Journal of Autonomous Intelligence (2024) Volume 7 Issue 1
doi: 10.32629/jai.v7i1.648

ORIGINAL RESEARCH ARTICLE

Assessment of first-phase COVID-19 pandemic in Europe using hierarchical clustering based on principal components analysis

Sanjay Kumar[1,*], Evrim Oral[2]

[1] *Department of Statistics, Central University of Rajasthan, Bandarsindri, Kishangarh, Ajmer 305817, Rajasthan, India*

[2] *Biostatistics Program, School of Public Health, LSU Health Sciences Center, New Orleans, LA 70112, USA*

*** Corresponding author:** Sanjay Kumar, sanjay.kumar@curaj.ac.in

ABSTRACT

It is of great interest for researchers to assess the COVID-19 pandemic in Europe. Grouping of COVID-19-affected regions is an effective way to monitor and optimize planning to combat the disease. This paper applied hierarchical clustering based on principal components analysis (HCPCA) to COVID-19 data from affected European countries. Considering several attribute indices, we obtained a new set of indicators using principal components analysis to aggregate and reduce the dimension of attribute indices of affected countries. Further, we obtained groups of affected countries subject to their similarity using hierarchical clustering to the reduced observations of new attributes indices of these countries. This study aims to group European countries with similar epidemic severity using some presumed attribute indices. The study is limited up to 24 May 2020, to assess if the outputs of the study could help governments, administrators, World Health Organization (WHO), healthcare service professionals, and other decision-makers to optimize their policies and plan their regulations in the country level requirements so that transmission of infections, deaths, critical conditions of patients could be minimized. For this purpose, we used hierarchical clustering using principal components analysis to obtain better clusters of countries with similar epidemic severity.

Keywords: principal components analysis; dendrogram; hierarchical clustering; data science; data mining

ARTICLE INFO

Received: 21 May 2023
Accepted: 6 August 2023
Available online: 16 October 2023

1. Introduction

Coronavirus disease-2019 (COVID-19) is caused by a severe acute respiratory syndrome coronavirus 2 (SARS COV-2). It was first reported in December 2019 by the Municipal Health Commission in Wuhan, China. Further, it started spreading in other countries in the world. To date, millions of confirmed positive cases and deaths were reported. Liu et al.[1] studied the disease transmission pattern in China in various age-group populations. Based on the seriousness of the transmission of infections, and the occurrence of deaths, WHO declared it to be an outbreak in January 2020. As of 13 March 2020, many new cases were reported in European countries, so WHO considered Europe to be the epicenter of the COVID-19 pandemic, where the infection was aggressively spreading through Italy.

Data science has become a leading field over the last few years, which covers almost every industry sector and is gradually growing. Tools related to data science play a vital role in assessing the status of pandemics in a specific region, such as COVID-19, and can be used to optimize policies and planning to stop the spreading of such infections and reduce deaths and critical cases. The COVID-19 pandemic provides big data with different structures, which must be

analyzed quickly and effectively. Healthcare practitioners can access such big data of vital patients to optimize treatment facilities using data mining, machine learning, etc. Data science is very valuable in tracing hot spots and controlling the spread of COVID-19 infections by analyzing multi-parametric data through numerous methods like principal components analysis (PCA), hierarchical clustering (HC), and neural network analysis (NNA)[2–4].

When we deal with a multidimensional dataset with several variables, PCA is an appropriate technique that can be used to reduce the dimension of the dataset into a few important variables containing the most information in the dataset. Therefore, in this paper, we applied PCA to obtain new sets of indices having a linear combination of the original pieces of information. Further, we applied hierarchical clustering using the PCA approach to group 44 European countries affected by COVID-19 infections. Then, we applied the K-means algorithm to have better groups of similar countries. Our analysis is limited up to 24 May 2020, to assess if such analysis could help control the severe situation in the first phase of COVID-19 in European countries. The major objective of this study is to help governments, administrators, and healthcare service professionals optimize their monitoring techniques, making appropriate policies and planning in these affected European countries according to their country-level regulations. These regulations will help reduce the transmission of COVID-19 infections, cases, and deaths.

2. Materials and methods

2.1. Study area

We included 44 European countries in the study listed in **Table 1**.

Table 1. A list of names of European countries and their assumed codes.

Country	Code	Country	Code	Country	Code	Country	Code
Russia	C1	Poland	C12	Hungary	C23	Latvia	C34
Spain	C2	Ukraine	C13	Greece	C24	Albania	C35
Italy	C3	Romania	C14	Bulgaria	C25	Andorra	C36
France	C4	Austria	C15	Bosnia and Herzegovina	C26	San Marino	C37
Germany	C5	Denmark	C16	Croatia	C27	Malta	C38
Belgium	C6	Serbia	C17	North Macedonia	C28	Channel Islands	C39
Belarus	C7	Czechia	C18	Estonia	C29	Isle of Man	C40
Sweden	C8	Norway	C19	Iceland	C30	Montenegro	C41
Switzerland	C9	Moldova	C20	Lithuania	C31	Faeroe Islands	C42
Portugal	C10	Finland	C21	Slovakia	C32	Gibraltar	C43
Ireland	C11	Luxembourg	C22	Slovenia	C33	Liechtenstein	C44

We did not include the United Kingdom, Netherlands, Monaco, and Vatican City in our study because information on a few of the indices used in this study was not available for these countries.

