Factor Analysis Approach to Classify COVID-19 Datasets in Several Regions

Mohammad Reza Mahmoudi [1], Dumitru Baleanu [2], Shahab S. Band [3], Amir Mosavi [4,*]

[1] Department of Statistics, Faculty of Science, Fasa University, Fasa, Fars, Iran; mahmoudi.m.r@fasau.ac.ir

[2] Institute of Space Sciences, Magurele-Bucharest, Romania

[3] Future Technology Research Center, College of Future, National Yunlin University of Science and Technology, Taiwan

[4] John School of the Built Environment, Oxford Brookes University, Oxford OX3 0BP, UK

* Corresponding: a.mosavi@brookes.ac.uk

Abstract. The aim of this research is to investigate the relationship between the counts of cases with Covid-19 and the deaths due to it in seven countries that are severely affected by the pandemic. First, the Pearson's correlation is used to determine the relationships among these countries. Then, the factor analysis is applied to categorize these countries based on their relationships.

Keywords: Covid-19, Coronavirus; Correlation, Factor Analysis

1. Introduction

In the winter months of 2019-2020, another type of coronavirus, Covid-19, has been reported in Wuhan [1,2]. This virus has severe destructive effects on the respiratory system. From January to now (April 18, 2020), this epidemic has become epidemic all over the world, and day by day the cases with Covid-19 and the deaths due to Covid-19 are extremely increasing in most countries [4-6]. There are many techniques to analyze natural phenomena including artificial intelligence [21], mathematical and statistical methods [22] such as optimization, deep learning, time series analysis, machine learning, regression modeling, clustering, and numerical analysis [23-25]. Forecasting epidemiology of a specific type of virus can help to obtain a proper estimation of effects and survival of the virus in the environment and providing appropriate management solutions to control it [21,26]. Using machine learning (ML) based techniques can help provide an accurate estimator for outbreak prediction or the mortality rate of a specific virus [2,22,25].

Since Covid-19 has many impacts on the environment, health, society, and economy, the study of the rate of spread of this disease and the comparison of its rate in different countries is essential. There are some researches about the classification of Covid-19 datasets [27,28]. These researches are based on time series analysis, principal component analysis, and fuzzy clustering [30].

This research aims to study the relationships between the counts of the cases with Covid-19 and the deaths due to it in seven countries that are severely affected by this pandemic disease. First, the coefficients of correlation are computed to determine the relationships between these countries. Then, the factor analysis is applied to categorize these countries using the counts of cases and deaths.

2. Material and Method

This section is devoted to studying the research's dataset and to and introducing the factor analysis.

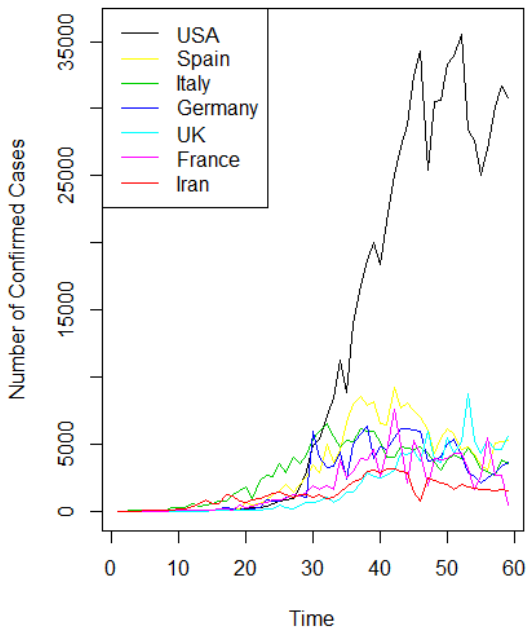
2.1. Dataset

In this work, the counts of the cases with Covid-19 and the deaths due to it in United States America, United Kingdom, Spain, Italy, Iran, Germany, and France from February 22 to April 18 of 2020, are considered from [31] extracted from R-Shiny. Table 1 summarizes the descriptive statistics of dataset containing the mean and the standard deviation. It can be observed that Iran and the United States of America have the minimum and the maximum counts of the cases with Covid-19. Besides, Germany and the United States of America have the minimum and the maximum of deaths due to Covid-19. The plots for the counts of the cases with Covid-19 and the deaths due to it are also demonstrated in Figure 1 (a, b, c and d).

