TEACHING TIME SERIES ANALYSIS USING COVID-19 DATA IN SOUTHEAST EUROPE

Zagorka Lozanov-Crvenković, Emilija Nikolić Đorić, Ksenija Dumičić, Blagica Novkovska
Faculty of Sciences, University of Novi Sad
Faculty of Agriculture, University of Novi Sad
Faculty of Economics and Business, University of Zagreb
Faculty of Economics, University of Tourism and Management Skopje
zlc@dmi.uns.ac.rs

COVID-19 pandemic overwhelmed our world affected the health, economy, and all activities of the world population and yet it produced interesting time series, suitable as examples in teaching statistics. Relevant and real-world data such as these can be used in a classroom to stimulate the learning of important statistical concepts, such as time series analysis, and make statistics more interesting and exciting. Here, we consider an approach to guide some stages of time series learning. With COVID-19 data from three countries from Southeast Europe we illustrate several stages of time series analysis.

INTRODUCTION

The purpose of studying time series is understanding their structure, measuring the similarity between two or more time series, and forecasting. If students understand the underlying concepts of the model of their time series, they can then make sensible statements about their predictions. When learning about time series analysis, basic *descriptive statistics* and *densities* of distribution should be calculated and interpreted. Rolling or *moving averages* should be used to reduce noise and smooth time series data. Deeper analysis can be done by *cross correlation* of two series. The *forecasting* can be done by neural network models and ARIMA models. The *dissimilarity* between different time series can be measured by dissimilarity index. This index combines the dissimilarity between the raw values and the dissimilarity between the temporal correlation behavior. We illustrate these stages of time series analysis with COVID-19 data from three countries from Southeast Europe.

Since COVID-19 pandemic is now part of our everyday life, analysis of these time series may be more interesting to students. Therefore, data on the daily number of COVID-19 confirmed cases and number of deaths, from March 6, 2020, to April 30, 2021, for Serbia, Croatia and North Macedonia are used as the examples for mentioned analysis. Data are obtained from John Hopkins coronavirus resource center data base - https://coronavirus.jhu.edu/. Statistical package R was used in analysis.

RESULTS

Table 1. present basic descriptive statistics and robust summary statistics of observed series. Figure 1 presents density functions of observed time series.

Table 1. Summary and robust summary statistics for time series of daily data

N=421	Country	Mean	Max	Std.Dev.	Coef.Var. (%)	Median	rCV*	Skewness-Bowley	Kurtosis-Moor
Confirmed cases	**Croatia**	789.01	4620	1093.28	138.56	255.0	96.86	0.63	1.86
	North Mac.	361.91	1511	388.50	107.34	166.0	87.35	0.63	1.53
	Serbia	1637.83	7999	2030.62	123.98	351.0	91.09	0.81	1.07
Deaths	**Croatia**	16.82	92	21.49	127.75	5.0	100.00	0.71	1.13
	North Mac.	11.53	51	11.56	100.25	7.0	71.43	0.47	1.43
	Serbia	15.11	69	16.64	110.07	7.0	85.71	0.57	1.15

Distributions of all three series are right skewed, as it can be seen in Figure 1., especially for Serbia. To reduce the influence of extreme values on the summary statistics, robust summary statistics are calculated and presented in Table 1. * $rCv = \frac{MAD}{m} \cdot 100\%$, $m = median(x)$, $MAD = median\left(|m - x|\right)$.

Right skewness is also confirmed, with robust Bowley coefficient of skewness, which is highest for Serbia. Because of right skewness of series, robust Moor kurtosis is calculated, and for all series are higher than 1.23. However, they do not deviate much from this value, and therefore, these distributions are not heavy tailed.