2.2. Methodology

We divided this sub-section into three stages (I–III). Collection of data, summary statistics, and graphical representations are given in Stage I; execution of data mining techniques to the data set is provided in Stage II, and interpretations on variations among clusters using box plots are given in Stage III.

Stage I: Data collection and graphical representation of its characteristics

The samples on COVID-19 cases from European countries were collected till 24 May 2020 (05:02 GMT) via Worldometer (https://www.worldometers.info/coronavirus[5]). In our study, we included six

attribute indices: tested cases (T_d), confirmed cases (C_n), deaths (D_t), recovered cases (R_e), critical cases (C_r), and the percentage of the population (P_p) since the severity of situations of COVID-19 in the countries mostly depend upon these attribute indices. A summary of basic statistics, such as values of minimum, maximum, median, mean, first and third quartiles, and standard deviation (SD) for each of the indices, are given in **Table 2**. We also showed the nature of data points through graphical representations using box plots and histograms in **Figure 1**. We did not remove extreme observations from the data set because these observations may help administrators, healthcare service professionals, and others understand severe situations at the country level.

Table 2. Summary statistics of COVID-19 status of European per one million (1 M) population.

Summary	Tests	Confirmed	Deaths	Recovered	Critical	Percentage of population
Min.	4563.0	275.0	0.0	116.0	0.00	0.005
1st Qu.	19,575.0	678.0	22.5	498.0	1.75	0.154
Median	42,682.0	1782.0	49.0	928.0	4.00	0.728
Mean	52,927.0	2740.0	162.1	1836.0	9.50	2.014
3rd Qu.	61,200.0	3746.0	185.5	2453.0	11.00	1.372
Max.	187,393.0	19,603.0	1238.0	8453.0	71.00	19.520
SD	45,969.86	3321.36	258.1	2056.1	13.47	3.728

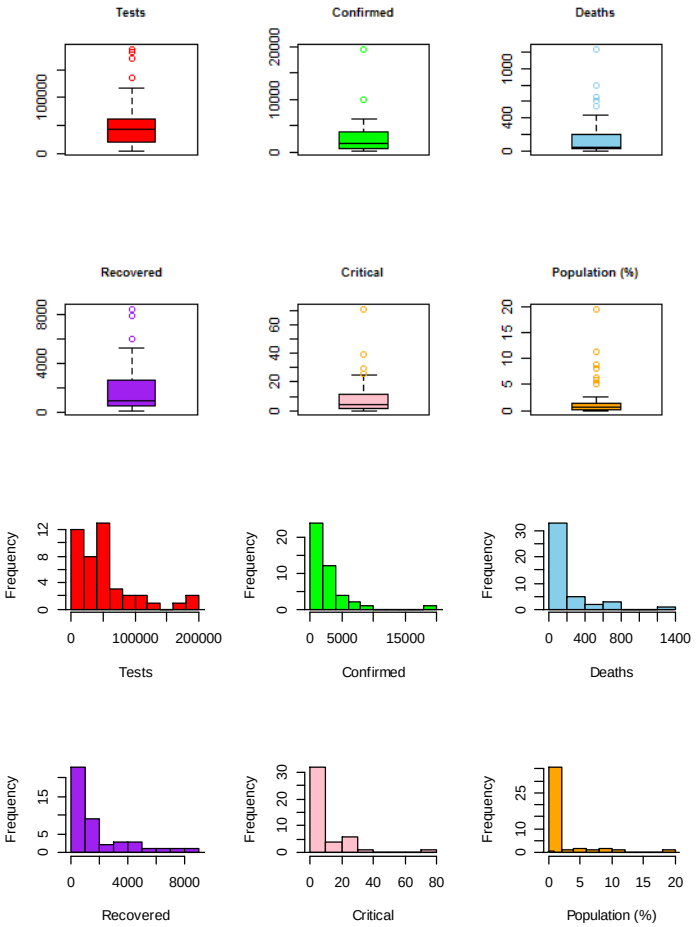

Figure 1. Box plots and histograms for the status of the four cases of COVID-19 in 44 affected countries in Europe (red, green, sky blue, purple, light red & orange colors represent the cases related to tests, confirmed, deaths, recovered, critical and population (%), respectively).

Stage II: Analysis using data mining techniques

Data mining is the process of finding hidden features of big data. We executed some data mining techniques, such as Pearson's correlation coefficient matrix, PCA and HCPCA, to get valuable information.

1) Pearson correlation matrix

A correlation matrix helps to understand the strength of a linear relationship. We obtained a correlation (ρ) matrix of attribute indices to assess the relationships between them and to evaluate their significance; we obtained corresponding p-values for testing the hypothesis $H_0: \rho = 0$ against $H_A: \rho \neq 0$ (**Table 3**). We followed Ratner[6], who has studied the nature of correlation and showed that values between 0 and 0.30 (0 and −0.30), between 0.30 and 0.70 (−0.30 and −0.70), and between 0.70 and 1.0 (−0.70 and −1.0) indicate a weak positive (negative), moderately positive (negative) and strongly positive (negative) linear relationship, respectively.