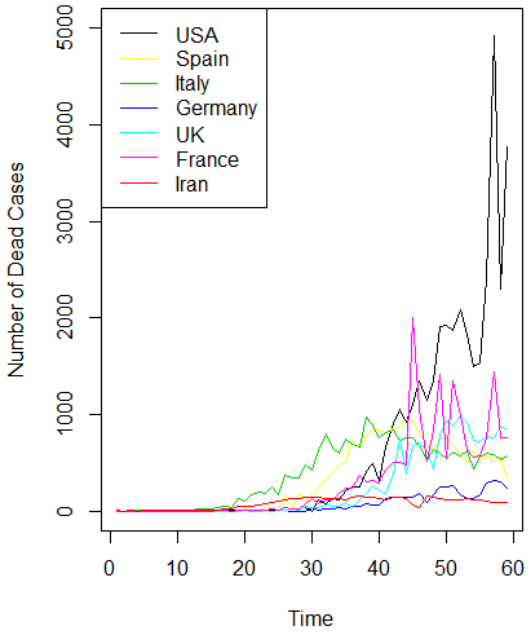
Table 1: The mean and standard deviation for the counts of the cases with Covid-19 and the deaths due to this pandemic disease

	Country	**Mean ± Standard Deviation**
Cases	United States America	11900.8 ± 13327.5
	Spain	3187.6 ± 3016.8
	Italy	2922.6 ± 2056.1
	Germany	2329.2 ± 2245.2
	France	1851.5 ± 1876.0
	United Kingdom	1842.1 ± 2207.7
	Iran	1294.5 ± 921.5
Deaths	United States America	628.0 ± 1013.4
	Italy	385.5 ± 309.5
	Spain	330.1 ± 340.5
	France	316.6 ± 447.3
	United Kingdom	247.1 ± 342.0
	Iran	80.7 ± 55.6

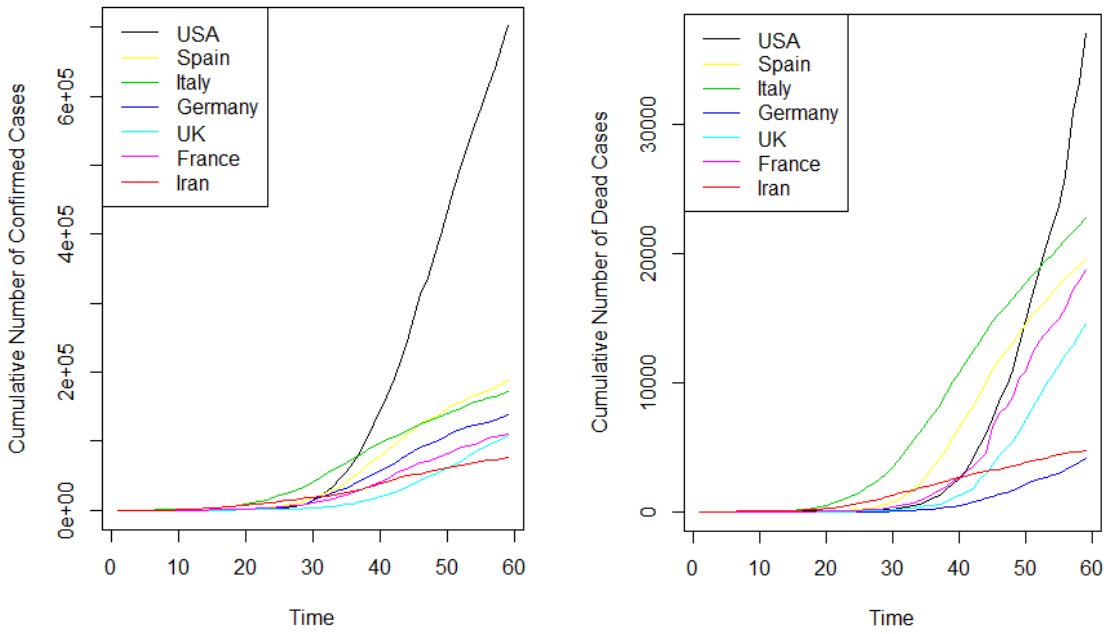
	Germany	69.7 ± 94.9



(a)



(b)



(c) (d)

Figure 1: Counts of the cases (a), counts of the deaths (b), cumulative counts of the cases (c), and cumulative counts of the deaths (d)

The relationships between the rates of the spread of Covid-19 among these countries have been studied using Pearson's coefficient of correlation. As it can be seen in Table 2, all of the values are more than 0.5 and significant, and consequently, there are strong positive relationships between the rates of spread of Covid-19 in all of the countries.

Table 2: Pearson's coefficient of correlation between the rates of spread of Covid-19

		United States America	Spain	Italy	Germany	United Kingdom	France	Iran
Patients	France	1	0.586	0.850	0.766	0.851	0.856	0.673
	Germany	0.586	1	0.654	0.514	0.562	0.550	0.942
	Iran	0.850	0.654	1	0.848	0.884	0.866	0.718
	Italy	0.766	0.514	0.848	1	0.842	0.879	0.567
	Spain	0.851	0.562	0.884	0.842	1	0.929	0.666
	United Kingdom	0.856	0.550	0.866	0.879	0.929	1	0.654
	United States America	0.673	0.942	0.718	0.567	0.666	0.654	1
Deaths	France	1	0.802	0.621	0.677	0.815	0.804	0.716
	Germany	0.802	1	0.565	0.629	0.764	0.927	0.885
	Iran	0.621	0.565	1	0.934	0.794	0.555	0.449
	Italy	0.677	0.629	0.934	1	0.892	0.626	0.508
	Spain	0.815	0.764	0.794	0.892	1	0.743	0.627
	United Kingdom	0.804	0.927	0.555	0.626	0.743	1	0.889
	United States America	0.716	0.885	0.449	0.508	0.627	0.889	1

