Journal of Turkish

Operations Management

Investigating Covid 19 data for G20, EU and OECD countries via using time series analysis & cluster analysis

Mehmet Güray ÜNSAL[1*], Reşat KASAP[2]

[1]Department of Statistics, Faculty of Art and Science, Usak University, Uşak, Turkey
e-mail: mgunsal@gazi.edu.tr, ORCID No: http://orcid.org/0000-0001-7081-9060
[2]Department of Statistics, Faculty of Science, Gazi University, Ankara, Turkey
e-mail: rkasap@gazi.edu.tr, ORCID No: http://orcid.org/0000-0002-9306-3101
*Corresponding Author

Article Info

Article History:

Received: 02.11.2020
Revised: 21.11.2020
Accepted: 22.11.2020

Keywords

Covid 19,
G20,
EU,
OECD,
Statistical analysis

Abstract

Shortly after Covid 19 virus first appeared, it turned into an epidemic that threatened the health of the world. Many countries have difficulties to control the spread of the virus. It causes many casualties and also affects the countries from a socio-economic aspect. Therefore, it is an important issue that needs to be examined in detail. In this study, the detailed research for the Covid 19 data of G20 (Group of 20), EU (European Union) and OECD (Organisation for Economic Cooperation and Development) countries is carried out by using different statistical methods. It is aimed to present different perspectives for researchers by discussing important findings and results.

1. Introduction

2019 novel coronavirus (Covid 19) is officially appeared at the end of 2019 and spread to Europe, Asia, America, Africa and the other continents in a short time. It is a kind of virus that settles in the lungs and shows signs of difficulty in breathing, it is also highly contagious (Medica News Today, 2020). In a short time, it becomes one of the major health problems in the world. Many developed and developing countries have been adversely affected by this disease. Member countries of Group of 20 (G20), European Union (EU) or Organisation for Economic Cooperation and Development (OECD) consisted a large part of the world economy and population have a hard time struggling with Covid 19 pandemic.

Because of being a pandemic worldwide, Covid 19 became a very important issue that attracted researchers in multi-disciplinary fields. One of these fields is Statistical Science which let the researchers work on the data of the novel coronavirus infections. It is clear that Covid 19 causes both health problems and socio-economic effects for many developed and developing countries in Asia, Europe, America and Africa. From this aspect, Covid 19 has several effects on the organizations such as G20, EU and OECD. Economical developments, social developments, and health care systems of the world countries will be obviously affected at the end of this pandemic. Because of these reasons, it is important to classify the affected countries. Furthermore, making comments on their time series data structure is very crucial issue to understand the spreading of the virus and this kind of analysis has a vital role for forecasting its final effects.

In this study, the member countries of G20, EU or OECD are investigated in terms of Covid 19 effect by using important statistical indicators (total number of cases, total number of recovered cases and total number of deaths) of the pandemic. Different statistical methods such as time series analysis and cluster analysis are used separately to provide different perspectives for the researchers who work related to the pandemic.

The paper is structured by the following sections: Section 2 investigates the statistical methods (time series analysis and cluster analysis, respectively) used in this study. The results of both time series analysis and cluster analysis are given in Section 3. Section 4 implies the results and makes comments on them to evaluate different perspectives for the researchers worked on the pandemic.

2. Methodology

In this section, the statistical methods used in the study are presented. The methods investigated in the analysis part of the study are time series analysis and cluster analysis, respectively. The part of time series analysis contains both linear and non-linear models such as Box-Jenkins, Holt and Brown. The part of cluster analysis contains both hierarchical and non-hierarchical clustering methods such as linkage method and K-means, respectively. Research and publication ethics were followed in this study.

2.1. Time Series Analysis

The time series is a collection of random variables, which are measurements of variables with data obtained in chronological order over time. A time series is generally shown in the format Z_t, $t = 1,2,\ldots,n$ with n sample sizes. Thus, t^{th} observed data over time is expressed by Z_t (Box, 1994; Ünsal and Kasap, 2014).

Box-Jenkins modeling process include three different stages. The first one of these stages is model selection which is based on consideration of sequence graphs. The second one is the parameter estimation process which includes the maximum likelihood estimation method in Statistical Theory. The last one is model diagnostic checking which includes statistical F test for considering whether the model is appropriate or not. The degree of the model is determined by past observation values (p) and past error values (q). Time series which is an encounter in real life are often not stable or nonstationary which is the model of the degree of the autoregressive parameter is p, the degree of the moving average parameter is q and d is the number of differences received, this model is called the autoregressive integrated moving average model and is called $ARIMA(p,d,q)$. The model is $\phi_p(B)(1-B)^d Z_t = \theta_0 + \theta_q(B) + A_t$.Where ϕ(B) and θ(B) are polynomials of p and q degrees, respectively (Box, 1994).

