Forecasting of COVID-19 Confirmed Cases in Different Countries with ARIMA Models

Tania Dehesh[1], H.A.Mardani-Fard[2], Paria Dehesh[3*]

[1] Department of Biostatistics and Epidemiology, Kerman University of Medical Sciences, Kerman, Iran

[2] Department of Mathematics and Statistics, Yasuj University, Yasuj, Iran

3 Department of Epidemiology, School of Public Health, Iran University of Medical Sciences, Tehran, Iran

Email address: Tania_dehesh@yahoo.com,

Email address: h_mardanifard@yahoo.com

Email address:Paria_dehesh@yahoo.com

Forecasting of COVID-19 Confirmed Cases in Different Countries with ARIMA Models

Abstract

Background: The epidemic of a novel coronavirus illness (COVID-19) becomes as a global threat. The aim of this study is first to find the best prediction models for daily confirmed cases in countries with high number of confirmed cases in the world and second to predict confirmed cases with these models in order to have more readiness in healthcare systems.

Methods*:* This study was conducted based on daily confirmed cases of COVID-19 that were collected from the official website of Johns Hopkins University from January 22th, 2020 to March 1th, 2020. Auto Regressive Integrated Moving Average (ARIMA) model was used to predict the trend of confirmed cases. Stata version 12 were used.

Results: Parameters used for ARIMA were (2,1,0) for Mainland China, ARIMA (2,2,2) for Italy , ARIMA(1,0,0) for South Korea, ARIMA (2,3,0) for Iran, and ARIMA(3,1,0) for Thailand. Mainland China and Thailand had almost a stable trend. The trend of South Korea was decreasing and will become stable in near future. Iran and Italy had unstable trends.

Conclusions: Mainland China and Thailand were successful in haltering COVID-19 epidemic. Investigating their protocol in this control like quarantine should be in the first line of other countries' program

Keyword: COVID-19; Coronavirus; ARIMA; Trend; Predict; Forecast; Epidemic

Background

A novel Virus belongs to Corona virus's family that has been transmitter from animal to human and was recognized in Wuhan, China, in December 2019. This can cause serious illness and death (1). It has since been identified as a zoonotic coronavirus, similar to severe acute respiratory syndrome coronavirus (SARS-CoV) and Middle East Respiratory Syndrome Coronavirus (MERS-CoV), and was named 2019-nCoV (2). A total number of 4515 cases including 106 deaths were confirmed on 27th of January 2020(3). A local seafood market in Wuhan was visited by many cases in the initial research, and it is indicated that a common-source zoonotic exposure may cause this new illness (4) The prevalence scope of this disease is unclear, because at present the prevalence of this disease is so dynamic(1). There is obvious variation among countries in epidemiological surveillance and detection capacity for suspected cases. (5) Several cases of COVID-19 infections were also reported outside China, in other Asian countries, the United States, France, Australia, and Canada. In this situation when the illness does not have any specific treatment, the prevention of disease and preparation in healthcare services is very important .Modeling and future forecast of daily number of confirmed cases can help the treatment system in providing services for the new patients .The statistical prediction models could be helpful in forecasting and controlling this global epidemic threat. Here in this study, Auto Regressive Integrated Moving Average (ARIMA) model could be useful to predict the confirmed cases of COVID-2019. This model has more ability compared to some prediction models such as wavelet neural network (WNN) and the support vector machine (SVM) in prediction of natural disasters(6). The global geographic regions in this study are according to World Health Organization (WHO) classification of regions. The data of countries with high number of confirmed cases analyzed in this study was based on WHO regions. For each country,

the best ARIMA model is identified, and then 17 future days is predicted. The daily confirmed cases data of COVID-2019 from January 22th, 2020 to March 1th, 2020 were collected from the official website of Johns Hopkins University and were used to build these models. The aim of this study is first to find the best predicting models for confirmed cases in countries with high number of confirmed cases in the world and second to predict confirmed cases with these models. These models can help in predicting patients in near future to have better treatment staff preparedness in these countries.

Methods

Data source

The daily confirmed cases of COVID-2019 from January 22th, 2020 to March 1th, 2020 were collected from the official website of Johns Hopkins University (https://gisanddata.maps.arcgis.com/apps/ opsdashboard/index.html), and Microsoft Excel 2019 was used to build a time-series database

ARIMA models

A total number of 41 (from January 22th, 2020 to March 1th, 2020) days were collected to develop ARIMA model.