* p-value < 0.001

2.2. Principles of Factor Analysis

Factor analysis (FA) as a popular multivariate statistical technique transforms some dependent features into some other features called factors such that the first factors of this transformation have the main information of the first dataset [32,33]. In other words, FA is used to convert a dataset with high dimensions to a dataset with lower dimensions, by considering minimum factors such that the dimension of the converted dataset is decreased. FA focuses on the correlations of variables such that the variables in a factor are highly correlated with each other and the variables in different factors are highly uncorrelated with each other. In applications, the number of the main factors in FA is usually considered as the number of the eigen-values of the correlation's matrix with the values larger than one. To investigate the suitability of FA, the Kaiser-Meyer-Olkin (KMO) index is used. The KMO>0.8 verifies the accuracy of FA.

Assume $\boldsymbol{X} = \left(X_1, \dots, X_p\right)^T$ is a random vector. Denote

$$\boldsymbol{\mu} = E(\boldsymbol{X}) = \left(\mu_1, \dots, \mu_p\right)^T, \qquad (1)$$

and

$$\Sigma = Var(\boldsymbol{X}) = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1p} \\ \cdots & \cdots & \cdots \\ \sigma_{p1} & \cdots & \sigma_{pp} \end{bmatrix}, \qquad (2)$$

as the mean vector and covariance matrix of $\boldsymbol{X}$.

The equation of factor analysis with m factors ($m \leq p$) is presented by

$$\boldsymbol{X} - \boldsymbol{\mu} = L\boldsymbol{F} + \boldsymbol{\varepsilon}, \qquad (3)$$

such that

$$\mathrm{L} = \begin{bmatrix} l_{11} & \cdots & l_{1m} \\ \cdots & \cdots & \cdots \\ l_{p1} & \cdots & l_{pm} \end{bmatrix}, \quad (4)$$

$$\boldsymbol{F} = (F_1, \ldots, F_m)^T, \quad (5)$$

and

$$\boldsymbol{\varepsilon} = \left(\varepsilon_1, \ldots, \varepsilon_p\right)^T. \quad (6)$$

The matrix L and the vectors $\boldsymbol{F}$ and $\boldsymbol{\Psi}$ are called the factor loading matrix, the factors and errors, respectively.

This model can be rewritten by

$$X_i - \mu_i = \sum_{j=1}^{m} l_{ij} F_j + \varepsilon_i, \quad i = 1, \ldots., p, \quad (7)$$

such that l_{ij} is named as the loading of X_i on the factor F_j.

In orthogonal factor analysis, we have

$$Cov(\boldsymbol{X}, \boldsymbol{F}) = L, \quad (8)$$

and

$$\Sigma = \mathrm{L}\mathrm{L}^T + \Psi, \quad (9)$$

where

$$\Psi = Var(\boldsymbol{\varepsilon}). \quad (10)$$

Consequently,

$$Var(X_i) = \sum_{j=1}^{m} l_{ij}^2 + \Psi_i, \quad (11)$$

and

$$Cov(X_i, X_j) = \sum_{k=1}^{m} l_{ik}\, l_{jk}. \quad (12)$$

$\sum_{j=1}^{m} l_{ij}^2$ is determines the proportion of $Var(X_i)$ that can be explained by the factors F_1, …, F_m.

The main aim of factor analysis is to find the values of the loadings. To compute the matrices L and Ψ, different approaches such as principal component and maximum likelihood can be applied. The principal component approach uses eigen-values and eigen-vectors to decompose the matrix Σ to find the matrix L. Maximum likelihood approach computes the likelihood and then optimizes it to find the matrices L and Ψ.

When the loading values are estimated, we can consider loading plots. Loading plots can be used to

- Study the correlations between variables
- Categorize and Classify the variables
- Detect the number of factors

In the loading plot, the angle between two variables (θ) determines the correlation (r) between them (for example, see Figure 2). $\theta = 90^0$ verifies that two variables are uncorrelated ($r = 0$). The cases $\theta = 0^0$ and $\theta = 180^0$ refer to exact positive and negative linear relationships, respectively.

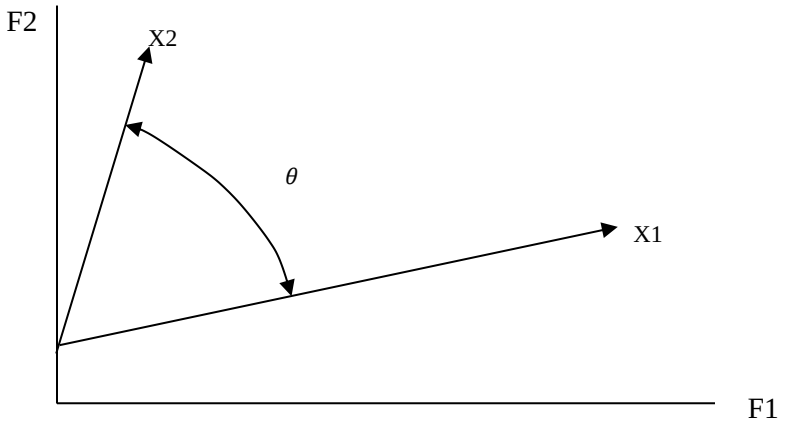


Figure 2: Loading plot in FA technique

3. Results

This section reports the results of the FA approach to classify the countries based on research's variables. It should be noted that the number of the main factors in FA was considered as the number of the eigen-values of the correlation's matrix with the values larger than one. Moreover, the KMO values were more than 0.8 that verify the accuracy of FA approach.