Table 3. Correlation matrix among variables and respective *p*-values.

	ρ					
Variables	**Tests**	**Confirmed**	**Death**	**Recovered**	**Critical**	**Population (%)**
Tests	1.00	-	-	-	-	-
Confirmed	0.45	1.00	-	-	-	-
Death	0.14	0.84	1.00	-	-	-
Recovered	0.58	0.86	0.61	1.00	-	-
Critical	–0.12	0.43	0.52	0.31	1.00	-
Population (%)	–0.14	–0.06	0.06	–0.14	0.14	1.00
	***p*-values**					
Variables	**Tests**	**Confirmed**	**Death**	**Recovered**	**Critical**	**Population (%)**
Tests	0.0000	-	-	-	-	-
Confirmed	0.0020	0.0000	-	-	-	-
Death	0.3718	0.0000	0.0000	-	-	-
Recovered	0.0000	0.0000	0.0000	0.0000	-	-
Critical	0.4564	0.0039	0.0003	0.0393	0.0000	-
Population (%)	0.3667	0.7072	0.6873	0.3519	0.3490	0.0000

2) Principal components analysis (PCA)

PCA[7–9] is used to process a data set for dimensionality reduction for further data mining. It is an important data mining technique for data reduction without substantial loss of information. It forms new sets of uncorrelated indices of linear combinations of the original correlated indices. These new uncorrelated indices are called principal components.

Before execution of the PCA, we performed the Kaiser-Mayer-Olkin (KMO) test[10,11] and Bartlett test[11,12] to determine if the PCA is suitable for the data set. By performing the KMO and Bartlett's tests in PCA, one can assess the suitability of the dataset and ensure that the variables exhibit sufficient inter-correlation for meaningful extraction. If the tests indicate a good fit (high KMO value and significant Bartlett's test), it provides confidence in proceeding with PCA. Otherwise, it may be necessary to reevaluate the dataset or consider alternative techniques for data analysis.

When variables are measured in different scales (e.g., kilograms, kilometers, centimeters, …) or when the mean and/or the standard deviation of variables are largely different, scaling the dataset is recommended; otherwise, the dissimilarity measures obtained will be severely affected. Thus, we scaled the data set using R

software (version R i386 3.6.3) for further analysis. We obtained the desired number of principal components based on scree plots (**Figure 2**).

Further, we extracted the principal components and provided the corresponding components scores and related explanation of PCA in **Table 4**. In **Table 4**, RC1 and RC2 represent rotating component loadings, h2 represents common component variance.

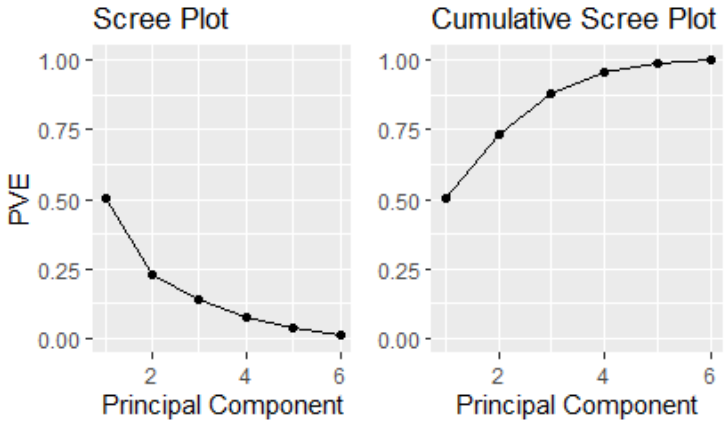

Figure 2. Scree plots for obtaining an optimum number of principal components.

Table 4. Standardized loadings based on correlation matrix and related explanation of PCA.

Indices	RC1	RC2	h2	Aspects	RC1	RC2
T_d	0.44	–0.72	0.70			
C_n	0.96	–0.12	0.94			
D_t	0.87	0.22	0.81			
R_e	0.87	–0.34	0.87	SS loadings	2.99	1.41
C_r	0.60	0.58	0.69	Proportion var	0.50	0.23
P_p	0.00	0.62	0.38	Cumulative var	0.50	0.73

3) Validation of clustering in the data set

Cluster analysis (CA) is one of the important data mining techniques to discover hidden predictive knowledge from big data. It identifies clusters of similar observations within a data set of interest[7,13,14].

Before executing CA, we performed the Hopkins test for validation. We also supported this validation with the visual assessment of the cluster tendency (VAT) approach in **Figure 3**.

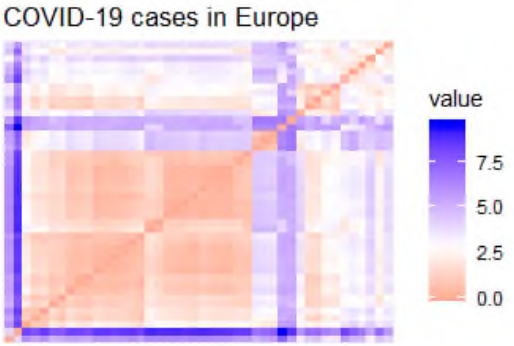

Figure 3. Dissimilarity matrix image for VAT.