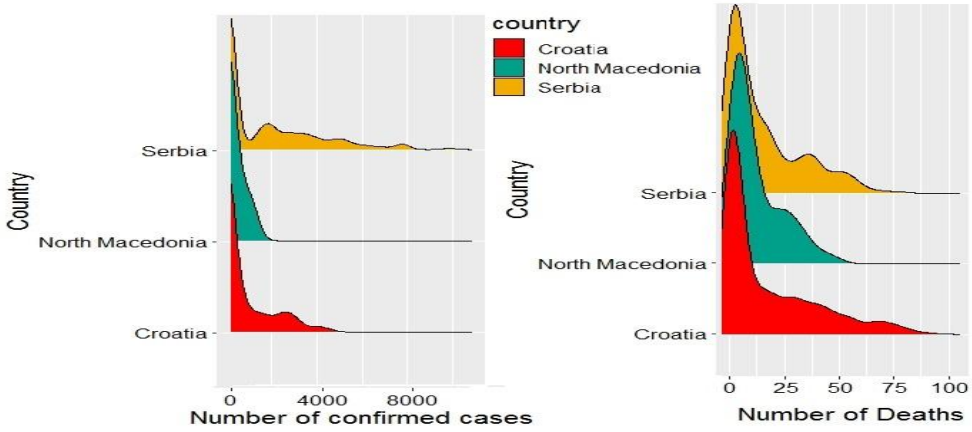

Figure 1. Density functions of number of confirmed cases and number of deaths 6.3.2020-30.4.2021

To be able to compare time series of confirmed and death cases, values are calculated per 100,000 population. According to UN data, at midyear 2020, population for Croatia is estimated at 4,105,267, for Serbia 6,944,975, and North Macedonia 2,083,374. In Table 2. descriptive statistics of daily data per 100,000 population are given.

Table 2. Descriptive statistics of time series of daily data calculated per 100,000 population

N=421	State	Mean	Median	Max	Std.Dev.
Confirmed	**Croatia**	19.22	6.21	112.54	26.63
cases	**North Macedonia**	17.37	7.97	72.53	18.65
	Serbia	23.58	5.05	143.74	30.10
Deaths	**Croatia**	0.41	0.12	2.24	0.52
	North Macedonia	0.55	0.34	2.45	0.55
	Serbia	0.22	0.10	1.15	0.25

Presenting time series as data per 100,000 population enables us to compare mean, median maximal value, and standard deviation. The largest mean value of confirmed cases per 100,000 population was in Serbia, and smallest in North Macedonia. On the other hand, largest mean value of deaths per 100,000 population was in North Macedonia, and smallest in Serbia.

Moving average used for analyses of time series by creating a time series of averages of different subsets of the full data set. By calculating the moving average, the impacts of random, short-term fluctuations are mitigated. Rolling or moving averages are a way to reduce noise and smooth time series data. In Figures 2. and 3. original and 7 day moving average series of confirmed cases and deaths in three observed countries are presented. Smoothed time series of COVID-19 time series clearly show similar behavior of time series from three countries, with peaks occurring at the same time. All three series have nonlinear trend. As an exercise, students can do some other methods of

smoothing out the noise and extracting the trend as: locally weighted scatter plot smoothing (LOESS), exponentially weighted moving average (EWMA), the triangular moving average.

Cross correlation of confirmed versus death cases, obtained using 7 day moving average series shows the lag between confirmed cases and death cases. Figure 4. displays cross correlations of number of confirmed cases and number of deaths (7-day moving average). Here, cross correlations show that the highest correlation was after 15 days for Croatia, 11 days for North Macedonia and 10 days for Serbia, see Table 3. Rami Krispin, author of *coronavirus* R-package, was the first who explored cross correlation of number of confirmed cases and deaths until September 2020) and found that for Italy the highest correlation was after 4 days and for Belgium 6 days.

