Clustering framework to cope with COVID-19 for cities in Turkey

Marco de agrupamiento para hacer frente a COVID-19 para ciudades en Turquía

Estrutura de agrupamento para lidar com COVID-19 para cidades na Turquia

Didem Guleryuz[1]
Erdemalp Ozden[2]

Received: June 15th, 2021
Accepted: August 23th, 2021
Available: September 6th, 2021

How to cite this article:
D. Guleryuz, E. Ozden, “Clustering framework to cope with COVID-19 for cities in Turkey,” *Revista Ingeniería Solidaria,* vol. 17, no. 3, 2021.
doi: https://doi.org/10.16925/2357-6014.2021.03.06

Research article. https://doi.org/10.16925/2357-6014.2021.03.06

[1] Department of Management Information Systems, Bayburt University, Bayburt, Assistant Professor, Turkey,

E-mail:dguleryuz@bayburt.edu.tr

ORCID: https://orcid.org/0000-0003-4198-9997

[2] Department of Economics, Bayburt University, Bayburt, Turkey, Research Assistant,

E-mail: eozden@bayburt.edu.tr

ORCID: https://orcid.org/0000-0001-5019-1675

Abstract

Introduction: This article is the product of the research “Clustering Framework to Cope with COVID-19 for Cities in Turkey", developed at Bayburt University in 2021.

Problem: Turkey's risk map, presented in January 2021, to take local decisions in tackling the COVID-19 pandemic, was based on confirmed cases only. Health, socio-economic and environmental indicators are also important for management decisions of COVID-19. The risk map to be designed by adding these indicators will support more effective decisions.

Objective: The research aims to propose a clustering scheme to design a risk map of cities for Turkey.

Methodology: The unsupervised clustering algorithm suggested dividing the cities of Turkey into clusters, considering health, socio-economic, environmental indicators, and the spread pattern of COVID-19.

Results: We found that cities are clustered into five groups while megacity Istanbul alone formed a cluster, three of Turkey's largest cities formed another cluster. Other clusters consist of 19, 26, and 32 cities, respectively. The most important determinants which have predictive power are identified.

Conclusion: The suggested clustering method can be a decision support system for policymakers to determine the differences and similarities of cities in quarantine decisions and normalization phases for the following periods of the pandemic.

Originality: To the best of our knowledge, this study differs from previous studies because countries were grouped in previous studies by only considering the confirmed cases. In this study, cities were clustered in terms of the health, socio-economic, and environmental indicators to make decisions locally.

Limitations: The distribution of confirmed cases by age could be added, especially to make decisions about education, but this data is not officially announced.

Keywords: COVID-19, Unsupervised Learning, Clustering Algorithm, Decision Support System

Resumen

Introducción: este artículo es producto de la investigación “Modelo de agrupación para hacer frente al COVID-19 en ciudades de Turquía", desarrollado en la Universidad de Bayburt en 2021.

Problema: el mapa de riesgo de Turquía, presentado en enero de 2021 con el fin de tomar decisiones para abordar la pandemia de COVID-19, solo se basó en casos confirmados. Indicadores de salud, ambientales y socioeconómicos también son relevantes para la toma de decisiones sobre el manejo del la pandemia. El mapa de riesgos diseñado, teniendo en cuenta estos indicadores, puede soportar la toma de decisiones más efectivas.

Objetivo: se propone un esquema de agrupamiento con el fin de diseñar un mapa de riesgos para las ciudades de Turquía.

Metodología: el algoritmo de agrupamiento no supervisado sugirió dividir la ciudades turcas en grupos, considerando indicadores de salud, ambientales y socioeconómicos, además del patrón de propagación del COVID-19.

Resultados: se descubrió que las ciudades se agrupan en cinco. La megaciudad de Estambul conformó un solo grupo, mientras que tres de las ciudades más grandes de Turquía formaron otro. Otros grupos quedaron conformados por 19, 26 y 32 ciudades, respectivamente. Se identificaron los determinantes con poder predictivo más importantes.

Conclusión: el método de agrupamiento sugerido puede ser un sistema de soporte a la decisión para los hacedores de política, que les permitirá determinar las diferencias y similitudes de las ciudades en relación con la toma de decisiones para la cuarentena y las fases de normalización para los periodos posteriores a la pandemia.

Originalidad: hasta ahora, este estudio se distingue de trabajos previos en virtud de que los países se agruparon considerando solo los casos confirmados. En este estudio, las ciudades se agruparon, teniendo en cuenta indicadores de salud, ambientales y socioeconómicos para tomar decisiones localmente.

Limitaciones: la distribución de casos confirmados por edad pudo añadirse, especialmente para tomar decisiones sobre educación, pero estos datos no fueron públicamente divulgados.

Palabras clave: COVID-19, aprendizaje no supervisado, algoritmo de agrupamiento, sistema de soporte a la decisión.

Resumo

Introdução: Este artigo é produto da pesquisa "Modelo de cluster para lidar com o COVID-19 em cidades turcas", desenvolvida na Bayburt University em 2021.

Problema: o mapa de risco da Turquia, apresentado em janeiro de 2021 para tomar decisões para enfrentar a pandemia de COVID-19, foi baseado apenas em casos confirmados. Indicadores de saúde, ambientais e socioeconômicos também são relevantes para a tomada de decisões sobre o gerenciamento da pandemia. O mapa de risco elaborado, levando em consideração esses indicadores, pode subsidiar tomadas de decisão mais efetivas.

Objetivo: é proposto um esquema de agrupamento para projetar um mapa de risco para cidades na Turquia.

Metodologia: O algoritmo de agrupamento não supervisionado sugeriu dividir as cidades turcas em agrupamentos, considerando indicadores de saúde, ambientais e socioeconômicos, bem como o padrão de disseminação do COVID-19.

Resultados: verificou-se que os municípios estão agrupados em cinco. A megacidade de Istambul formou um grupo, enquanto três das maiores cidades da Turquia formaram outro. Outros grupos foram formados por 19, 26 e 32 cidades, respectivamente. Foram identificados os determinantes com maior poder preditivo.

Conclusão: o método de agrupamento sugerido pode ser um sistema de apoio à decisão para os formuladores de políticas, que lhes permitirá determinar as diferenças e semelhanças das cidades em relação à tomada de decisão para a quarentena e as fases de normalização para os períodos pós-pandemia.

Originalidade: Até agora, este estudo difere de trabalhos anteriores pelo fato de os países terem sido agrupados considerando apenas os casos confirmados. Neste estudo, as cidades foram agrupadas, levando em consideração indicadores de saúde, ambientais e socioeconômicos para tomar decisões localmente.

Limitações: A distribuição de casos confirmados por idade pode ser agregada, especialmente para decisões de educação, mas esses dados não foram divulgados publicamente.