One can perform hierarchical clustering (HC) using various methods such as average-linkage, single-linkage, complete-linkage, and ward methods. It is very important to verify which method best fits the data

set, so we obtained agglomerative coefficients for average linkage, single-linkage, complete linkage, and ward methods.

4) Hierarchical clustering based on principal components (HCPC)

We observed from **Table 3** that significant correlations exist among attribute indices. The results may not be very beneficial if we perform only CA for clustering or grouping the affected European countries. The mixed algorithms of PCA and CA can optimize the results of the clustering process. So, in this paper, using the R software, we performed hierarchical clustering using Ward's method based on the principal components (HCPC).

We obtained a dendrogram (**Figure 4**) for the HCPC using Ward's algorithm to show the clustering of affected countries of Europe. Further, these clusters were used to obtain final and robust results (clusters) after consolidation of K-means. The final and robust result was represented on the factor map obtained from the first two principal components (**Figure 5**). We also showed a three-dimensional graphical representation of the dendrograms on the factor map obtained from the first two principal components in **Figure 6**.

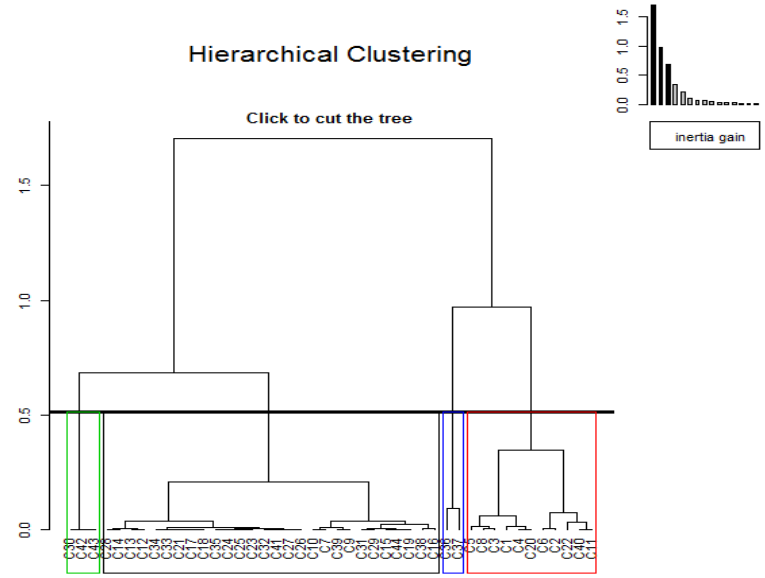

Figure 4. Hierarchical clustering based on principal components analysis (black, red, green & blue colors show cluster 1, 2, 3 & 4, respectively).

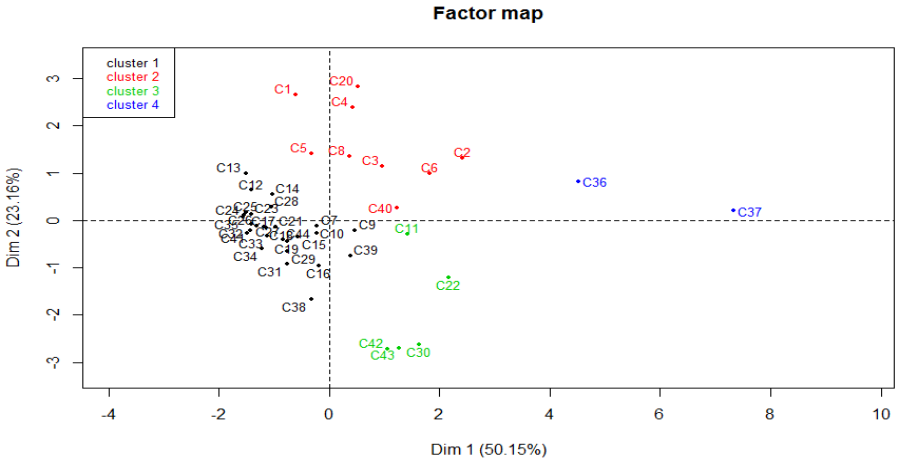

Figure 5. Clusters on factor map (black, red, green & blue colors represent the clusters 1, 2, 3 & 4, respectively).

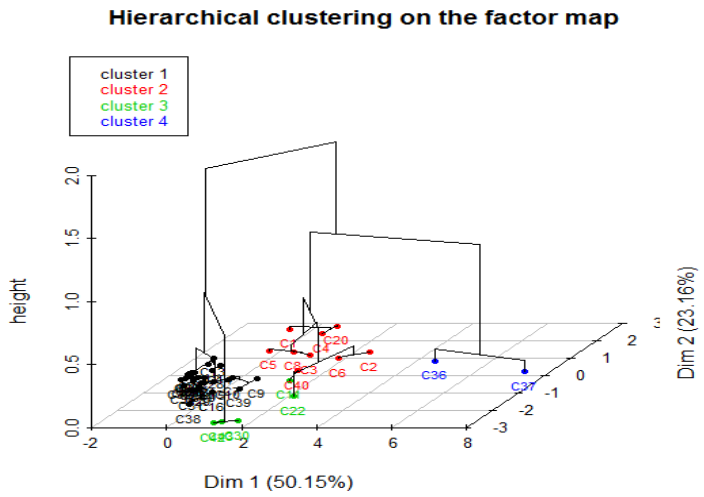

Figure 6. A dendrogram, clusters, and principal components map showing the clustering of countries for the status of COVID-19 (black, red, green & blue colors represent clusters 1, 2, 3 & 4, respectively).