3.1. Counts of Cases with Covid-19

The results of FA technique to categorize the research countries, on basis of the counts of the cases with Covid-19, are provided in Figure 3. The outputs demonstrate the statistical differences between the relationships among the countries and we can categorize the countries into the following classes:

First-class: Iran, France, Spain, Germany, Italy.

Second class: United Kingdom, United States America.

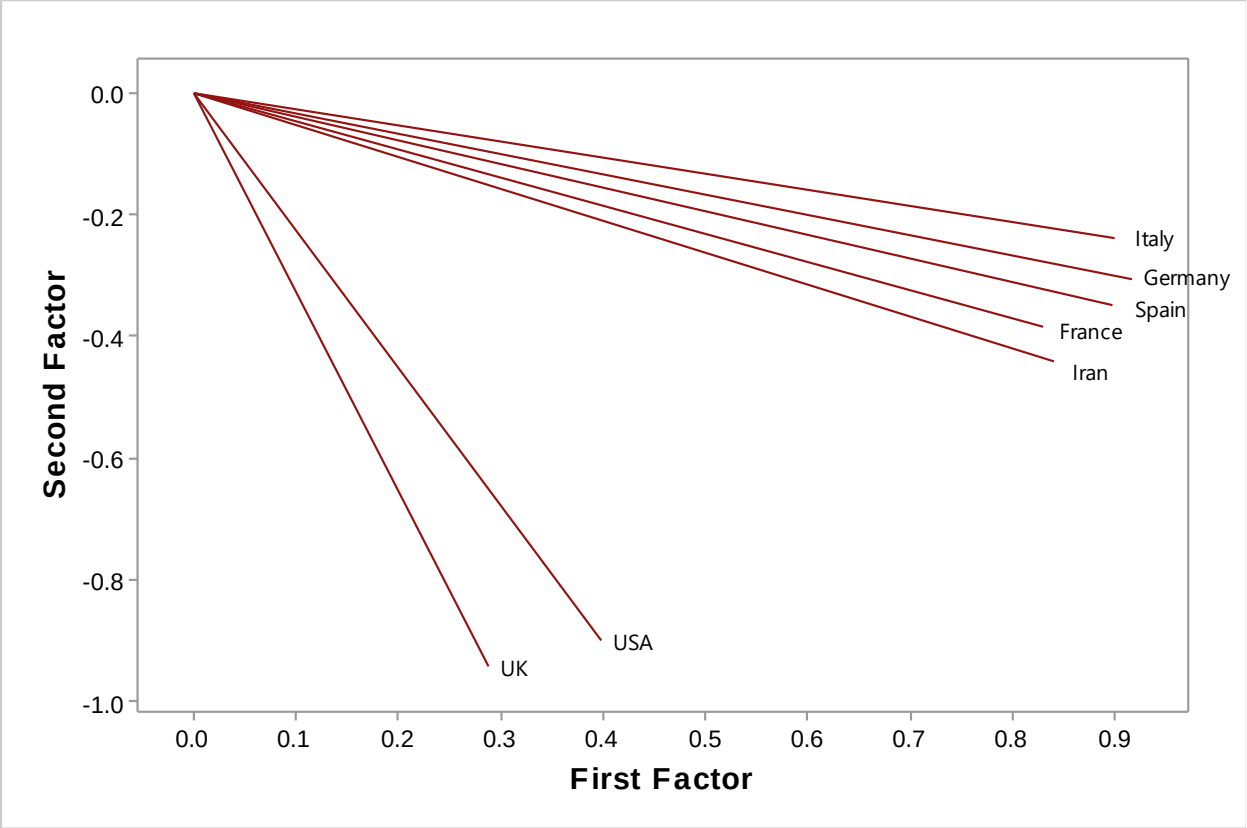


Figure 3: FA technique to categorize the countries in basis of the counts of the cases with Covid-19

3.2. Counts of the Deaths Due to Covid-19

The results of FA technique to categorize the research countries, on basis of the counts of the deaths due to Covid-19, are provided in Figure 4. The outputs demonstrate the statistical differences between the relationships among the countries and we can categorize the countries into the following classes:

First-class: France, United Kingdom, Germany and United States America.

Second class: Iran, Italy and Spain.