Palavras-chave: COVID-19, aprendizado não supervisionado, algoritmo de agrupamento, sistema de apoio à decisão.

1. INTRODUCTION

COVID-19, which emerged with the first case in Wuhan, China, on December 31st, 2019, was declared a pandemic by the World Health Organization (WHO) on March 11th, 2020, when the first case was seen in Turkey [1]. The coronavirus spread rapidly worldwide, and the outbreak was seen in Korea, Japan, Europe, and America, respectively, after China. In the first six months of 2020, Europe, Italy, and Spain were severely affected by the virus, encountering many confirmed cases and deaths. After that, with the spread of the virus in the ABD, the United States was also heavily hit.

Humanity has faced many epidemics throughout history, but it seems that the epidemics in history did not spread as quickly as COVID-19. The biggest reason for this is that human mobility has increased, with globalization dramatically affecting transmission. Due to human mobility, the epidemic that started in China spread very quickly everywhere globally. As a result, states have made radical decisions that restrict global mobility and affect the global economy and people to a great extent, such as national border restrictions, quarantines for cities, and lockdown policies. These decisions for the effective management of the epidemic almost caused life to cease. During the one-year epidemic period, many countries have carried out vaccine development studies; China, Germany, England, and Russia have been the pioneers in this process. With the late arrival of the epidemic in Turkey and the rapid intervention of decision-makers, the first wave of the epidemic was lighter than in European countries. However, in the new process called the third wave, it lagged behind the European countries. The main reason for this situation is that there is a shortage of vaccine access, as in other developing countries; whereas the vaccination started in Europe in December 2020, it started in March 2021 in Turkey.

The COVID-19 pandemic has caused significant challenges to local healthcare systems, global healthcare systems, and the global economy. As of May 6th, 2021, there were a total of 156,164,449 confirmed cases worldwide, while the number of death cases reached 3,260,489. Turkey, one of the countries that COVID-19 has attacked, reported 4,955,594 confirmed cases and 41,883 death cases [2].

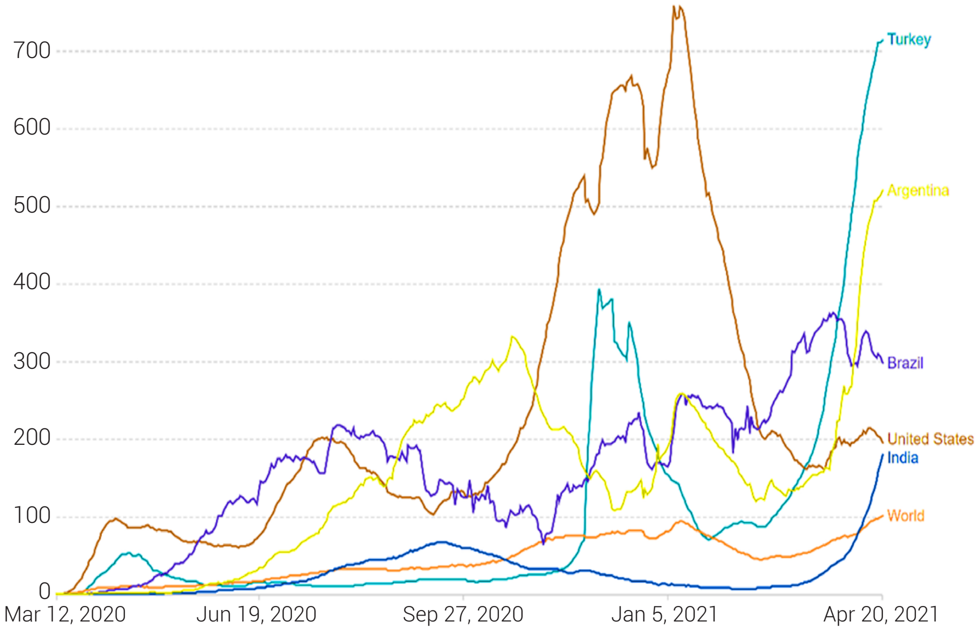

Figure 1. Daily new confirmed COVID-19 cases (per million people)
Source: [3].

Fig. 1 shows that the controlled process in the first stages of the epidemic in Turkey has turned into a more complex process to control due to the late start of vaccination and the disruptions in the decision-making process. With the rapid increase in covid cases, it has been observed that restrictions across the country are not effective. Instead, both academic and public spheres have begun to debate whether regional or city-based decisions will allow faster and more effective results.

The short- and long-term impact of the epidemic on human health is still in the process of being discovered. Despite the positive regional news with the increase in the vaccination rate, a widespread global solution seems unlikely in the short term. The precautions applied for the effective management of the epidemic, such as restrictions, quarantines, and lockdown policies, negatively affect the sustainability of operations in all sectors.

Global outbreaks differ from other risks depending on their source, and these differences require countries to go beyond their current management strategy and plans to have effective epidemic management. Therefore, this epidemic is not just an epidemic, and it requires an effective management plan.

The United Nations has described this epidemic as a humanitarian, economic and social crisis. There is a global economic recession depending on the duration and impact of the epidemic. Based on the data of this recession, global trade is estimated to fall by 13% to 32%. Although COVID-19 still continues to spread with various effects in many countries worldwide, China and some Asian countries have controlled the epidemic [4].

The spread of COVID-19 affects the management policies of countries and, therefore, many countries need particular policies and plans. In general, the privatized policies implemented by the country administrators, considering the situation they are in, led the administrators to local policies within their countries with the prolongation of the epidemic period. Turkey published a risk situation map by province in February 2021 and announced that the outbreak management decision would be taken locally based on risk situations. In this period, with the entry into the period called the third wave of the epidemic, risk maps were updated every week, and decisions were made on a regional basis for epidemic management. The weekly number of cases in the provinces is given in Fig. 2.

