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, Aynur Yonar, Mustafa Agah Tekindal, Melike Tekindal

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

Abstract

Objectives: The aim of this study is to present 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 the selected countries of G8 countries, Germany, United Kingdom, France, Italy, Russian, Canada, Japan, and Turkey 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.

Results: Japan (Holt Model), Germany (ARIMA (1,4,0)) and France (ARIMA (0,1,3)) provide statistically significant but not clinically qualified results in this data set. UK (Holt Model), Canada (Holt Model), Italy (Holt Model) and Turkey (ARIMA (1,4,0)) and in the results are more reliable. Specified for the particular model used in this case Turkey.

Conclusion: In future studies, 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.

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

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 older people. While the COVID-19 coronavirus is spreading worldwide, besides health risks, it is also forcing economies of the nations. COVID-19 pandemic has affected unprecedentedly communities and economies everywhere around the world. The countries are late in taking a series of measures to stop the epidemic, and their current healthcare capacities are insufficient to treat patients. Although countries follow different strategies to prevent the epidemic, they have started...
to take various measures to prevent spreading as soon as possible by following in each other’s strategies. This study will be a guide in what direction the existing healthcare services (personnel, equipment, etc.) should increase their capacity in the coming days in the face of the expected number of cases. Taking into consideration the expected numbers, it is very important in terms of giving pain to the studies that will increase their health service capacities. The epidemic’s rapid spreading reveals the necessity of doing what should be done immediately and by taking the right steps.

This paper is designed to give communities and also the government a sense of how fast this pandemic is progressing and to inform them of necessary precautions. For this purpose, the number of the COVID-19 epidemic cases of the selected countries of G8 countries, Germany, United Kingdom, France, Italy, Russian, Canada, Japan, and Turkey 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 start date of the epidemic varies in countries, and thus the models are evaluated and installed separately for each country. The estimates show how the course of the epidemic will be in the following days, taking into account increase rates of the current cases.

The rest of the paper is organized as follows. In Section 2, the data is introduced and some parametric and curve estimation models used in this study, the Box-Jenkins and Brown/Holt linear exponential smoothing methods are explained. In Section 3, application results are given. Finally, the conclusions are given in Section 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 modellings, different time-dependent policies can be developed by obtaining estimated figures in 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 country with a variety of statistical methods. The modeling will be guiding both in health and social terms. Of course, there will be slight differences in the modeling after the specified section time. However, if the study with the specified models finds the opportunity to be published early, it will be a very serious guide for policymakers.

**Methods**

**Data Set:** The data in this study sets involve the number of positive COVID-19 pandemic cases belonging to the between 1/22/2020 and 3/22/2020 in selected G-8 countries: Germany, United Kingdom, France, Italy, Russian, Canada, Japan, and Turkey. In this study, the data is modeled via some curve estimation models to estimate the number of positive COVID-19 cases. Then, the forecasts of the COVID-19 positive cases are made by using the Box-Jenkins and Brown and Holt linear exponential smoothing methods which are the linear exponential smoothing methods. The analyses are conducted by 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:**

1. Linear: $\hat{y} = b_0 + b_1x$
2. Logarithmic: $\hat{y} = b_0 + (b_1 \ln(x))$
3. Inverse: $\hat{y} = b_0 + (\frac{b_1}{x})$
4. Quadratic: $\hat{y} = b_0 + b_1x + b_{11}x^2$
5. Cubic: $\hat{y} = b_0 + b_1x + b_{11}x^2 + b_{111}x^3$

**Time series:** It is a series derived from the observations made at periodical time intervals. This series enables to improve a proper model and to make prospective estimations by using statistical methods. However, stationary series are required for estimating the values which they will take prospectively by using the past values for any series. Since non-stationary series contain up-and-down values exhibiting variance at high level, margin of error in the possible estimates is quite high. 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 period for which this covariance is calculated" methods are used for searching the stationarity. Those that are most common among these methods are ACF (Autoregressive Correlation Function) and PACF (Partial Autoregressive Correlation Function) graphics and Augmented Dickey Fuller (ADF) unit root test.

**Box-Jenkins Method (ARIMA):** Box-Jenkins method proposed by Box, Jenkins is widely used for time series analysis. This method includes ARIMA models applied to the series that are non-stationary but are made stationary with the operation of difference of the series. The base of
the Box-Jenkins method is to choose an ARIMA model that includes the most suitable but limited parameter among the various model options, depending on the nature of the considering data.

ARIMA \((p, d, q)\) models are obtained by taking the difference of series from \(d\) degree and adding to ARIMA \((p, q)\) model for the stabilizing 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 \(d\) stands 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\(^{[11]}\)

ARIMA \((p, q)\) model is shown as follows\(^{[10]}\)

\[
Y_t = \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{i=1}^{q} \Theta_i \varepsilon_{t-i} + \epsilon_t \tag{6}
\]

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

\[
\nabla Y_t = Y_t - Y_{t-1} = Y_t' \tag{7}
\]

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:

\[
\nabla^d Y_t = \nabla^{d-1} Y_{t-1} \tag{8}
\]

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

\[
Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \epsilon_1 - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \cdots - \theta_q \epsilon_{t-q} \tag{9}
\]

Here: \(\phi_p\) are the parameter values for autoregressive operator, \(\theta_q\) are the error term coefficient, \(\Theta_q\) 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.

