphs and Augmented Dickey-Fuller (ADF) unit root test.[9]

Box-Jenkins Method (ARIMA): The Box-Jenkins method suggested by Box and Jenkins[10] is widely used for time series analysis. This method includes non-stationary ARIMA models applied to series, but made stationary by the operation of the difference of the series. The basis of the Box-Jenkins method is to select an ARIMA model that includes the most suitable but limited parameter from a variety of model options, depending on the nature of the data being considered.

ARIMA *(p, d, q)* models are obtained by taking the difference of the series from d degree and adding them to the ARMA *(p, q)* model for the stabilization process. In the ARIMA *(p, d, q)* models, p is the degree of the Autoregressive (AR) model, q is the degree of the moving average (MA) model and d indicates how many differences are required to make the series stationary. ARIMA model becomes AR *(p)*, MA *(q)*, or ARMA *(p, q)* if the time series is stationary[11]

ARMA *(p, q)* model is shown as follows[10]

$$Y_t = \sum_{i=1}^{p} \alpha_i Y_{t-i} + \sum_{i=1}^{q} \beta_i \varepsilon_{t-i} + +\varepsilon_t \quad (6)$$

First, the difference of the non-stationary Y_t time-series is obtained by equation (2).

$$\nabla^1 Y_t = Y_t - Y_{t-1} = Y_t' \quad (7)$$

If Y_t' series is not still stationary, difference taking process is repeated for the d times until being stationary. The general form for the difference taking process is given as follows:

$$\nabla^d Y_t = \nabla^{d-1} Y_t - \nabla^{d-1} Y_{t-1} \quad (8)$$

The expression of ARIMA (p, d, q) model can be defined as follows:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \alpha_1 - \theta_1 \alpha_{t-1} - \alpha_2 - \theta_2 \alpha_{t-2} - \cdots - \alpha_q - \theta_q \alpha_{t-q} \quad (9)$$

Here: ϕ_p are the parameter values for autoregressive operator, α_q are the error term coefficient, Θ_q are the parameter values for moving average operator, Y_t is the time series of the original series differenced at the degree *d*.[7,12]

Linear Exponential Smoothing Methods: Exponential smoothing was introduced in the late 1950s.[13-15] Forecasts produced using exponential smoothing methods weighted averages of past observations. These methods give decreasing weights to past observations and thus the more recent the observation the higher the associated weight. This framework enables reliable estimates to be produced quickly in most applications. In this study, Brown and Holt linear exponential smoothing methods which are the most widely used in the literature are utilized.

Holt Linear Exponential Smoothing Method: This model is appropriate for a series with a linear trend and no seasonality. Its relevant smoothing parameters are level and trend, and, in this model, they are not constrained by each other's values. Holt's exponential smoothing is most similar to an ARIMA with zero degree of autoregression, two degrees of differencing, and two degrees of moving average.

In this method, estimates are made using the equations below.

$$Y_t' = \alpha Y_t + (1-\alpha)(Y_{t-1}' + B_{t-1}) \quad (10)$$

$$b_t = \gamma(Y_t' - Y_{t-1}') + (1-\gamma)B_{t-1} \quad (11)$$

$$\hat{Y}_{t+m} = Y_t' - B_t m \quad (12)$$

where α and γ are the smoothing constants in the range of [0,1].

Brown Linear Exponential Smoothing Method: This model is a special case of Holt linear exponential smoothing method. In this model, they are assumed that level and trend which are the smoothing parameters are equal.

In this method, estimates are made using the equations below.

$$Y_t' = \alpha Y_t + (1-\alpha)(Y_{t-1}') \quad (13)$$

$$Y_t'' = \alpha Y_t' + (1-\alpha)(Y_{t-1}'') \quad (14)$$

$$a_t = Y_t' + (Y_t' - Y_t'') = 2Y_t' - Y_t'' \quad (15)$$

$$b_t = \frac{\alpha}{1-\alpha}(Y_t' - Y_t'') \quad (16)$$

$$\hat{Y}_{t+m} = a_t + b_t m \quad (17)$$

where α is the smoothing constant in the range of [0,1].