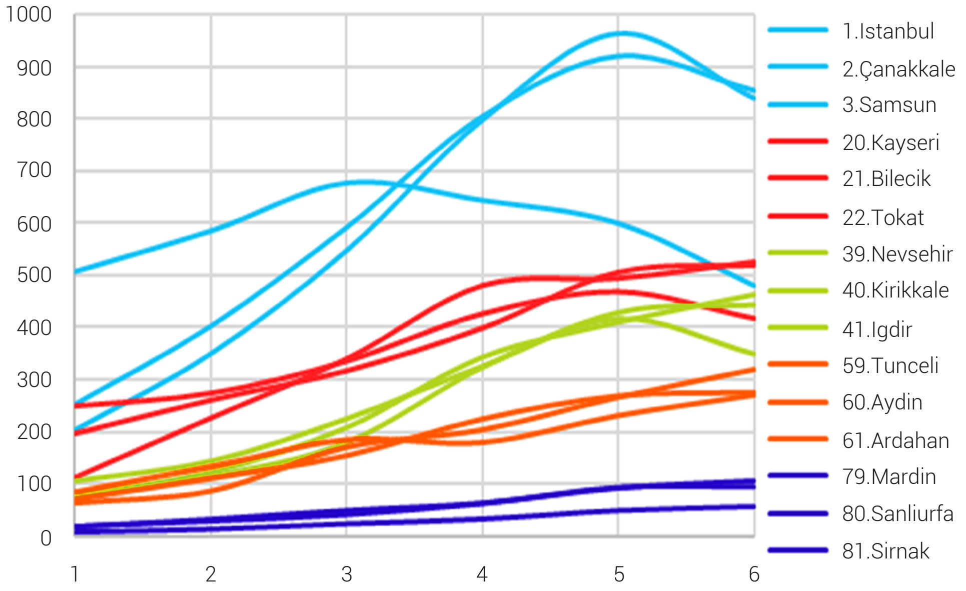

Figure 2. Number of cases by weekly provinces (most cases to least)
Source: [1]

Factors such as the rate of spread of the virus in cities, the number of patients, and the number of deaths affect health systems regionally. Therefore, it is critical to examine the relationships between the rate of spread of the epidemic, health services, and demographic indicators. The risk map created by the Turkish Ministry of Health only takes into account the daily number of cases in cities.

In this study, the risk status of COVID-19 cities in Turkey was evaluated and clustered using the k-means technique using health indicators, demographic structure, economic indicators of the cities, and cases. First, COVID-19 data sets of indicators selected for all cities were created with the most up-to-date data. Then, correlations between these data sets were calculated and analyzed. Observed data values were rescaled to make more robust analyses for data clusters in different scales. Cities have been clustered using the K-means technique. With the clustering of cities, administrators will be able to make similar local decisions for similar cities. Urban administrations in the same cluster will follow common strategies regarding the restrictions and normalization process.

The main contribution of this study is the clustering of cities, not only by considering the indicators related to COVID-19, but also by considering various indicators related to economic, health, and environment. The difference of this study, is that in previous studies, countries were grouped by considering only the number of COVID-19 cases to see the impact of COVID-19 on countries. However, in this study, cities were clustered in terms of health indicators, weekly case numbers, and other economic and environmental indicators to make decisions locally. According to the clustering results, the city governments authorized to take decisions on pandemic management can develop similar management strategies and follow common decision procedures. The reason for using k-means in clustering applications is that employing the method is simple and gives effective results. In the literature review, to the authors' knowledge, this study is the first to cluster cities according to different indicators using the k-means method and provide a support system for COVID-19 management decisions on a local basis.

1.1 Literature review

Since the epidemic has affected everywhere to a great extent from the beginning of the epidemic, studies have shown the epidemic's effect in many areas; such as monitoring the spread of the epidemic, outbreak management, medical scope of the epidemic, the effects of the epidemic on social life. Previous studies used methods such as artificial intelligence-based methods, machine learning, deep learning algorithms, and mathematical and statistical methods. Some of the studies of the COVID-19 outbreak prediction and clustering are summarized below.

Guleryuz (2021) developed a prediction model to estimate the total case, the growth rate of total cases, the growth rate of total case, the number of total deaths, the growth rate of total deaths, the number of total cases, the number of new cases, and the number of new deaths for Turkey. Exponential Smoothing, Long Short Term Memory, and Box Jenkins methods are employed. As a result of the study, the daily number of cases will not show an increasing trend [5].

Nikolopoulos et al. (2021) emphasized that policymakers should harness the power of science to support the decision-making process in the outbreak. For this purpose, they proposed a new hybrid prediction method based on COVID-19 growth rates and closest neighbors and clustering with machine learning and deep learning models. Thus, dependent short-term supply chain disruptions are modeled, and it is predicted that the proposed method will support decision-makers in planning activities. As a result, it can instantly help decision-makers have optimum decisions for the COVID-19 pandemic management and possible future pandemics [6].

Melin et al. (2020) has proposed a model that uses the self-organizing maps method to group similar countries based on coronavirus case numbers. Using the clustering capability of this method, the number of cases was grouped spatially for similar countries, and it was seen which countries had similar characteristics. The study aims to show that countries with similar characteristics can effectively fight an epidemic if they follow similar management strategies. While the studies carried out until the time of the conflict generally include prediction models showing the time-dependent spread of the virus, the spatial dimension has been added to the model proposed in this study. As a result of the study, it was seen that self-organizing maps known as unsupervised learning could be used to cluster similar countries in their fight against the coronavirus pandemic [7].

Mahmoudi et al. (2020) proposed a model that examines the relationships between the distributions of the spread of coronavirus into countries. The spread distribution of the virus was monitored for seven countries determined in the study, and countries were clustered using the fuzzy clustering technique. In the first stage, the high population of the USA, especially among the selected countries, affected the clustering. This problem was solved by scaling the time series. The scaled datasets are clustered using fuzzy clustering. As a result of the study, it was observed that the COVID-19 spread distribution for Spain and Italy was similar, and other countries were different [8].

Olivieri et al. (2021) proposed a hierarchical clustering method to cluster the regions of Italy. Italy is a country that suffered highly in the first period of the epidemic among European countries. There has been a great crisis in the health system in the country, and the intensive care units have been insufficient. For this reason, clustering has been made based on the number of COVID-19 cases to use intensive care units effectively. As a result of the study, it is seen that cluster analysis can be used for preventive actions and optimized health systems in such epidemics [9].

Kücükefe (2020) examined OECD countries and China, considering that the COVID-19 epidemic affected the country's economies on a global scale. In the study, countries have been clustered using the k-means algorithm. The indicators used are current account balances, GDP growth rate, and deaths per million population. The proposed method divided the countries into 3 clusters. It has been observed that countries with a current account surplus above 2.5% of GDP managed to limit their GDP decline to less than -15%. In addition, countries with high mortality rates and current account deficits are included in another cluster [10].

Azarafza et al. (2020) proposed a clustering method to provide the COVID-19 disease spread pattern in Iran at the local level. K-means and geographical information system mapping were used in the study. After the probability density of the infection maps was prepared, the infection pattern between provinces was clustered with k-means. The QOM region is the main point of the coronavirus spread for Tehran, but the city of Tehran is the region responsible for spreading the virus across Iran [11].

Rizvi et al. (2021) aim to cluster selected countries using social, economic, health, and environmental indicators that affect the spread of the disease. Thus, countries will be able to implement policies to control the prevalence of the disease. Eighteen features were used in the study, and the k-means method divided the countries into four groups. In addition, the results of correlation analysis between the features selected in the study and the number of cases and deaths of the countries are also presented. According to the analysis results, the most effective factor in mortality rates is the prevalence of underlying diseases, while ineffective and weak factors are environmental health indicators [12].