The ARIMA models are important techniques in time series analysis that could be used in auto correlated data analysis. These models include autoregressive (AR) model, moving average (MA) model, and seasonal autoregressive integrated moving average (SARIMA) model (7) .

The parameters of ARIMA model is as follows: (p, d, q)(P,D,Q)S generally, p refers to the order of auto-regression, d refers the degree of trend difference, q refers to the order of moving average, P refers to the seasonal auto-regression lag, D refers to the degree of seasonal difference, Q refers to the seasonal moving average, S refers to the length of the cyclical pattern(8).

Before analyzing, time series must become station-based on mean and variance. The Augmented Dickey-Fuller (ADF) is used (9)in recognizing stationary in the mean and Box Cox test in recognizing whether the time series is stationary based on variance or not. Log transformation and differences are remedy approaches to stabilize the time series for variance and mean, respectively (10). Seasonal differences were used to stabilize the series from seasonality trend. The initial number of ARIMA model was guessed through autocorrelation function (ACF) graph and partial autocorrelation (PACF) graph.

All the models that passed the residual test (normality and stationary in variance) were compared using Akaike information criterion (AIC). The model which has the least AIC was selected as the best model. The methodology of current study was based on a previous study as reference (11). Microsoft Excel 2016 was used to build the database of daily COVID-19 in the world and STATA version 12 software was adopted to develop the ARIMA model. The statistical significance level was set at 0.05.

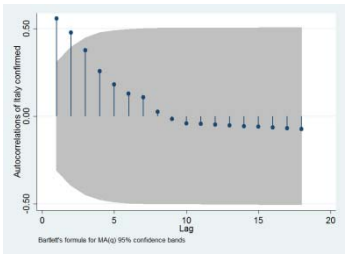
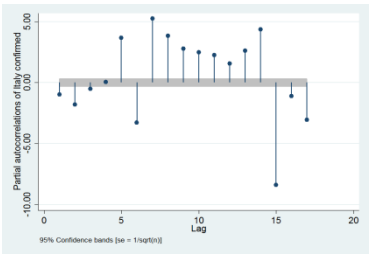
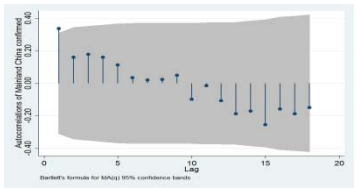
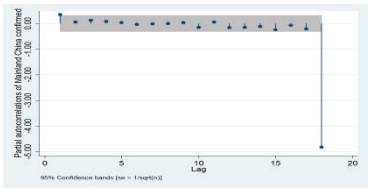
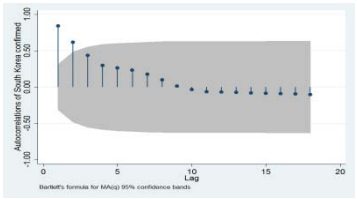
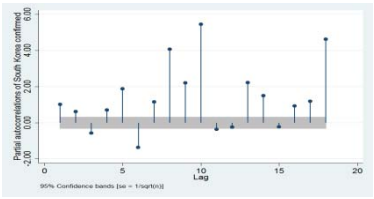
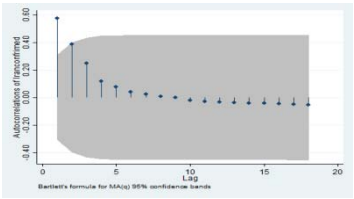
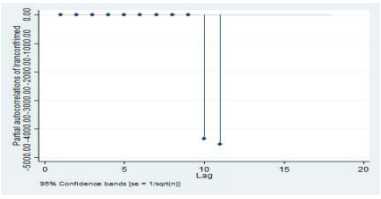
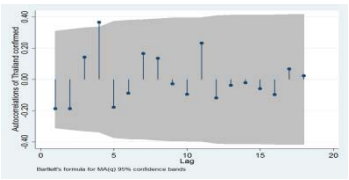
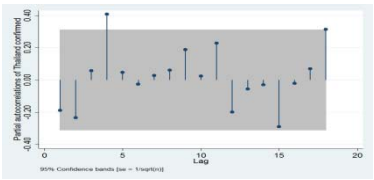
Ethics

Since no primary data collection was undertaken, no patient or public was involved; no formal ethical assessment or informed consent was required. All data were collected from the official website and all data were fully anonymized.