Results

Some parametric and non-parametric models have been used to model the number of cases suffering from the COVID-19 epidemic depending on the days in countries. Among these models, the model with the highest R2 value was determined as cubic and the results are given in Table 1. Also, curve estimation graphs are given in Figure 1 to determine which model fits the data better. It is also observed from these graphs that the cubic model is the best for all countries.

The stationarity of the residuals is examined and the ACF and PACF graphics of the series for countries are given in Figure 2. When the graphs are analyzed, there are only a few values that exceed the confidence limit, so the series can be evaluated as stationary.

Table 2 shows the goodness of fit criteria values of the Box-Jenkins and exponential smoothing models. Generally, the models have high R2 values except for Japan. Furthermore, these models can be used since the MAPE values are less than 10%.

The fitting of the models and the forecast values for the number of COVID-19 cases can be seen in Figure 3.

As can be seen from Figure 1 and Figure 3, Japan, Germany, and France provide statistically significant but not clinically qualified results in this data set. The results are more reliable for the UK, Canada, Italy, and Turkey. This is specified for the particular model used in this case Turkey.

Conclusion

Different results have been observed for the countries due to their different epidemic exposure dates and social,

Table 1. Summary models of the regression models for the countries

Country	Methods	Summary Model					Estimation of parameters				
		R^2, %	F	df1	df2	p	Constant	β1	β2	β3	
Germany	Cubic	0.853	110.548	3	57	0.000	-373.559	97.687	-5.378	0.078	$\hat{y}$ = -373.56 - 97.69 x - $5.38x^2+0.078x^3$
Italy	Cubic	0.914	202.894	3	57	0.000	-356.594	102.404	-6.335	0.106	$\hat{y}$ = -356.594 - 102.404 x - $6.34x^2+0.106x^3$
Japan	Cubic	0.521	20.705	3	57	0.000	4.796	-0.798	0.047	0.002	$\hat{y}$ = 4.796 - 0.798 x + $0.047x^2+0.002x^3$
Canada	Cubic	0.789	71.176	3	57	0.000	-30.389	7.670	-0.404	0.006	$\hat{y}$ = -30.389 - 7.670 x - $0.404x^2+0.006x^3$
Russia	Cubic	0.887	149.052	3	57	0.000	-7.839	1.982	-0.104	0.001	$\hat{y}$ = -7.839 + 1.982 x - $0.104x^2+0.001x^3$
UK	Cubic	0.707	45.946	3	57	0.000	-113.254	28.480	-1.509	0.021	$\hat{y}$ = -113.254 + 28.480 x - $1.509x^2+0.021x^3$
Turkey	Cubic	0.836	33.769	3	58	0.000	-61.393	14.429	-0.708	0.009	$\hat{y}$ = -61.393 + 14.429 x - $0.708x^2+0.009x^3$
France	Cubic	0.825	89.422	3	57	0.000	-159.081	43.741	-2.543	0.039	$\hat{y}$ = -159.081 + 43.741 x - $2.543x^2+0.039x^3$