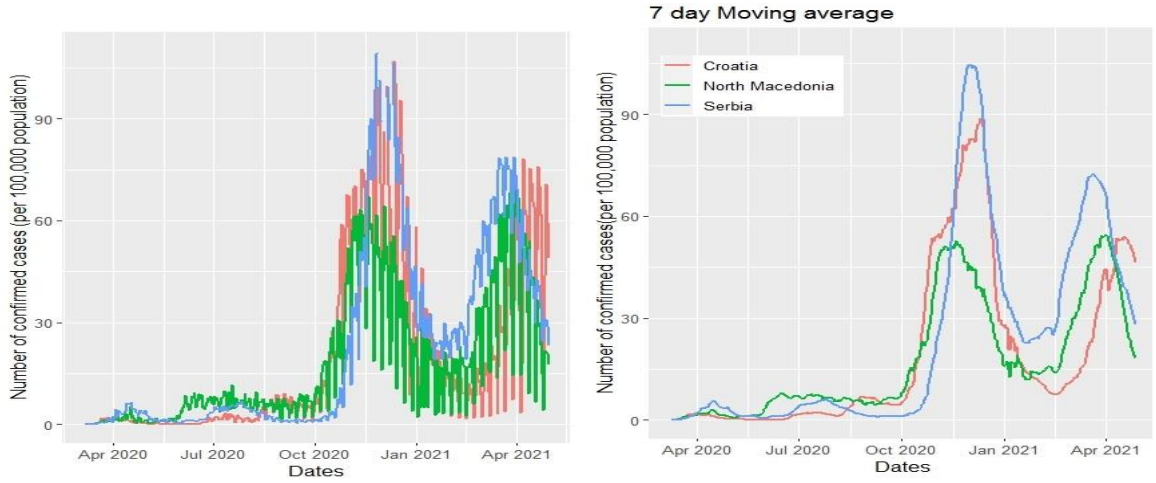

Figure 2. Number of confirmed cases and 7 day moving average of number of confirmed cases (per 100,000 population) 6.3.2020-30.4.2021

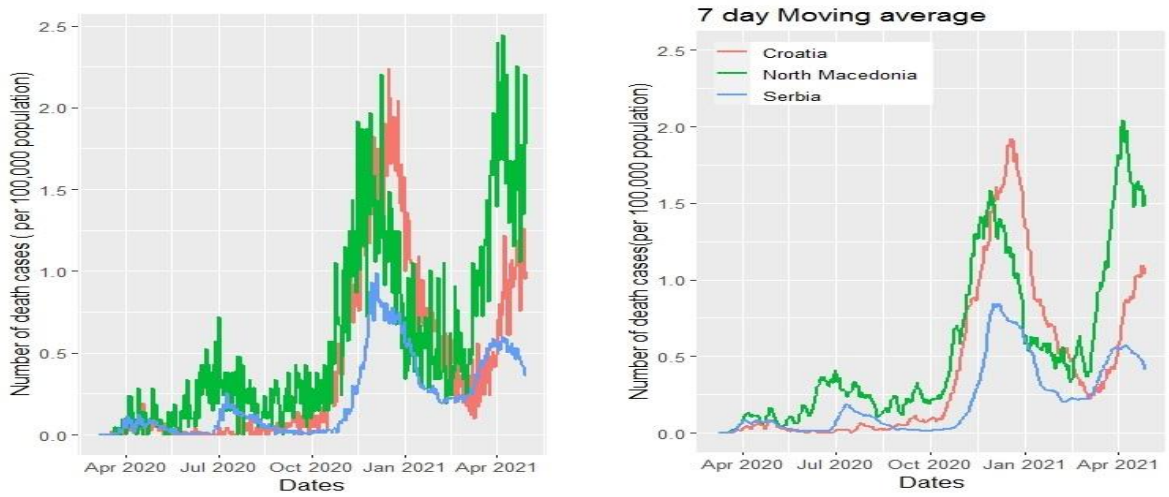

Figure 3. Number of death cases and 7 day Moving average of number of death cases (per 100,000 population) 6.3.2020-30.4.2021

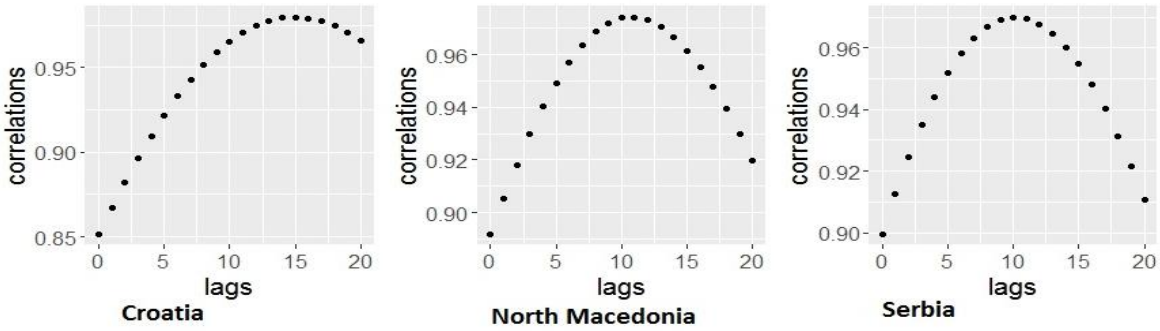

Figure 4. Cross correlation of number of confirmed cases (7-day moving average) and number of deaths (7-day moving average)

A longer period of delay in considered countries may be a consequence of improved treatment of patients, increasing hospital capacity in the second half of the year.