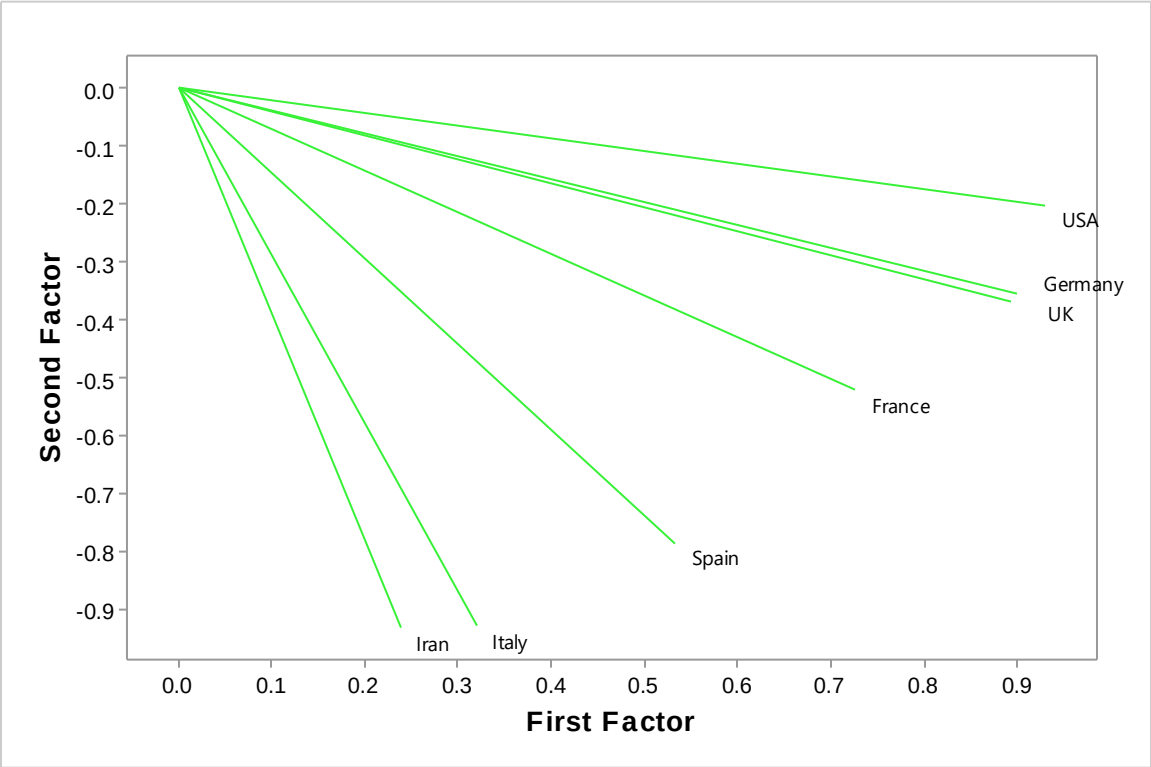


Figure 4: FA technique to categorize the countries in basis of the counts of the deaths due to Covid-19

3.3. Cumulative Counts of the Cases with Covid-19

The results of FA technique to categorize the research countries, on basis of the cumulative counts of the cases with Covid-19, are provided in Figure 5. The outputs demonstrate the statistical differences between the relationships among the countries and we can categorize the countries into the following classes:

First class: France, Spain, Germany, Iran and Italy.

Second class: United Kingdom and United States America.