Stage III: Analysis using box plot

Here, we made inferences regarding the variations among clusters using box plots. Here, we considered the median to be the best appropriate measure of central tendency because the distributions of the indices were skewed (**Figure 1**)[15]. Box plots are preferred in such cases for better visual comparisons between clusters and to draw inferences. At this stage, all clusters (C-I, C-II, C-III, and C-IV) of the robust result were compared for each of the indices (except P_p) (**Figure 7**). In **Figure 7**, a long red dotted horizontal line shows the overall median, and a bold black horizontal line indicates the median of data points in each cluster for each index. Our main focus was on the median values of clusters in each case for making inferences. Here, we converted values of the indices related to tested, confirmed, death, recovered, and critical cases into percentages using the formulas: $T_d = \frac{T_d}{Total\ sum\ of\ T_d\ cases} \times 100\%$, $C_n = \frac{C_n}{T_d} \times 100\%$, $D_t = D_t/C_n \times 100\%$, $R_e = R_e/C_n \times 100\%$, and $C_r = C_r/C_n \times 100\%$.

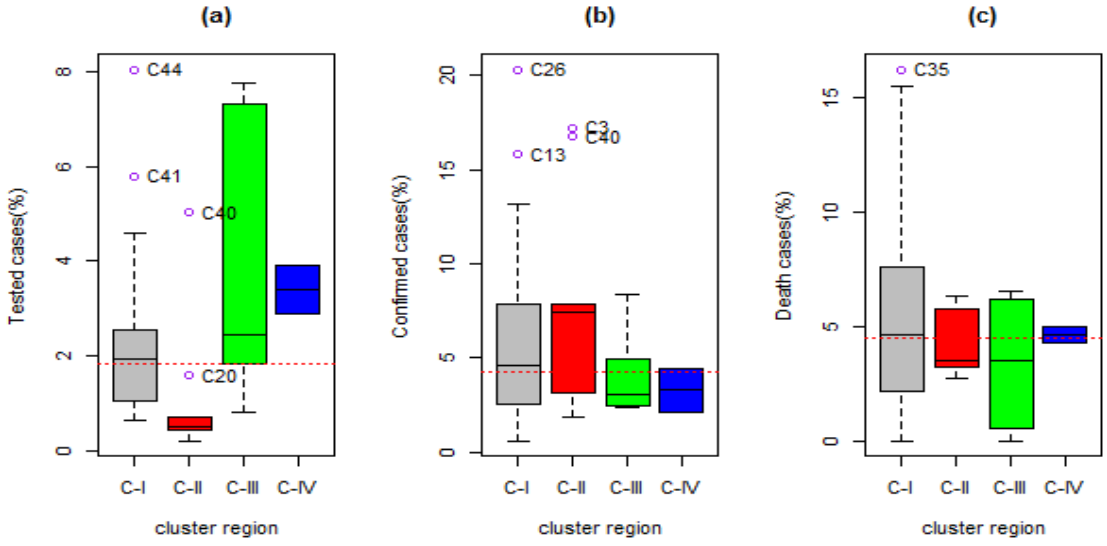

Figure 7. (*Continued*).

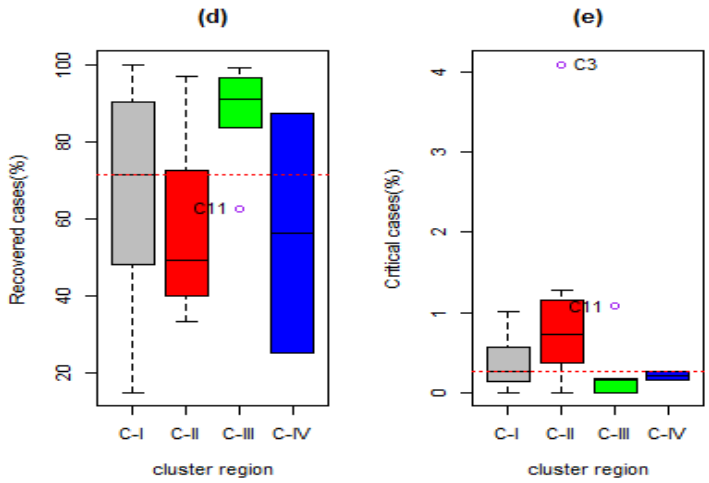

Figure 7. Box plots showing variation among clusters in all cases of COVID-19 (black, red, green & blue colors represent the clusters 1, 2, 3 & 4, respectively).