The forecasts of considered time series are obtained by means of ARIMA models and univariate neural network autoregression models (NNAR) that are suitable in the case of noisy data (Hyndman & Athanasopoulos, 2021). Estimation of models and computing forecasts were done applying forecast package (Version 8.14) of statistical software R version 4.0.4. and are presented in Figures 5., 6. and 7. The blue lines present forecasted values, 20 days ahead, and grey areas are corresponding 95% confidence intervals.

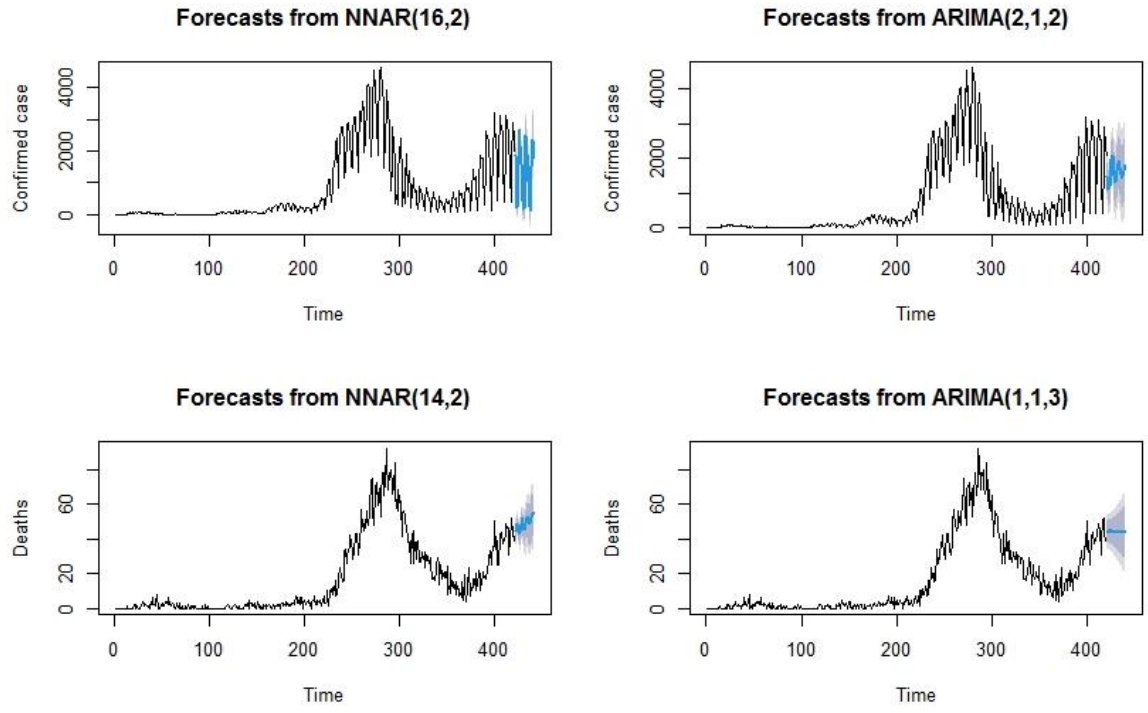

Figure 5. The forecasts of time series for Croatia by means of ARIMA and NNAR models

The forecast for all three time series in three countries show that stagnation of confirmed cases and number of deaths can be predicted. In Table 3. neural network autoregression model and ARIMA model used for forecasting are reported.