Finding the best model by using the AIC (Akaike's Information Criterion). The AIC is formally defined as below (Box, 1994): (M is the number of the estimated parameters in the model)

$$AIC(M) = n \ln \hat{\sigma}_A^2 + 2M \quad (1)$$

In this paper, also used exponential smoothing methods for model identification. is a procedure for continually revising an estimate in the light of more recent experiences. Exponential smoothing method is a methodology that uses the average (smoothing) past values of the data of time series in a decreasing (exponential) manner (Montgomery, 2015):

Simple exponential smoothing has the following equations:

$$L(t)=\alpha Y(t)+(1-\alpha)L(t-1) \quad (2)$$

$$\hat{Y}(k)=L(t)$$

Brown's exponential smoothing has the following equations:

$$L(t)=\alpha Y(t)+(1-\alpha)L(t-1) \quad (3)$$

$$T(t)=\alpha(L(t)-L(t-1))+(1-\alpha)T(t-1)$$

$$\hat{Y}_t(k)=L(t)+((k-1)+\alpha^{-1})T(t)$$

Holt's exponential smoothing has the following equations:

$$L(t)=\alpha Y(t)+(1-\alpha)(L(t-1)+T(t-1)) \quad (4)$$

$$T(t)=\gamma(L(t)-L(t-1))+(1-\gamma)T(t-1)$$

$$\hat{Y}(k)=L(t)+kT(t)$$

Then the best model has the smallest Mean Absolute Error (MAE) (Montgomery, 2015).

$$\text{MAE}=\frac{1}{F}\sum_{l=1}^{F}\left|e_l\right| \quad (5)$$

2.2. Cluster Analysis

Cluster analysis is a multivariate statistical method that is used for grouping the units by considering the variables representing important properties of them. The units that are similar in terms of related variables are assigned to the same group. The measure of similarity is calculated by a distance function such as Euclidean, Mahalanobis, Manhattan, Minkowski distance fonctions etc.(Johnson and Wichern, 2007). One of the most popularly used distance function, Euclidean can be calculated as following:

$$d_{AB}=\sqrt{\sum_{i=1}^{n}\left(x_{Ai}-x_{Bi}\right)^2} \quad (6)$$

Here, d_{AB} is Euclidean distance value between unit A and unit B, n is the number of variables in the analysis, x_{Ai} and x_{Bi} are the values of i th variable of unit A and unit B, respectively (Johnson and Wichern, 2007). K-means and hierarchical clustering approaches used in this study are important methods in cluster analysis. Now, we discuss basic definitions of these two popular methods.

2.2.1. K-means

K-means is a non-hierarchical clustering method proposed by MacQueen (1967). K-means describe an algorithm to assign each unit to the cluster having the nearest centroid (Johnson and Wichern, 2007). This algorithm can be explained in a few steps. Firstly, the optimum value of the number of clusters (K) is determined. K number of units randomly chosen and each of them are assigned for each cluster. Secondly, the remaining units are assigned to the relevant clusters by considering the minimum distance function value (usually by using Euclidean distance), and after each assignment, the new cluster centers are updated. Then, repeat the second step until no more reassignments for units take place .(Johnson and Wichern, 2007).

2.2.2. Linkage Method

In this sub-section, the linkage methods (agglomerative hierarchical procedures) are considered. The linkage methods can be categorized into three main types as single-linkage (based on minimum distance), complete-linkage (based on maximum distance), and average linkage (based on average distance) (Johnson and Wichern, 2007). The agglomerative hierarchical procedures for a group of items can be described in the following steps: Firstly, each item is considered as a seperate clusters (in other words, for N items, we start with N clusters), secondly, the distance matrix (start with the dimensions NxN) is calculated to consider the nearest pair of clusters, and these clusters are merged. Thirdly, update the distance matrix by considering the distance between new clusters. Then, repeat the second and third steps for N-1 times to finalize the algorithm .(Johnson and Wichern, 2007). In this study, the average linkage method is used to evaluate hierarchical clustering results. In average linkage methodology, the distance between two clusters is calculated as the average distance between all pairs of items (Johnson and Wichern, 2007).

3. Results

The data used in this study are obtained in DataHub (2020) which is frequently updated and used as an official data source by researchers during the pandemic. The data source is public and does not require any permission.

The first part of this section is about the results of time series analysis and the second part has the results of cluster analysis. The daily data in April and May are used for time series analysis. The data used for cluster analysis is collected on the 20th days of April and May, respectively.

