Big Data to monitor the epidemiological situation of COVID-19 (Corona virus): Application of Cluster analysis for Algerian provinces.

توظيف البيانات الضخمة لمراقبة الحالة الوبائية لفيروس COVID-19 فيروس كورونا): تطبيق التحليل العنقودي للولايات الجزائرية.

Dr. DELMADJI Ahlam

Higher School of Commerce HSC, Economic Reforms, Development and Integration into the World Economy laboratory –Algiers (Algeria), a_delmadji@esc-alger.dz

Received: 12/04/2022 **Accepted:** 29/06/2022 **Published:** 30/06/2022

Abstract:

The purpose of this paper is to analyze the similarities of 48 provinces of Algerian government where Corona-virus pandemic (covid19), which has outset from China and infected thousands of deaths around the world .It has been profoundly affecting almost all countries at all levels and this situation placed researchers before a biggest challenge to confront. Using Big Data is one of the most known tools to monitor the spread of this virus .In this regards, the similarities of the 48 provinces were investigated using the Hierarchical Cluster Analysis based on 7 variables (Cumulative confirmed cases, New confirmed case, Incidence rate, Cumulative deaths cases, New death case, Mortality rate and Fatality rate %). As a result, the 48 provinces have been grouped in 7 clusters in which high densely populated provinces were the most effected like Alger and Oran. Whilst, the low densely populated provinces constituted the lowest incidence of disease.

Key words: Big data, Covid-19, Algeria, Hierarchical cluster analysis.

JEL Classification Codes: C8,I1,I3,G02

الملخص:

تهدف هذه الورقة البحثية إلى تحليل أوجه التشابه بين 48 ولاية في الجزائر من حيث نسبة تفشى جائحة فيروس كورونا (covid19)، الذي انطلق من الصين متسببا في آلاف الوفيات و تاركا وراؤه نتائج عكسية بشكل عميق على جميع البلدان تقريبًا و على جميع المستويات. في هذا الصدد, يعد استخدام البيانات الضخمة من أكثر الأدوات المعروفة لرصد ومتابعة انتشار هذا الفيروس ، لذلك ومن اجل تحقيق هذا الهدف، تم الاعتماد على طريقة التحليل الهرمي العنقودي باستخدام سبعة متغيرات وتشمل (الحالات المؤكدة التراكمية ، الحالات المؤكدة الجديدة ، معدل الحدوث ، حالات الوفيات التراكمية ، حالات الوفيات الجديدة ، معدل الوفيات ونسبة٪ الوفيات). وتوصلت هذه الدراسة إلى تجميع الـــ 48 ولاية جزائرية في 7 مجموعات حيث كانت الولايات ذات الكثافة السكانية العالية هي الأكثر تضررًا مثل الجزائر ووهران. في حين شكلت الولايات ذات الكثافة السكانية المنخفضة أقل معدلا للإصابة بالأمراض وشملت مناطق الهضاب والجنوب عموما.

الكلمات المفتاحية: البيانات الضخمة ، كوفيد -19 ، الجزائر ، التحليل الهرمي العنقودي.

تصنيفات JEL : **C8 ، I1، I3، G02**

Corresponding author: DELMADJI Ahlam, **e-mail :** a_delmadji@esc-alger.dz

INTRODUCTION:

COVID-19 is the global crisis of our time and the utmost human challenge. This pandemic is much more than a health crisis, it is also a socioeconomic crisis that putted all countries in suffering and under pressure which will have devastating social, economic and political impacts and will take time and disappear (WHO, 2020, pp. 2-3) . It has started in China, the first case was detected on December 1st, 2019 by Chinese doctors which have warn about an unknown new virus named covide-19 and appeared in people working in the wholesale seafood market (Lu, Stratton, & Tang, 2020, p. 401) from HUANAN to Wuhan in central China's Hubei province. Meanwhile, they have been quickly kept a secret by the Chinese government hoping to solve this crisis internally (Yu & Li, 2021, p. 348), but it turns out that this virus is similar to the SARS epidemic that arose on 2002 in China and which was very deadly.

The virus has spread rapidly throughout Chinese territory during the following months and has expand to the rest of the world in early 2020.Consequently, The World Health Organization declared an international concern emergency on January, 30th, 2020 (Dhama *& al.*, 2020, p. 02).

On the basis of the reports that the spread of this virus due to citizens transmission between cities, most of governments have adopted several restriction and preventive measures to limit its spread such as(Uddin, Imam, & Ali Moni, 2021, pp. 1-8) :travel and transportation restrictions, shutting down border gates, halting the international flights, controlling the transit within the provinces..These Government’s decisions aimed to safeguard people’s lives and health and to relieve the economic impact of it.

Two years have passed and till today (28/01/2022), there is no specific good treatments for COVID-19 exist for the time. The number of dead has reached approximately 5 651 185 people worldwide. Scientists and researchers are testing a variety of possible medications and vaccines like : Pfizer-BioNTech,Moderna (Pilishvili *& al.*, 2021, p. 753) , and working hard right till now to develop an effective one.

For Algerian government, the first case has declared by the Pasteur Institute on February 25, 2020 when an Italian is tested positive for covid19. And from March 1, 2020, sixteen members of the same family, in the Blida province, have been infected during a wedding party. Since then, Blida province becomes the corona-virus epidemic epicentre in Algeria and infections number was increasing(Leveau, Aouissi, & Kebaili, 2022, pp. 1-6). Till today (28/01/2022), a total of cases 247 568 have been detected according to the Ministry of health.

In the course of this crisis, we are seeing the role played by the Big Data technology, data mining, data science and artificial intelligence to response and monitor of the corona virus spread (Alsunaidi *& al.*, 2021, pp. 1-24), in which computer engineers in any country have developed digital platforms to collect the Data. In this regards, using Big Data analytics examines the hidden correlations and provides answers almost immediately which help governments to combat the outbreak of covid19 and figure out best responses(Sheng, Amankwah-Amoah, Khan, & Wang, 2021, pp. 1-20).

Researchers in this area are nowadays increasing, in this context the purpose of this paper is to provide firstly a literature review in which this subject has been treated and secondly to use the cluster analysis technique in order to asses and monitor the infections number in

Algeria which will help to reach and understand the virus transmission among the populations and affect positively on the governments decisions regarding its legal actions.

For this reason, this research will focus on the Algerian government and cover its trends and actions in terms of Covid-19 and fighting this disease. According to the preceding background, this work is based on, and tries to answer, the major following question:

What is the role played by Big data and Hierarchical cluster analysis to analyze and monitor the Corona-virus pandemic (Covid-19) in Algeria in order to rationalize its decisions and legal actions?

Given the main question, this research aims to answer the following two sub-questions:

- How effective is the political decisions and legal actions to fight Covid_19 in Algeria?
- What are the Big Data analytics and how can we monitor the epidemiological situation of COVID-19 (Corona virus) with cluster analysis method?

In this context, the research methodology adopted in this study is based on the descriptive and analytical approaches in the theoretical framework. On the other hand, the quantitative approach will also be employed in the empirical framework based on the multivariate cluster observations using the Hierarchical cluster analysis to analyze the similarities of 48 provinces of the Algeria.

To carry out and conduct this study, the rest of this research paper is structured as follows. First, we introduce the literature review regarding the previous studies in this area as well as developing the hypotheses. We then highlight the research methodology and empirical framework. After that we will discuss the main findings and provide recommendations and suggestions.

