sciendo

**ECONOMETRICS. EKONOMETRIA**
**Advances in Applied Data Analysis**
**Year 2023, Vol. 27, No. 2**
ISSN 2449-9994

# Cluster Analysis and Visualisation Describing the Phenomenon of the Covid-19 Virus Pandemic

**Grażyna Trzpiot**
University of Economics in Katowice, Katowice, Poland
e-mail: grazyna.trzpiot@ue.katowice.pl
ORCID: 0000-0002-5129-5764

**Zuzanna Krysiak**
University of Economics in Katowice, Katowice, Poland
e-mail: zuzanna.krysiak@edu.uekat.pl
ORCID: 0009-0002-5187-4296



*Quote as:* Trzpiot, G., and Krysiak, Z. (2023). Cluster Analysis and Visualisation Describing the Phenomenon of the Covid-19 Virus Pandemic. *Econometrics. Ekonometria. Advances in Applied Data Analysis*, *27*(2).

DOI: 10.15611/eada.2023.2.03
JEL Classification: C38, I15, J11

**Abstract:** The article refers to the topic of the SARS CoV-2 virus pandemic and focuses on the effect of vaccines against this virus. The relation between the administered vaccines and the development of the global pandemic is very pertinent as the problem is being faced by the whole world. The difficulty lies in the fight against the pandemic, which is the cause of the very high death rate due to the virus, and has caused a global economic crisis. Demonstrating patterns and possible anomalies between data on the number of people vaccinated and the course of the disease and the number of deaths is an important factor in raising awareness of the risk of spreading the virus. The methods presented in the second chapter are data agglomeration and the k-means method. The study compared the results obtained in six selected countries from different regions of the world and presented the most important factors influencing the development of the pandemic. The presented methodology was also the basis for a deeper discussion of the factors determining the spread of the virus and can be an introduction to the analysis of time series. At the same time, it enabled the creation of patterns related to the studied phenomenon (for selected countries) defining local factors contributing to the spread of the disease and determining the effectiveness of the vaccines administered in them. The empirical analysis was conducted on the basis of data available in the electronic scientific publication https://ourworldindata.org/. The visualisations were made in the Tableau program, and the cluster analysis was carried out using the Statistica package.

**Keywords:** Covid-19, virus, vaccinations, morbidity cluster analysis coronavirus.

## 1. Introduction

An analysis was conducted for countries located in Europe (Poland, Italy), Americas (Chile, Mexico) and Asia (India, Israel). The countermeasure to the rapid outbreak of the pandemic has been the introduction of vaccines against the Covid-19 virus.

The countries were chosen because of their geographical location and diversity. The division by continents was intended to compare two countries that were connected by continent and differed, for example, in terms of population, number of people over 65 years of age, the country's approach to recommendations regarding actions against the virus, the number of people vaccinated, the number of people who died because of the virus. This division was introduced in order to show the common patterns relating to the course of the pandemic and the effectiveness of vaccination in such different environments where the virus is present. The innovativeness of these studies lies precisely in the identification of patterns occurring between the analysed data on various demographic environments susceptible to the spread of the same disease. The results obtained from this analysis can bring important conclusions for social institutions.

However, there are many determinants of the course of the disease, which are defined later in this work. It is important to notice the problem of the dependence of the number of people vaccinated in a given society, and the course of the disease and the number of deaths and morbidity among the elderly, or the effectiveness of the vaccinations introduced. Each of the analysed countries was selected due to the different methods of dealing with the pandemic, and were significantly different, e.g. in terms of the population ageing, geographic location and/or restrictions related to the pandemic introduced by a given country. The effectiveness of vaccines was questioned both because of the large number of choices given by producers and the development and flare-ups of the pandemic over repeated periods of time.