3.1. Results of Time Series Analysis

In this section, we try to find out the model structure of each country and consider the variability of the model structure from April to May. To see this, the modeling processes for each country are considered by using the time series models mentioned in Section 2. Time series models obtained for the data of April and May in terms of daily number of confirmed cases, recovered cases and deaths of the fifty member countries can be seen in Table 1.
When the time series models obtained for each of April and May in terms of three variables (confirmed, recovered, deaths) given in the table are carefully examined, it is observed that the majority of countries have the models Holt, Brown and Simple in terms of these three variables in April. However, it can be said that the majority of countries have ARIMA (p, d, q) models in terms of these three variables in May.

Table 1. Time series models of the countries for the daily data obtained in April and May

Countries	April Confirmed	April Recovered	April Deaths	May Confirmed	May Recovered	May Deaths
Argentina	Holt	ARIMA(0,1,0)	Holt	ARIMA(0,2,0)	ARIMA(0,1,0)	ARIMA(0,1,0)
Australia	Holt	Holt	Holt	ARIMA(0,1,0)	ARIMA(0,1,0)	Simple
Austria	Brown	Brown	Brown	ARIMA(0,1,0)	ARIMA(0,1,0)	Holt
Belgium	ARIMA(0,2,1)	ARIMA(0,1,0)	Brown	Brown	ARIMA(0,1,0)	ARIMA(0,1,0)
Brazil	ARIMA(1,1,0)	Brown	Brown	Brown	ARIMA(0,1,0)	ARIMA(0,1,0)
Bulgaria	ARIMA(0,1,0)	Brown	Holt	Brown	ARIMA(0,1,0)	ARIMA(0,1,0)
Canada	Holt	Brown	Holt	Brown	ARIMA(0,1,0)	Holt
Chile	Brown	ARIMA(0,1,0)	Holt	Brown	Brown	Brown
China	Simple	Simple	Simple	ARIMA(0,1,0)	ARIMA(0,0,0)	ARIMA(0,0,0)
Colombia	Brown	ARIMA(0,1,0)	ARIMA(0,1,0)	ARIMA(0,1,0)	ARIMA(0,1,0)	ARIMA(0,1,0)
Croatia	Holt	Brown	Holt	Brown	Holt	ARIMA(0,1,0)
Cyprus S.	Brown	Holt	ARIMA(0,1,0)	D. Trend	Holt	ARIMA(0,1,0)
Czechia	Holt	Brown	Holt	ARIMA(0,1,0)	Holt	ARIMA(0,1,0)
Denmark	Holt	ARIMA(0,1,0)	Brown	Holt	ARIMA(0,1,0)	Holt
Estonia	Holt	Holt	Holt	ARIMA(0,1,0)	ARIMA(0,1,0)	ARIMA(0,1,0)
Finland	Holt	ARIMA(0,1,0)	Brown	Holt	ARIMA(0,1,0)	Holt
France	ARIMA(0,1,0)	ARIMA(0,1,0)	Holt	ARIMA(0,1,0)	ARIMA(0,1,0)	ARIMA(0,1,0)
Germany	Brown	Holt	Holt	ARIMA(0,1,0)	ARIMA(0,1,0)	ARIMA(0,1,0)
Geece	ARIMA(0,1,0)	ARIMA(0,1,0)	Holt	ARIMA(0,1,0)	ARIMA(0,1,0)	ARIMA(0,1,0)
Hungry	Holt	ARIMA(0,1,0)	Brown	ARIMA(0,1,0)	Brown	Holt
Iceland	D. Trend	Holt	Holt	ARIMA(0,1,0)	Holt	ARIMA(0,1,0)
Indonesia	Holt	Brown	Brown	ARIMA(0,1,0)	Holt	ARIMA(0,1,0)
India	Holt	Holt	Holt	ARIMA(0,2,1)	Brown	Holt
Ireland	Brown	Simple	Holt	Holt	ARIMA(0,1,0)	Brown
Isreal	Holt	Brown	ARIMA(0,2,0)	Holt	Holt	Brown
Italy	Brown	Brown	D. Trend	Brown	Holt	Brown
Japan	Holt	ARIMA(0,1,0)	Brown	ARIMA(0,2,0)	Holt	Holt
Latvia	Holt	Brown	Holt	Brown	ARIMA(0,1,0)	ARIMA(0,1,0)
Lithuania	Holt	Brown	Holt	ARIMA(0,1,0)	Holt	ARIMA(0,1,0)
Luxembourg	ARIMA(0,2,1)	Simple	Brown	Holt	Brown	ARIMA(0,1,0)
Malta	Holt	ARIMA(0,1,0)	Holt	Brown	Holt	ARIMA(0,1,0)
Mexico	Brown	ARIMA(0,1,0)	Brown	Brown	ARIMA(0,1,0)	Brown
Netherland	Brown	ARIMA(0,0,0)	Holt	Holt	ARIMA(0,0,0)	Holt
New Zeland	Brown	Brown	Brown	ARIMA(0,1,0)	Brown	Simple
Norway	D. Trend	ARIMA(0,0,0)	Holt	Brown	Brown	ARIMA(0,1,0)
Poland	Holt	Brown	Brown	ARIMA(0,1,0)	Holt	Brown
Portugal	Holt	Holt	Holt	Brown	ARIMA(0,1,0)	ARIMA(0,1,0)
Romania	ARIMA(0,1,0)	Brown	Holt	Holt	ARIMA(0,1,0)	Holt
Russia	Brown	ARIMA(0,1,0)	Brown	Holt	Brown	Brown
SaudiArabia	Brown	ARIMA(0,1,0)	Holt	Brown	Brown	Brown
Slovakia	Brown	Brown	ARIMA(0,1,0)	ARIMA(0,1,0)	Holt	Holt
Slovenia	Brown	Holt	Holt	Holt	Simple	Holt
SouthAfrica	Brown	Holt	Holt	ARIMA(0,1,0)	Holt	ARIMA(0,1,0)