3. Results

From the ρ-matrix and corresponding p-values in **Table 3**, it was observed that the tested cases showed p-values tend to zero with confirmed and recovered cases which indicates that correlation coefficients are significant while it is much different from zero with critical cases and percentage of the population which indicates that the correlation coefficients are not significant. It suggests that if a country had a high number of tested cases, confirmed and recovered cases were also high. The confirmed cases showed the p-values tend to be zero with the death, recovered, and critical cases, which indicates that correlation coefficients are significant, while it was much different from zero with the percentage of the population, which suggests that the correlation coefficient is not significant. The deaths showed the p-values tend to be zero with the recovered and critical cases, which indicates that the correlation coefficients are significant, while it showed the p-value much different from zero with the population percentage, which indicates that the correlation coefficient is not significant.

The KMO test showed that the overall measure of sampling adequacy (MSA) value was 0.7, and the individual MSA values for T_d, C_n, D_t, R_e, C_r and P_p were 0.6, 0.6, 0.6, 0.7, 0.8 and 0.6, respectively. Thus, the KMO test indicated sampling adequacy for the model and for each of the indices used in the analysis. Further, we performed the Bartlett test, which resulted in $\text{Chi} - \text{square} = 151.93$ with a p-value of 9.997×10^{-25}, indicating that the correlation coefficient matrix was not an identity matrix and hence the execution of PCA for the given data set is suitable.

In **Figure 2**, we examined the elbow point where the proportion of variance explained (PVE) decreased significantly. The scree plots showed the PVE by each principal component, indicating that two principal components are sufficient for the analysis, which explained about 73% (**Table 3**) of the total variance.

From **Table 4**, we observed that the first principal component (PC1) was mainly explained by the indices C_n, D_t, R_e and C_r which explained 50% of the total variance. The PC1 was also positively correlated with all the index variables. The second principal component (PC2) was mainly explained by the indices T_d and P_p which explained 23% of the total variance. The PC2 was negatively correlated with indices T_d, Cn, Re and positively correlated with index variables D_t, C_r and P_p.

The expressions of principal components are given below according to the results obtained in **Table 4**.

$$\text{PC1} = 0.44T_d + 0.96C_n + 0.87D_t + 0.87R_e + 0.60C_r$$

and

$$\text{PC2} = -0.72T_d - 0.12C_n + 0.22D_t - 0.34R_e + 0.58C_r + 0.62P_p$$

For the validation of CA, the Hopkins test yielded a value of 0.22 of Hopkins statistic that was below the threshold value of 0.5. It showed that there exist meaningful clusters in the data set. **Figure 3** confirmed high similarity in red color while low similarities in blue color and showed about four clusters for the data set. The agglomerative coefficients for average-linkage, single-linkage, complete-linkage, and Ward methods were 0.83, 0.77, 0.88, and 0.90, respectively. It can be seen that the Ward method provided the highest value of the agglomerative coefficient. Hence, it is best-fit relative to the average, single, and complete linkage. We summarize our findings from **Figure 4** and **Figure 5** in **Table 5** given below:

Table 5. Summarization of findings from **Figure 4** and **Figure 5**.

From Figure 5				**From Figure 6**			
Cluster I	**Cluster II**	**Cluster III**	**Cluster IV**	**Cluster I**	**Cluster II**	**Cluster III**	**Cluster IV**
Belarus	Russia	Iceland	Andorra	Belarus	Russia	Ireland	Andorra
Switzerland	Spain	Faeroe-Islands	San-Marino	Switzerland	Spain	Luxembourg	San-Marino
Portugal	Italy	Gibraltar	-	Portugal	Italy	Iceland	-
Poland	France	-	-	Poland	France	Faeroe-Islands	-
Ukraine	Germany	-	-	Ukraine	Germany	Gibraltar	-
Romania	Belgium	-	-	Romania	Belgium	-	-
Austria	Sweden	-	-	Austria	Sweden	-	-
Denmark	Ireland	-	-	Denmark	Moldova	-	-
Serbia	Moldova	-	-	Serbia	Isle of Man	-	-
Czechia	Luxembourg	-	-	Czechia	-	-	-
Norway	Isle of Man	-	-	Norway	-	-	-
Finland	-	-	-	Finland	-	-	-
Hungary	-	-	-	Hungary	-	-	-
Greece	-	-	-	Greece	-	-	-
Bulgaria	-	-	-	Bulgaria	-	-	-
Bosnia and Herzegovina	-	-	-	Bosnia and Herzegovina	-	-	-
Croatia	-	-	-	Croatia	-	-	-
North Macedonia	-	-	-	North Macedonia	-	-	-
Estonia	-	-	-	Estonia	-	-	-
Lithuania	-	-	-	Lithuania	-	-	-
Slovakia	-	-	-	Slovakia	-	-	-
Slovenia	-	-	-	Slovenia	-	-	-
Latvia	-	-	-	Latvia	-	-	-
Albania	-	-	-	Albania	-	-	-
Malta	-	-	-	Malta	-	-	-
Channel Islands	-	-	-	Channel Islands	-	-	-
Montenegro	-	-	-	Montenegro	-	-	-
Liechtenstein	-	-	-	Liechtenstein	-	-	-