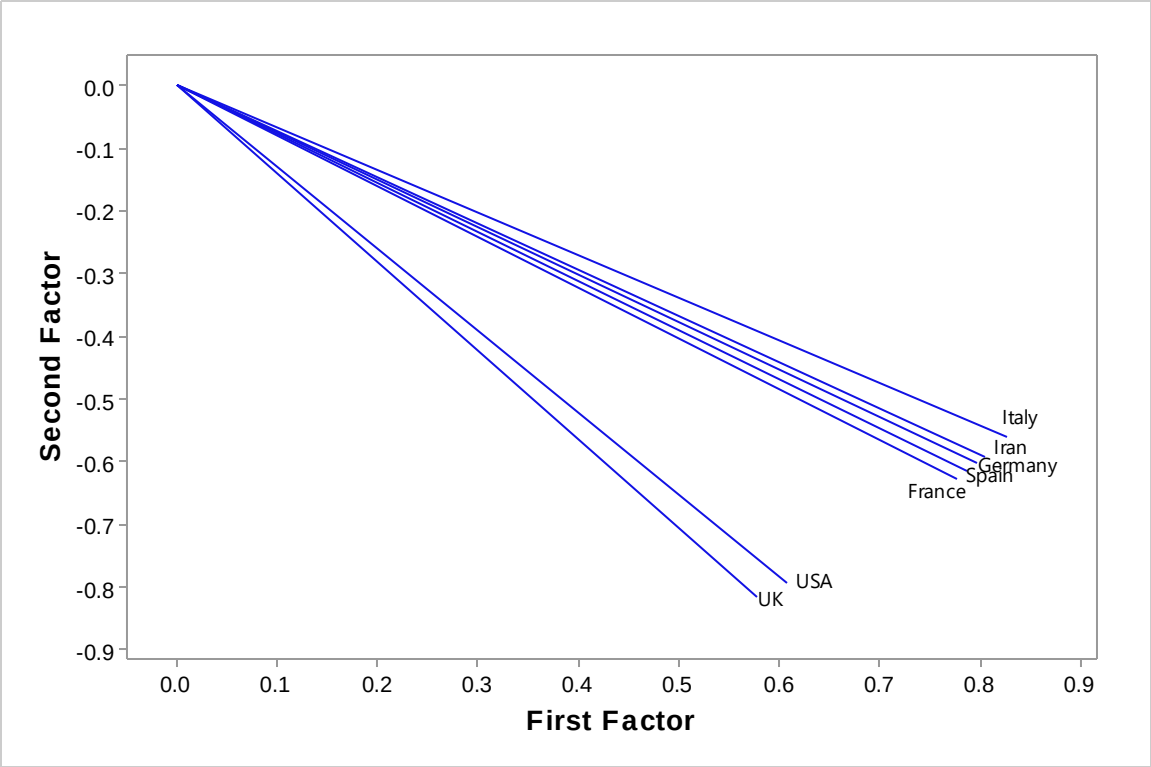


Figure 5: FA technique to categorize the countries on basis of the cumulative counts of the cases with Covid-19

3.4. Cumulative Counts of the Deaths Due to Covid-19

The results of FA technique to categorize the research countries, in basis of the cumulative counts of the deaths due to Covid-19, are provided in Figure 6. The outputs demonstrate the statistical differences between the relationships among the countries and we can categorize the countries into the following classes:

First-class: France, United Kingdom, Germany and United States America.

Second class: Iran, Italy and Spain.

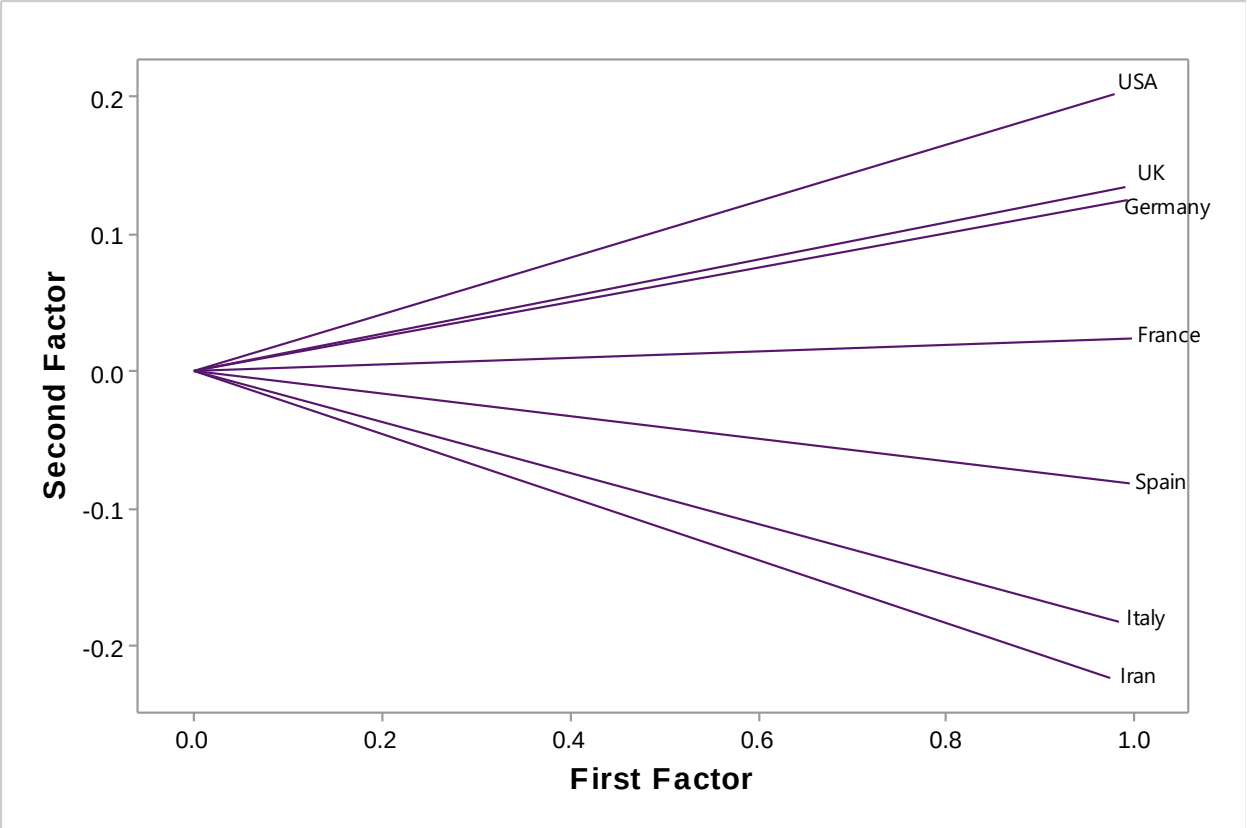


Figure 6: FA technique to categorize the countries on basis of the cumulative counts of the deaths due to Covid-19

