HEDG

HEALTH, ECONOMETRICS AND DATA GROUP

THE UNIVERSITY of York

WP 20/07

An ARIMA model to forecast the spread and the final size of COVID-2019 epidemic in Italy

Gaetano Perone

March 2020

http://www.york.ac.uk/economics/postgrad/herc/hedg/wps/

An ARIMA Model to Forecast the Spread and the Final Size of COVID-2019 Epidemic in Italy

Gaetano Perone[†]

Abstract: Coronavirus disease (COVID-2019) is a severe ongoing novel pandemic that is spreading quickly across the world. Italy, that is widely considered one of the main epicenters of the pandemic, has registered the highest COVID-2019 death rates and death toll in the world, to the present day. In this article I estimate an autoregressive integrated moving average (ARIMA) model to forecast the epidemic trend over the period after April 4, 2020, by using the Italian epidemiological data at national and regional level. The data refer to the number of daily confirmed cases officially registered by the Italian Ministry of Health (www.salute.gov.it) for the period February 20 to April 4, 2020. The main advantage of this model is that it is easy to manage and fit. Moreover, it may give a first understanding of the basic trends, by suggesting the hypothetic epidemic's inflection point and final size.

Keywords: COVID-2019; infection disease; pandemic; Italy; time series; ARIMA; forecasting models.
JEL Codes: C22; C53; I18

Highlights:

- ARIMA models allow in an easy way to investigate COVID-2019 trends, which are nowadays of huge economic and social impact.
- These data may be used by the health authority to continuously monitor the epidemic and to better allocate the available resources.
- The results suggest that the epidemic spread inflection point, in term of cumulative cases, will be reached at the end of May.
- Further useful and more precise forecasting may be provided by updating these data or applying the model to other regions and countries.

[†] Gaetano Perone.
Department of Management, Economics and Quantitative Methods, University of Bergamo, Via dei Caniana 2, 24127, Bergamo, Italy.
e-mail: gaetano.perone@unibg.it.

1. Introduction

Coronavirus disease (COVID-2019) is a severe ongoing novel pandemic that has emerged in Wuhan, the capital city of China's Hubei province, in December 2019. In few months, it has spread quickly across the world, and at the time of writing has affected more than 200 countries and has caused about tens of thousands of deaths. The most affected countries are China, France, Germany, Italy, Spain, and USA. Italy is considered one of the main epicentres of the pandemic due to its pretty high death rates (12.33%) and death toll (15,362),[2] and it represents the nucleus of this short paper.

When an epidemic occurs, one crucial question is to determine its evolution and inflection point. So, the main aim of this paper is to provide a short-term forecast of the spread of COVID-2019 in Italy, by using an autoregressive integrated moving average (ARIMA) model on national and selected regional data, over the period after April 4, 2020.[3] The paper is organized as follows. In section 2, I will introduce the data used in the econometric analysis. In section 3 I will discuss the empirical strategy. In section 4, I will present the main findings. Finally, in section 5 I will stress the possible meaning and consequences of the results.

2. Data description

The data used in this analysis refer to the number of new daily COVID-2019 confirmed cases from February 20, 2020 to April 4, 2020, and are extracted from the official website of the Italian Ministry of Health (www.salute.gov.it). They include the overall national trend and five selected Italian regions: Emilia Romagna, Lombardy, Marche, Tuscany, and Veneto. Marche and Tuscany belong to the centre of Italy, while Emilia Romagna, Lombardy, and Veneto belong to the north of Italy. These regions have been chosen because of their centrality in the Italian outbreak; in fact, they are characterized by the highest number of COVID-2019 confirmed cases on April 4, 2020. Lombardy is the country's leading region, with a mortality rate of 17,62% and 49,118 confirmed cases, the 39.41% of the overall Italian cases, followed by Emilia Romagna (16,540 cases), Veneto (10,824 cases), Tuscany (5,671 cases), and Marche (4,341).[4] About 79.4% of COVID-2019 cases are concentrated in the north of the country. This clearly shows that COVID-2019 has especially affected the north of the country.

The descriptive analysis of the overall and regional data shows that the new daily COVID-2019 confirmed cases have increased approximately until the 37^{th}-38^{th} day since the start of the epidemic. Then, they have showed a gradual decreasing trend, by suggesting a possible epidemic stabilization and slowdown (Figure 1). I will try to deepen the implications of this trend in the next sections.

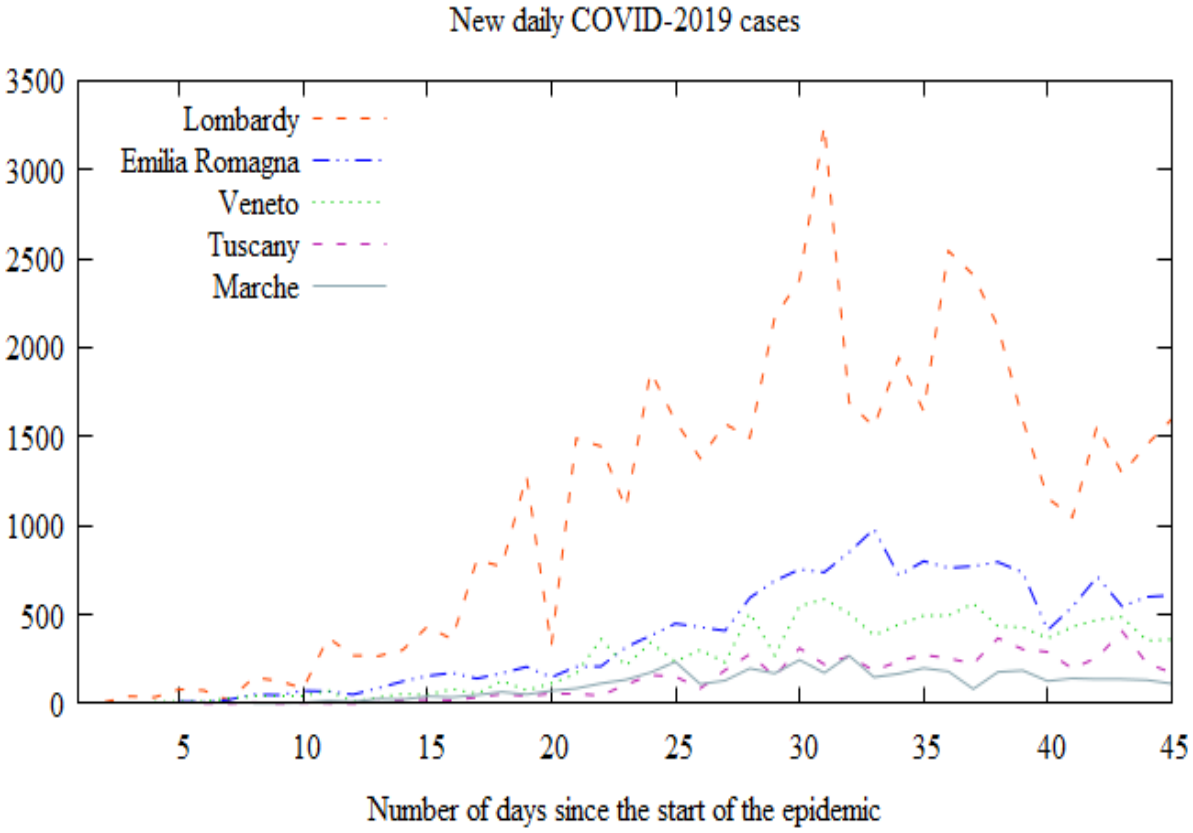
---

[2] On April 4, 2020, Italy had the world's highest death rates and death toll due to COVID-2019.