1- Literature review and hypotheses developing

Till today (28/01/2022), Studies about this pandemic are increasing over time. In this context, we tried through this paper giving an overview of previous studies and researches that tackle this issue:

Sanjay Kumar published a research paper about Monitoring Novel Corona Virus (COVID‑19) Infections in India by Cluster Analysis. The objective of this study was to use hierarchical cluster analysis, Box plot, Dendrograms and Data mining to classify Indian states based on COVID-19 status and to optimize monitoring tools in affected provinces which will help the government to understand and follow accurately the spread of this pandemic (Kumar, 2020, pp. 417-425). The results obtained indicated that the hierarchal cluster analysis diagram showed that the 27 states and 5 UTs have been grouped into six optimum numbers of clusters, all provinces were affected with the pandemic except some provinces under cluster II, and t Maharashtra state had the high number of confirmed cases. On the other hand, Box plot results shows that provinces under clusters III and VI needed monitoring techniques whilst provinces under clusters I, IV and VI required more medical facilities.

Besides, another study was conducted by **Mehdi Azarafza** with other researchers to analyse the spread pattern of corona-virus (COVID-19) infection in Iran using clustering analysis method for time series modelling. Based on national data sources, the study was spanning from the period February 19 to March 22, 2020 and covered Iran as the provincial level. According to the results obtained, tow provinces were the most cities located within the crisis areas of IRAN **(Akgün, Mehda, & Azarafza, 2021, pp. 1-6)**.

Based on the above results, we frame our first hypothesis as follows:

H01: Clustering the similar provinces in Algeria will help to determine the most affected areas which lead to make better trends in monitoring the pandemic.

As well, a research paper delivered by **Syeda Amna Rizvi and all** covered data of 79 countries and 18 social, economic, health and environmental indicators that are associated with COVID-19 spread. It aimed to cluster similar countries according these factors. Thus, it will be possible to take proactive decisions to fight against the Covid- 19 **(Rizvi, Umair, & Cheema, 2021, pp. 1-12)**. Using k-means algorithm, the results has unveiled that the model was able to group the countries into 4 clusters based on cluster mean percentage of COVID-19 confirmed cases and COVID-19 death cases. Cluster 1 includes 33 between developed and developing countries showed third highest percentage. Cluster 2 contains only developed countries with second highest cluster mean percentage. The third Cluster consists the two countries "China and India" and showed the highest percentage .Whilst; cluster 4 contains 23 developing countries, and has the least percentage of COVID-19 confirmed cases and COVID-19 death cases. These results lead the policy makers to make better decisions in monitoring the pandemic.

Accordingly, the second hypothesis is framed as follows:

H 02: Developing management and monitoring tools based on Big data analytics and statistics diagrams will help the Algerian government and decision makers to fight against the covid19.

Furthermore, **Vasilios Zarikas and all** published a data article entitled "Clustering analysis of countries using the COVID-19 cases dataset ".The study aimed to use Johns Hopkins epidemiological data in order to cluster countries with respect to active cases, active cases per population and active cases per population and per area(Zarikas, Poulopoulos, Gareiou, & Zervas, 2020, pp. 1-8) . The processing of the data has been done using a new specially designed clustering algorithm adapted to compare the various COVID time-series of different countries. Further, the results obtained indicate that the disease spreads more easily in countries that have dense big cities. In this context, the proposed of the third hypothesis will be as follows:

H03 : The densest populated areas are the hotspots for the spread of the disease in Algeria

Moreover, the researchers **Özlem, Murat and Hikmet Şevgin** in their recent paper, examined the similarities of 50 countries using multivariate statistical analysis techniques including Hierarchical Cluster Analysis and Multi -dimensional Scaling Analysis in terms of covid-19 indices (Kayri & Sevgin, 2021, pp. 308-315). Seven variables used in this analysis and as a conclusion of this research, results in both methods have been found very close to each other. Developed and developing countries showed similarities in combating of the epidemic, this result went beyond what was expected in terms of developed countries will be more effective in controlling the covid-19.

Besides, **Narayana Darapaneni** and his team published a study about "Machine learning approach for clustering of countries to identify the best strategies to combat Covid-19" **(Darapaneni & *al.*, 2021, pp. 1-7).** The purpose of this study is to determine the impact of the governments' measures on covid-19 consequences and identify the adequate strategies they can adopt for policy makers. This global study used Machine Learning techniques to classify countries based on "Demographic, economic, health, and weather conditions with Covid-19 epidemiology data. The results and observations obtained indicate that countries which started a stricter containment measures quickly and eased out it gradually rather like Argentina ,were more in control then countries which eased out the restrictions within a month or two and had to go back to a strict restrictions such as France, Israel, Jordan, Trinidad and Tobago.

2-The study area

2-1-Overview of Algeria's case study

Algeria is an African northern country located in the centre of Maghreb region ,it is a largest one in the continent, Its limits includes the Mediterranean Sea , Morocco , Western Sahara and Mauritania , Niger and Mali , Libya and Tunisia .It is inhomogeneous and deeply dissymmetrical area, it is divided into 3 parts (Lebbihiat, Atia, Arıcı, & Meneceur, 2021, p. 03): the Tell Mountains near the sea, the steppe-like high plateaus to their south, and the Saharan Atlas farther down that abuts the desert which stretches till the south for 990 miles.

According to the **STATISTA** rapports on 2020, the Algerian population amounted to approximately 45.02 million inhabitants(STATISTA, 2022) who are still sparsely populated overall and the majority is concentrated on the coast in Mediterranean climate.

2-2-The Algerian measures to fight COVID-19

In the respect of the ongoing outbreak of Corona-virus pandemic (COVID-19) and its impact on the worldwide health threat of citizens, the Algerian government implemented a set of measures to contain the spread of this virus and mitigate the negative impact of Covid-19. To this end, we present these measures in the following (Ministry of Health, 2020):

- the closure of all land borders;
- The suspension of international air and maritime travel;
- the suspension of all domestic flights;
- the suspension of the Cultural and sporting events;
- The closure of all the schools and universities as well as the restaurants, cafes and public baths;
- A curfew has been imposed in several provinces according to the spread percent;

- the wearing of protective mask has been imposed in the public places;
- Establishing a center in Bourj Al-Kefan on the toll-free number 30-30 to receive calls from people who suspect they are symptomatic;

3- Research Methodology

3-1-Theoretical framework

Thanks to Big data analytics, data science, AI and ML, researchers around the world contributed and provided, through their studies and researches papers ,a great efforts to understand ,control and monitor the pandemic Covid -19 (Bragazzi *& al.*, 2020, p. 02). Big data includes a complex and large data that are very difficult to be deal with any Software. It always defined by its 5-V which are(Kapil, Agrawal, & Khan, 2016, p. 01) :"Volume, Velocity, Veracity, Variety, and Value". Besides, Big data analytics is the complex process to examine big data using many methods(Rajaraman, 2016, p. 02) where machine learning plays a crucial role nowadays. One of its techniques called hierarchical cluster analysis is used in this study.

Hierarchical clustering analysis, also known as Hierarchical clustering is one of the cluster analysis methods and a Machine learning technique. It is an algorithm works via grouping similar observations or variables into groups called mostly clusters or tree of clusters(Murtagh & Contreras, 2012, p. 87), each cluster is distinct from another and at the same, data within one cluster are broadly similar among them. This study will be conducted using this technique and main steps will be clarified in the next sections.

3-2- Empirical framework

The methodology followed to deal with this thematic includes 3 steps:1/ Dataset used for the Analysis,2/ Graphical data representation and description,3/ Procedure and technique: The hierarchal cluster analysis

3-2-1 - Data set used for the Analysis

Based on the Algerian Ministry of Health documents and the publications of official website for COVID-19 epidemiological situation in Algeria delivered by the National institute of public health: "https://www.insp.dz/index.php/publications/situation-epidemiologique-covid19.html" ,we have collected data for 48 provinces arranged in alphabetical order: *Adrar, Ain Defla, Ain Temouchent, Algiers, Annaba, Batna, Béchar,Béjaïa, Biskra, Blida, Bordj Bou Arréridj, Bouira ,Boumerdès, Chlef, Constantine, Djelfa, El Bayadh, El Oued, El Taref, Ghardaïa, Guelma, Illizi, Jijel, Khenchela, Laghouat, M'Sila, Mascara, Médéa, Mila, Mostaganem, Naâma, Oran, Ouargla, Oum El Bouaghi, Relizane, Saïda, Sétif, Sidi Bel Abbes, Skikda, Souk Ahras, Tamanrasset, Tebessa, Tieret, Tindouf, Tipaza, Tissemsilt, Tizi Ouzou and Tlemcen.*

3-2-2 Graphical data representation and description

Graphical representation is a way to present the data and make clearly the information. In this following, we will look at the charts presentation of the covid-19 dataset along with its merits.