4. Conclusion

Since Covid-19 has many impacts on the environment, health, society, and economy, the study of the rate of spread of this disease and the comparison of its rate in different countries is essential. This research aimed to study the cases with Covid-19 and the deaths due to this pandemic disease in seven countries that are severely affected by this pandemic disease. The cases and the deaths in United States America, United Kingdom, Spain, Italy, Iran, Germany, and France from February 22 to April 18 of 2020, were considered. First, the coefficients of correlation were computed to determine the relationships among these countries. The outputs showed that there were strong positive relationships between the rates of spread in all the countries. Then, the factor analysis was applied to categorize the countries on basis of the counts and deaths. For the cases with Covid-19, the United Kingdom and United States America were similarly distributed to each other and were differently distributed from other countries. Also, for the deaths, Iran, Italy and Spain were similarly distributed to each other and were differently distributed from other countries. For future works, the authors suggest classifying the Covid-19 datasets of more regions

based on FA technique or apply this technique to classify the regions for other epidemic or pandemic diseases.

Authors Contribution: Mohammad Reza Mahmoudi: Conceptualization, Investigation, Data curation, Validation, Methodology, Software, Writing- Original draft preparation; Dumitru Baleanu: Conceptualization, Supervision, Visualization, Writing - review editing; Shahab S. Band: Validation, Visualization, Software, Writing - review editing; Amir Mosavi: Visualization, Writing - review editing.

Conflicts of Interest: The authors declare no conflict of interest.

Funding: This research received no funding.

References

1. Pinter, G.; Felde, I.; Mosavi, A.; Ghamisi, P.; Gloaguen, R.J.M. COVID-19 pandemic prediction for Hungary; a hybrid machine learning approach. **2020**, *8*, 890.
2. Mahmoudi, M.R.; Heydari, M.H.; Qasem, S.N.; Mosavi, A.; Band, S.S.J.A.E.J. Principal component analysis to study the relations between the spread rates of COVID-19 in high risks countries. **2021**, *60*, 457-464.
3. Lu, R.; Zhao, X.; Li, J.; Niu, P.; Yang, B.; Wu, H.; Wang, W.; Song, H.; Huang, B.; Zhu, N.J.T.l. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. **2020**, *395*, 565-574.
4. Sun, J.; He, W.-T.; Wang, L.; Lai, A.; Ji, X.; Zhai, X.; Li, G.; Suchard, M.A.; Tian, J.; Zhou, J.J.T.i.m.m. COVID-19: epidemiology, evolution, and cross-disciplinary perspectives. **2020**, *26*, 483-495.
5. Park, M.; Cook, A.R.; Lim, J.T.; Sun, Y.; Dickens, B.L.J.J.o.c.m. A systematic review of COVID-19 epidemiology based on current evidence. **2020**, *9*, 967.
6. Bulut, C.; Kato, Y.J.T.j.o.m.s. Epidemiology of COVID-19. **2020**, *50*, 563-570.
7. Burke, R.M.; Midgley, C.M.; Dratch, A.; Fenstersheib, M.; Haupt, T.; Holshue, M.; Ghinai, I.; Jarashow, M.C.; Lo, J.; McPherson, T.D.J.M., et al. Active monitoring of persons exposed to patients with confirmed COVID-19—United States, January–February 2020. **2020**, *69*, 245.
8. Hunter, D.J.J.N.E.J.o.M. Covid-19 and the stiff upper lip—the pandemic response in the United Kingdom. **2020**, *382*, e31.
9. Razai, M.S.; Doerholt, K.; Ladhani, S.; Oakeshott, P.J.B. Coronavirus disease 2019 (covid-19): a guide for UK GPs. **2020**, *368*.
10. Lillie, P.J.; Samson, A.; Li, A.; Adams, K.; Capstick, R.; Barlow, G.D.; Easom, N.; Hamilton, E.; Moss, P.J.; Evans, A.J.J.o.I. Novel coronavirus disease (Covid-19): the first two patients in the UK with person to person transmission. **2020**, *80*, 578-606.
11. Legido-Quigley, H.; Mateos-García, J.T.; Campos, V.R.; Gea-Sánchez, M.; Muntaner, C.; McKee, M.J.T.l.p.h. The resilience of the Spanish health system against the COVID-19 pandemic. **2020**, *5*, e251-e252.
12. Lazzerini, M.; Putoto, G.J.T.L.G.H. COVID-19 in Italy: momentous decisions and many uncertainties. **2020**, *8*, e641-e642.
13. Onder, G.; Rezza, G.; Brusaferro, S.J.J. Case-fatality rate and characteristics of patients dying in relation to COVID-19 in Italy. **2020**, *323*, 1775-1776.
14. Remuzzi, A.; Remuzzi, G.J.T.l. COVID-19 and Italy: what next? **2020**, *395*, 1225-1228.