Zarikas et al. (2020) has clustered 30 countries based on active cases, active cases per population, and active cases per population and per area. The results of this cluster analysis at the early phase of the epidemic will be helpful for decision-makers in rapid decision-making. Hierarchical clustering analysis was used in the study. In addition, a specially designed new clustering algorithm has been proposed that compares various time series of COVID-19 cases from different countries [13].

Carrillo-Larco and Castillo-Cara (2020) propose a model to cluster countries based on prevalence estimates of the selected diseases, socio-economic status, air pollution, and health system coverage using machine learning algorithms. The data were determined from different sources by performing a PCA features analysis. The k-means method was used for cluster analysis. As a result of the study, 155 countries were divided into five different clusters [14].

The rest of the paper is composed as follows: Section 2 includes the materials and methods and introduces the developed clustering methodology, Section 3 represents the results and discussion, Section 4 and Section 5 include the conclusion and references, respectively.

2. MATERIALS AND METHODS

2.1 Data sources

After the literature review, 11 different variables, including demographic, economic, health, and environmental indicators of cities, were used for COVID-19 clustering. The variables are shown in detail in Table 1. Accordingly, the population over 65 years, the young population between 0-14 years, life expectancy at birth and population density per km^2 were used as demographic variables. Employment rate and GDP per capita were selected for economic indicators, while hospital beds and application per doctor variables were selected for health services. Finally, air pollution was chosen as an environmental indicator.

Table 1. Description of Variables

Notion	Variable Name	Description (Date)	Data Sources
COVID-19 Cases	Covid Cases	Average Weekly Number of Cases by Cities (per 100 thousand) between March 13 - April 23	MinHealth [1]
Demographic	65+	65 and overpopulation (2020)	TurkStat [15]
	0-14	0-14 age group population (2020)	TurkStat [15]
	Life expectancy	Life expectancy at birth (2020)	TurkStat [15]
	Population density	Population density (people per square kilometer) (2020)	TurkStat [15]
Economic	Employment rate	The employment rate (%) (2020)	ILO [16]
	GDP	GDP per capita (Turkish Liras)	Worldbank [17]
	Nurses	Total number of nurses (2020)	MinHealth [1]
Health Services	Hospital Beds	Total number of hospital beds per 100 thousand people (2018)	MinHealth [1]
	Applications per doctor	Number of applications per doctor (2019)	MinHealth [1]
Environmental	Air Pollution	Average of PM10 values of the stations (air pollution) (μg/m³)	TurkStat [15]

Source: Own work

In Table 1, the weekly numbers of COVID-19 cases per hundred thousand people are collected by provinces. For this data, the data of the provinces published weekly from the Ministry of Health were collected for the period between March 13 and April 23, and this variable was created by taking the average values of these data.

2.2 Standardization

When the statistical values of the prepared data set for cluster analysis are examined, it can be seen in Table 2 that the indicators belonging to different determinants are in very different scales. For instance, air pollution, which is used as an environmental indicator, takes values between 18 and 113 μg / m³, while economic and demographic indicators take much higher values with various standard deviations.

Table 2. Descriptive Statistics of Variables

Variables	Mean	Std. Dev.	Min	Max	Skewness	Kurtosis
Covid Cases	277.64	142.07	29.998	637.41	0.4779476	2.77093
65+	98192	147724	9676	1137610	4.966	32.756
0-14	235410	408151	12492	3312147	5.6257	41.39
Life expectancy	78.135	1.0371	74.955	80.504	-0.088	3.7825

(continúa)

(viene)

Variables	Mean	Std. Dev.	Min	Max	Skewness	Kurtosis
Population density	132.82	332.53	11.23	2975.8	7.8859	67.667
Employment rate	46.219	6.2032	27.8	59.1	-1.0842	4.422
GDP	39596	13645	16727	86798	1.0914	4.4856
Nurses	2351.8	4270.4	203	34502	5.8182	42.042
Hospital Beds	276.19	82.35	120	502	0.741	3.2889
Applications per doctor	5834.4	1245.2	2763.3	8067.4	-0.1652	2.2219
Air Pollution	55.334	20.293	18	113	0.6434	3.2422

Source: Own work

The fact that the units in which the variables are measured differ significantly (different averages and variances) may cause biased results by affecting a variable in the clustering algorithm more than it should be. Therefore, a data set must be transformed appropriately to make clustering more robust. There are many standardization methods available. In this study, variables in the data set are transformed so that the mean of the variables, also called z-score standardization, is 0, and the standard deviation is 1 [18].

$$x_i' = \frac{x_i - \mu}{\sigma} \tag{1}$$

Here x_i represents the raw data, while μ and σ show the mean and standard error, respectively. Average data is subtracted from each raw value and divided by the standard deviation to reach the standardized values according to Eq. 1. After standardization, the negative values in the data set show the observations below the average, while the positive values show the values higher than the average.

2.3 Number of clusters

Many methods are used to determine the appropriate number of clusters in non-hierarchical clusters. The appropriate number of clusters was determined in this study according to the frequenctly used Elbow and Average Silhouette Index method.

2.3.1 Elbow method

The Elbow method aims to determine a suitable K value where the variance in the total set is minimum. WCSS (Within Cluster Sum of Square) is calculated by taking the sum of the square of the distance of each point from the center of the cluster [19]. The Elbow method says that the point where the amount of change in WCSS decreases, that is, the elbow point, is the optimum point.

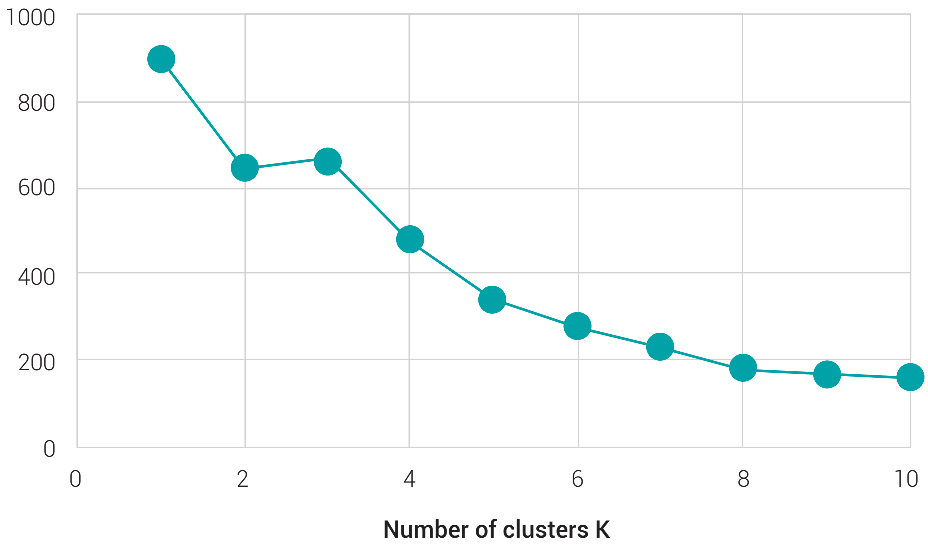

Figure 3. The optimal number of clusters - Elbow Method
Source: Own work

As seen in Fig. 3, the appropriate clustering number can be chosen as 4 or 5 according to the elbow method. In order to better determine what the cluster number will be, the silhouette method, which is another method frequently used in the literature, has been applied.