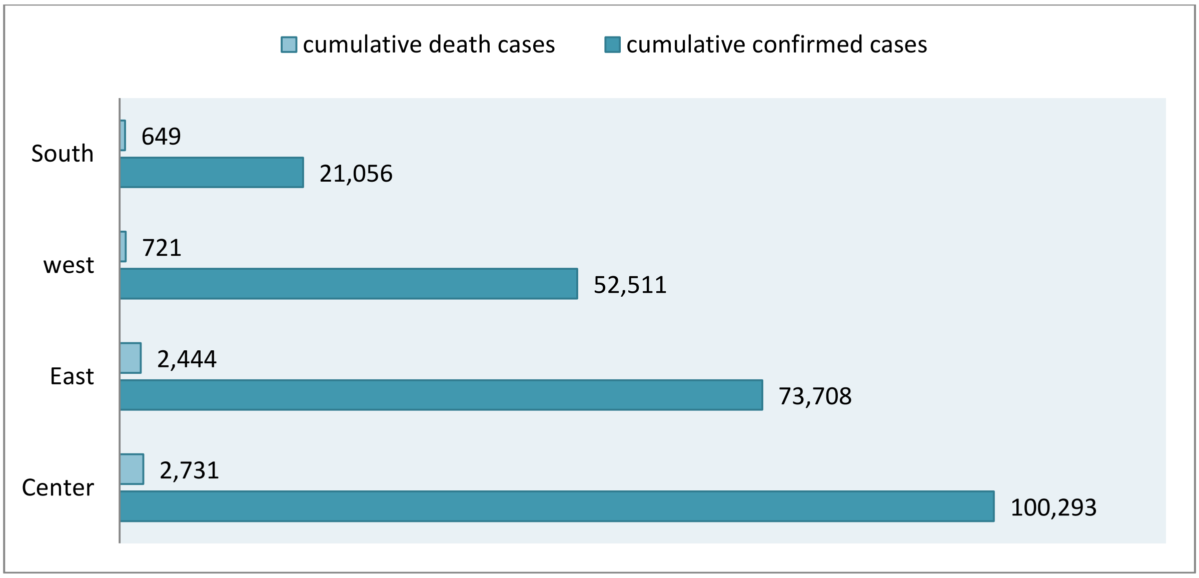
According to the data delivered by the Corona Virus Monitoring and Follow-up Committee of the Ministry of Health and Population, The first outbreak was recorded in Blida province on the 12 March 2020. It has taken just one month to covers all the Algerian provinces(Boukhatem, 2020, p. 03) .

The following charts present the covid-19 epidemiological situation per region and per provinces, as well as a summary of the data set of 2 years till 28 January, 2022 depending upon three parameters: Total number of confirmed cases, Total number of death cases and the Incidence rate.

According the figure 01, the total numbers of confirmed cases per the “center, east, west and south” regions are: 100 293, 73708, 52511 and 21056 cases respectively. Whilst, the total number of deaths cases of these four regions are: 795 , 532, 420 and 123 cases respectively. However, figure 02 shows the Evolution of the cumulative incidence rate through the four regions till January 28, 2022. Based on data set of National institute of public health (NIPH, 2022, pp. 2-3), the Center region crossed the threshold of 100000 cases "100 293 cumulative confirmed cases” with an incidence of 654.46 cases per 100,000 inhabitants compared to previous months.

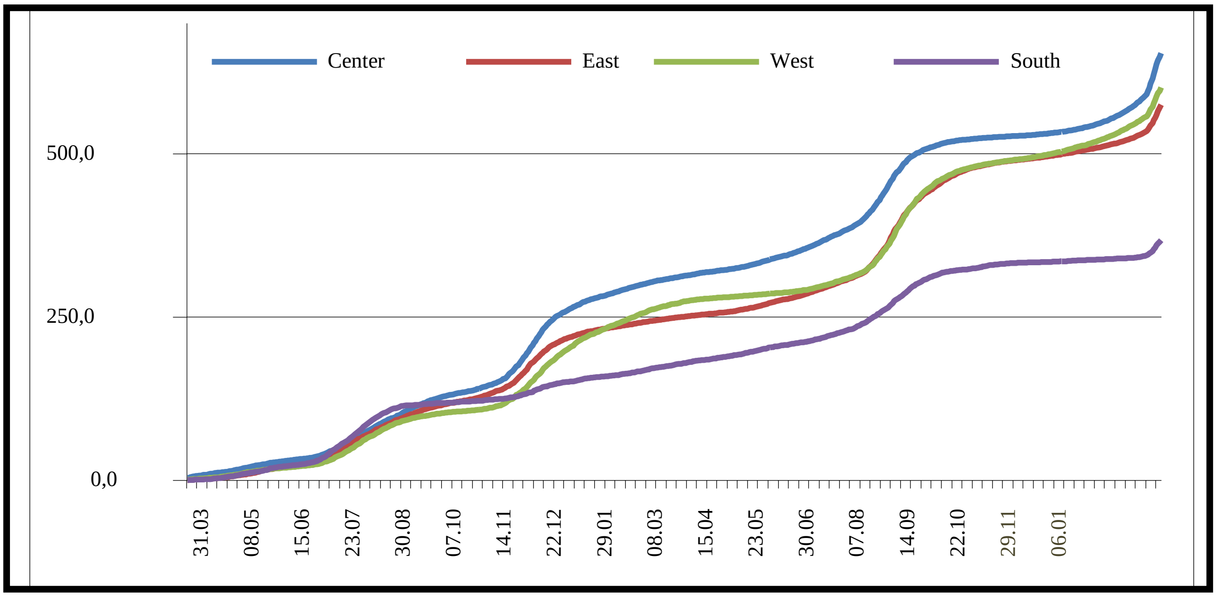
For the East and West regions, the incidence rate recorded 574.95 for the East and 601.41 cases per 100,000 inhabitants for the West. While, in the South region the incidence rate recorded 367,85 cases per 100 000 inhabitants.

Fig (1): Distribution of confirmed and deaths cases by region in Algeria till January 28, 2022



Source: **Computed by the researcher based on INSP data.**

Fig (2): Evolution of the cumulative incidence rate by health region till January 28, 2022

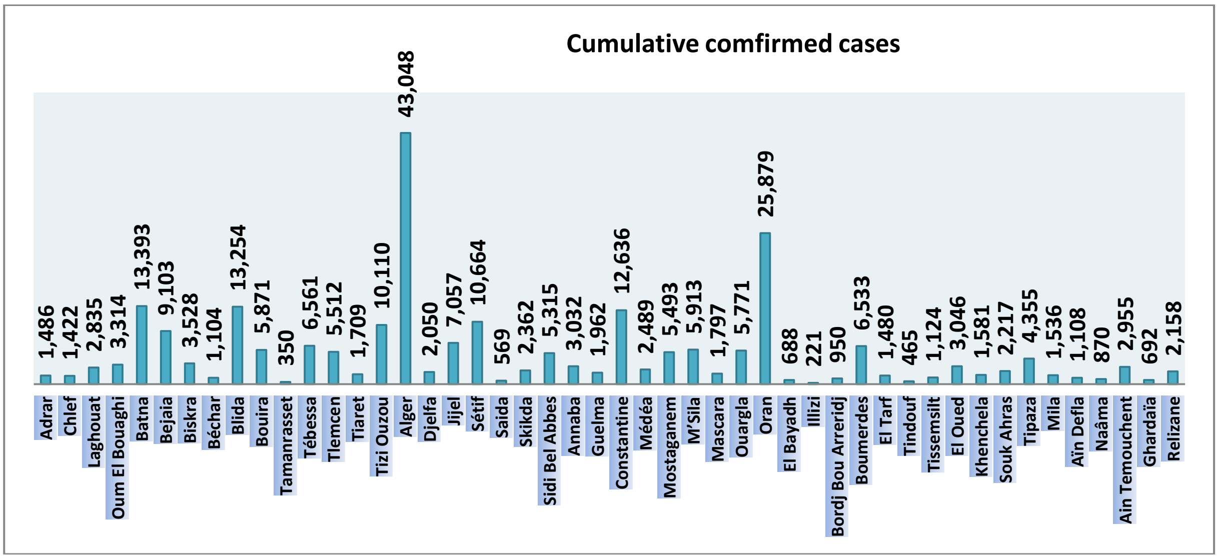


Source: Compiled by the researcher based on INSP data.