SouthKorea	ARIMA(0,2,0)	ARIMA(1,2,0)	Holt	Brown	ARIMA(0,1,0)	ARIMA(0,1,0)
Spain	Holt	Holt	D. Trend	ARIMA(0,1,0)	Holt	ARIMA(0,1,0)
Sweden	Brown	Holt	Holt	ARIMA(0,1,0)	ARIMA(0,1,0)	ARIMA(0,1,0)
Switzerland	Brown	Holt	D. Trend	Brown	Holt	ARIMA(0,1,0)
Turkey	Brown	ARIMA(0,1,0)	Brown	Brown	ARIMA(0,2,0)	ARIMA(0,2,0)
UK	Holt	Simple	Holt	Holt	ARIMA(0,1,0)	ARIMA(0,1,0)
USA	ARIMA(0,1,1)	ARIMA(0,1,0)	Holt	Holt	ARIMA(0,1,0)	ARIMA(0,1,1)

3.2. Results of Cluster Analysis

In this section, the data of April and May (collected on the 20th of April and May, respectively) are considered. The same three variables (confirmed, recovered, deaths) are used in cluster analysis, as well. According to K-means results, the total within clusters sum of squares stop the decreasing rapidly, after the value of 3 as seen in Figure 1 and 2. Thus, the optimum value of the number of clusters can be determined as 3 in April and May.