As can be seen from **Table 5**, the countries Ireland and Luxembourg moved from cluster II (C-II) to cluster III (C-III). Here individuals were represented with different colors according to the cluster in which they lie. A square showed the centroid of each cluster. There was no change found in cluster I (C-I) and cluster IV (C-IV). From **Figure 5**, it can be seen that all four clusters are separated very well on the first two principal components. We found refined results of Ward's method after the K-means consolidation process. However, only two (Ireland and Luxembourg) out of 44 countries moved from one cluster (C-II) to another cluster (C-III), and this proves the stability of our results. In **Figure 6**, the principal components map, dendrograms, and the clustering issue from these dendrograms yield different information that are

superimposed for better visualization of the data set. It shows Ireland and Luxembourg are oriented toward the other components of C-III.

Figure 7a showed that the median value of tested cases of C-IV was greater than that of C-I, C-II, and C-III, while this value was least in C-II among all the clusters. Similarly, it can be seen that the median value of C-III was greater than that of C-I. The C-II was least dispersed, while C-III was more dispersed among all the clusters. That is, many countries in C-II and C-III had similar situations of testing percentages at certain parts of the scale, while in other parts of the scale, the values were more variable in percentage. The median value of tested cases was the lowest in C-II and the highest in C-IV than that of all the clusters' overall median value. Further, **Figure 7a** showed Montenegro and Liechtenstein to be outliers in C-I and Moldova and Isle of Man in C-II, which means that these countries had a high percentage of tested cases.

Figure 7b showed that the median value of confirmed cases of C-II was greater than that of C-I, C-III, and C-IV, while this value was least in C-III among all the clusters. Similarly, it can be seen that the median value of C-I was greater than that of C-III. The C-III and C-IV were approximately equally dispersed, while C-I and C-II were more but reasonably similar dispersed; however, the overall range in C-I was greater than C-II. Many countries in C-III and C-IV had similar situations of confirmed cases. The median value of confirmed cases was lower in C-II and C-IV while high in C-I and C-II than that of all the clusters' overall median value. Further, **Figure 7b** showed Ukraine, Bosnia, and Herzegovina as outliers in C-I and Italy and Isle of Man in C-II, meaning that these countries had a high percentage of confirmed cases.

Figure 7c showed that the median value of death cases in C-I and C-IV were at the same level, but the distribution of the values differed in these two clusters. Similarly, the median value of death cases in C-II and C-III were at the same level, but the distribution of the values differed in these two clusters. The C-IV was least dispersed, while C-I and C-III were more dispersed among all the clusters. That is, countries in C-IV had similar situations of deaths, while in C-I and C-III, these values were variable in percentage. The median value of death cases was approximately equal in C-I and C-IV compared to the overall median value of all the clusters, while it was less in C-II and C-III. Further, **Figure 7c** showed that Albania is an outlier in C-I, which means that this country had a high percentage of deaths.

Figure 7d showed that the median value of recovered cases of C-III was greater than that of C-I, C-II, and C-III, while this value was least in C-II among all the clusters. Similarly, it can be seen that the median value of C-III was greater than that of C-IV. The C-III was least dispersed, while C-IV was more dispersed among all the clusters. Many countries in C-II had similar situations of recovery percentage while it was more variable in C-IV. The value of recovery cases in C-I and C-II was approximately equally distributed; however, it was more in C-I than in C-II. The median values of recovered cases in C-II and C-IV were less than that of the overall median value of all the clusters; however, it was approximately the same in C-I as the overall median and more in C-III than the overall median of all the clusters. Further, **Figure 7d** showed Ireland to be an outlier in C-III, meaning that this country had a low percentage of recovered cases.

Figure 7e showed that the median value of critical cases of C-II was greater than that of C-I, C-II, and C-IV, while this value was least in C-III among all the clusters. The median value of C-I was equal to the overall median of all the clusters. The C-III and C-IV were least dispersed, while C-I and C-II were more dispersed. Countries in C-II and C-IV had similar situations of critical cases, while it was more variable in C-I and C-II. Further, **Figure 7e** showed Italy to be an outlier in C-II and Ireland in C-III, meaning that these countries had a high percentage of critical cases.

4. Discussion

Our findings provide valuable insights into the diverse impact of the first phase of the COVID-19 epidemic in European countries. For this study, we defined a set of six attribute indices. We obtained

Pearson's correlation coefficients among attribute indices which showed that all of them were positively correlated. We identified clusters of countries based on their similarity using HCPC. This technique uses mixed algorithms of clustering and principal components methods. Since the results from the application of CA alone might not be very good, the application of clustering based on PCA can overcome the analysis-related complications due to the presence of a large number of indices as well as due to the presence of highly correlated attribute indices in the clustering process and thus HCPC provided improved clusters. Here, the mixed algorithms of the Ward's and the K-means methods provided final robust results. The consolidation of K-means clustering validated the stability of our results in C-I and C-IV, while only 2 out of 44 countries moved from C-II to C-III.