Results

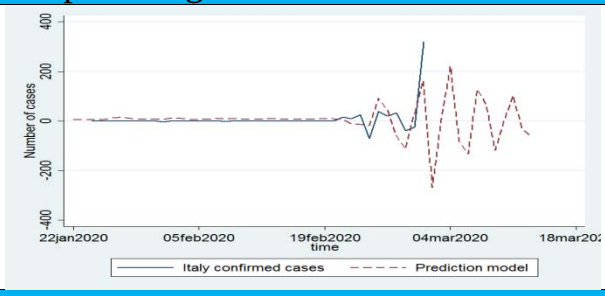
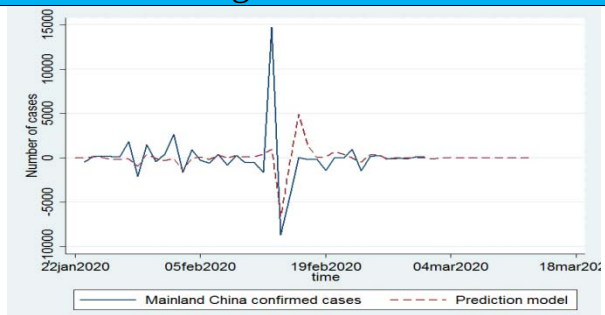
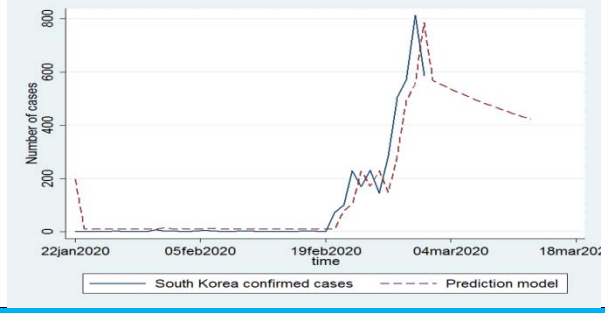
Table 1 shows the ACF and PACF plots used for choosing the model parameters. The reporting of these ACF and PACF showed that confirmed cases of COVID-2019 were not influenced by the seasonality. According to these plots the P and Q parameters of ARIMA models were guessed. Then the guess models were compared according to AIC value.

Table 1 The best ARIMA models for forecasting number of daily confirmed cases according to ACF and PCF plots

Country	Model	ACF	PACF
Italy	ARIMA (2,2,2)		
China	ARIMA (2,1,0)		
South Korea	ARIMA (1,0,0)		
Iran	ARIMA (2,3,0)		
Thailand	ARIMA (3,1,0)		

The final models that were reported in table 2 had the lowest AIC values. Table 2 shows the forecast plots of ARIMA models for different countries with high number of confirmed cases according to WHO regions. The closes of predicted plots with actual confirmed data could be observed in these plots. This shows the precision of models in forecasting.

Table 2 The plot of actual confirmed cases with the prediction result of ARIMA models in countries with high confirmed cases in different world regions (WHO regions)

Country	Model	Plot
Eroupean Region		
Italy	ARIMA (2,2,2)	
Western Pacific Region		
China	ARIMA (2,1,0)	
South Korea	ARIMA (1,0,0)	
Eastern Mediterranean Region		

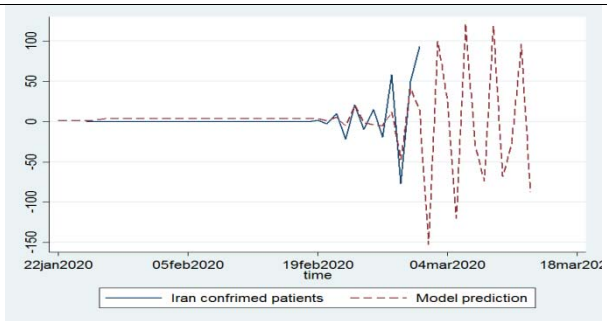
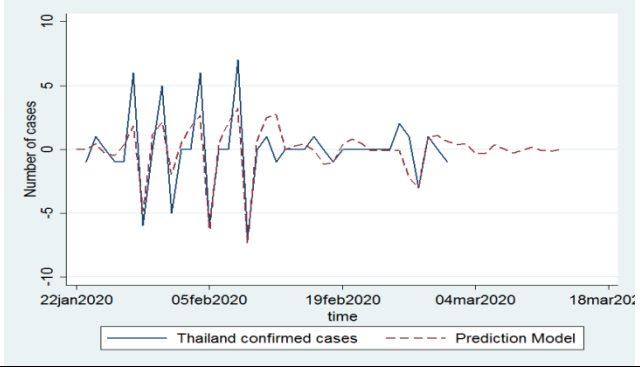
Iran	ARIMA (2,3,0)	
South-East Asia Region		
Thailand	ARIMA (3,1,0)	

Table 3 shows the forecast for 17 days (from 5th of March until 21th of March) with 95% confidence interval for different countries. This table shows that China may have stable trend after 6th of March. South Korea presented a downward trend after 15th of March. This hope is strong that this trend remains stable in near future in this country. Thailand almost controlled the epidemic and had zero or one confirmed cases daily. Iran and Italy did not have a stable trend in these 17 days.