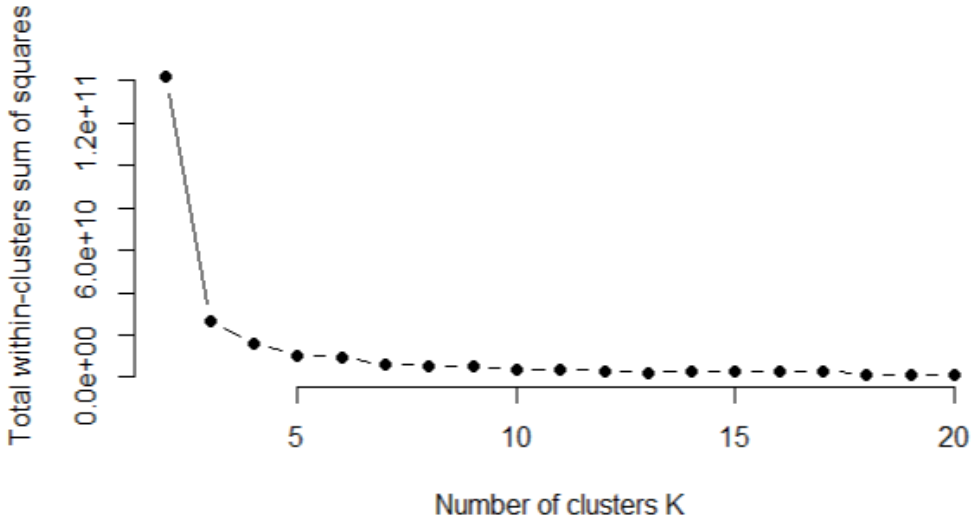

Figure 1. Total Within Clusters Sum of Squares in April

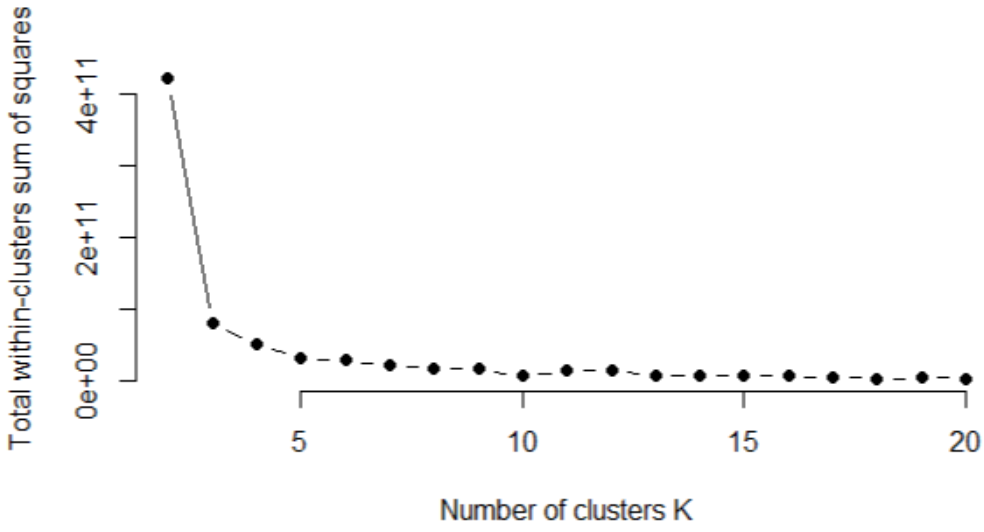

Figure 2. Total Within Clusters Sum of Squares in May

Table 2 shows clusters of the countries for April and May. Here, we list the clusters of the countries ascending order. In other words, USA stands out in the third cluster as the country which is most affected by the pandemic. The countries affected in the second degree are in the second cluster and the least affected countries are in the first cluster. According to the K-means results in April, China, France, Germany, Italy, Spain, Turkey and United Kingdom (UK) form a cluster. United States of America (USA) creates a separate cluster, individually. The other remaining countries are in the same cluster. In May, Russia and Brazil change their cluster and move to the cluster which contains France, Germany, Italy, Spain, Turkey and United Kingdom (UK). Furthermore, China moves to

the first cluster which consist of the least affected countries. All the other countries stay in the same clusters. K-means results for three clusters can be seen visually in Figure 3 (results in April) and 4 (results in May).

Table 2. Clusters for the countries in April and May

No	Countries	Cluster in April	Cluster in May	No	Countries	Cluster in April	Cluster in May
1	Argentina	1	1	26	Italy	2	2
2	Australia	1	1	27	Japan	1	1
3	Austria	1	1	28	Latvia	1	1
4	Belgium	1	1	29	Lithuania	1	1
5	Brazil	1	2	30	Luxembourg	1	1
6	Bulgaria	1	1	31	Malta	1	1
7	Canada	1	1	32	Mexico	1	1
8	Chile	1	1	33	Netherland	1	1
9	China	2	1	34	New Zeland	1	1
10	Colombia	1	1	35	Norway	1	1
11	Croatia	1	1	36	Poland	1	1
12	Cyprus S.	1	1	37	Portugal	1	1
13	Czechia	1	1	38	Russia	1	2
14	Denmark	1	1	39	SaudiArabia	1	1
15	Estonia	1	1	40	Slovakia	1	1
16	Finland	1	1	41	Slovenia	1	1
17	France	2	2	42	SouthAfrica	1	1
18	Germany	2	2	43	SouthKorea	1	1
19	Greece	1	1	44	Spain	2	2
20	Hungry	1	1	45	Sweden	1	1
21	Iceland	1	1	46	Switzerland	1	1
22	Indenosia	1	1	47	Turkey	2	2
23	India	1	1	48	USA	3	3
24	Ireland	1	1	49	UK	2	2
25	Isreal	1	1	50	Romania	1	1