2.3.2 Average Silhouette Index

Rousseeuw (1987) proposed an Average Silhouette Index that would define each unit's suitability for its own cluster [20]. Let us show $a(i)$: the average distance (similarity) of $unit\ i$ to all points in its set and $b(i)$: the minimum of the average distances of $unit\ i$ to all points in the other sets. Accordingly, the equation is as follows:

$$Sil\ (i) = \frac{b\,(i) - a\,(i)}{max\ [a\,(i), b\,(i)]} \tag{2}$$

Eq. 2 shows the Average Silhouette Index for $unit\ i$. If the $Sil(i)$ value approaches 1, it is concluded that the $unit\ i$ is more suitable for the cluster to which it is assigned, and if the $Sil(i)$ value approaches 0 or negative, the *unit* i is not suitable for the cluster to which it is assigned. Negative values occur only when a unit cannot be assigned to its optimal set.

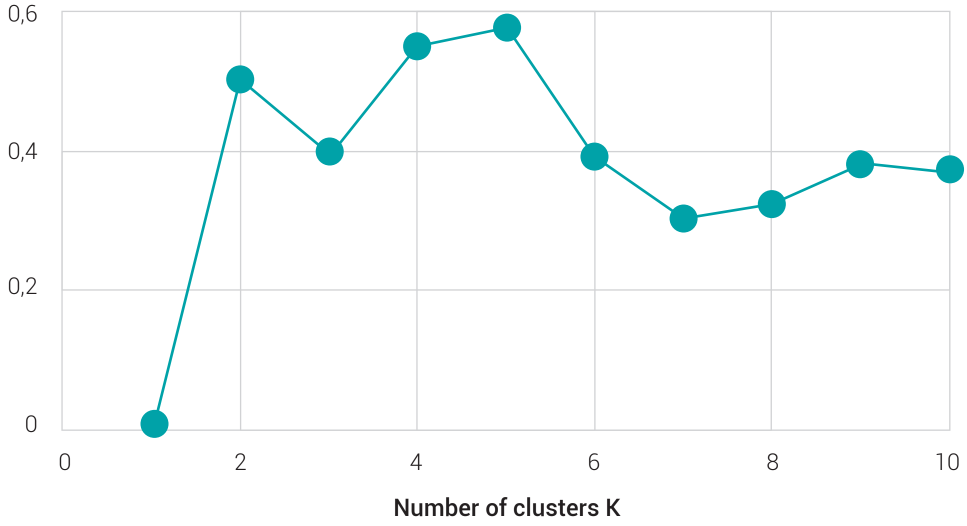

Figure 4. The optimal number of clusters - Average Silhouette Index
Source: Own work

As can be seen in Fig.4, the silhouette value was found to be Cluster 5 as the highest value. When evaluated with the Elbow method, the optimum number of clusters was determined to be 5.

2.4 Clustering methodology

We have many data on major diseases that have spread worldwide, such as COVID-19. This large amount of data contains many dimensions and fields and, therefore, has a very complex structure. Even today's data mining methods have difficulty producing meaningful results from this complex database. In order to solve significant problems containing such multidimensional and complex data, it is necessary to divide them into smaller and more easily solvable sub-problems. After solving each sub-problem, inferences can be drawn for the larger problem by combining the solutions. However, in some cases with many dimensions, such as COVID-19, the data is so dispersed that it is impossible to predict where the data will be divided and how it can be divided into subgroups. In these cases, using unsupervised methods can help. The non-hierarchical k-means clustering method is among the most popular of these unsupervised methods.

Cluster analysis is a collection of methods in the data matrix that helps to divide natural groups into uncertain units, variables, or subsets (groups, classes) that are similar to each other. Clustering analysis tries to form homogeneous groups using some measures whose units are calculated based on the similarity or distances between variables. Cluster analysis does not differentiate between dependent and independent variables.

K-means is one of the most widely used non-hierarchical methods. Theoretically, in this method, the number of clusters is determined to be at least two and less than or equal to the maximum number of observations [18].

In this study, the number of cluster centers was previously determined according to the Elbow and Silhouette method. After the determined number of clusters, the average of the cluster centers in each iteration is updated to be at the minimum distance from the respective cluster centers. Here, the Euclidian distance, one of the most frequently used distance measures, was used to determine the distance. Basically, this distance can be defined as the geometric distance in multidimensional space. As a result of these iterations, optimum clusters are formed. Therefore, this study leads to the clustering of cities that may be similar in line with the indicators determined in terms of the provinces of Turkey. It is valuable in guiding policymakers to see which common policies should be established for clusters of these cities and which indicators are effective.

3. RESULTS AND DISCUSSIONS

There are many determinants behind the spread of the COVID-19 outbreak. In this respect, it is essential to look at the correlation between the determinants (and indicators) selected for the study and the COVID-19 cases belonging to the cities. In this section, firstly, this relationship will be looked at, and then the results of the k-means algorithm will be analyzed.

3.1 Correlation of indicators

It is a statistical method that checks the existence of a linear relationship between two numerical measurements and determines the direction and density of this relationship. There are two different coefficients for normally distributed datasets, and the non-normally distributed datasets are the Pearson correlation coefficient and Spearman rank-order correlation coefficient, respectively [21], [22]. First of all, it was analyzed whether the data were normally distributed or not.

Table 3. Normality Test

Variable	Obs	Pr (Skewness)	Pr (Kurtosis)	adj chi^2	Prob>chi^2
Covid Cases	81	0.0693	0.8735	3.44	0.1788
65+	81	0.0000	0.0000	.	0.0000
0-14	81	0.0000	0.0000	.	0.0000
Life expectancy	81	0.7281	0.1180	2.65	0.2656
Population density	81	0.0000	0.0000	.	0.0000
Employment rate	81	0.0003	0.0262	14.64	0.0007
GDP	81	0.0002	0.0226	14.91	0.0006
Nurses	81	0.5154	0.0460	4.48	0.1063
Hospital Beds	81	0.0000	0.0000	.	0.0000
Applications per doctor	81	0.0072	0.3830	7.23	0.0269
Air Pollution	81	0.0175	0.4266	5.95	0.0511

Source: Own work

In the analysis, the null hypothesis is "the data follows a normal distribution". If p-value of chi^2 is greater than 0.05 implying its significance at a 5% level, it means the null hypothesis cannot be rejected. Table 3 shows that covid cases, life expectancy, application per doctor, and air pollution have a normal distribution; the rest are not normally distributed. Therefore, Spearman rank's correlation analysis was applied instead of Pearson. These correlation results are given in Table 4.