As stated by the Corona Virus Monitoring and Follow-up Committee of the Ministry of Health and Population on January 28, 2022, the total number of confirmed and deaths cases from the beginning of the appearance of the first case till this date are 247 568 and 6 545 respectively.

A maximum number of confirmed cases as shown in **Figure 03**, have been recorded in Oran and Algiers provinces which have exceed the 20000 cases, followed by Batna ,Blida, Tizi Ouzou, Constantine , Setif with number cases ranging between 10000 and 20000 but the rest of the provinces was in close proportions less than 10000 cases.

Fig (3): Cumulative confirmed cases per province till January 28, 2022

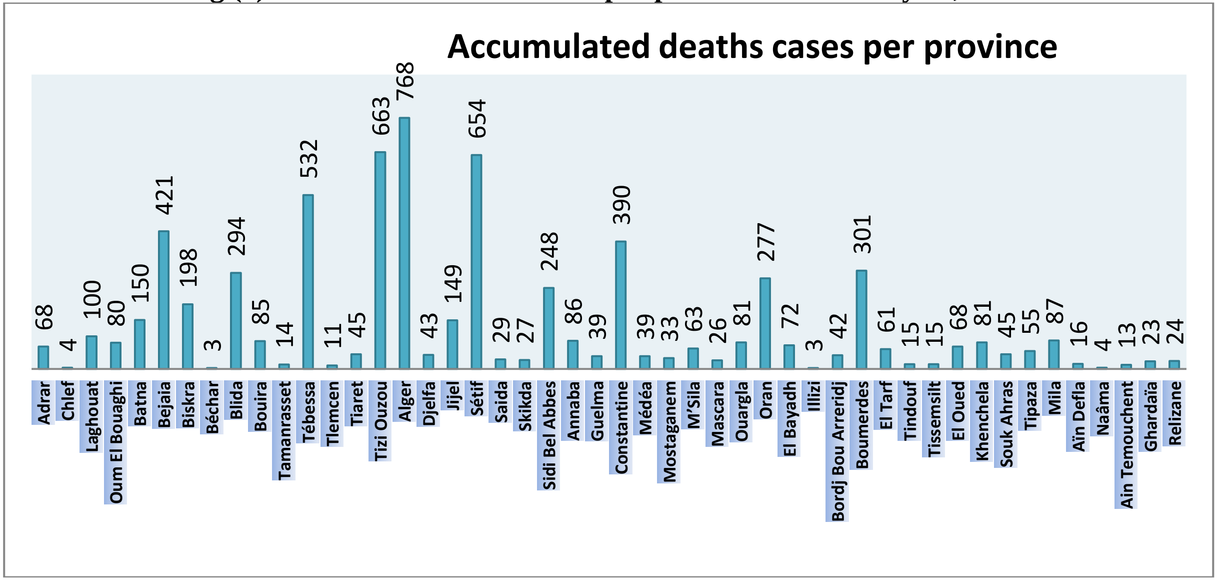


Source: Computed by the researcher based on INSP data.

Regarding the number of patients who died during this period, it has a close relationship with the number of confirmed cases as shown on the Figure 04 .

A highest number of deaths cases have been recorded in Algiers, Setif ,Tizi ouzou and Tébessa provinces which have exceed the 500 cases, followed by bejaia ,Blida, Sidi belabas, Constantine , Oran and Boumredes with number cases ranging between 200 and 500 . While the rest of the provinces was in close proportions less than 200 cases.

Fig (4) : Cumulative deaths cases per province till January 28, 2022



Source: computed by the researcher based on INSP data.

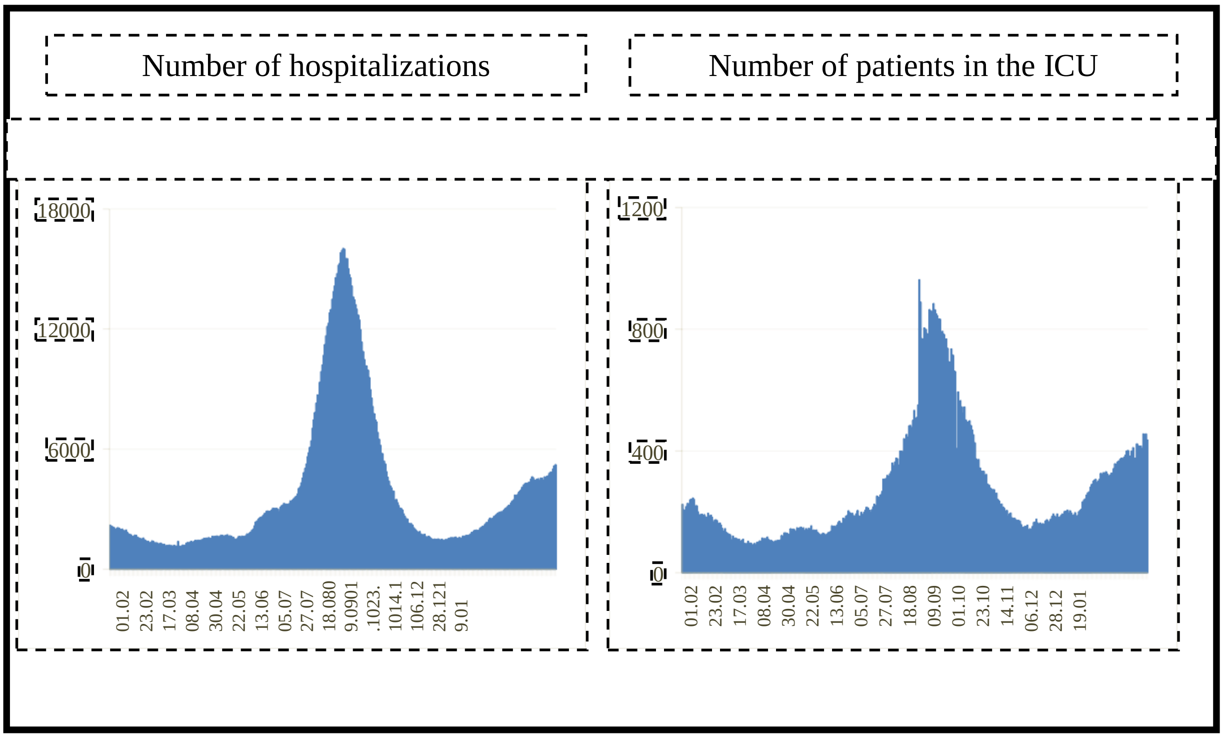
The following graph **"Figure 05"** visualizes the evolution of the number of patients hospitalized and the number of patients in intensive care between February 01, 2021 and January 28, 2022.