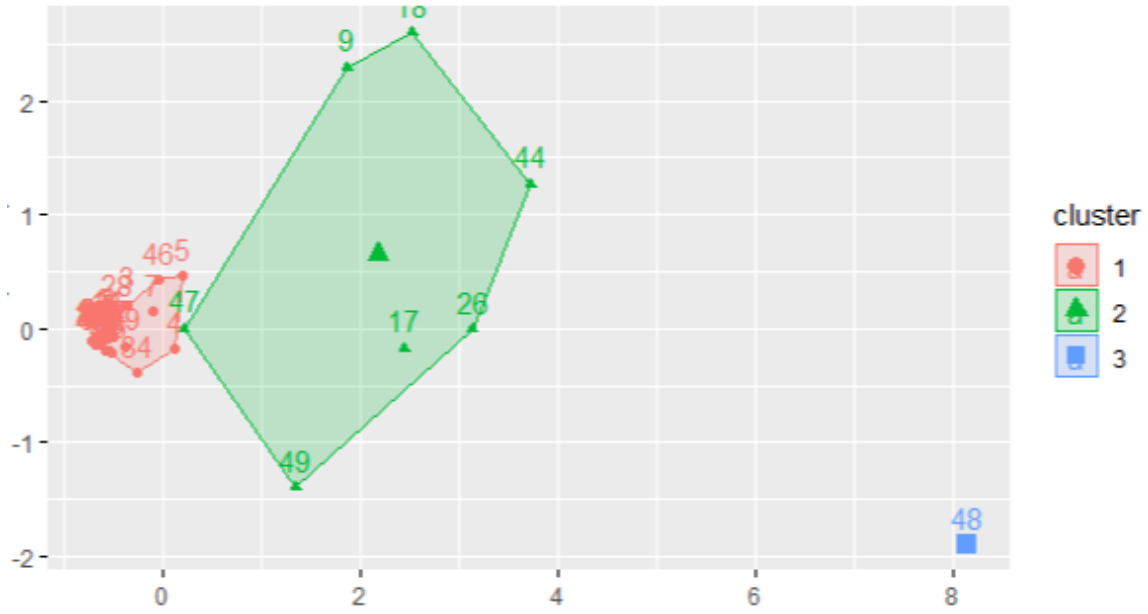

Figure 3. Cluster Plot in April

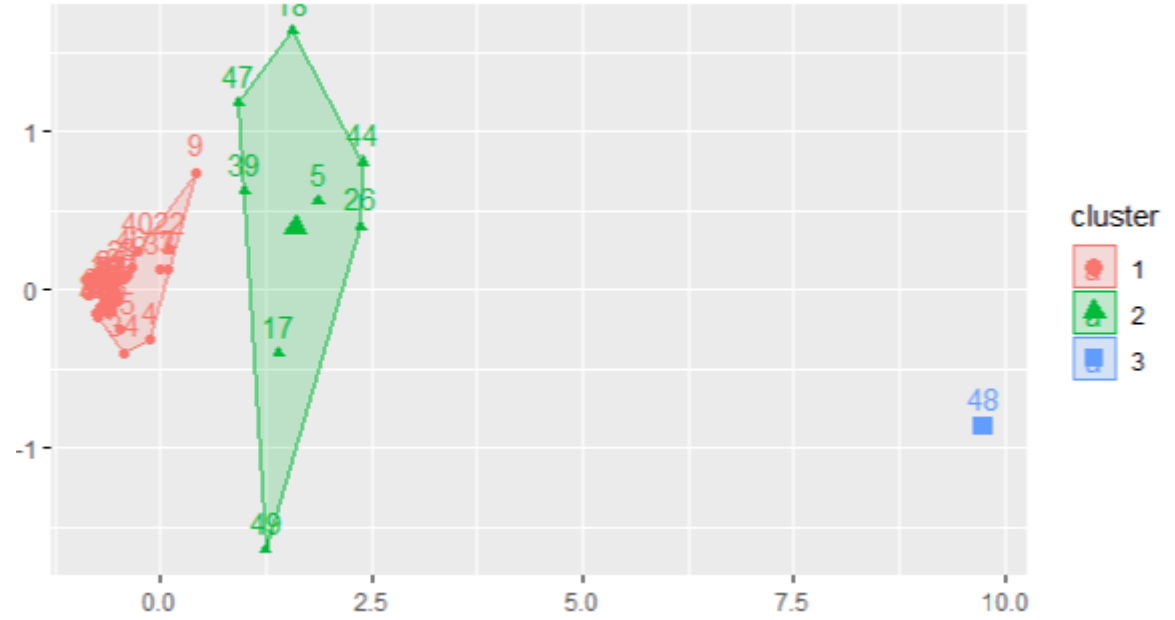

Figure 4. Cluster Plot in May

Secondly, the hierarchical clustering method is investigated for the countries. The results for April and May can be seen visually in Figure 5 and 6, respectively.