Table 3 Prediction of daily confirmed cases for 17 days according to ARIMA models with 95% CI

Date	European Region	Western Pacific Region		Eastern Mediterranean Region	South-East Asia Region
	Italy	China	South Korea	Iran	Thailand
2020-03-05	198.15 (125.87,270.42)	107.05 (-436.30,860.40)	521.37 (247.41,795.32)	514.54 (458.72,570.35)	0.65 (-4.26,5.55)
2020-03-06	473.36 (362.59,584.14)	120.91 (-446.30,884.67)	506.96 (207.09,806.84)	756.28 (690.46,822.10)	1.34 (-3.60,6.28)
2020-03-07	686.06 (566.99,805.12)	129.14 (-442.85,894.82)	493.20 (171.48,814.92)	605.97 (535.35,676.60)	1.05 (-4.16,6.27)
2020-03-08	383.12 (253.99,512.25)	120.73 (-436.53,889.29)	480.05 (139.60,820.49)	560.92 (490.27,631.57)	0.71 (-4.55,5.98)
2020-03-09	331.45 (191.31,471.59)	122.19 (-447.84,890.90)	467.48 (110.79,824.17)	753.28 (678.04,828.51)	0.95 (-4.33,6.23)
2020-03-010	593.67 (452.30,735.03)	124.16 (-446.53,893.00)	455.47 (84.57,826.37)	566.66 (486.04,647.29)	1.18 (-4.11,6.47)
2020-03-11	532.58 (382.36,682.80)	122.74 (-444.68,891.67)	443.99 (60.58,827.40)	610.34 (529.04,691.64)	0.91 (-4.41,6.23)
2020-03-12	346.92 (196.63,497.21)	122.80 (-446.19,891.73)	433.03 (38.54,827.52)	730.97 (648.42,813.53)	0.85 (-4.51,6.20)
2020-03-13	468.21 (313.02,623.40)	123.22 (-446.13,892.16)	422.55 (181.20,826.89)	547.47 (460.70,634.23)	1.04 (-4.32,6.39)
2020-03-14	569.20 (412.85,725.54)	123.00 (-445.71,891.94)	412.54 (60,825.67)	653.73 (565.02,742.43)	1.05 (-4.31,6.41)
2020-03-15	432.86 (274.76,590.96)	122.97 (-445.94,891.91)	402.97 (181.04,823.97)	697.22 (608.48,785.97)	0.89 (-4.47,6.26)
2020-03-16	405.40 (245.44,565.36)	123.06 (-445.96,891.99)	393.82 (34.24,821.89)	547.41 (456.36,638.47)	0.94 (-4.43,6.31)
2020-03-17	525.05 (364.83,685.28)	123.02 (-445.88,891.96)	385.09 (49.32,819.49)	684.98 (591.41,778.55)	1.03 (-4.34,6.40)
2020-03-18	500.28 (338.40,662.16)	123.01 (-445.91,891.95)	376.74 (63.38,816.86)	660.19 (566.37,754.01)	0.98 (-4.40,6.35)
2020-03-19	414.35 (252.47,576.24)	123.03 (-445.92,891.97)	368.76 (76.51,814.03)	562.48 (467.91,657.05)	0.92 (-4.45,6.30)
2020-03-20	467.69 (304.83,630.56)	123.03 (-445.91,891.96)	361.14 (88.79,811.06)	701.33 (604.59,798.07)	0.98 (-4.40,6.35)
2020-03-21	515.52 (352.45,678.59)	123.02 (-445.91,891.96)	353.85 (100.28,807.99)	626.87 (529.24,724.50)	1.00 (-4.38,6.38)

Discussion

The confirmed cases trend in China may become stable. South Korea will have a stationary trend in near future. Thailand almost controlled and the confirmed cases for this country were frequently between zero and one. Iran and Italy had unstable trend.