The average daily number of hospitalized patients is 4927,4 during the last week. According to the Ministry of Health(NIPH, 2022, p. 05), during the week of January 22 to 28, sixteen provinces recorded an increase in the daily number of hospitalizations. These include: Mascara, Souk Ahras, Ouargla, Boumerdès, Aïn Defla, Biskra, Tizi Ouzou, Tiaret, Mila, Algiers, M'Sila, Chlef, Blida, Médéa, Batna and Sétif.

Regarding the intensive care units ICU, the estimated average number was 432,3 patient per day during the week of January 22 to 28. Twenty-four provinces recorded an increase in intensive care hospitalizations, seven of which have a growth rate more than 50% compared to the previous week. These provinces are Souk Ahras, Laghouat, El Oued, Saïda, M'Sila Oum El Bouaghi and Médéa.

On the other hand, eight provinces recorded no case in the intensive care unit ICU, these provinces was mainly the southern provinces and some northern provinces including: Béchar, Tebessa, El Bayadh, Illizi, Tindouf, Tissemsilt, Ain Defla and Ghardaia.

Fig (5) : Daily evolution of the number of hospitalizations and the number of patients in the ICU between February 01, 2021 and January 28, 2022



Source: Compiled by the researcher based on INSP data.

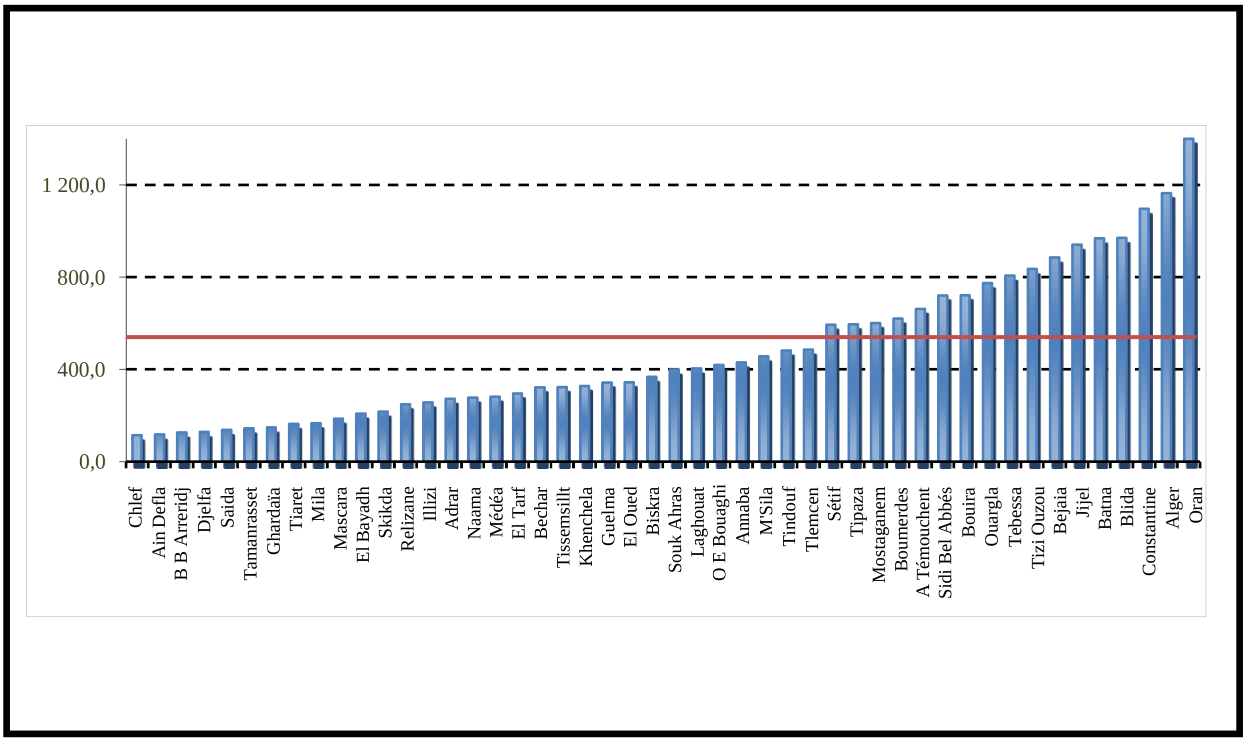
The graph below “**Figure 06**” illustrates the incidence rates of confirmed cases per province till January 28.

Based on National Institute of Public Health declaration(NIPH, 2022, p. 03) and as of 28th January, 247 568 cumulative cases have been declared at national level, with an incidence of 581,15 cases per 100 000 inhabitants versus 545,36 on January 21. The increase is estimated at 6.6%. In this context, seventeen provinces recorded an incidence rate higher than the national rate, three of which exceeded the threshold of 1000 cases per 100 000 inhabitants. These are, in descending order, Oran (1399, 35 cases per 100,000 inhabitants), Algiers (1164, 11) and Constantine (1095, 91).

In one week, between January 21 and 28, nineteen provinces recorded a growth rate above 5.0%, seven of which had a growth rate above 10%; these are, Tlemcen , Skikda , Algiers , Sidi Bel Abbes , Batna , Médéa and Tipaza .

On the other hand, few provinces have a growth rate less than 1.0% including: Djelfa, Guelma, Khenchela and Aïn Defla. In Tindouf, no new cases were declared, the incidence rate remained stable, it is 481.70 cases per 100,000 inhabitants on the two aforementioned dates.

Fig (6) :Distribution of the incidence rate per provinces till January 28, 2022



Source: Compiled by the researcher based on INSP data.

3-2-3 Procedure and technique: The hierarchal cluster analysis

In this study, the Minitab software was used to carry out the cluster analysis. We aim to cluster the sample observations "48 provinces" into groups based on the similarities found in the data sets "**See the Annexe**". These groups are depending upon seven variables including: 1/Cumulative confirmed cases, 2/New confirmed case, 3/Incidence rate, 4/Cumulative deaths cases, 5/New death case, 6/Mortality rate and 7/Fatality rate %.

Based on the multivariate cluster observations, in which a squared “Euclidean Distance” and “complete Linkage” are used to compute the similarities within the data set. The **DINDOGRAM** of the hierarchical cluster analysis was obtained **“Figure 07”.**

This method is the most commonly used technique that uses by several researchers in case of a large number of observations.

We present in the following the main results obtained using Minitab software:

a- Amalgamation Steps for one cluster

In this results (**Table 01)** ,the data contain a total of 48 observations : in step one: two clusters observations 01 and 36 in the worksheet joined the form of new cluster, this steps created 47 clusters in the data with the similarity level of 99,9362 and distance level of 27,3.

All the similarity level is high and the distance level is low ,so the number of cluster is high to be useful.

At each subsequence step, as new clusters are formed, the similarity level decreases and the distance level increases.

At the final step, all the observations are joined into a single cluster.

This table shows the clusters that were joined at each step, the distance between the clusters and the similarity.