Based on the robust results obtained from HCPC, we grouped 44 affected countries into four clusters (C-I, C-II, C-III & C-IV). All countries under C-II and a few under C-I had very low testing that needs to be increased. The countries under C-I and C-II were highly affected by confirmed cases. So, these countries need screening, lockdown, and legal action optimization. It is also required to optimize treatment facilities so that deaths which were very high in some C-I countries, could be reduced. There was a large variation in the distribution of recovery cases in C-I, C-II, and C-IV countries. The critical cases were very high in countries under C-II and a few under C-I, which suggested management optimization for the timely admission of patients in hospitals and their proper treatments. Such actions could help governments, doctors, healthcare service providers, etc., in controlling the cases in all aspects, and in turn, it would reduce the seriousness of COVID-19.

Our study has some limitations. First, this study is limited up to 24 May 2020, to assess if such analysis could help controlling the severe situation in the first phase of COVID-19 in European countries. Further, this study considers only endogenous variables. Exogenous variables are not included here. For example, do countries with strict lockdown orders tend to fall in a given cluster? Do countries with substantial international travel tend to do worse? Does the country's Gross Domestic Product (GDP) affect which cluster a country is in? So, the study could further be modified with such exogenous variables and some other undocumented factors.

5. Conclusions

Our study focused on identifying clusters of similar epidemic patterns across European countries. We utilized a combined hierarchical clustering approach on principal components to achieve satisfactory clustering based on a set of epidemic indicators. This approach integrated three data mining methods: PCA, hierarchical clustering, and the K-means algorithm. A defined set of attribute indices guided the clustering process. Our study's findings help understand the severity of the epidemic in European countries, which might help related officials to adopt appropriate management in accordance with the severity at the country level. This study illustrates the use of HCPC to identify better clusters of affected countries for potential targeted care management like increasing testing, optimization of screening, lockdown, legal actions, optimization of management for timely admission of patients in hospitals and their proper treatments, etc.

However, our study has some limitations. First, this study is limited up to 24 May 2020, to assess if such analysis could help control severe situations in the first phase of COVID-19 in European countries. Further, the study could be modified with some exogenous variables and some other undocumented factors.

Author contributions

Conceptualization, SK; methodology, SK; software, SK; validation, SK and EO; formal analysis, SK; investigation, SK and EO; resources, SK and EO; data curation, SK and EO; writing—original draft preparation, SK; writing—review and editing, EO; visualization, SK and EO; supervision, SK. All authors have read and agreed to the published version of the manuscript.

Compliance with ethical standards

This article does not contain any studies with human participants.

Acknowledgments

The authors are grateful to the editor and referees for their valuable suggestions, which led to improvements in the article.

Conflict of interest

The authors declare no conflict of interest.

References

1. Liu Y, Gu Z, Xia S, et al. What are the underlying transmission patterns of COVID-19 outbreak? An age-specific social contact characterization. *eClinicalMedicine* 2020; 22: 100354. doi: 10.1016/j.eclinm.2020.100354
2. Olson DL, Shi Y. *Introduction to Business Data Mining*. McGraw-Hill/Irwin; 2007.
3. Shi Y, Tian YJ, Kou G, et al. *Optimization Based Data Mining: Theory and Applications*. Springer; 2011.
4. Kumar S. Monitoring novel corona virus (COVID-19) infections in India by cluster analysis. *Annals of Data Science* 2020; 7(3): 417–425. doi: 10.1007/s40745-020-00289-7
5. Available online: https://www.worldometers.info/coronavirus (accessed on 24 May 2020).
6. Ratner B. The correlation coefficient: Its values range between +1/−1, or do they? *Journal of Targeting, Measurement and Analysis for Marketing* 2009; 17(2): 139–142. doi: 10.1057/jt.2009.5
7. Johnson RA, Wichern DW. *Applied Multivariate Statistical Analysis*, 6th ed. Pearson Education, Inc.; 2007.
8. Husson F, Josse J, Pages J. Principal component methods-hierarchical clustering-partitional clustering: Why would we need to choose for visualizing data? *Applied Mathematics Department* 2010; 17.
9. Maugeri A, Barchitta M, Basile G, Agodi A. Applying a hierarchical clustering on principal components approach to identify different patterns of the SARS-CoV-2 epidemic across Italian regions. *Scientific Reports* 2021; 11(1): 7082. doi: 10.1038/s41598-021-86703-3
10. Kaiser HF. An index of factorial simplicity. *Psychometrika* 1974; 39: 31–36. doi: 10.1007/BF02291575
11. Hair JF Jr, Black WC, Babin BJ, Anderson RE. *Multivariate Data Analysis*, 7th ed. Prentice Hall; 2010.
12. Bartlett MS. A note on the multiplying factors for various χ^2 approximations. *Journal of the Royal Statistical Society: Series B (Methodological)* 1954; 16(2): 296–298. doi: 10.1111/j.2517-6161.1954.tb00174.x
13. Romesburg HC. *Cluster Analysis for Researchers*. Lifetime Learning Publications; 1984.
14. Ward JH Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* 1963; 58(301): 236–244. doi: 10.1080/01621459.1963.10500845
15. Goon AM, Gupta MK, Dasgupta B. *Fundamentals of Statistics*, 8th ed. The World Press, Kolkata; 2002.