Cluster analysis and visualisation can be a powerful tool to understand the spread and impact of the Covid-19 pandemic. The research included in this article allowed to identify high-risk areas. Cluster analysis was used to identify the areas most affected by the pandemic, enabling policy makers to target resources and interventions where they were most needed. By understanding the factors affecting the spread of the virus, and by analysing Covid-19 case patterns using cluster analysis, researchers can gain insight into how the virus spreads and identify potential causes or contributing factors. Cluster analysis was also used to compare the spread and impact of the virus across regions, helping to identify similarities and differences in response and make policy decisions. Due to the shortage of data, the completeness of the databases, countries representative for the given continents were selected due to the best quality of data in terms of their completeness.

The study also presented visualisations of complex data. Cluster analysis was combined with data visualisation techniques to create visual representations

of Covid-19 data that are easy to understand and communicate to a wider audience, helping decision makers and the public to make informed decisions.

The aim of the article was to study and analsze patterns and clusters of Covid-19 infections, deaths and recoveries in different regions and countries, using statistical methods such as cluster analysis and visualisation tools. The article also sought to identify factors related to the spread of Covid-19, such as demographic, environmental and behavioural factors, and investigated the effectiveness of various public health interventions such as social distancing, and wearing face masks.

## 2. Vaccine production worlwide

The infectious disease Covid-19 has been developing since November 2019, and was declared a global pandemic by the World Health Organization (WHO) on 11 March 2020. Since the outbreak of the epidemic, data have been collected from around the world providing information on minimum mortality. As of 10 May 2021, confirmed cases of virus infection in the world amounted to 158, 334, 639; 3, 294, 120 people have died so far, and 94, 372, 990 have been cured. The drastic changes introduced by the ongoing pandemic became the impulse to find a solution to end it. The result of the work of pharmaceutical companies from all over the world was about 100 vaccines in various phases of implementation as part of the development plan activated to accelerate diagnosis, vaccines and therapy for the new coronavirus strains (*WHO coronavirus*…, 2021).

The pandemic has created major economic challenges going back to early 2020, including a short but very steep recession. Starting from the second half of last year, the economy began to recover and continues to show improvement. The rapid development of vaccines that demonstrated the ability to provide a high degree of immunity to Covid-19 likely played a significant role. Versions of the vaccine developed by several pharmaceutical companies received immediate approval and have been actively distributed. It is estimated that at least 70 to 80 percent of the population needs to be vaccinated to control the virus in the wider population, it is believed the markets will watch closely the rate of vaccine rollout: "The sooner we vaccinate a large portion of the population, the sooner we will return to some form of economic normality" (Haworth, 2021) – Senior Director of Investment Strategy at US Bank.

There are now more than 242 vaccines proposed worldwide, 821 vaccine trials underway and 80 Covid-19 vaccines approved by at least one country (Covid19 track vaccines, 2023). Published studies, mainly in high-income countries, cite concerns about the safety of Covid-19 vaccines, including the rapid pace of vaccine development, as one of the main reasons for the pandemic fluctuating (Wouters et al., 2021). Due to the pace of work and the global pandemic problem, which was increasing seasonally, the number of vaccine-producing companies was increasing

all the time. The largest manufacturers include: Moderna, Astrazeneca, Pfizer, Sinovac, J&J, Sputnik V and Novavax. Work on inventing a vaccine against Covid-19 gathered unprecedented and rapid pace (*WHO coronavirus*…, 2021). Pharmaceutical companies were bombarded with information about the successive milestones achieved in the race to find a remedy for the pandemic. The pace of the work caused some concern in both the public and medical communities. The whole world was looking forward to a return to normality (*Pharmaceutical technology…*, 2022).

In clinical trials, there are five categories of vaccines: the whole virus, protein subunit, viral vector and nucleic acid (RNA and DNA). Some of them try to smuggle the antigen into the body, others use the body's own cells for the production of viral anti-gene (CEFARM24, 2021). The following coronavirus vaccines are currently available in the EU (Kamińska, 2021):

- Pfizer/BioNTech mRNA vaccine,
- mRNA vaccine against Covid-19 by Moderna,
- AstraZeneca coronavirus vector image,
- Janssen Single-Dose Vaccine (Johnson & Johnson Group),
- Nuvaxovid protein vaccine against Covid-19 by Novavax (approved in the EU on 20 December 2021).