[3] In the Appendix B (Figure B1), I will also include an updated estimation on national data, for the period after April 9, 2020.

[4] It's necessary to stress that Piedmont has registered 11,709 cases at the same date. However, I decided not to consider this region in the analysis due to the presence of an outlier on April 4, 2020.

Figure 1. New daily COVID-2019 confirmed cases in the Italian regions since the start of the epidemic.



Notes: author's elaboration on Italian Ministry of Health data (www.salute.gov.it).[5]

3. Empirical strategy

In the last few months an increasing body of literature has attempted to forecast the trend and the final size of the COVID-2019 pandemic by using different approaches (Batista 2020; Benvenuto et al. 2020; Fanelli and Piazza 2020; Giordano et al. 2020; Gupta and Pal 2020; Kumar et al. 2020; Read et al. 2020; Wu et al. 2020; Zhao et al. 2020; Zhou et al. 2020). The autoregressive integrated moving average (ARIMA) model is one of them (Benvenuto et al. 2020; Gupta and Pal 2020; Kumar et al. 2020). ARIMA model could be considered one of the most used prediction models for epidemic time series (Rios et al. 2000; Li et al. 2012; Zhang et al. 2014). It is frequently used with non-stationary time series in order to capture the linear trend of an epidemic or disease. In particular, it allows to predict a given time series by considering its own lags, i.e. the previous values of the time series, and the lagged forecast errors.

[5] The data are available at URL: https://github.com/pcm-dpc/COVID-19/tree/master/dati-regioni.

The optimal ARIMA model parameters have been chosen *i*) by using the Akaike's information criterion (AIC) and a measure of forecast accuracy, i.e. the mean absolute error (MAE);[6] *ii*) by investigating the autocorrelation function (ACF) and the partial autocorrelation function (PACF) of the residuals;[7] and *iii*) by testing the fundamental statistical assumptions about residuals, i.e. normality, heteroskedasticity, and independence.

Specifically, I follow the approach of He and Tao (2018), Wang et al. (2018), and Benvenuto et al. (2020).[8] In first instance, I check if the Italian regional and national time series are stationary by using the Augmented Dickey-Fuller (1981) (ADF) test and the modified ADF-GLS (or ERS) test for unit root developed by Elliott, Rothenberg and Stock (1992).[9] The tests (table 1) show that all variables, with the only exception of Tuscany, have a unit root and need to be transformed into a stationary process.

Then, I use AIC and MAE to identify ARIMA lag order (p), degree of differencing (d), and order of moving average (q). The parameters of ARIMA, that minimize the information lost by models and the measure of forecast accuracy, are reported in Table 2. They are the following: Emilia Romagna (0, 2, 1), Marche (0, 2, 2), Lombardy (1, 2, 1), Tuscany (3, 2, 1), Veneto (0, 2, 2), and Italy (4, 2, 2). So, in the case of Emilia Romagna, Marche, and Veneto, ARIMA models assumes the form of linear exponential smoothing models.

Finally, I implement three different tests to perform diagnostic cheeks on the residuals: *i*) the Doornik and Hansen's (1994) test for normality; *ii*) the Engle's (1982) Lagrange Multiplier test for the ARCH (autoregressive conditional heteroskedasticity) effect; and *iii*) the Ljung-Box test for autocorrelation. All tests allow to accept the null hypothesis of normality, homoskedasticity, and autocorrelation of the residuals (Table 3).[10] The basic estimated equation is the following:

$$\Delta\hat{y}_t = k + \phi_1 \Delta y_{t-1} + \cdots + \phi_p \Delta y_{t-p} + \omega_1 \Delta e_{t-1} + \cdots + \omega_q \Delta e_{t-q} + e_t \qquad [1]$$

Where Δ denotes the non-seasonal differences, $\hat{y}$ means the time series that will be explained in time *t*, *k* is the constant term, *p* is the lag order, ϕ is the coefficient of each parameter *p*, *q* is the order of moving average, ω is the coefficient of each parameter *q*, and e_t means the residuals of errors in time *t*.

[6] According to Hyndman and Athanasopoulos (2018), MAE is one of the most commonly used scale-dependent measure to assess the forecast accuracy. Moreover, it is easy to interpret, such as AIC value. In both cases, the lowest values indicate the best ARIMA model.

[7] ACF and PACF correlograms are showed in the Appendix A (Figures A1 to A3).

[8] To carry out the econometric analysis I used Gretl (version 2020a), and R (version 3.6.3)

[9] I use two different approaches because, as stated by Gujarati and Porter (2009), there is no a recognized uniformly powerful test for detecting unit root.

[10] The only exceptions are Emilia Romagna and Lombardy, that are affected by non-normality and autocorrelation, respectively. If the normality is not a necessary condition for forecasting, the violation of the independence assumption may generate some problems, by suggesting greater prudence when interpreting the results.

4. Results

Table 4 reports the summary of the results of ARIMA models for the overall and regional data (Figures 2 to 7). The forecast algorithm seems to indicate that the national new daily COVID-2019 cases are largely stabilized and will probably drop near to zero (local cases) in the next 38 days, at least (Figure 2). The hypothetic inflection point[11] will be reached after May 12, 2020. And the final epidemic size could be between 194,000 and 206,000 cases.[12] A similar downward trend is obtained by fitting a specific ARIMA model for the single regions. Specifically, Emilia Romagna requires 42 days to come closer to zero local new cases (Figure 3), Lombardy needs 32 days (Figure 4), Tuscany requires 56 days (Figure 6), and Veneto needs 28 days to significatively flatten the COVID-2019 curve (Figure 7), at least. Marche needs only 12 days (Figure 5), but it does not seem a very prudent estimation. So, in Table 4 I show the average mean between the first and the second best ARIMA model for Marche. It indicates that Marche needs on average 40 days to reach the hypothetic inflection point.

Moreover, these models also allow, indirectly, to provide an approximation of the total number of deaths due to COVID-2019. In fact, by multiplying the estimated total national cases for the current Italy's mortality rate (on April 4), I obtain a number of deaths between 23,920 and 25,400.

The absence of significant residual spikes in ACF and PACF correlograms, shows that the models are a good fit (Figures A1, A2, and A3 in the Appendix A).[13]

Table 1. Results of ADF and ERS test for unit root.

Regions	Daily cases		At first difference	
(constant + trend)	ADF	ERS	ADF	ERS
Emilia Romagna [1]	-1.7823	-1.9214	-6.0472***	-7.4615***
Lombardy [1]	-3.0634	-2.1693	-9.2458***	-9.4513***
Marche [3]	-0.754	-1.0985	-7.091***	-10.3264***
Tuscany [1]	-4.5378***	-4.4087***	-6.9036***	-9.3454***
Veneto [2]	-1.473	-1.7598	-11.6103***	-11.8692***
Italy [1]	-1.547	-1.6586	-5.8425***	-5.9628***

Notes: lags in brackets. For lag length selection I used AIC approach. Significance level: 0.01***; 0.05**; 0.1*.