Table 3. Univariate neural network autoregression and ARIMA models for considered time series

Time series	**Croatia**	**North Mac.**	**Serbia**	**Croatia**	**North Mac.**	**Serbia**
Model	**Neural network autoregression**			**ARIMA models**		

Confirmed cases	NNAR (16,2)	NNAR (26,2)	NNAR (15,2)	ARIMA (2,1,2)	ARIMA (4,1,2)	ARIMA (2,1,1)
$\hat{\sigma}^2$	35346	5468	33849	105173	19197	104553
Deaths	NNAR (14,2)	NNAR (10,2)	NNAR (10,2)	ARIMA (1,1,3)	ARIMA (1,1,3)	ARIMA (0,1,1)
$\hat{\sigma}^2$	15.88	16.79	4.00	22.75	20.05	5.64

Estimated variance is smaller for neural network autoregression models, which indicates better fit to original data. Although both models can be applied for short-term forecasting earlier research of other authors suggest that the NNAR model is more suitable in modeling complex nonlinear time series. Additional research will be conducted to compare out-of-sample evaluation of forecasting performance of two applied models.

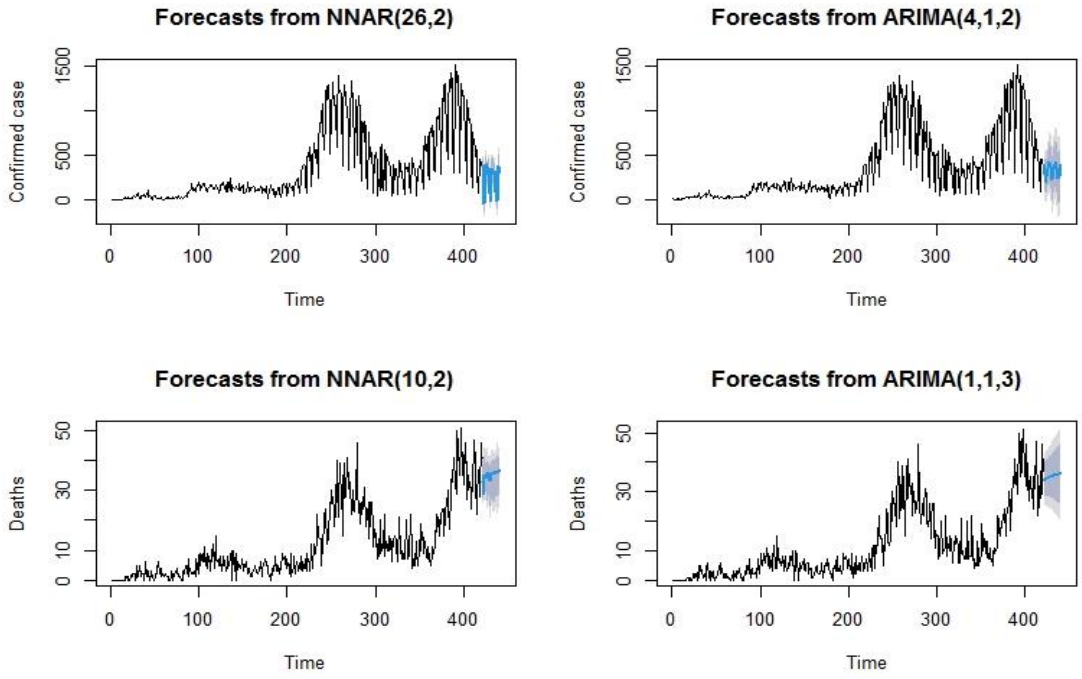

Figure 6. The forecasts of time series for North Macedonia by means of ARIMA and NNAR models