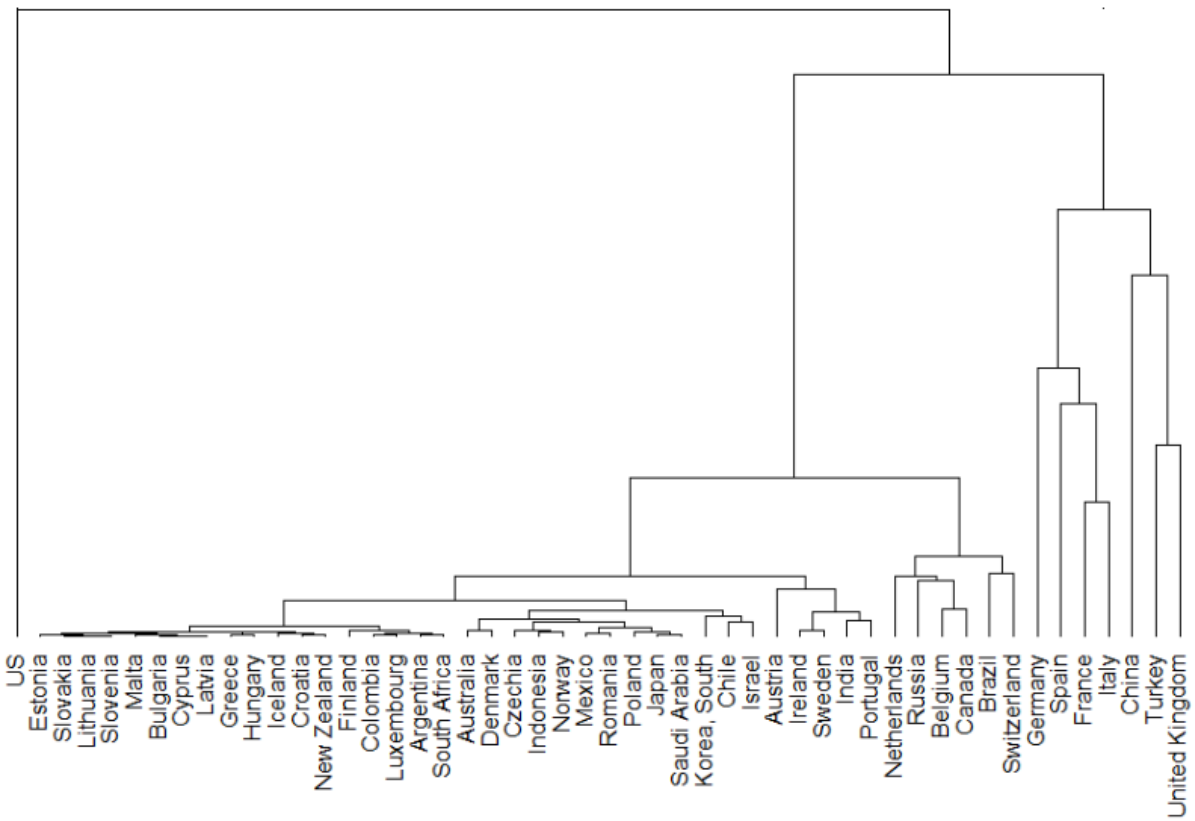

Figure 5. Dendrogram in April

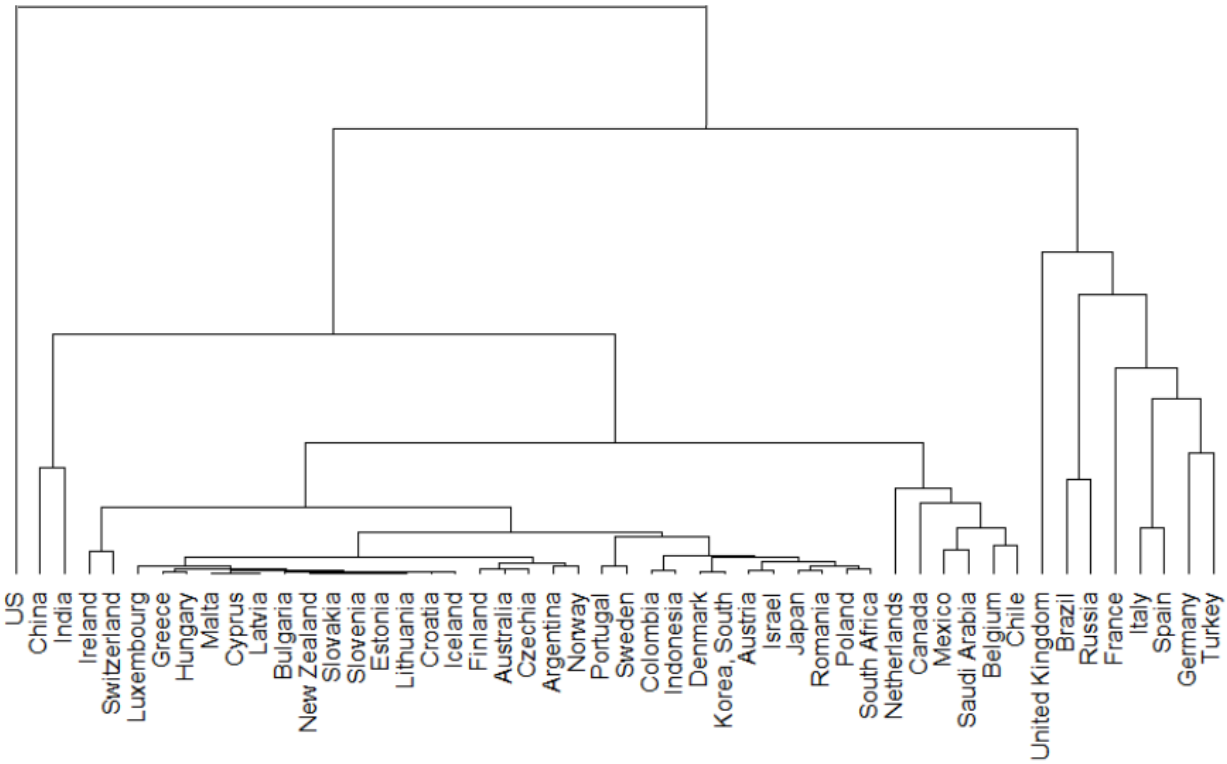

Figure 6. Dendrogram in May

According to the dendrograms**,** in April, France and Italy are linked together, then Spain and Germany are joined them. On the other hand, Turkey and the UK are linked together and China is linked with Turkey and the UK. Finally, all these countries construct the second cluster which consist the second degree affected countries. USA forms a cluster individually, which is most affected country. In May, in the second cluster, Italy and France are linked firstly. Then Russia and Brazil are linked together, Germany and Turkey show similar features according to the dendrogram. Afterward, Germany and Turkey are linked with Italy and Spain, then France is joined to these four countries. Russia, Brazil and the UK are joined to Turkey, Germany, Italy, Spain and France. In May, these eight countries construct the second cluster. All the linkages can be seen in Figure 5 and 6 for April and May, respectively.

4. Discussions and Conclusions

According to the results of the time series analysis, the time series models for fifty countries are obtained in terms of the daily number of cases, recovered cases and deaths for April and May. According to the time series models obtained in April in terms of these three variables given, most of the countries are according to the Holt, Brown and Simple models. However, in May, it can be said that most of the countries have ARIMA (p, d, q) models in terms of these three variables. This can be said that the processes of the series included different variability in April and May.

In the results of cluster analysis, the clusters are obtained in terms of the total number of cases, recovered cases and deaths for April and May. USA shows different positions from the other countries in the analysis in both April and May. Because, USA is the most affected country from Covid 19 disease. In April, China, France, Germany, Italy, Spain, Turkey and the UK show similar features and take part in the same cluster (the second one) as the other affected countries in the world. It is a very interesting point that while Brazil and Russia are in the first cluster. Then, they move to the second cluster which contains more affected countries. This indicates that the pandemic process for