[11] I mean the inflection point of the cumulative number of COVID-2019 confirmed cases.

[12] The epidemic final size is obtained by summing the original values until April 4, 2020, and the forecast values for the period after April 4, 2020, minus and plus the mean standard deviation calculated for forecast values.

[13] The only exception is Lombardy and Veneto, that have two (lag 4 and 11) and one (lag 3) significant spikes, respectively. However, these remain the best possible models.

Table 2. The optimal parameters for ARIMA models.

Regions	AR-I-MA parameters	AIC value	Mean absolute value
Emilia Romagna	(0, 2, 1)	476.9949	64.995
Lombardy	(1, 2, 1)	642.7858	306.25
Marche	(0, 2, 2)	380.2007	27.491
Tuscany	(3, 2, 1)	428.1028	38.716
Veneto	(0, 2, 2)	477.2967	60.88
Italy	(4, 2, 2)	650.9967	334

Notes: for parameter selection I used AIC approach.

Table 3. The results of normality, ARCH, and autocorrelation tests for the ARIMA models (Figures 2-7).

Regions	Doornik-Hansen test for normality		Engle's LM test for ARCH effect		Ljung Box test for autocorrelation	
	Value	p-value	Value	p-value	Value	p-value
Emilia Romagna	12.467	0.002	0.8611	0.973	13.1893	0.1542
Lombardy	0.966	0.6168	14.9047	0.1356	14.783	0.0388
Marche	4.148	0.1257	5.2826	0.8715	5.6599	0.6853
Tuscany	1.645	0.4393	5.7887	0.8327	5.9615	0.4275
Veneto	2.103	0.3493	3.4188	0.9698	13.0181	0.1112
Italy	2.775	0.2496	5.008	0.8906	4.9098	0.2967

Notes: for lag selection I followed Hyndman and Athanasopoulos (2018), that suggest a value of 10.

Table 4. Summary of the results of ARIMA models (Figure 2-7).

Regions	Inflection point (days since April 4)	Inflection point (date)	Epidemic final size (approximate)
Emilia Romagna	42	May 16, 2020	32,600 to 33,400
Marche (mean)*	40	May 14, 2020	6,300 to 7,100
Lombardy	32	May 6, 2020	77,000 to 80,000
Tuscany	56	May 30, 2020	14,400 to 15,200
Veneto	28	May 2, 2020	15,300 to 16,000
Italy	38	May 12, 2020	194,000 to 206,000

Notes: *this is the average mean between the first and the second best ARIMA model for Marche.

Figure 2. Results of ARIMA forecast approach for overall national data.

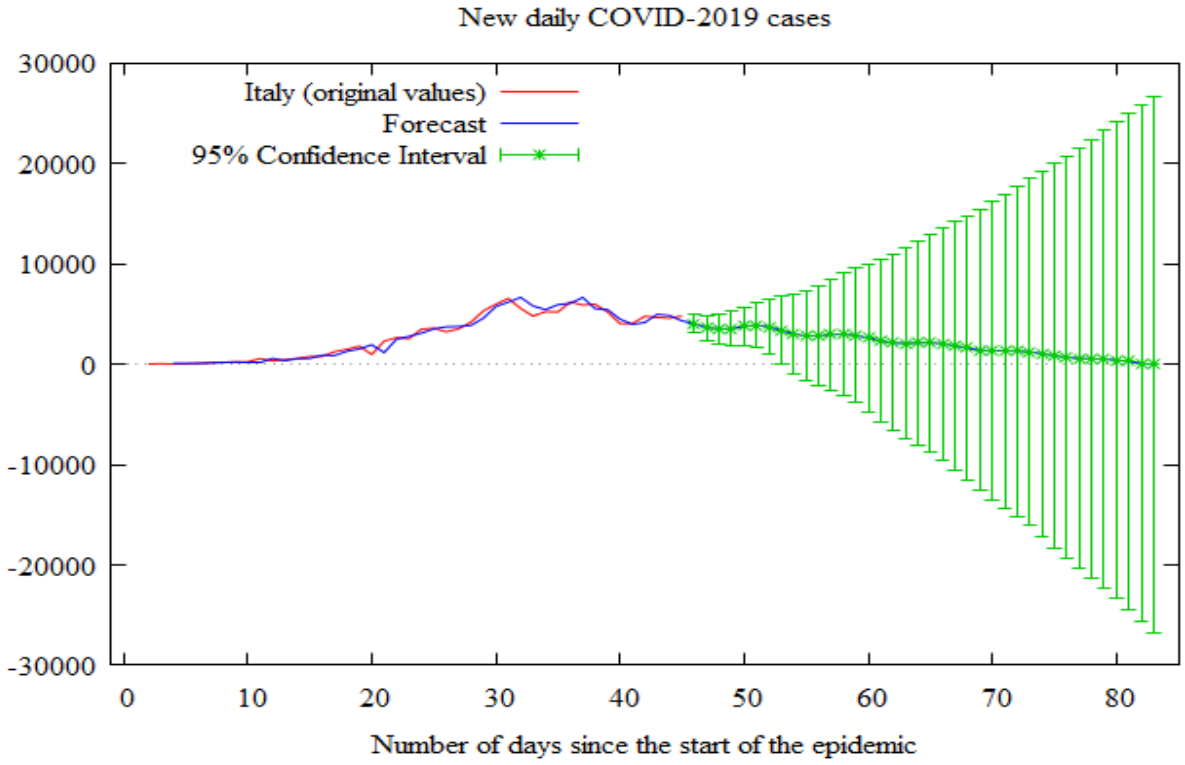


Figure 3. Results of ARIMA forecast approach for Emilia Romagna.