China's confirmed cases showed stationary trend after 6th of March. This result is in accordance with the result of previous study that predicted China's epidemic to end up after March 10th, 2020(12). This may be due to sever control and quarantine that were performed in China. This fact shows that quarantine worked well to reduce human exposure and control this epidemic. In another study on Chinese data the trend of confirmed cases in China during 10th–24th of January, 2020 was followed by exponential function (13).

This study confirmed that South Korea had an almost moderate trend that became decreasing after 15th of March.

Iran did not have stationary trend even after 15th of March. In a study according to the passengers that came from other countries such as UAE, Lebanon, and Canada the prevalence of cases in Iran estimated about 18,300 cases assuming an outbreak duration of 1.5 months in the country(14). This paper was published online on 22.2.2020. Therefore Iran is at the beginning of the outbreak and the unstable trend was acceptable.

Thailand had the best control system in the exposure of this epidemic. Checking their program could be the best guide for future epidemics.

Although more data are needed to have a more detailed prevision, these models could be helpful in predicting future confirmed cases if the spread of the virus did not change very strangely. As we know this virus is novel and have the ability to be transmitted severely. This ability may affect all the predictions, but to our knowledge and for the time of writing, these models are the best.

Conclusions

Mainland China and Thailand were successful in haltering COVID-19 epidemic. South Korea was also shown decreasing trend and may control this epidemic in near future. Investigating their protocol in this control like quarantine should be in the first line of other countries' program

References

1. Paules CI, Marston HD, Fauci AS. Coronavirus infections—more than just the common cold. JAMA. 2020;323(8):707-8.
2. Liu Y, Gayle AA, Wilder-Smith A, Rocklöv J. The reproductive number of COVID-19 is higher compared to SARS coronavirus. Journal of Travel Medicine. 2020.
3. Jung S-m, Akhmetzhanov AR, Hayashi K, Linton NM, Yang Y, Yuan B, et al. Real-Time Estimation of the Risk of Death from Novel Coronavirus (COVID-19) Infection: Inference Using Exported Cases. Journal of Clinical Medicine. 2020;9(2):523.
4. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The Lancet. 2020;395(10223):497-506.
5. Niehus R, De Salazar PM, Taylor A, Lipsitch M. Quantifying bias of COVID-19 prevalence and severity estimates in Wuhan, China that depend on reported cases in international travelers. medRxiv. 2020.
6. Zhang Y, Yang H, Cui H, Chen Q. Comparison of the Ability of ARIMA, WNN and SVM Models for Drought Forecasting in the Sanjiang Plain, China. Natural Resources Research. 2019:1-18.
7. Fattah J, Ezzine L, Aman Z, El Moussami H, Lachhab A. Forecasting of demand using ARIMA model. International Journal of Engineering Business Management. 2018;10:1847979018808673.
8. Wei W, Jiang J, Liang H, Gao L, Liang B, Huang J, et al. Application of a combined model with autoregressive integrated moving average (ARIMA) and generalized regression neural network (GRNN) in forecasting hepatitis incidence in Heng County, China. PloS one. 2016;11(6).
9. Cao S, Wang F, Tam W, Tse LA, Kim JH, Liu J, et al. A hybrid seasonal prediction model for tuberculosis incidence in China. BMC medical informatics and decision making. 2013;13(1):56.
10. Cheung Y-W, Lai KS. Lag order and critical values of the augmented Dickey–Fuller test. Journal of Business & Economic Statistics. 1995;13(3):277-80.
11. Wang Y-w, Shen Z-z, Jiang Y. Comparison of ARIMA and GM (1, 1) models for prediction of hepatitis B in China. PloS one. 2018;13(9).

12. Li Q, Feng W. Trend and forecasting of the COVID-19 outbreak in China. arXiv preprint arXiv:200205866. 2020.
13. Lai C-C, Shih T-P, Ko W-C, Tang H-J, Hsueh P-R. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and corona virus disease-2019 (COVID-19): the epidemic and the challenges. International journal of antimicrobial agents. 2020:105924.
14. Tuite AR, Bogoch I, Sherbo R, Watts A, Fisman DN, Khan K. Estimation of COVID-2019 burden and potential for international dissemination of infection from Iran. medRxiv. 2020.

Data references

Johns Hopkins University Center for Systems Science and Engineering, 2019. https://github.com/CSSEGISandData/COVIDQ319/blob/master/time_series/time_series_2019-ncov-Confirmed.csv