Brazil and Russia in May deteriorate, considerably. Additionally, China move to the first cluster from the second cluster in May. It shows that China improves its status from April to May. The other countries do not change their clusters. It is clear that Covid 19 deeply effects the world economy and health system. The countries in the 3rd and 2nd clusters have a large part of the world economy and population. As mentioned above, Russia and Brazil join the countries in the second group in May, this group consists of the negatively affected countries.

Briefly, while the time series data of the countries give mostly non-linear form based models in April, the data in May are generally modeled on the linear form. Furthermore, according to the results in cluster analysis, the countries in the third and the second clusters have a significant portion of the world economy and population, and the countries in these clusters are negatively affected countries in the pandemic.

Contribution of Researchers

Mehmet Güray ÜNSAL designed Introduction and Conclusion parts, analyzed cluster analysis part of the paper, and gave contribution on collection of the data. Reşat KASAP designed time series analysis part of the paper and gave contribution on collection of the data.

Conflict of Interest

The authors declared that there is no conflict of interest.

References

Medical News Today (2020) Retrieved May 2020, from https://medicalnewstoday.com/articles/COVID-19#incubation-period

Box, G. E. P., (1994) *Time series analysis : forecasting and control*. Englewood Cliffs, N.J. Prentice Hall.

Ünsal, M.G., & Kasap, R., (2014) Cases of Residual Types in Diagnostic Checking for ARMA Model. *Hacettepe Journal of Mathematics and Statistics* 43(3):1-10. http://hjms.hacettepe.edu.tr/uploads/7cc26d6d-b394-4c77-933e-24fd1b92796e.pdf

Montgomery, D.C., (2015*) Introduction to time series analysis and forecasting*. New Jersey Wiley.

Johnson, R.A., & Wichern, D.W., (2007) *Applied Multivariate Statistical Analysis*. 6th Edition, Pearson Prentice Hall, Upper Saddle River.

MacQueen, J.B., (1967) Some Methods for Classification and Analysis of Multivariate Observations. *In: Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1: Statistics, University of California Press, Berkeley, 281-297.

Data Hub (2020) Retrieved May 2020, from https://datahub.io/core/covid-19#data