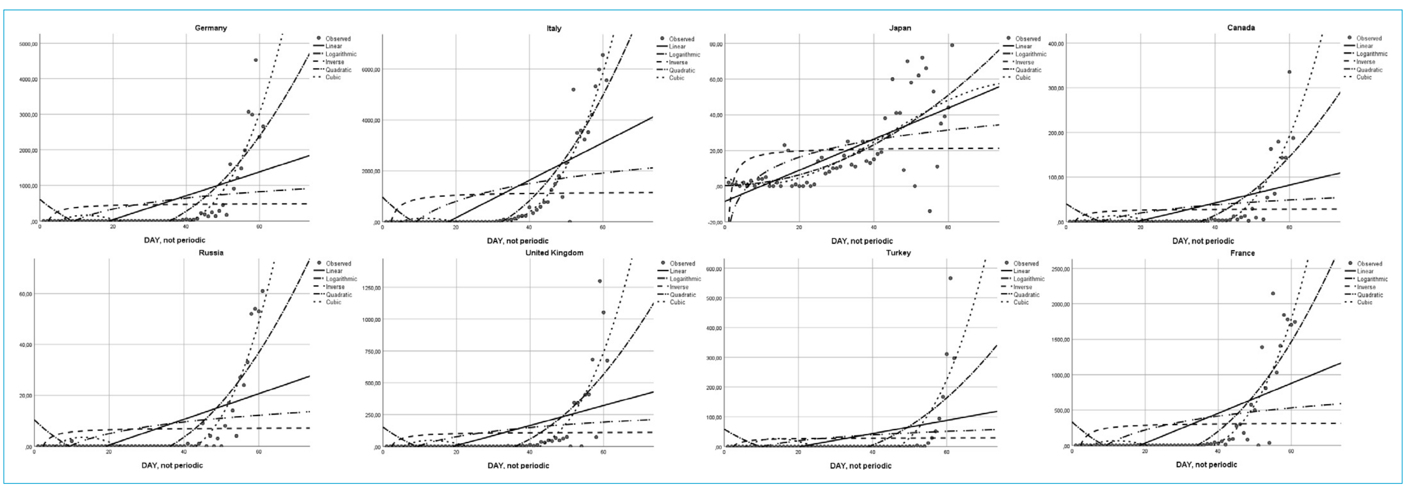

Figure 1. Curve estimates for the countries.

cultural and technological developments such as health policies, preliminary measures, average age and economic levels. Countries caught late in the epidemic can trace the natural history of the infection spread cases previously seen in other countries, and thus they are more successful in combating the epidemic taking various measures.

In this study, the model established using the COVID-19 pandemic case numbers of the countries provides information about the estimated number of cases that may occur in the future. days. The measures taken by countries such as the individual attitudes of the societies towards the specified measures and the number of virus tests to be performed are factors that may affect the number of cases. Since this study was conducted with the current measures, the forecasts obtained may differ from the number of cases that occur in the future. The more precautions are taken, the fewer the number of cases.

Discussion

In future studies with more data and healthier evaluations can be made as a matter of course. However, since this study provides information about the levels that the number of cases can reach if the course of the current situation cannot be intervened, it can guide countries to take the necessary measures and to intervene early.

Disclosures

Peer-review: Externally peer-reviewed.