Table (1): Amalgamation Steps “”Euclidean Distance, Complete Linkage”

Step	Number of clusters	Similarity level	Distance level	Clusters joined		New cluster	Numberof obs. in new cluster
1	47	99,9362	27,3	1	36	1	2
2	46	99,9346	28,0	8	38	8	2
3	45	99,8139	79,7	32	47	32	2
4	44	99,7860	91,7	23	39	23	2
5	43	99,7775	95,3	14	29	14	2
6	42	99,7359	113,2	1	40	1	3
7	41	99,7219	119,1	10	30	10	2
8	40	99,7157	121,8	13	27	13	2
9	39	99,6659	143,2	21	26	21	2
10	38	99,6592	146,0	20	32	20	3
11	37	99,6487	150,5	2	43	2	2
12	36	99,6237	161,2	34	44	34	2
13	35	99,6197	162,9	17	48	17	2
14	34	99,6003	171,3	11	33	11	2
15	33	99,4821	221,9	3	23	3	3
16	32	99,4614	230,8	17	24	17	3
17	31	99,4147	250,8	4	7	4	2
18	30	99,3970	258,3	8	45	8	3
19	29	99,3548	276,4	1	2	1	5
20	28	99,3424	281,7	5	9	5	2
21	27	99,3302	287,0	8	34	8	5
22	26	99,3075	296,7	21	41	21	3
23	25	99,2886	304,8	12	35	12	2
24	24	99,2174	335,3	3	46	3	4
25	23	99,1841	349,6	10	28	10	3
26	22	99,1677	356,6	11	37	11	3
27	21	99,1066	382,8	1	14	1	7
28	20	99,0528	405,8	13	22	13	3
29	19	98,8710	483,7	11	20	11	6
30	18	98,7615	530,6	17	21	17	6

31	17	98,5870	605,4	15	19	15	2
32	16	98,5000	642,7	12	18	12	3
33	15	98,3705	698,2	10	13	10	6
34	14	98,3640	700,9	3	4	3	6
35	13	98,0867	819,8	5	25	5	3
36	12	97,8865	905,5	8	11	8	11
37	11	97,4781	1080,5	1	17	1	13
38	10	96,4214	1533,3	3	42	3	7
39	9	96,2537	1605,1	6	15	6	3
40	8	95,8763	1766,8	10	12	10	9
41	7	94,7055	2268,5	1	8	1	24
42	6	90,3175	4148,5	1	3	1	31
43	5	89,9550	4303,8	5	6	5	6
44	4	83,9616	6871,8	1	10	1	40
45	3	69,2049	13194,4	1	5	1	46
46	2	59,8988	17181,7	16	31	16	2
47	1	0,0000	42845,8	1	16	1	48

Source: Computed by the researcher using Minitab Software

b- Final Partition

Based on the distance and the similarity results above of "**table 01**", it indicates that 07 clusters are reasonably sufficient for the final partition

After rerunning the analysis and specify 7 clusters, additional tables that describe the characteristics of each cluster that is included in the final partition "**Table 03**". The decision about the final grouping is also called cutting the dindogram.

For more information about Cluster Centroids and Distances Between Cluster Centroids, see appendix 02 and 03.

Table (2) : Final Partition for one cluster

	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance from centroid
Cluster1	48	2524763189	4372,53	37904,0

Source: Computed by the researcher using Minitab Software

Table (3) : Final Partition for 7 cluster

	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance from centroid
Cluster1	24	10386703	571,889	1140,31
Cluster2	7	1745219	399,629	1069,65
Cluster3	3	386791	337,882	481,04

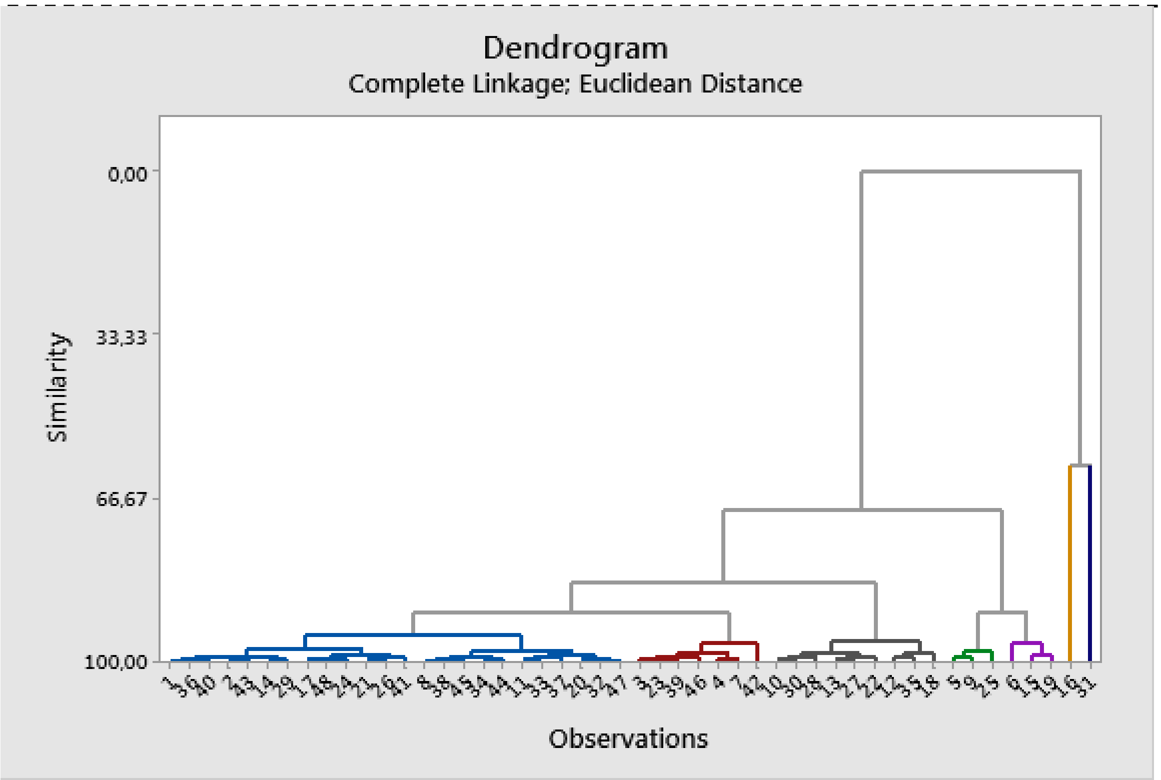
Cluster4	3	1339441	597,941	877,94
Cluster5	9	3202282	533,773	1086,58
Cluster6	1	0	0,000	0,00
Cluster7	1	0	0,000	0,00

Source: Computed by the researcher using Minitab Software

c- DENDROGRAM of hierarchal cluster analysis.

At final stage, a DENDROGRAM was created using a final partition of 7 clusters; each cluster joins the provinces that share common characteristics into a group depending upon on seven variables of covid-19 in Algeria. The following figure presents this DENDROGRAM:

Fig(07) : Euclidean distances “DENDROGRAM” after infection clustering in Algeria



Note that: 1/Adrar .2/Chlef.3/Laghouat.4/Oum El Bouaghi.5/Batna.6/Bejaia.7/Biskra.8/Béchar. 9/Blida.10/Bouira.11/Tamanrasset.12/Tébessa.13/Tlemcen.14/Tiaret.15/Tizi Ouzou.16/Alger.17/Djelfa.18/Jijel.19/Sétif.20/Saida.21/Skikda.22/Sidi Bel Abbes.23/Annaba.24/Guelma.25/Constantine.26/Médéa.27/Mostaganem.28/M’Sila.29/Mascara.30/Ouargla.31/Oran.32/El Bayadh.33/Illizi.34/Bordj Bou Arreridj.35/Boumerdes.36/El Tarf.37/Tindouf.38/Tissemsilt.39/El Oued.40/Khenchela.41/Souk Ahras.42/Tipaza.43/Mila.44/Aïn Defla.45/Naâma.46/Ain Temouchent.47/Ghardaïa.48/Relizane

Source: Computed by the researcher using Minitab Software

3.2.4. Results and observations

Depending upon the figure 07 above, the Hierarchical cluster analysis DENDOGRAM grouped the 48 Algerian provinces into seven clusters:

Cluster 01: It is composed of 24 provinces that is in observations in bleu one corresponded to: Adrar, El Tarf, Khenchela, Tiaret Mila Mascara Djelfa, Relizane ,Guelma ,Skikda , Médéa Souk Ahras, Chlef,Ghardaïa, El Bayadh, Tindouf, Saida, Illizi, Tamanrasset, Aïn Defla, Bordj Bou Arreridj, Naâma, Tissemsilt and Béchar.

Cluster 02: It is directly to the right with red color ,it is composed of the following provinces: Laghouat .Annaba ,El Oued ,Ain Temouchent ,Tipaza ,Biskra and Oum El Bouaghi.