\[
Y_t' = a Y_t + (1 - a)(Y_t'_{-1} + B_{t-1}) \tag{10}
\]

\[
b_t = \gamma (Y_t' - Y_{t-1}') + (1 - \gamma)B_{t-1} \tag{11}
\]

\[
Y_{t+m}' = Y_t' - B_t m \tag{12}
\]

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

where \(a\) and \(\gamma\) 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.

\[
Y_t' = a Y_t + (1 - a)(Y_t'_{-1}) \tag{13}
\]

\[
Y_{t''}' = a Y_t' + (1 - a)(Y_{t-1}''') \tag{14}
\]

\[
a_t = \gamma (Y_t' - Y_{t-1}'') + 2Y_t' - Y_{t''} \tag{15}
\]

\[
b_t = \frac{\alpha}{1-\alpha} (Y_t' - Y_{t''}) \tag{16}
\]

\[
Y_{t+m}' = a t + b t m \tag{17}
\]

where \(\alpha\) 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 the countries. Among these models, the model with the highest \(R^2\) value is 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 examined, there are only a few values that exceed the confidence limit, thus 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 \(R^2\) values except for Japan. Furthermore, these models can be used because the MAPE values are less than 10%.

The fitting of the models and the forecast values for the number of the 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. UK, Canada, Italy and Tur-
key and in the results are more reliable. Specified for the particular model used in this case Turkey.

**Conclusion**

Differences results for the countries have been observed since they have different epidemic exposure dates and social, cultural and technological developments such as health policies, preliminary measures, average ages and economic levels. Countries caught late in the epidemic can monitor the natural history of the country previously seen cases of spread of infection and thereby taking various measures are more successful in combating the epidemic. In this study, the models, which are established by using the number of COVID-19 pandemic cases of the countries, provide information about the estimated number of cases that may be for 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, 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

---

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

<table>
<thead>
<tr>
<th>Countries</th>
<th>Methods</th>
<th>Summary Model</th>
<th>Estimation of parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>R², %</td>
<td>F</td>
<td>df1 df2 p</td>
<td>Constant β1 β2 β3</td>
</tr>
<tr>
<td>Germany</td>
<td>Cubic</td>
<td>0.853 110.548 3 57 0.000</td>
<td>-373.559 97.687 -5.378 0.078</td>
</tr>
<tr>
<td>Italy</td>
<td>Cubic</td>
<td>0.914 202.894 3 57 0.000</td>
<td>-356.594 102.404 -6.335 0.106</td>
</tr>
<tr>
<td>Japan</td>
<td>Cubic</td>
<td>0.521 20.705 3 57 0.000</td>
<td>4.796 0.798 0.047 0.002</td>
</tr>
<tr>
<td>Canada</td>
<td>Cubic</td>
<td>0.789 71.176 3 57 0.000</td>
<td>-30.389 7.670 -0.404 0.006</td>
</tr>
<tr>
<td>Russia</td>
<td>Cubic</td>
<td>0.887 149.052 3 57 0.000</td>
<td>-7.839 1.982 -0.104 0.001</td>
</tr>
<tr>
<td>UK</td>
<td>Cubic</td>
<td>0.707 45.946 3 57 0.000</td>
<td>-113.254 28.480 -1.509 0.021</td>
</tr>
<tr>
<td>Turkey</td>
<td>Cubic</td>
<td>0.836 33.769 3 58 0.000</td>
<td>-61.393 14.429 -0.708 0.009</td>
</tr>
<tr>
<td>France</td>
<td>Cubic</td>
<td>0.825 89.422 3 57 0.000</td>
<td>-159.081 43.741 -2.543 0.039</td>
</tr>
</tbody>
</table>

**Figure 1. Curve estimates for the countries.**
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

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

Conflict of Interest: None declared.

Supervision – M.A.T.; Data collection &/or processing – H.Y., A.Y.;
Analysis and/or interpretation – M.A.T.; Literature search – H.Y.,

References

1. WHO. 2020 [World Health Organization ]. Available from:
   https://www.who.int/emergencies/diseases/novel-coronavi-
   rus-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
3. Rao CR, Toutenburg H. Linear models. Linear models: Spring-
   er; 1995. p. 3-18.
   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
6. Fischer B. Decomposition of time series: comparing different
7. Gujarati DN, Porter DC. Basic econometrics (ed.). New York:
8. Yenice S, Tekindal MA. Forecasting the stock indexes of frag-
   ile five countries through Box-Jenkins methods. International
9. Dickey DA, Fuller WA. Likelihood ratio statistics for autoregres-
   sive time series with a unit root. Econometrica: journal of the
10. Box GE, Jenkins GM, Reinsel GC, Ljung GM. Time series analy-
11. Wickramarachchi A, Herath H, Jayasinghe-Mudalige U, Ed-
    behavior of major poultry products in Sri Lanka. Journal of
12. Brockwell PJ, Davis RA. Introduction to time series and fore-
    casting: springer; 2016.
    cambridge, mass., arthur d. little. Inc; 1956.
14. Holt CC. Forecasting trends and seasonals by exponentially
    weighted averages. carnegie institute of technology. Pitts-
    burgh ONR memorandum; 1957.
15. Winters PR. Forecasting sales by exponentially weighted mov-