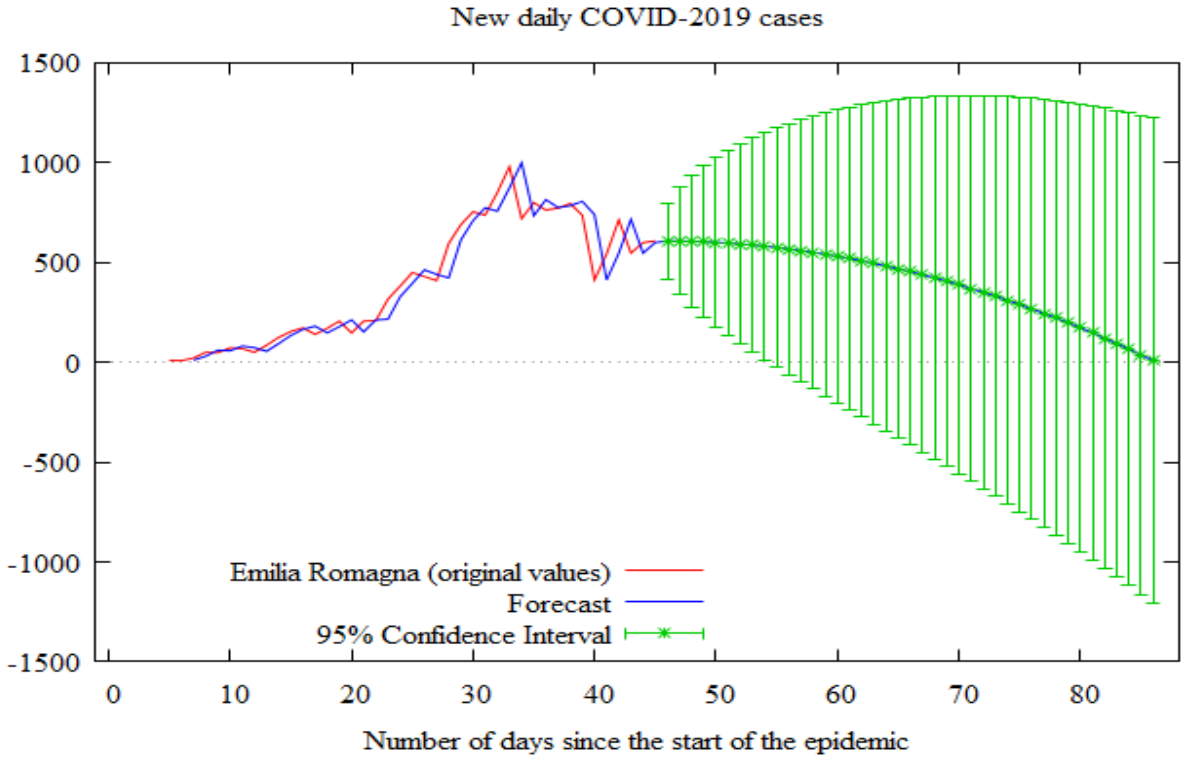


Figure 4. Results of ARIMA forecast approach for Lombardy.

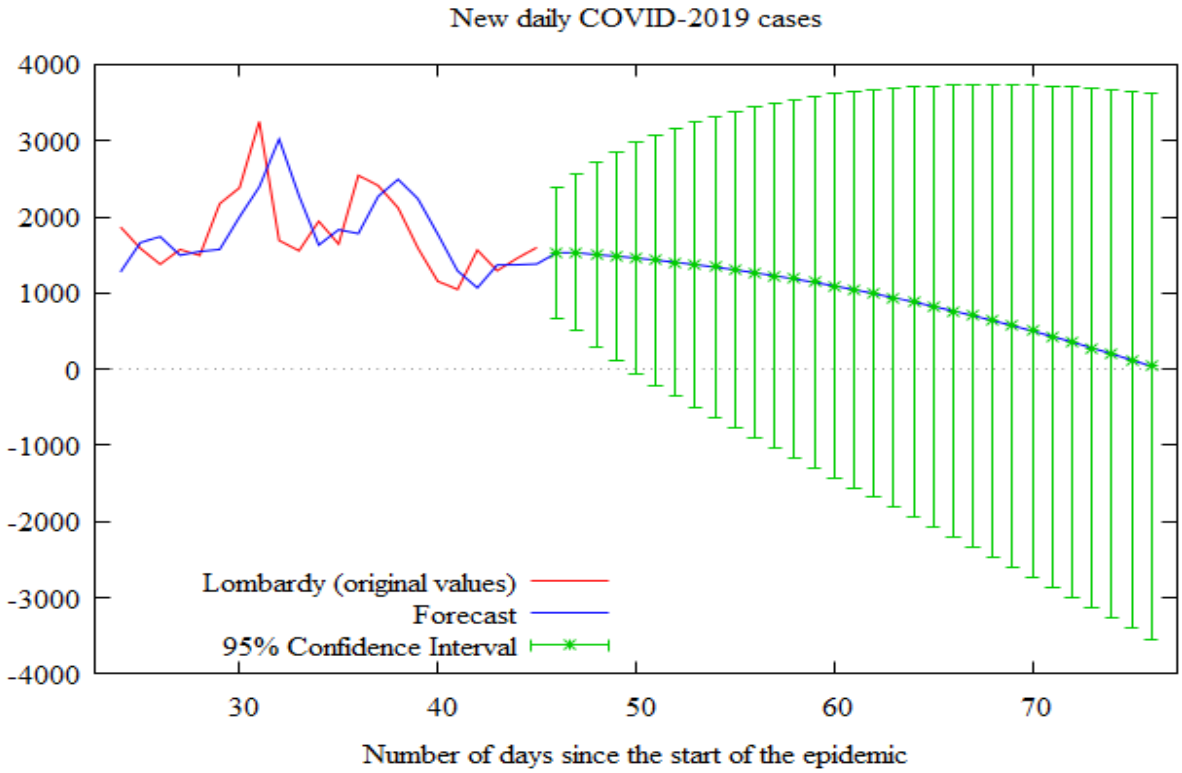


Figure 5. Results of ARIMA forecast approach for Marche.

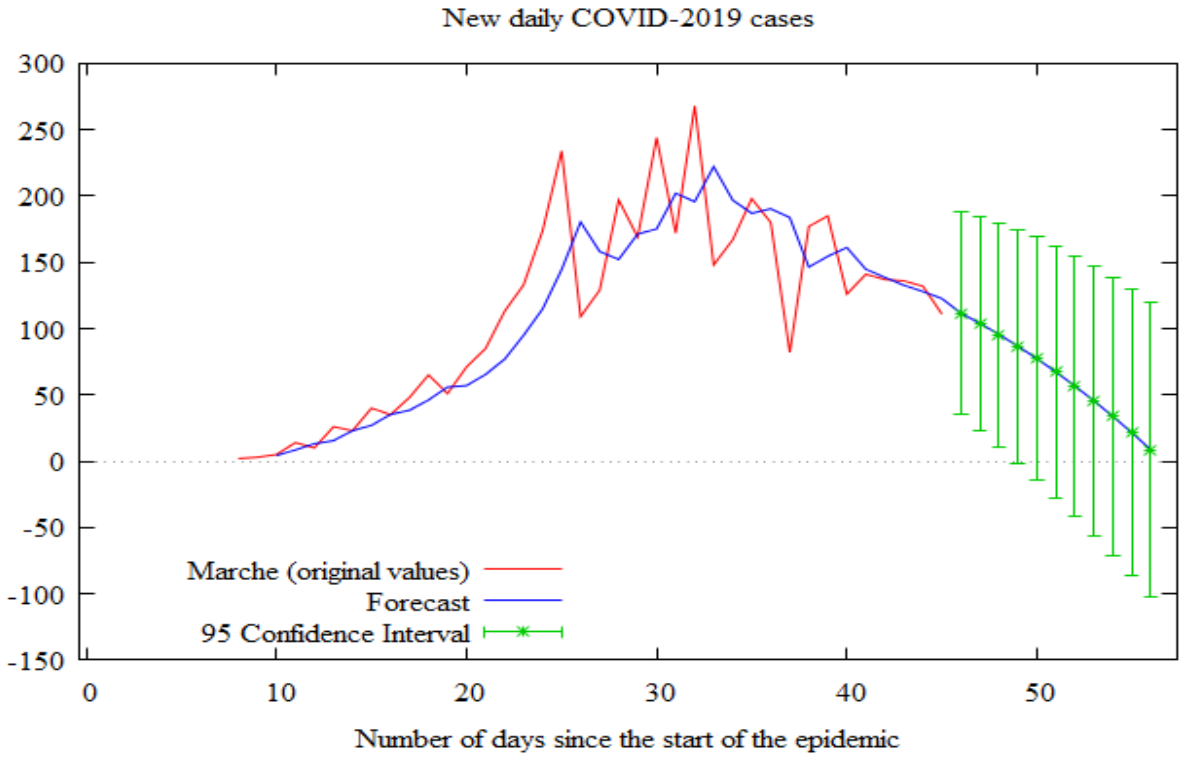


Figure 6. Results of ARIMA forecast approach for Tuscany.