Cluster 03: It corresponded to the gray one and composed of 9 provinces: Bouira ,Tlemcen ,Mostaganem, M’Sila ,Ouargla Sidi Bel Abbes,Tébessa ,Jijel and Boumerdes.

Cluster 04: It is on green color and corresponded to three provinces: Blida ,Constantine and Batna.

Cluster 05: The pink one composed to 3 provinces: Sétif , Tizi Ouzou and Bejaia

And finally, cluster **06 and 07** with yellow and black color and composed just one province Alger and Oran respectively.

4- Discussion and hypotheses testing

According to the results obtained in our research, as well as the findings of previous studies cited in the literature review, it is too necessary than ever for policy-makers and governments to be up to date with digitalization and the uninterrupted new technologies . Furthermore, developing and establishing management tools based on Big data analytics, statistics diagrams, artificial intelligence and machine learning will be important to lead to make rational resolves and take proactive decisions to fight against the covid19. These findings are consistent with those of the study delivered by **Syeda Amna Rizvi** with his team treated above. As stated by them, clustering the similar countries help governments to make better decisions in monitoring the pandemic(Rizvi *&* *al.*, 2021). **So the first and second hypotheses are confirmed**

Moreover, our results support the findings reached by **Vasilios Zarikas and his team** in their study "Clustering analysis of countries using the COVID-19 cases dataset" discussed above in literature review(Zarikas *&* *al.*, 2020) **in which confirmed the third hypothesis**.. Our findings indicate that high densely populated provinces are considered the most vulnerable to the spread of the disease like Alger and oran seted out in figure 07 under the cluster 6 and 7 respectively. And Sétif , Tizi Ouzou and Bejaia corresponded to the cluster 5 and finally Blida ,Constantine and Batna under the cluster 04. Whilst, provinces with low population density are the lowest incidence of disease which mostly constitute the regions of the south and the plateaus in Algeria, and corresponded mostly to the clusters 1,2 and 3 as shown also apparently in figure 03 concerning the confirmed cases.

Conclusion

The present study aimed to identify, during the period «From 25 February 2020 to 28 January 2022, the reach and spread of corona-virus pandemic in Algeria. Therefore, a dataset regarding seven variables corresponded to «total confirmed cases, total deaths and incidence rate » have been collected based on the national sources (The Algerian Ministry of Health documents and the official website for COVID19 epidemiological map in Algeria as well as INSP documents).

In this regards, we used the Hierarchical clustering technique in attempt to classify the 48 provinces of Algeria. According to the results, It was found that the provinces under the clusters 01 and 02 need to enhancing the monitoring techniques like "evacuations, closedown, curfews" in order to help decision makers to understand and monitor the COVID-19.Wilst ,provinces under clusters 03,04,05,06 and 07 need intensive care to reduce the number of deaths.

Despite the widespread use of this method and its importance to help decision-makers to make rational and effective resolves, this study highlights some limitations that must be addressed to provide properly outcomes. Findings of this Cluster analysis method depend on the chosen variables. Subsequently adding or removing any variable will be lead to other results and clusters. As shortcoming example, the variable related to the number of recoveries cases has a great importance, which will add value to the results, but unfortunately due to the data lack regarding this variable in Algeria, thence our study did not include it.

As recommendation for future studies on this topic, these findings will open several new horizons for other further researches, both at micro and macro level. The aspect that requires the most attention from researchers is how to make assessment of the impact of Corona virus disease (COVID-19) pandemic on different sectors in Algeria, studying possible solutions and establish potential plans.

Bibliography List:

Akgün, H., Mehda, A., & Azarafza, M. (2021). Clustering method for spread pattern analysis of corona-virus (COVID-19) infection in Iran. doi: https://doi.org/10.35877/454RI.asci31109

Alsunaidi, S. J., Almuhaideb, A. M., Ibrahim, N. M., Shaikh, F. S., Alqudaihi, K. S., Alhaidari, F. A., . . . Alshahrani, M. S. (2021). Applications of big data analytics to control COVID-19 pandemic. *Sensors, 21*(7), 2282. doi: https://doi.org/10.3390/s21072282

Boukhatem, M. (2020). Novel coronavirus disease 2019 (COVID-19) Outbreak in Algeria: a new challenge for prevention. *Community Med. Health Care, 5*(1), 1035.

Bragazzi, N. L., Dai, H., Damiani, G., Behzadifar, M., Martini, M., & Wu, J. (2020). How big data and artificial intelligence can help better manage the COVID-19 pandemic. *International journal of environmental research and public health, 17*(9), 3176. doi: https://doi.org/10.3390/ijerph17093176

Darapaneni, N., Venkatasubramanian, J., Paduri, A. R., Kumar, P., Vigneswar, S., Thangeda, K. C., & Thakur, A. (2021). Machine Learning Approach For Clustering Of Countries To Identify The Best Strategies To Combat Covid-19. *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS),* 1-7. doi: 10.1109/IEMTRONICS52119.2021.9422621

Dhama, Kuldeep, Patel, S. K., Sharun, K., Pathak, M., Tiwari, R., . . . Panwar, P. K. (2020). SARS-CoV-2 jumping the species barrier: zoonotic lessons from SARS, MERS and recent advances to combat this pandemic virus. *Travel medicine and infectious disease, 37*, 101830. doi: https://doi.org/10.1016/j.tmaid.2020.101830

Kapil, G., Agrawal, A., & Khan, R. (2016). *A study of big data characteristics.* Paper presented at the 2016 International Conference on Communication and Electronics Systems (ICCES).

Kayri, M., & Sevgin, H. (2021). Investigation of Coronavirus Pandemic Indicators of the Countries with Hierarchical Clustering and Multidimensional Scaling. *Eastern Journal of Medicine, 26*(2), 308-315. doi: DOI: 10.5505/ejm.2021.72681

Kumar, S. (2020). Monitoring novel corona virus (COVID-19) infections in India by cluster analysis. *Annals of Data Science, 7*(3), 417-425. doi: https://doi.org/10.1007/s40745-020-00289-7

Lebbihiat, N., Atia, A., Arıcı, M., & Meneceur, N. (2021). Geothermal energy use in Algeria: A review on the current status compared to the worldwide, utilization opportunities and countermeasures. *Journal of cleaner production, 302,* 126950. doi: https://doi.org/10.1016/j.jclepro.2021.126950

Leveau, C. M., Aouissi, H. A., & Kebaili, F. K. (2022). Spatial diffusion of COVID-19 in Algeria during the third wave. *Geojournal*, 1-6. doi: https://doi.org/10.1007/s10708-022-10608-5

Lu, H., Stratton, C. W., & Tang, Y. W. (2020). Outbreak of pneumonia of unknown etiology in Wuhan, China: The mystery and the miracle. *Journal of medical virology, 92*(4), 401. doi: 10.1002/jmv.25678

Ministry of Health, P. a. H. R. (2020, 07/01/2021). Actions taken by the Algerian government, from https://covid19.sante.gov.dz/mesures-prise-par-le-gouvernement/#pll_switcher

Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2*(1), 86-97. doi: https://doi.org/10.1002/widm.53

NIPH, N. i. o. p. h. (2022). Epidemiological situation of January 28, 2022, from https://www.insp.dz/images/evenements/Coronavirus/Bulletin%20epidemiologique%20N258du%2028%20janvier%202022.pdf

Pilishvili, T., Fleming-Dutra, K. E., Farrar, J. L., Gierke, R., Mohr, N. M., Talan, D. A., . . . Hou, P. C. (2021). Interim estimates of vaccine effectiveness of Pfizer-BioNTech and Moderna COVID-19 vaccines among health care personnel—33 US sites, January–March 2021. *Morbidity and Mortality Weekly Report, 70*(20), 753. doi: 10.15585/mmwr.mm7020e2

Rajaraman, V. (2016). Big data analytics. *Resonance, 21*(8), 695-716.