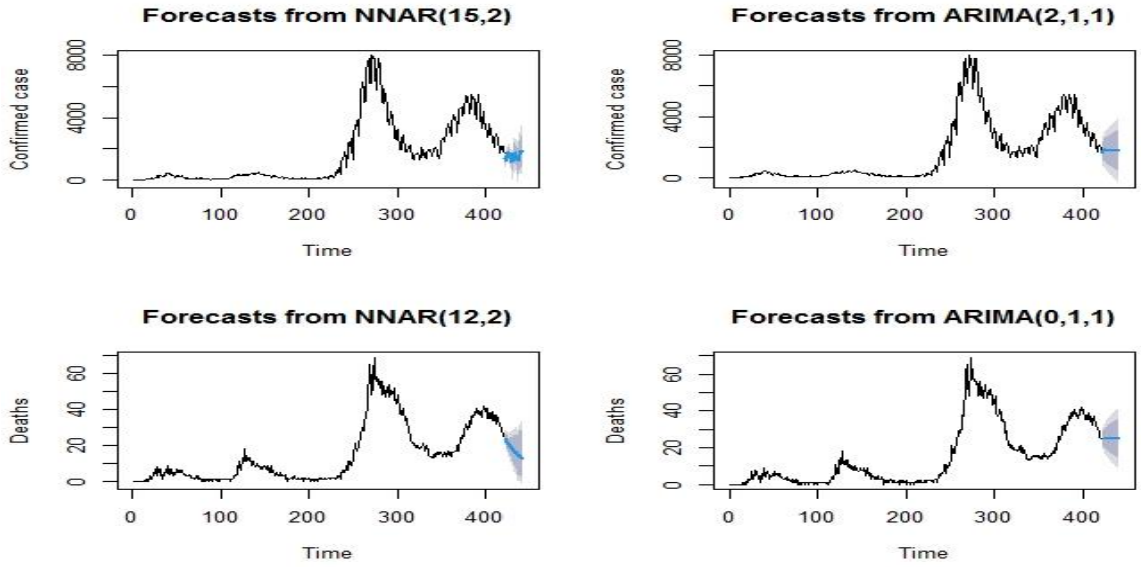

Figure 7. The forecasts of time series for Serbia by means of ARIMA models and NNAR models

The dissimilarity between time series for Croatia, North Macedonia, and Serbia COVID-19 time series (per 100,000 population) was measured by *dissimilarity index*. Index values were calculated using the package TSdist (ver 3.7) of R and presented in Table 4. This index combines the dissimilarity between the raw values and the dissimilarity between the temporal correlation behavior of the series. Computation of these measures for larger number of time series allows clustering by using conventional clustering algorithms (Montero & Vilar, 2014).

Table 4. Dissimilarity index of time series (per 100,000 population)

Confirmed cases	**Croatia**	**North Macedonia**
North Macedonia	146.79	
Serbia	264.18	211.29

Number of deaths	**Croatia**	**North Macedonia**
North Macedonia	7.19	
Serbia	8.14	9.92

There is the greatest degree of dissimilarity between the number of confirmed cases in Croatia and Serbia and the number of deaths in Serbia and North Macedonia.

In the analysis of observed time series and the comparison of individual countries, the complexity of the data collection problem should also be considered. Acording to a new research by the Institute for Health Metrics and Evaluation (IHME) at the University of Washington School of Medicine COVID-19 deaths are significantly underreported in almost every country and a new approach of the estimation of total mortality should be accepted http://www.healthdata.org/news-release/covid-19-has-caused-69-million-deaths-globally-more-double-what-official-reports-show).

CONCLUSION

The purpose of studying time series is understanding their structure, forecasting, and measuring the similarity between two or more time series. Using COVID-19 data in the learning process of basic time series has been very successful, because this awakes the students to the learning approach of time series. Here, we illustrate several stages of time series analysis with COVID-19 data from three countries from Southeast Europe. Results were based on original research and give interesting way to teaching statistics. Students can learn about something which is interesting, meaningful, and worth knowing, about phenomenon affecting their everyday life.

ACKNOWLEDGMENT

The corresponding author acknowledges financial support of the Ministry of Education, Science and Technological Development of the Republic of Serbia (Grant No. 451-03-9/2021-14/ 200125)

REFERENCES

Hyndman, R.J.& Athanasopoulos, G. (2014). *Forecasting:Principles and Practice*. OTexts

Hyndman, R.J. & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R, *Journal of Statistical Software, 26*(3).

Hyndman, R.J., & Athanasopoulos, G. (2021). *Forecasting: principles and practice*, 3rd edition, OTexts: Melbourne, Australia. https://otexts.com/fpp3/

Montero, P. & Vilar, J. A. (2014). TSclust: An R Package for Time Series Clustering. *Journal of Statistical Software, 62*(1), 1-43. http://www.jstatsoft.org/v62/i01/

R Core Team. (2021). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

https://predictivehacks.com/covid19-correlation-between-confirmed-cases-and-deaths/