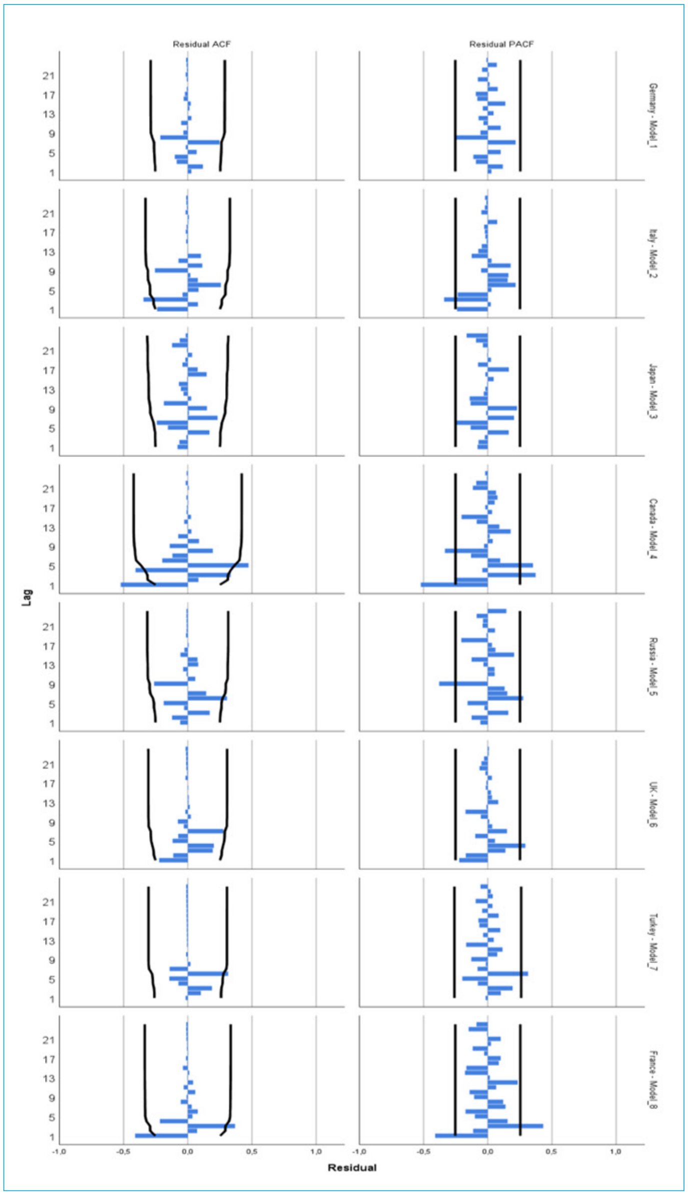

Figure 2. The graphs of ACF and PACF of residuals.

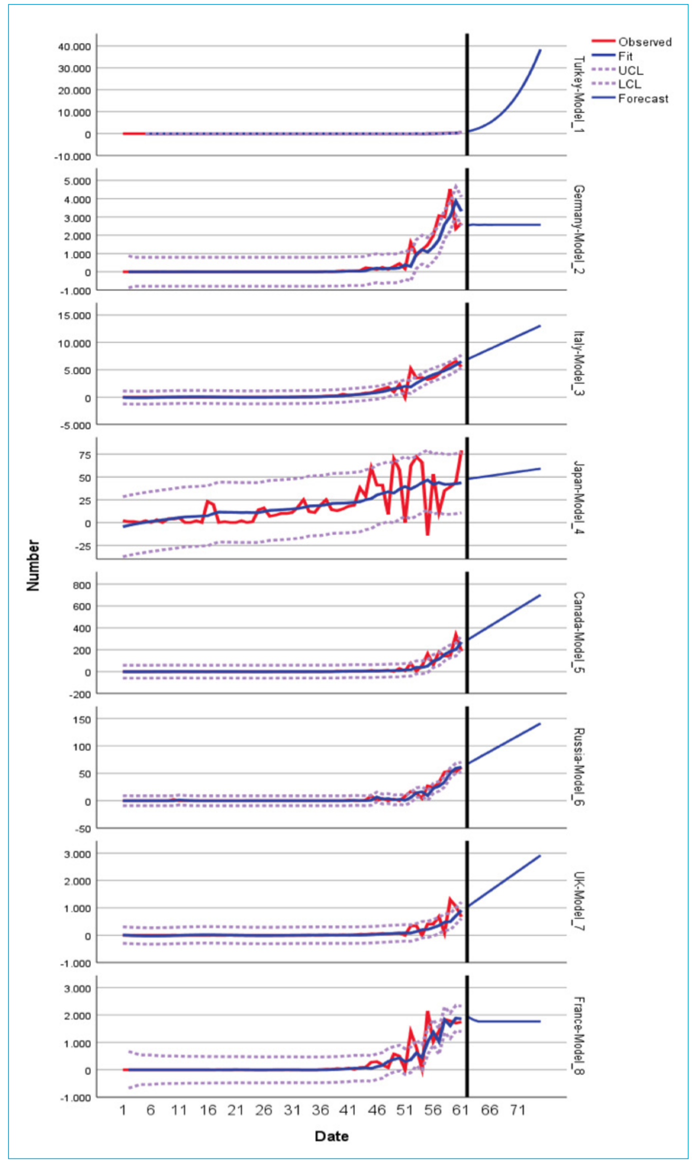

Figure 3. The fitting the models and forecast graphs of the number of positive COVID-19 cases.