The rapid development, testing and production of multiple effective SARS--CoV-2 vaccines was a breakthrough achievement in 2020. Never before has a vaccination campaign started so soon after a new pathogen has been identified. In many cases, developing a vaccine took many years or even decades, yet for Covid-19, scientists have developed several highly effective vaccines over a very few years. At the same time, there is a problem related to the effectiveness of vaccines that concerns people in many societies. The lack of trust and knowledge in the subject of vaccination translates significantly into the global introduction of vaccines.

The reluctance of people to receive safe and recommended available vaccines, known as "vaccination reluctance" was a growing problem even before the Covid-19 pandemic (Mathieu, 2021). In turn, reporting on such vaccine efficacy studies is essential to build public confidence and reduce reluctance to vaccinate (Pierre Vergera, 2020). Vaccine studies found that perceived vaccine efficacy was the strongest predictor of Covid-19 vaccine uptake in the United States (Kreps, 2020). S.M. Sherman examined to what extent people are ready to be vaccinated against the Covid-19 virus. From the careful observations on this subject, the consensus was that an extremely important point and an argument for getting vaccinated is the effectiveness of vaccines, that is, to what extent the vaccine gives certainty about not getting sick, and the second point was its safety (Sherman, n.d.) Conclusive studies confirming the effectiveness of vaccines in educating the public and in presenting tangible evidence are becoming extremely important.

Covid-19 vaccines are recognised as a way to stop the ongoing pandemic. However, the effectiveness of this method requires a high level of public approval.

Studies conducted prior to the Covid-19 pandemic showed that factors such as perceived risk group and safety concerns, especially those related to side effects, can influence vaccine acceptance (Pogue et al., 2020). Other research conducted by Ł. Jach showed a model of attitudes towards the Covid-19 vaccine taking into account four predictors: fear of Covid, attitudes towards science, beliefs that Covid-19 is a hoax, and that the SARS-CoV-2 virus was man-made. Research results have shown that the decision to vaccinate may be related mainly to current factors regarding the presentation of the disease (e.g. in the media), its causes and the actual level of risk (Jach, 2021).

In turn, research conducted based on the analysis of comments in Polish society showed that the most common argument justifying reluctance or even opposition to vaccination against Covid-19, was the lack of trust in the intentions of representatives of the authorities, as well as doctors and pharmaceutical companies (approximately 77% of those questioned) (Ciesek-Ślizowska, 2021).

Data are needed to understand the magnitude and origin of changes in vaccine effectiveness and other potential impacts on Covid-19 morbidity and mortality. In order to fill these gaps, the research described in this article, based on cluster analysis, was conducted to define and describe the phenomena occurring during the pandemic.

## 3. Discussion of the results of data cluster analyses regarding the impact of vaccines on the course of the pandemic

Data on the Covid-19 pandemic were used to analyse this phenomenon in terms of finding a link between the use of the vaccine and the course of the disease. The course of the pandemic is described by the variables described in the next section, which were selected for the study due to occurrence of many cases of variables and the possibility of describing the current reality with them. The variables also show the effect of vaccines against the virus. Thanks to their division into clusters and observation, it was possible to analyse patterns and anomalies in the data, which allow the development of research in this topic and are an introduction to data forecasting. This is an important point of reference in possible future pandemic threats.

All analyses were carried out in the Statistica package, which enables advanced data analysis. The methodology chosen in the next chapter was based on the analysis of basic statistics, cluster analysis and multiple regression showing the exact relations between statistical data.