Table 4. Spearman Rank's correlation matrix for all determinants

	Covid Cases	65+	0-14	Life exp	Pop. density	Emp. rate	GDP	Nurses	Hospital Beds	App.Pe doctor	Air Pollution
Covid Cases	1										
65+	0.2627	1									
0-14	-0.1548	0.761	1								
Life expectancy	0.0119	0.0788	-0.0553	1							
Population density	0.1053	0.6249	0.6785	0.0662	1						
Employment rate	0.168	0.0291	-0.25	0.0324	-0.109	1					
GDP	0.4346	0.4338	0.0409	0.1409	0.2267	0.3723	1				
Nurses	-0.069	-0.3729	-0.3118	-0.0119	-0.1165	-0.0938	-0.2241	1			
Hospital Beds	0.0701	0.9314	0.9062	0.0201	0.6596	-0.101	0.252	-0.4595	1		
Applications per doctor	0.2819	0.5315	0.2143	0.012	0.1317	0.1244	0.1885	-0.5853	0.5241	1	
Air Pollution	-0.1911	0.1607	0.3125	0.0937	0.2916	0.0576	0.0461	0.0079	0.2305	-0.0731	1

Source: Own work

When the table is examined, the highest positive relationship with the covid case is GDP per capita (0.4346), while relatively less strong and positive relationships are respectively Hospital beds (0.2819), 65+ (0.2627), the Employment rate (0.1680), and population density (0.1053). Spearman rank's correlation shows that negative associations exist for air pollution (-0.1911), under14 (-0.1548), and application per doctor (-0.0690).

3.2 Cluster analysis

The cluster was built for 81 cities in Turkey. Accordingly, cities divided into 5 clusters are listed in Table 5.

Table 5. Clusters based on Weekly COVID-19 Cases

Clusters	Cities
Cluster 1	Adana, Afyonkarahisar, Aydin, Bolu, Bursa, Denizli, Düzce, Edirne, Elazig, Erzurum, Eskisehir, Gaziantep, Isparta, Kars, Kayseri, Kirikkale, Konya, Kütahya, Malatya, Manisa, Samsun, Sivas, Tokat, Trabzon, Yozgat, Zonguldak
Cluster 2	Istanbul
Cluster 3	Adiyaman, Batman, Bitlis, Diyarbakir, Gümüshane, Hakkari, Hatay, Igdir, Kahramanmaras, Karaman, Mardin, Mersin, Mugla, Mus, Osmaniye, Sanliurfa, Siirt, Sirnak, Van
Cluster 4	Ankara, Antalya, Izmir
Cluster 5	Agri, Aksaray, Amasya, Ardahan, Artvin, Balikesir, Bartin, Bayburt, Bilecik, Bingöl, Burdur, Çanakkale, Çankiri, Çorum, Erzincan, Giresun, Karabük, Kastamonu, Kilis, Kirklareli, Kirsehir, Kocaeli, Nevsehir, Nigde, Ordu, Rize, Sakarya, Sinop, Tekirdag, Tunceli, Usak, Yalova

Source: Own work

Table 5 shows that Cluster 1 contains 26, Cluster 2 contains 1, Cluster 3 contains 19, Cluster 4 contains 3, and Cluster 5 contains 32 cities. Moreover, the mapping of the cities according to clusters is shown in Figure 5.

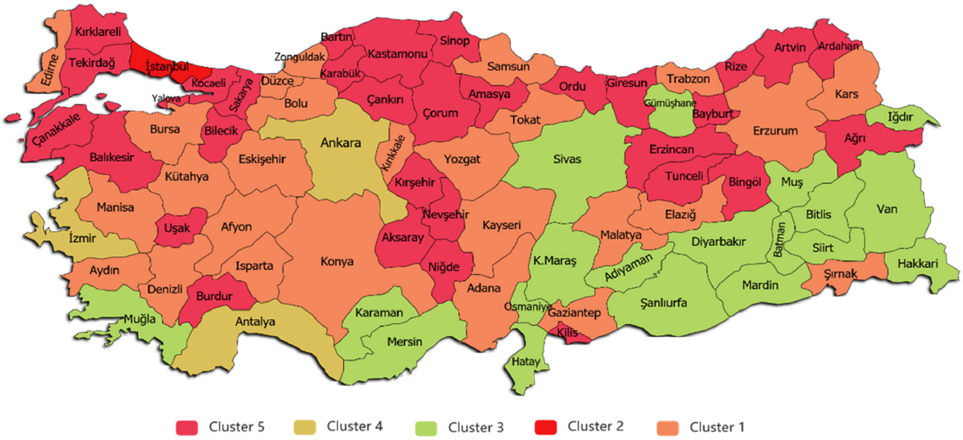

Figure 5. Map of Turkey by clustered cities
Source: Own work

In Fig. 5, cities in the same cluster are painted in similar colors. This means, according to determinants, city authorities can prepare and implement similar applications for cities of similar color. Variables were selected for different determinants in the study. While determining the clusters of these variables, their order of importance is shown in Fig. 6. Accordingly, the most important variable is Population density, while the least important one is life expectancy. These results are also meaningful. Because it is known that the places where this virus spreads the most are crowded areas. However, since life expectancy is determined at birth, it is essential for measuring the general health of the region, but it is not very important in terms of the spread of the disease.