Table 2. The goodness of fit criteria of the Box-Jenkins and exponential smoothing models

		Model Fit Statistics								Ljung-Box Q (18)		
	Model Type	Stationary R-squared	R-squared	RMSE	MAPE	MAE	MaxAPE	MaxAE	Normalized BIC	Statistics	DF	p
Turkey	ARIMA(1,4,0)	0.205	0.995	6.171	2.578	2.176	1692.189	28.823	3.711	12.477	17	0.045
Germany	ARIMA(1,1,0)	0.188	0.826	394.416	6.389	147.556	653.246	1501.333	12.023	10.541	17	0.049
Italy	Holt	0.843	0.892	589.553	7.878	254.685	2719.729	3334.104	12.894	24.690	16	0.075
Japan	Holt	0.821	0.462	16.403	3.682	10.864	1032.755	60.545	5.730	19.645	16	0.037
Canada	Holt	0.841	0.777	29.212	8.001	11.152	692.214	133.395	6.884	60.943	16	0.000
Russia	Brown	0.646	0.908	4.474	6.595	1.908	308.867	17.851	3.064	20.297	17	0.032
United Kingdom	Holt	0.825	0.650	149.566	6.979	60.291	20263.194	805.081	10.150	16.715	16	0.040
France	ARIMA (0,1,3)	0.667	0.837	232.490	1.301	90.070	1148.194	1126.799	11.034	24.276	16	0.038

Conflict of Interest: None declared.

Authorship Contributions: Concept – M.A.T., H.Y.; Design – A.Y.; Supervision – M.A.T.; Data collection &/or processing – H.Y., A.Y.; Analysis and/or interpretation – M.A.T.; Literature search – H.Y., A.Y., M.T.; Writing – H.Y., A.Y., M.T.; Critical review – M.A.T.

References

1. WHO. 2020 [World Health Organization]. Available from: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/events-as-they-happen.
2. Farebrother R. Further results on the mean square error of ridge regression. Journal of the Royal Statistical Society Series B (Methodological). 1976;38(3):248-50.
3. Rao CR, Toutenburg H. Linear models. Linear models: Springer; 1995. p. 3-18.
4. Robinson PM. Root-N-consistent semiparametric regression. Econometrica: Journal of the Econometric Society. 1988:931-54.
5. Tekindal MA, Yazici AC, Yavuz Y. The modelling of time-series and the evaluation of forecasts for the future: the case of the number of persons per physician in turkey between 1928 and 2010. Biomedical Research. 2016;27(3).
6. Fischer B. Decomposition of time series: comparing different methods in theory and practice: Eurostat; 1995.
7. Gujarati DN, Porter DC. Basic econometrics (ed.). New York: McGraw-Hill. 2003.
8. Yenice S, Tekindal MA. Forecasting the stock indexes of fragile five countries through Box-Jenkins methods. International Journal of Business and Social Science. 2015;6(8):180-91.
9. Dickey DA, Fuller WA. Likelihood ratio statistics for autoregressive time series with a unit root. Econometrica: journal of the Econometric Society. 1981:1057-72.
10. Box GE, Jenkins GM, Reinsel GC, Ljung GM. Time series analysis: forecasting and control: John Wiley & Sons; 1976.
11. Wickramarachchi A, Herath H, Jayasinghe-Mudalige U, Edirisinghe J, Udugama J, Lokuge L, et al. An Analysis of price behavior of major poultry products in Sri Lanka. Journal of Agricultural Sciences–Sri Lanka. 2017;12(2).
12. Brockwell PJ, Davis RA. Introduction to time series and forecasting: springer; 2016.
13. Brown RG. Exponential smoothing for predicting demand. cambridge, mass., arthur d. little. Inc; 1956.
14. Holt CC. Forecasting trends and seasonals by exponentially weighted averages. carnegie institute of technology. Pittsburgh ONR memorandum; 1957.
15. Winters PR. Forecasting sales by exponentially weighted moving averages. Management science. 1960;6(3):324-42.