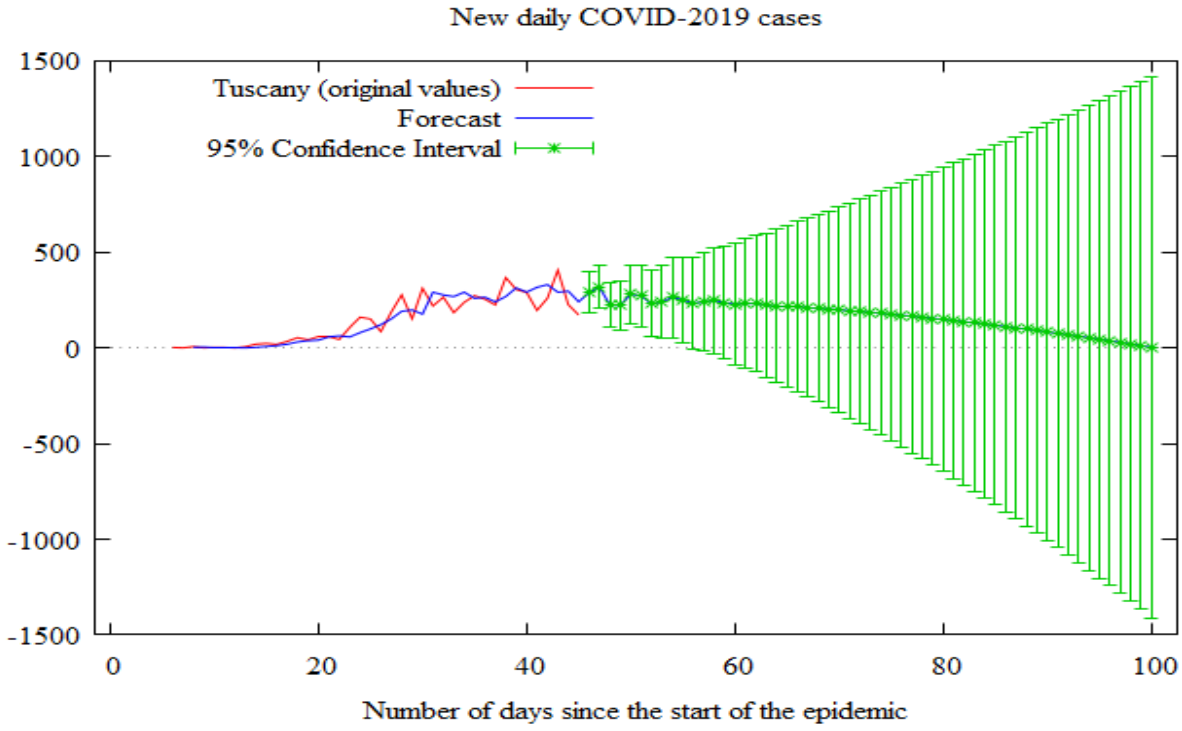
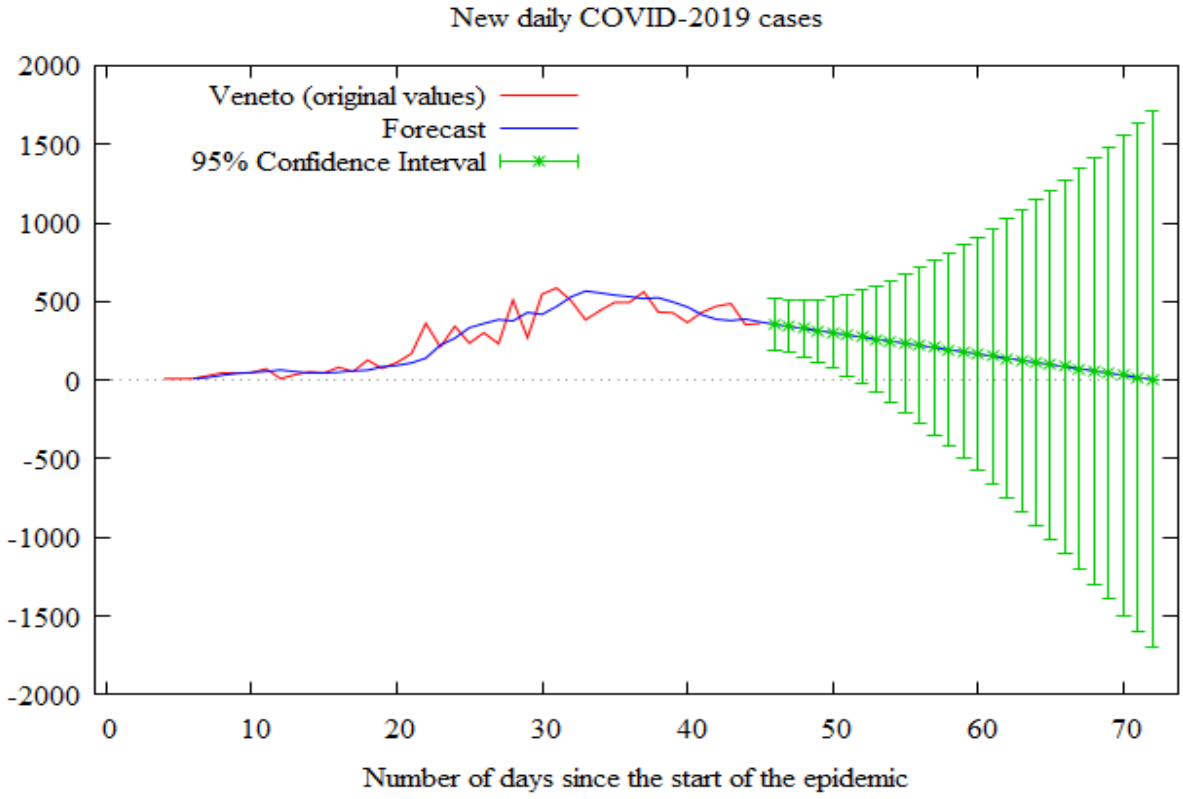


Figure 7. Results of ARIMA forecast approach for Veneto.