The limitation was the variable completeness of the data provided by official sources. The database does not contain complete information on each country, with many gaps in the data that prevented a thorough and reliable analysis. Some countries only provide data on cumulative doses administered so far, limits the comparability of the dataset across countries (Mathieu, 2021).

### 3.1. Databases and basic data statistics

The data used in the work describe the situation related to the global pandemic. The variables represent all major data related to confirmed cases, deaths, hospitalisations, and studies, as well as other variables.

The time range of the series is daily and ranges from 1 March 2020 until 22 January 2022. Thanks to the daily update of this database, it was possible to regularly observe the variables. The countries included in the analysis are Poland, Italy, Chile, Mexico, India and Israel. It should be added that the variables informing about the number of vaccinations have been studied since 28 December 2020.

In the paper, the explained variable concerns the current number of people vaccinated against the Covid-19 virus. The variable name is:

- new_vaccinations_per_milion.

Explanatory variables were used to describe the changes of the examined variables. For this study, the explanatory variables relate to the number of current people diagnosed with the disease, the number of current deaths from the disease, the number of people hospitalized, the number of tests performed daily, and the variables include the severity index, which is an indicator of the severity of the government's response. The measurement used is based on nine indicators including school closures, job closures and travel bans, scaled from 0 to 100 (100 = strongest answer). The names of the variables were:

- new_cases_per_milion,
- new_deaths_per_milion,
- new_tests_per_milion,
- stringency_index.

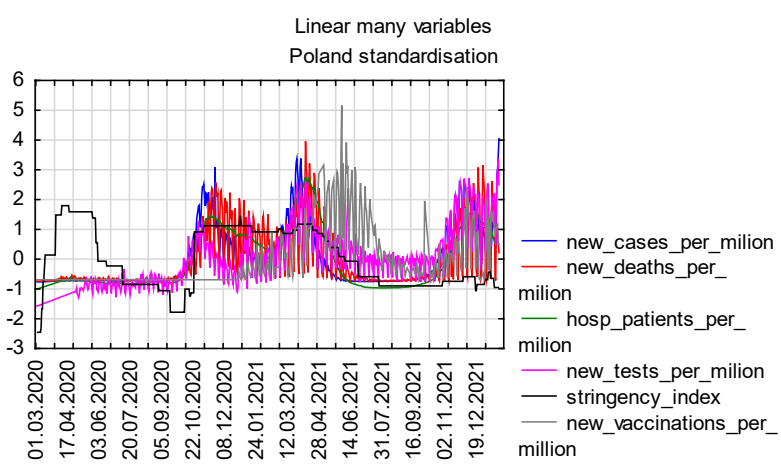


**Fig. 1.** Variables used in the analyses for Poland

Source: own elaboration.

The individual data used were ordered and then standardised, creating a homogeneous community. Selected data may show a trend and seasonality, which will be tested in subsequent stages of work (WHO, 2021).

The following data (Figure 1) show that there is a probability of correlation between the data. The variables may influence each other, in particular the variable concerning the current doses of vaccine administered on the number of deaths, morbidity and the number of hospitalised persons.

Subsequent analyses were carried out for the endogenous variable and the explanatory variable for subsequent countries compared in pairs: Poland – Italy, Chile – Mexico, and India – Israel. For each country there was a correlation between current vaccination doses and observations regarding the current number of deaths, new cases and tests performed. The dependence shows that with the increase in vaccinated persons, the number of current cases decreases. It may also result from the possible seasonality of data and external factors, such as climate.

Before the data analysis, the database was standardised and the missing data supplemented. The method of linear interpolation was used, thanks to which the missing data could be supplemented.

### 3.2. Categorisation of data

The classification algorithm used in this work was cluster analysis, which is a technique of grouping similar observations into several clusters based on the observed values of several variables. It aims to detect the natural division of objects, namely it groups similar observations into homogeneous subsets. The agglomeration method and the k-means method were used to group objects and their features.