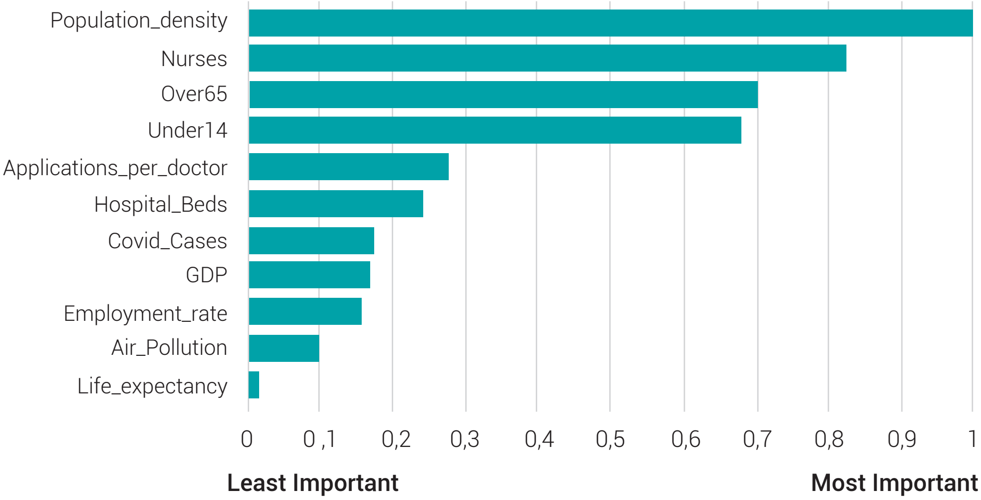

Figure 6. Predictor Importance of Indicators
Source: Own work

In order to interpret the data of the cities divided into these clusters, the averages of the clusters are shown in Table 6. The last column was added to compare the cluster with the country. That will allow us to compare clusters among themselves while at the same time making a comparison between that cluster and the country average.

Table 6. Cluster Mean of Variables

Indicators	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Turkey
Cities	26	1	19	3	32	81
Covid Cases	241	537	110	224	287	232
65+	105.651	1.137.610	57.150	421.222	53.735	98.192
0-14	217.174	3.312.147	250.383	839.584	88.548	235.410
Life expectancy	78	79	78	79	78	78
Population density	100	2.976	91	240	85	133
Employment rate	48	46	39	48	48	46
GDP	39.499	86.798	29.829	64.071	41.703	39.596
Nurses	2.464	34.502	1.601	10.426	945	2.352
Hospital Beds	358	261	217	302	243	276
Applications per D.	4.819	4.112	6.385	3.629	6.593	5.834
Air Pollution	59	55	70	52	44	55

The Best Cluster The Worst Cluster

🟩 🟨 🟧

Source: Own work

3.2.1 Analysis of Cluster 1

Table 7 shows that it has 26 cities in Cluster 1. While most of this cluster consists of cities in the central Anatolian region, it is seen that the cluster consists of cities from different parts of Turkey. It has the third-highest average among other groups in terms of weekly COVID-19 cases. It appears to have a similar average to Cluster 3 and 5 in terms of life expectancy. In terms of population density, cities in this cluster constitute the third most populous group. It can be said that it is one of the clusters with the highest average together with the other two clusters in terms of employment rate. When examined in terms of health services, it is again in third place in terms of the number of nurses, and it has been observed that it is the best group in terms of hospital beds. When examined in terms of application per doctor, it is observed that it is in third place, but it can be said that it has a better average than the average of

Turkey. When considered in terms of all indicators, it is seen that this cluster is closest to the average of Turkey.

3.2.2 Analysis of Cluster 2

There is only one city in Cluster 2, that is megacity Istanbul. 19% of the population of Turkey lives in Istanbul (The Results of Address Based Population Registration System, 2020, Turkish Statistical Institute, December 31st , 2020. Retrieved February 5th, 2021) [15]. Therefore, it is reasonable for Istanbul to form a cluster alone. In terms of demographic indicators, there is a serious gap between all other clusters. Again, in terms of GDP per person, it is seen that it has the highest value, but this rate is more than twice the average of Turkey. Although the number of nurses is still very high in health indicators, it is in third place in terms of hospital beds and below Turkey's average. When analyzed in terms of air pollution, it remained on average both among the clusters and compared to the Turkish average.

3.2.3 Analysis of Cluster 3

There are 19 cities in Cluster 3. Most cities in the southeastern Anatolian region of the country are in this group. Cities with the lowest number of cases per week are included in this cluster. This is because the average population over 65 years old is far behind the average of Turkey, and it is the second cluster where the population density is the lowest. Although it has a good average as a weekly covid case, it has been determined that it is not so in other indicators. The cluster is the lowest in terms of employment rate and income per capita, considering the economic indicators. According to health care, the average number of nurses is in the second-lowest cluster, while it is the last in terms of hospital bed per capita. It is the second most dense cluster as an application per doctor. The cities with the highest air pollution in terms of environmental factors are included in this cluster.

3.2.4 Analysis of Cluster 4

Cluster 4 includes three cities. In terms of population, Ankara, Izmir, and Antalya are the second, third, and fifth largest cities in Turkey, respectively. Therefore, it can be thought that these cities may have different averages when compared with other cities. When Table 7 is examined, it is close to the average of Turkey in terms of COVID-19 cases. In terms of demographics, it is the cluster with the highest average after Istanbul (Cluster 2). While the employment rate is one of the clusters in terms of economic determinants, it includes the cities with the second-highest average per capita income. This cluster is the second cluster in terms of nurses, the second cluster for hospital beds, and the first cluster in application per doctor in terms of healthcare services. For air pollution, it is the second-best cluster below the average for Turkey.

3.2.5 Analysis of Cluster 5

Cluster 5 contains the most cities among other clusters. The vast majority of 32 cities are located in the northern part of the country. The statistics show it is the second-highest cluster in terms of weekly covid cases (See Table 7). Although demographic determinants are relatively good, there is a negative picture in terms of health indicators. Looking at the details, the population over 65 is the lowest while the population density is again the lowest among others. While it is the third cluster in terms of GDP, this cluster has an average close to the Turkish average. Although it is the cluster with the least air pollution, the average number of nurses is the lowest, hospital beds are the second-lowest, and the cluster has the highest application per doctor. When Istanbul is excluded, the cities with the worst conditions in terms of covid data and health indicators are included.

4. CONCLUSIONS

It has been 16 months since the COVID-19 disease emerged, and during this time, it appeared in many variants. The virus can still be contained in very few parts of the world. Even these countries are in a state of alarm and continue to organize daily life by closely monitoring the number of cases. Countries made decisions by centralized methods when the number of cases exceeded certain threshold values in this process. Applying common rules to all cities with different geographical, economic, and health infrastructure features can sometimes put them in more troublesome situations than they have conditions.

When looking closely at Turkey, the properties of the cities are far from homogeneous. It is a country with megacities where millions of people live along with cities with a very low population. In addition, there is winter at one end and a summer climate at the other. While one city of Turkey is wholly engaged in agriculture, the income source of one city is tourism, and the income source of another city is industry. Therefore, we think that it may be more convenient for similar city managers to come together and share their knowledge and make quick and inclusive decisions instead of taking country-based policies.

Machine learning methods can be used to cope with all these challenges brought by COVID-19 and to group similar cities. Based on this idea, we clustered cities using the most important determinants, such as weekly COVID-19 cases of cities and the demographic, socio-economic, health, and environmental factors that will directly or indirectly affect the COVID-19 cases in cities by applying the unsupervised algorithm. Ultimately, the country is divided into five different clusters. Among these, while Istanbul alone formed a cluster, three of Turkey's largest cities formed another cluster. Other clusters consist of 19, 26, and 32 cities, respectively. While the three variables that are the most important predictive power in forming these clusters are population density, the total number of nurses, and population over 65+, the most insignificant indicators are life expectancy, air pollution, and employment rate.