4. Conclusion

ARIMA models can be viewed as a simple and immediate tool to program the health monitoring system at national and regional level. The main advantages of ARIMA forecasting approach are surely its ease of application and interpretation. By the contrary, it is sensitive to outliers in the data and, do not account for the noise, that is unknown by definition. For these reasons it may be considered a good model for short-term forecasting, but the results should be interpreted with thriftiness.

Results suggest that COVID-2019 epidemic in Italy will reach the inflection point, in term of cumulative cases, in the next 40-55 days, i.e. about the entire month of April and May 2020. Specifically, Lombardy and Veneto seem to require a lower number of days than the other regions, especially compared to Tuscany, that will need approximately 56 days to definitively flatten the COVID-2019 curve. The final epidemic size in Italy should be around 200,000 cases.

However, it is necessary to stress that this estimation is strongly related to the previous trend. The continuation of the restrictive measures and the strict compliance with the rules, such as traffic and travel restriction, ban on gatherings, and closure of commercial activities, may mitigate the size of the epidemic. Further useful and more precise forecasting may be provided by updating these data and applying the model to other regions and countries.

*The last update for Italy (Figure B1 in the Appendix B) substantially confirms the previous analysis, by indicating that the new daily cases will drop near to zero in the next 50 days, i.e. at end of May 2020, at least. However, the estimated total cases are significatively higher than the previous model (Figure 2) and should be between 254,000 and 272,000. The total deaths should range from 31,318 to 33,538.

References

[1] Batista M. (2020), Estimation of the final size of the COVID-19 epidemic. MedRxiv. doi:10(2020.02), 16-20023606.

[2] Benvenuto D., Giovanetti M., Vassallo L., Angeletti S., Picozzi S. (2020), Application of the ARIMA model on the COVID-2019 epidemic dataset, Data in Brief, 29, 105340. https://doi.org/10.1016/j.dib.2020.105340

[3] Dickey D. A., Fuller W. A. (1981), Likelihood ratio statistics for autoregressive time series with a unit root, Econometrica, 49 (4), pp. 1057-1072. DOI: 10.2307/1912517.

[4] Doornik J. A., Hansen H. (1994). An Omnibus Test for Univariate and Multivariate Normality, Working Paper, Nuffield College, Oxford University.

[5] Elliott G., Rothenberg T., Stock J. (1996), Efficient Tests for and Autoregressive Unit root, Econometrica, 64 (4), pp. 813-836. DOI: 10.2307/2171846.

[6] Engle R. F. (1982), Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation, Econometrica: Journal of the Econometric Society, 50 (4), pp. 987-1007. DOI: 10.2307/1912773.

[7] Fanelli D., Piazza F. (2020), Analysis and forecast of COVID-19 spreading in China, Italy and France, Chaos, Solitons & Fractals, 134, 109761. https://doi.org/10.1016/j.chaos.2020.109761

[8] Giordano G., Blanchini F., Bruno R., Colaneri P., Di Filippo A., Di Matteo A., Colaneri M. (2020), A SIDARTHE Model of COVID-19 Epidemic in Italy. ArXiv preprint. ArXiv:2003.09861.

[9] Gujarati D. N., Porter D. C. (2009), Basic Econometrics, 5th Edition, McGraw Hill Inc., New York.
[10] Gupta R., Pal S. K. (2020), Trend Analysis and Forecasting of COVID-19 outbreak in India. MedRxiv. https://doi.org/10.1101/2020.03.26.20044511
[11] He Z., Tao H. (2018), Epidemiology and ARIMA model of positive-rate of influenza viruses among children in Wuhan, China: A nine-year retrospective study, International Journal of Infectious Diseases, 74, pp. 61-70. doi: 10.1016/j.ijid.2018.07.003.
[12] Hyndman R. J., Athanasopoulos G. (2018), Forecasting: principles and practice, OTexts, Melbourne.
[13] Kumar P., Kalita H., Patairiya S., Sharma Y. D., Nanda C., Rani M., Rahmai J., Bhagavathula A. S. (2020), Forecasting the dynamics of COVID-19 Pandemic in Top 15 countries in April 2020 through ARIMA Model with Machine Learning Approach. MedRxiv. https://doi.org/10.1101/2020.03.30.20046227
[14] Li, Q., Guo N. N., Han Z. Y., Zhang Y. B., Qi S. X., Xu Y. G., ... & Liu Y. Y. (2012), Application of an autoregressive integrated moving average model for predicting the incidence of hemorrhagic fever with renal syndrome, The American journal of tropical medicine and hygiene, 87 (2), pp. 364-370.
[15] Ljung G. M., Box G. E. (1978), On a measure of lack of fit in time series models, Biometrika, 65 (2), pp. 297-303. DOI: 10.2307/2335207.
[16] Read J. M., Bridgen J. R., Cummings D. A., Ho A., Jewell C. P. (2020), Novel coronavirus 2019-nCoV: early estimation of epidemiological parameters and epidemic predictions. MedRxiv. https://doi.org/10.1101/2020.01.23.20018549
[17] Rios M., Garcia J. M., Sanchez J. A., Perez D. (2000), A statistical analysis of the seasonality in pulmonary tuberculosis, European Journal of Epidemiology, 16 (5), pp. 483-488. DOI: 10.1023/a:1007653329972
[18] Wang Y. W., Shen Z. Z., Jiang Y. (2018), Comparison of ARIMA and GM (1, 1) models for prediction of hepatitis B in China, PloS One, 13 (9): e0201987. doi:10.1371/journal.pone.0201987
[19] Wu J. T., Leung K., Leung G. M. (2020), Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study, The Lancet, 395 (10225), pp. 689-697. doi:10.1016/S0140-6736(20)30260-9.
[20] Zhang X., Zhang T., Young A. A., Li X (2014), Applications and Comparisons of Four Time Series Models in Epidemiological Surveillance Data, PLoS One, 9 (2), e88075.
[21] Zhao S., Lin Q., Ran J., Musa S. S., Yang G., Wan W., Lou Y., Gao, D., Yang L., He D., Wang M. H. (2020), Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak, International Journal of Infectious Diseases, 92, pp. 214-217. doi:10.1016/j.ijid.2020.01.050.
[22] Zhou T., Liu Q., Yang Z., Liao J., Yang K., Bai W., Xin L., Zhang, W. (2020), Preliminary prediction of the basic reproduction number of the Wuhan novel coronavirus 2019-nCoV, Journal of Evidence Based Medicine 1. DOI:10.1111/jebm.12376.

Appendix A

Figure A1. ACF and PACF correlograms for Italy (on the left) and Emilia Romagna (on the right).