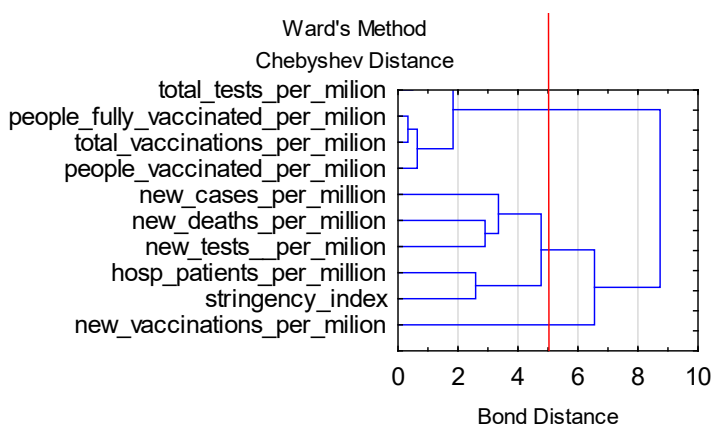


**Fig. 2.** Tree diagram for data from Poland

Source: own elaboration.

In the analysis presented below, concerning data for Poland, Ward's method was used, and the choice of the distance measure was the Chebyshev distance – a measure of the maximum difference distance which indicates during the analysis, objects divergent from each other in any one dimension (Aczel, 2000).

The distances between variables can be read in Figure 2. Note that the data are grouped into similar sets presented graphically below (Figure 2) using cluster analysis to group the data.

In this way, one link more and more objects together and aggregate them into ever greater clusters of elements that increasingly differ from each other. Eventually, in the final stage, all objects are linked together.

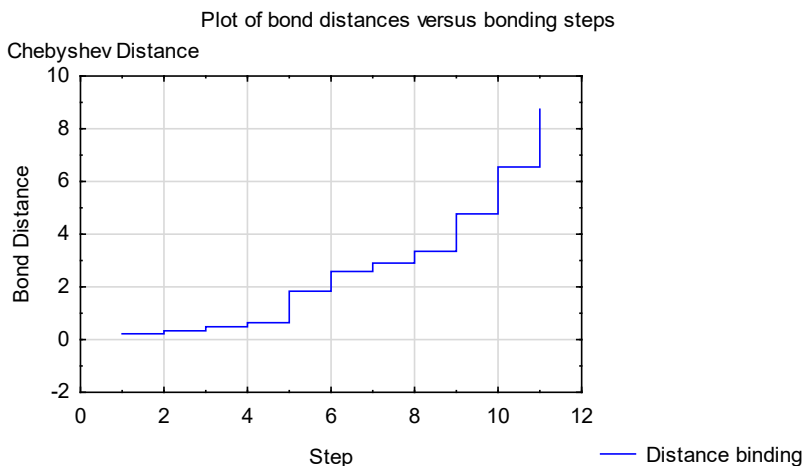


**Fig. 3.** Bond distance plot for Poland's data

Source: own elaboration.

As shown by the agglomeration diagram (Figure 3) the objects differ noticeably in steps 5 and 9 (Gan, 2007). There is a visible jump reflected in the tree diagrams, which divides the data into three groups, including one single (new_vaccinatios_ per_milion) and two larger groups.

An important piece of information is the existence of a single group describing the data on the current number of people vaccinated. This variable was entered into the database in the middle of the analysed period (Luszniewicz, 2001). For this reason, the distances between the current lesion data and the vaccination data are greater in the cluster analysis. However, when looking at the bond distance graph, the next step showing a distance jump is step 10 at a distance of 8. A tree diagram (Figure 4) at a distance of 8 creates two groups of variables similar to each other, and thanks to the distance matrix it is possible to read the distance between the vaccination

data and the data current changes (Everitt et al., 2011). As a result, these distances do not exceed 5, which indicates a high similarity and a relation between the data.