15. Takian, A.; Raoofi, A.; Kazempour-Ardebili, S.J.L. COVID-19 battle during the toughest sanctions against Iran. **2020**, *395*, 1035.
16. Rothe, C.; Schunk, M.; Sothmann, P.; Bretzel, G.; Froeschl, G.; Wallrauch, C.; Zimmer, T.; Thiel, V.; Janke, C.; Guggemos, W.J.N.E.j.o.m. Transmission of 2019-nCoV infection from an asymptomatic contact in Germany. **2020**, *382*, 970-971.
17. Amrane, S.; Tissot-Dupont, H.; Doudier, B.; Eldin, C.; Hocquart, M.; Mailhe, M.; Dudouet, P.; Ormières, E.; Ailhaud, L.; Parola, P.J.T.m., et al. Rapid viral diagnosis and ambulatory management of suspected COVID-19 cases presenting at the infectious diseases referral hospital in Marseille, France,-January 31st to March 1st, 2020: A respiratory virus snapshot. **2020**, *36*, 101632.
18. Stoecklin, S.B.; Rolland, P.; Silue, Y.; Mailles, A.; Campese, C.; Simondon, A.; Mechain, M.; Meurice, L.; Nguyen, M.; Bassi, C.J.E. First cases of coronavirus disease 2019 (COVID-19) in France: surveillance, investigations and control measures, January 2020. **2020**, *25*, 2000094.
19. Gautret, P.; Lagier, J.-C.; Parola, P.; Meddeb, L.; Mailhe, M.; Doudier, B.; Courjon, J.; Giordanengo, V.; Vieira, V.E.; Dupont, H.T.J.I.j.o.a.a. Hydroxychloroquine and azithromycin as a treatment of COVID-19: results of an open-label non-randomized clinical trial. **2020**, *56*, 105949.
20. Fanelli, D.; Piazza, F.J.C., Solitons; Fractals. Analysis and forecast of COVID-19 spreading in China, Italy and France. **2020**, *134*, 109761.
21. Ardabili, S.F.; Mosavi, A.; Ghamisi, P.; Ferdinand, F.; Varkonyi-Koczy, A.R.; Reuter, U.; Rabczuk, T.; Atkinson, P.M.J.A. Covid-19 outbreak prediction with machine learning. **2020**, *13*, 249.
22. Maleki, M.; Mahmoudi, M.R.; Wraith, D.; Pho, K.-H.J.T.m.; disease, i. Time series modelling to forecast the confirmed and recovered cases of COVID-19. **2020**, *37*, 101742.
23. Kavadi, D.P.; Patan, R.; Ramachandran, M.; Gandomi, A.H.J.C., Solitons; Fractals. Partial derivative nonlinear global pandemic machine learning prediction of covid 19. **2020**, *139*, 110056.
24. Heydari, M.; Avazzadeh, Z.; Mahmoudi, M.J.C., Solitons; Fractals. Chebyshev cardinal wavelets for nonlinear stochastic differential equations driven with variable-order fractional Brownian motion. **2019**, *124*, 105-124.
25. Ardabili, S.; Mosavi, A.; Band, S.S.; Varkonyi-Koczy, A.R.J.m. Coronavirus Disease (COVID-19) Global Prediction Using Hybrid Artificial Intelligence Method of ANN Trained with Grey Wolf Optimizer. **2020**.
26. Maleki, M.; Mahmoudi, M.R.; Heydari, M.H.; Pho, K.-H.J.C., Solitons; Fractals. Modeling and forecasting the spread and death rate of coronavirus (COVID-19) in the world using time series models. **2020**, *140*, 110151.
27. Muhammad, S.; Long, X.; Salman, M.J.S.o.t.t.e. COVID-19 pandemic and environmental pollution: A blessing in disguise? **2020**, *728*, 138820.
28. Gautam, S.; Hens, L. COVID-19: Impact by and on the environment, health and economy. Springer: 2020.
29. Mahmoudi, M.R.; Heydari, M.H.; Pho, K.-H.J.A.E.J. Fuzzy clustering to classify several regression models with fractional Brownian motion errors. **2020**, *59*, 2811-2818.
30. Mahmoudi, M.R.; Baleanu, D.; Mansor, Z.; Tuan, B.A.; Pho, K.-H.J.C., Solitons; Fractals. Fuzzy clustering method to compare the spread rate of Covid-19 in the high risks countries. **2020**, *140*, 110230.
31. Salehi, M.; Arashi, M.; Bekker, A.; Ferreira, J.; Chen, D.-G.; Esmaeili, F.; Frances, M.J.F.i.P.H. A synergetic R-Shiny portal for modeling and tracking of COVID-19 data. **2020**, *8*.
32. Martin, N.; Maes, H. *Multivariate analysis*; Academic press London: 1979.
33. Johnson, K.G.; Mollenhauer, K.; Tschöke, H. *Handbook of diesel engines*; Springer Science & Business Media: 2010.