The study results show that Cluster 1 includes the provinces closest to the German average when considering the determinants. As a megacity, Istanbul is included in Cluster 2; therefore, the highest numbers are reached in many indicators. Cluster 3 is a cluster where cities in the Southeast Anatolia region are densely located. It is one of the worst clusters in terms of economic and health indicators, although better than average in weekly covid cases. In the fourth cluster, Turkey's largest cities in terms of population after Istanbul are located here. Therefore, the policymakers of the cities in this cluster should make joint decisions since excessive mobilization is a compelling factor in coping with this disease. Finally, most of the cities in Cluster 5 are located in the northern region of the country. The weekly number of COVID-19 cases is the second highest in this cluster, and especially the health determinants are well below the Turkey average in this cluster. This clustering method shows that cities are similar to each other with different characteristics, so analysis can be done quickly to deal with these epidemics, and these results can help cities and make decisions to control the pandemic better. The analysis method applied in this study can be repeated as weekly data are updated, and different clusters can be created over time; thus, this process can be monitored dynamically.

In future studies, the analysis method in this study can be applied to other countries. Analyses could be extended using a similar machine learning method by adding other indicators, for instance, the distribution of confirmed cases by age, that are not readily available in Turkey. In addition, different clustering methods such as hierarchical, DBSCAN, and different unsupervised machine learning methods could be employed, and the results can be compared with this study.

5. REFERENCES

[1] MinHealth, “Republic of Turkey Ministry of Health,” 2021. [Online]. Available: https://covid19.saglik.gov.tr/TR-66494/pandemi.html.

[2] WHO, “WHO Coronavirus (COVID-19) Dashboard,” 2020. [Online]. Available: https://covid19.who.int/.

[3] JohnsHopkins, “Johns Hopkins University COVID-19 Data,” 2021. [Online]. Available: https://coronavirus.jhu.edu/map.html.

[4] M. Liu *et al.*, “The spatial clustering analysis of COVID-19 and its associated factors in mainland China at the prefecture level,” *Sci. Total Environ.*, vol. 777, p. 145992, 2021, doi: 10.1016/j.scitotenv.2021.145992.

[5] D. Guleryuz, “Forecasting Outbreak of COVID-19 in Turkey; Comparison of Box–Jenkins, Brown’s Exponential Smoothing and Long Short-Term Memory Models,” *Process Saf. Environ. Prot.*, p. 1, 2021, doi: https://doi.org/10.1016/j.psep.2021.03.032.

[6] K. Nikolopoulos, S. Punia, A. Schäfers, C. Tsinopoulos, and C. Vasilakis, “Forecasting and planning during a pandemic: COVID-19 growth rates, supply chain disruptions, and governmental decisions,” *Eur. J. Oper. Res.*, vol. 290, no. 1, pp. 99–115, 2021, doi: 10.1016/j.ejor.2020.08.001.

[7] P. Melin, J. C. Monica, D. Sanchez, and O. Castillo, “Analysis of Spatial Spread Relationships of Coronavirus (COVID-19) Pandemic in the World using Self Organizing Maps,” *Chaos, Solitons and Fractals*, vol. 138, p. 1, 2020, doi: 10.1016/j.chaos.2020.109917.

[8] M. R. Mahmoudi, D. Baleanu, Z. Mansor, B. A. Tuan, and K. H. Pho, “Fuzzy clustering method to compare the spread rate of Covid-19 in the high risks countries,” *Chaos, Solitons and Fractals*, vol. 140, pp. 1–9, 2020, doi: 10.1016/j.chaos.2020.110230.

[9] A. Olivieri, G. Palù, and G. Sebastiani, “COVID-19 cumulative incidence, intensive care, and mortality in Italian regions compared to selected European countries,” *Int. J. Infect. Dis.*, vol. 102, pp. 363–368, 2021, doi: 10.1016/j.ijid.2020.10.070.

[10] B. Kucukefe, “Covid-19’un OECD Ülkeleri ve Çin’de Makroekonomik Etkisinin Kümeleme Analizi,” *Ekon. Polit. Finans Araştırmaları Derg.*, vol. 5, pp. 280–291, 2020, doi: 10.30784/epfad.811289.

[11] M. Azarafza, M. Azarafza, and H. Akgün, “Clustering method for spread pattern analysis of corona-virus (COVID-19) infection in Iran,” *medRxiv*, p. 1, 2020, doi: 10.1101/2020.05.22.20109942.

[12] S. A. Rizvi, M. Umair, and M. A. Cheema, “Clustering of Countries for COVID-19 Cases based on Disease Prevalence, Health Systems and Environmental Indicators,” *medRxiv*, p. 1, 2021, doi: 10.1101/2021.02.15.21251762.

[13] V. Zarikas, S. G. Poulopoulos, Z. Gareiou, and E. Zervas, “Clustering analysis of countries using the COVID-19 cases dataset,” *Data Br.*, vol. 31, p. 105787, 2020, doi: https://doi.org/10.1016/j.dib.2020.105787.

[14] R. M. Carrillo-Larco and M. Castillo-Cara, “Using country-level variables to classify countries according to the number of confirmed COVID-19 cases: An unsupervised machine learning approach,” *Wellcome Open Res.*, vol. 5, pp. 1–22, 2020, doi: 10.12688/wellcomeopenres.15819.3.

[15] TurkStat, “Turkish Statistical Institute,” 2020. [Online]. Available: http://www.tuik.gov.tr/Start.do

[16] ILO, “International Labour Organization,” 2020. [Online]. Available: https://www.ilo.org/.

[17] Worldbank, “World Bank Open Data,” 2021. [Online]. Available: https://data.worldbank.org/.

[18] D. Guleryuz, “Evaluation of waste management using clustering algorithm in megacity Istanbul,” *Environ. Res. Technol.*, vol. 3, no. 3, pp. 102–112, 2020.

[19] B. Purnima and K. Arvind, “EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN,” *Int. J. Comput. Appl.*, vol. 105, no. 9, pp. 17–24, 2014, [Online]. Available: https://www.ijcaonline.org/archives/volume105/number9/18405-9674.

[20] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987, doi: https://doi.org/10.1016/0377-0427(87)90125-7.

[21] J. L. Myers and A. D. Well, *Research Design and Statistical Analysis*, 2nd Ed. Lawrence Erlbaum., 2003.

[22] S. Wright, “Correlation and causation,” *J. Agric. Res.*, vol. 20, no. 7, pp. 557–585, 1921.