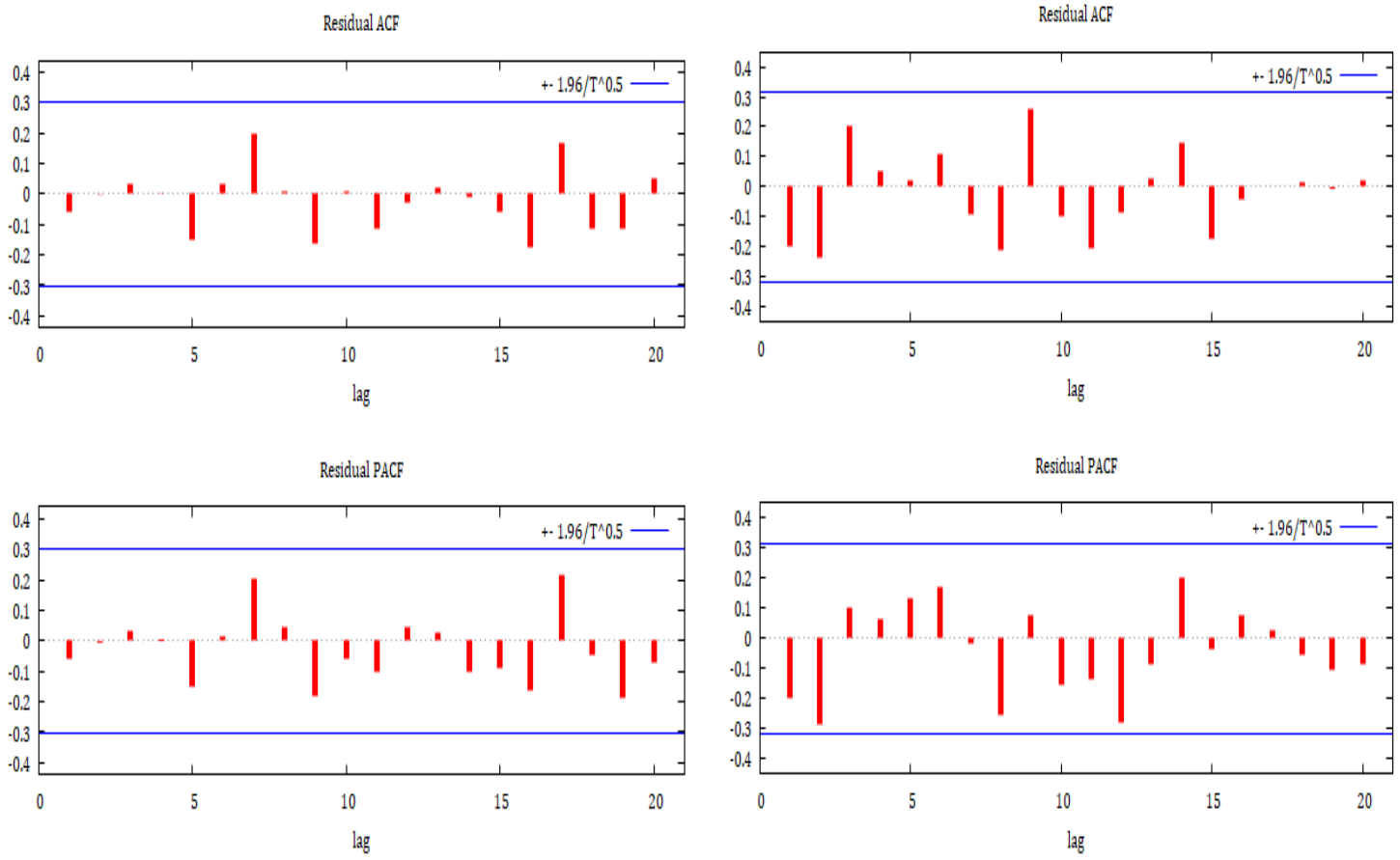


Figure A2. ACF and PACF correlograms for Lombardy (on the left) and Marche (on the right).

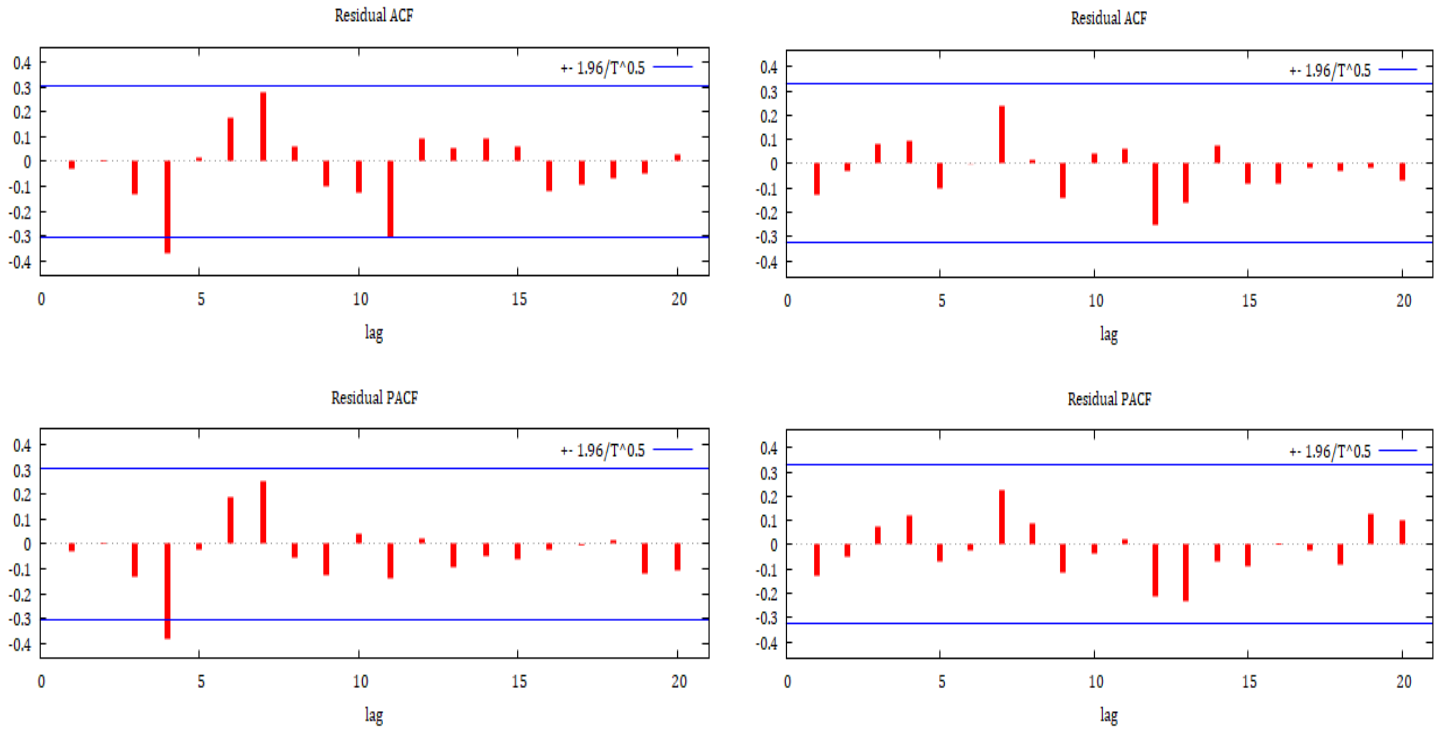
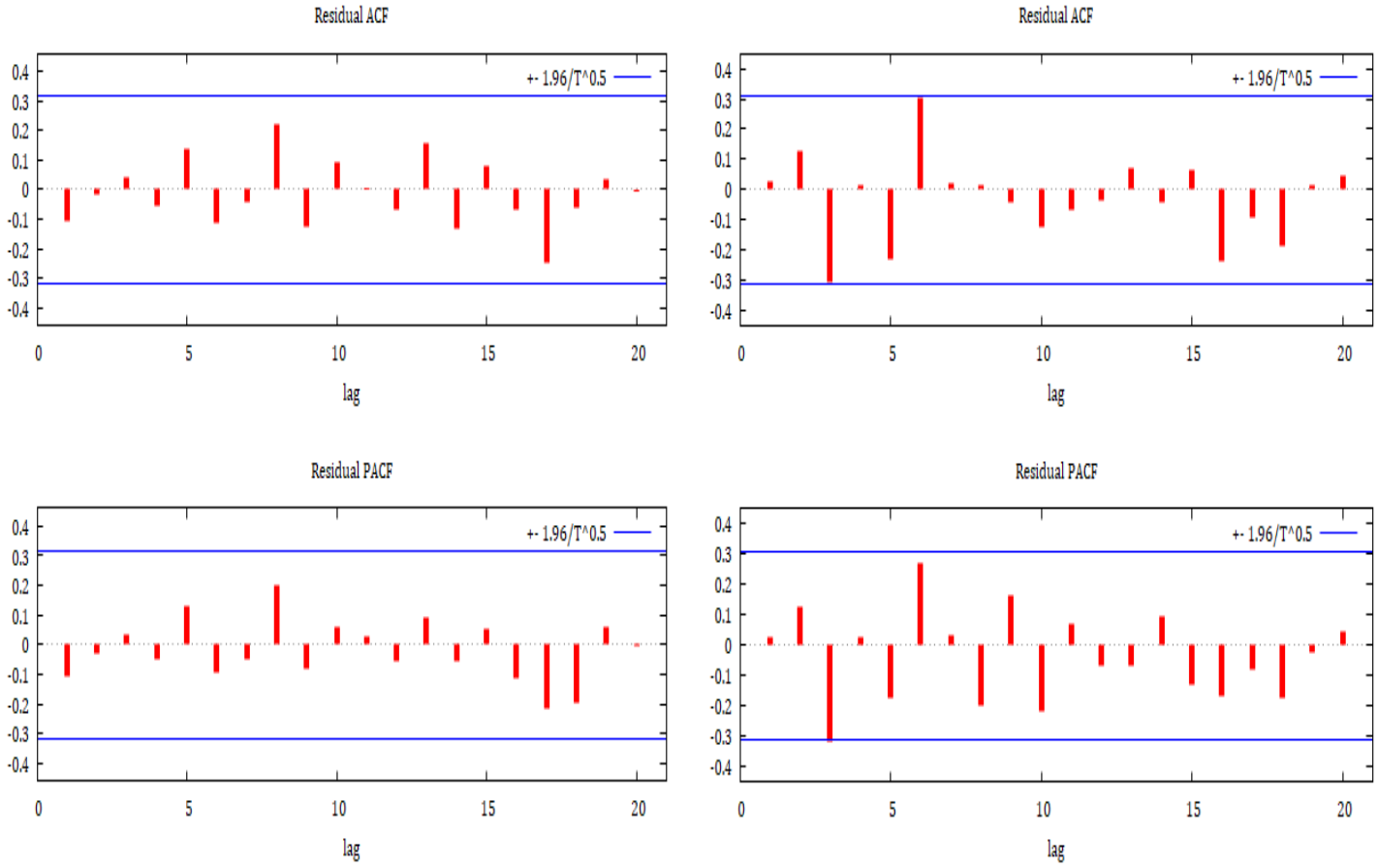