Rizvi, S. A., Umair, M., & Cheema, M. A. (2021). Clustering of countries for COVID-19 cases based on disease prevalence, health systems and environmental indicators. *Chaos, Solitons & Fractals, 151*, 111240. doi: https://doi.org/10.1101/2021.02.15.21251762

Sheng, J., Amankwah-Amoah, J., Khan, Z., & Wang, X. (2021). COVID-19 pandemic in the new era of big data analytics: Methodological innovations and future research directions. *British Journal of Management, 32*(4), 1164-1183. doi: https://doi.org/10.1111/1467-8551.12441

STATISTA. (2022). Algeria: Total population from 2016 to 2026(in million inhabitants), from https://www.statista.com/statistics/408028/total-population-of-algeria/

Uddin, S., Imam, T., & Ali Moni, M. (2021). The implementation of public health and economic measures during the first wave of COVID-19 by different countries with respect to time, infection rate and death rate. *2021 Australasian Computer Science Week Multiconference*, 1-8. doi: https://doi.org/10.1145/3437378.3437384

WHO. (2020). COVID‑19 strategy update (as of 14 April 2020) *Weekly Epidemiological Record* (Vol. 95, pp. 185-208).

Yu, X., & Li, N. (2021). Understanding the beginning of a pandemic: China’s response to the emergence of COVID-19. *Journal of Infection and Public Health, 14*(3), 347-352. doi: https://doi.org/10.1016/j.jiph.2020.12.024

Zarikas, V., Poulopoulos, S. G., Gareiou, Z., & Zervas, E. (2020). Clustering analysis of countries using the COVID-19 cases dataset. *Data in brief, 31*, 105787. doi: https://doi.org/10.1016/j.dib.2020.105787

Appendices:

Appendix 01: Distribution of confirmed and deaths cases per Algerian provinces till 28,january,2022.

Code	province	Cumulative confirmed cases	new confirmed case	incidence rate	Cumulative deaths cases	new death case	Mortality rate	Fatality rate %
1	**Adrar**	1 486	25	271,87	68	0	12,44	4,58
2	**Chlef**	1 422	6	114,5	4	0	0,32	0,28
3	**Laghouat**	2 835	11	403,94	100	0	14,25	3,53
4	**Oum El Bouaghi**	3 314	23	418,86	80	0	10,11	2,41
5	**Batna**	13 393	248	968,58	150	0	10,85	1,12
6	**Bejaia**	9 103	54	885,2	421	1	40,94	4,62
7	**Biskra**	3 528	5	366,6	198	0	20,57	5,61
8	**Béchar**	1 104	16	321,16	3	0	0,87	0,27
9	**Blida**	13 254	50	969,5	294	0	21,51	2,22
10	**Bouira**	5 871	17	720,95	85	0	10,44	1,45
11	**Tamanrasset**	350	4	144,33	14	0	5,77	4
12	**Tébessa**	6 561	18	805,71	532	0	65,33	8,11
13	**Tlemcen**	5 512	69	485,84	11	0	0,97	0,2
14	**Tiaret**	1 709	1	163,01	45	0	4,29	2,63
15	**Tizi Ouzou**	10 110	60	835,41	663	2	54,79	6,56
16	**Alger**	43 048	445	1 164,11	768	2	20,77	1,78
17	**Djelfa**	2 050	0	128,64	43	0	2,7	2,1
18	**Jijel**	7 057	19	940,76	149	0	19,86	2,11
19	**Sétif**	10 664	31	593,83	654	0	36,42	6,13
20	**Saida**	569	5	136,92	29	0	6,98	5,1
21	**Skikda**	2 362	21	216,69	27	0	2,48	1,14
22	**Sidi Bel Abbes**	5 315	187	719,75	248	0	33,58	4,67
23	**Annaba**	3 032	28	429,85	86	0	12,19	2,84

24	**Guelma**	1 962	1	341,89	39	0	6,8	1,99
25	**Constantine**	12 636	91	1 095,91	390	0	33,82	3,09
26	**Médéa**	2 489	13	281,13	39	0	4,4	1,57
27	**Mostaganem**	5 493	41	600,73	33	0	3,61	0,6
28	**M'Sila**	5 913	28	456,11	63	0	4,86	1,07
29	**Mascara**	1 797	22	186,17	26	0	2,69	1,45
30	**Ouargla**	5 771	54	773,94	81	0	10,86	1,4
31	**Oran**	25 879	73	1 399,35	277	0	14,98	1,07
32	**El Bayadh**	688	0	207,95	72	0	21,76	10,47
33	**Illizi**	221	0	256,31	3	0	3,48	1,36
34	**Bordj Bou Arreridj**	950	19	125,95	42	0	5,57	4,42
35	**Boumerdes**	6 533	69	619,16	301	0	28,53	4,61
36	**El Tarf**	1 480	13	294,62	61	0	12,14	4,12
37	**Tindouf**	465	0	481,7	15	0	15,54	3,23
38	**Tissemsilt**	1 124	1	323,12	15	0	4,31	1,33
39	**El Oued**	3 046	7	343,7	68	0	7,67	2,23
40	**Khenchela**	1 581	0	326,36	81	1	16,72	5,12
41	**Souk Ahras**	2 217	7	399,28	45	0	8,1	2,03
42	**Tipaza**	4 355	59	594,12	55	0	7,5	1,26
43	**Mila**	1 536	4	166,05	87	0	9,41	5,66
44	**Aïn Defla**	1 108	3	117,45	16	0	1,7	1,44
45	**Naâma**	870	1	277,36	4	0	1,28	0,46
46	**Ain Temouchent**	2 955	3	661,65	13	0	2,91	0,44
47	**Ghardaïa**	692	0	147,91	23	0	4,92	3,32
48	**Relizane**	2 158	18	247,79	24	0	2,76	1,11
	Total	**247 568**	**1 870**	**581,15**	**6 545**	**6**	**15,36**	**2,64**

Source : INSP documents.

Appendixe02 :Cluster Centroids

Variable	**Cluster1**	**Cluster2**	**Cluster3**	**Cluster4**	**Cluster5**	**Cluster6**
Cumulative comfirmed cases	1349,58	3295,00	13094,3	9959,00	6002,89	43048,0
New comfirmed case	7,50	19,43	129,7	48,33	55,78	445,0
Incidence rate	236,59	459,82	1011,3	771,48	680,33	1164,1
cumulative deaths cases	34,38	85,71	278,0	579,33	167,00	768,0
New death case	0,04	0,00	0,0	1,00	0,00	2,0
Mortality rate	6,56	10,74	22,1	44,05	19,78	20,8
Fatality rate	2,88	2,62	2,1	5,77	2,69	1,8

Variable	**Cluster7**	**Grand centroid**
Cumulative comfirmed cases	25879,0	5157,67
New comfirmed case	73,0	38,96
Incidence rate	1399,3	477,74
cumulative deaths cases	277,0	136,35
New death case	0,0	0,13
Mortality rate	15,0	13,43
Fatality rate	1,1	2,88

Source: Computed by the researcher using Minitab Software

Appendixe03 :Distances Between Cluster Centroids

	Cluster1	**Cluster2**	**Cluster3**	**Cluster4**	**Cluster5**	**Cluster6**	**Cluster7**
Cluster1	0,0	1958,9	11773,4	8643,4	4676,6	41717,5	24558,2
Cluster2	1958,9	0,0	9817,3	6689,7	2718,3	39767,4	22604,4
Cluster3	11773,4	9817,3	0,0	3160,0	7100,4	29959,7	12790,7
Cluster4	8643,4	6689,7	3160,0	0,0	3978,7	33094,3	15935,3
Cluster5	4676,6	2718,3	7100,4	3978,7	0,0	37055,2	19889,4
Cluster6	41717,5	39767,4	29959,7	33094,3	37055,2	0,0	17181,7
Cluster7	24558,2	22604,4	12790,7	15935,3	19889,4	17181,7	0,0

Source: Computed by the researcher using Minitab Software