Figure A3. ACF and PACF correlograms for Tuscany (on the left) and Veneto (on the right).



Appendix B

Figure B1. Results of ARIMA forecast approach for Italy. [Update: April 9, 2020]

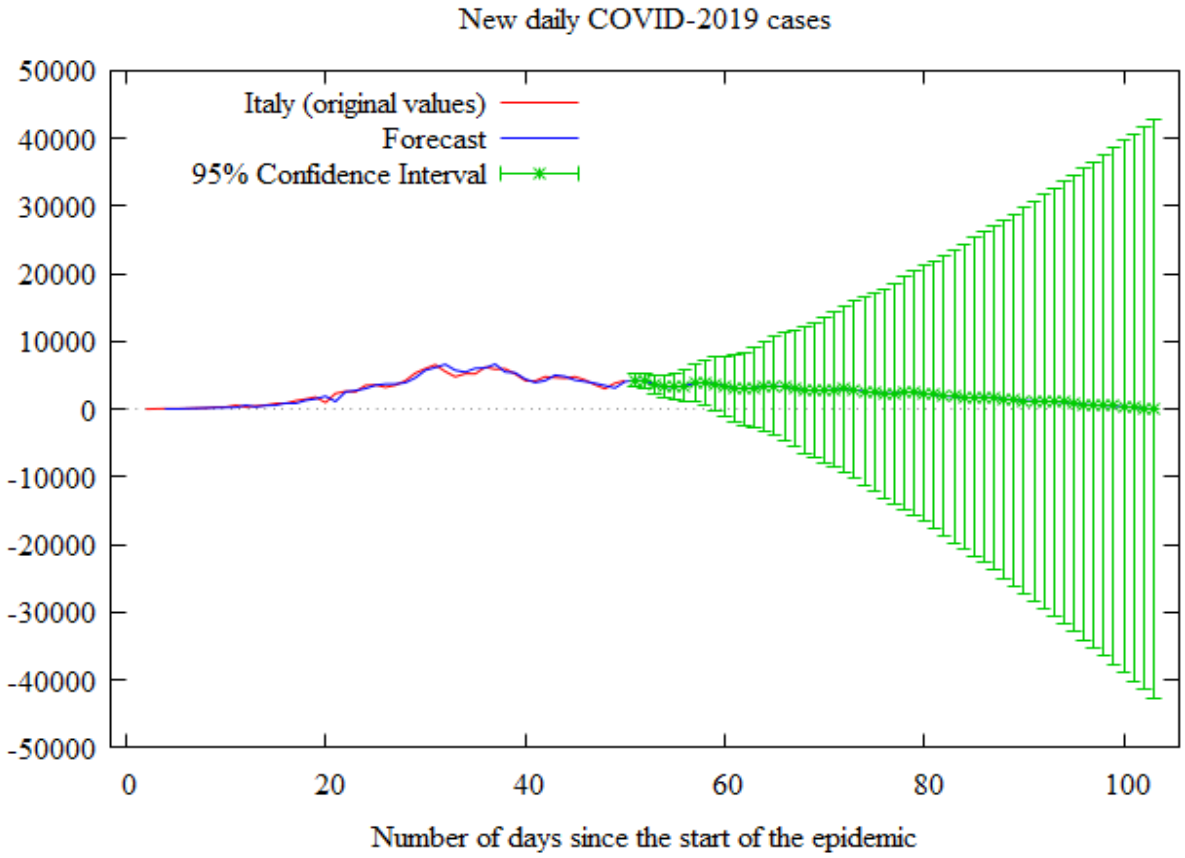


Figure B2. ACF and PACF correlograms for Italy. [Update: April 9, 2020]