The cluster analysis carried out for Italy reduced the distances between the data to similar measures as in the case of data for Poland. The similarities between the data are large and there are relations between them. The grouping was made by the Chebyshev distance, as a result of the selected maximum distances, even though such objects were excluded from the group, which differed from the others only on one value.

Summing up the analysis of agglomerations for the American countries, it was possible to notice a similarity broken down into categories, i.e. the number of clusters. The variables with the greatest distances were classified as current variables, e.g. current deaths. The remaining groups, however, indicated a strong relationship between the variables describing the vaccination doses administered and the number of infected, deceased or hospitalised people (Stanisz, 2006).

Summing up the analysis of agglomerations for Asian countries, significant differences in the breakdown into categories were noticeable. India as a country with an a significantly larger population shows similarity in the observations reporting the number of current cases dependent on the observations reporting the number of deaths and the number of vaccinations performed. On the other hand, Israel shows differences in the form of an observation concerning current cases, classified separately due to the lack of similarities with other data (StatSoft, 2011).

### 3.3. K-means method

K-means clustering is a vector quantisation method derived from signal processing that aims to break n observations into k clusters, where each observation belongs to the cluster with the closest mean (cluster centres or the centroid of the clusters) serving as a cluster.

In this part of the analysis, the numbers of clusters 2, 3 and 4 were compared. In each of them, the following elements were distinguished:

- Analysis of variance (Figure 3).
- Graph of mean values (Figure 4).
- Elements of clusters.

The algorithm grouped by the k-means method was aimed at creating clusters of the analysed observations that differ as much as possible from each other. For the data analysed for Poland, the clusters were divided into three. When divided into three clusters, the significance of $p$ in the analysis of variance (Figure 4) indicates the existence of significant differences between the groups. The analysis of variance shows that there are variables with a significance $p > 5\%$, but they were considered insignificant due to their very small number compared to observations with a significance value below 5%. K-means clustering indicated that the best solution was to divide the data into three clusters. Thanks to the use of both the hierarchical and k-means methods, the same variables were separated from a separate cluster when the data was divided into three groups. This means that the cluster elements for both analyses are the same (Trzpiot, 2017).

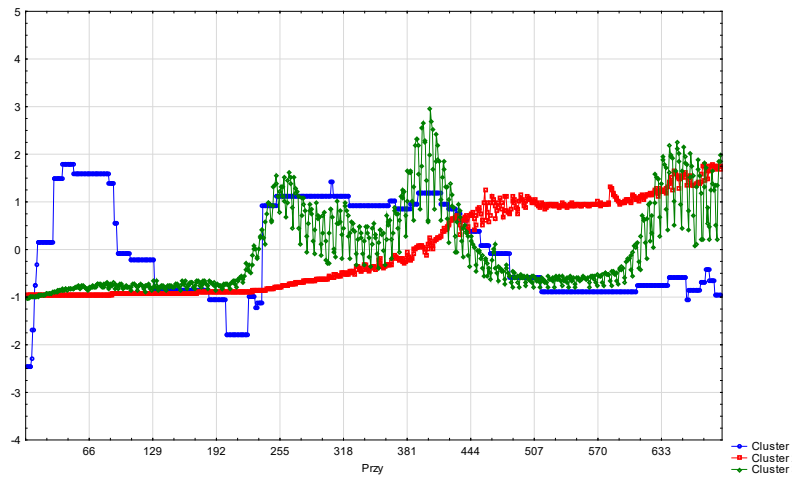


**Fig. 4.** Graph of mean values for each cluster for Poland data

Source: own elaboration.

When analysing the significance value, the Euclidean distances show differences between clusters, and the plot of mean values describes the correct choice of the number of clusters. The analysis of variance was satisfactory despite data with $p$ significance exceeding 5%, which were too rare to affect the analysis. Most of the data adopted a $p$ value of less than 5%. The significance of p indicates that there are still significant differences between the groups ($p < 0.05$).

**Table 1.** Euclidean distances

| Cluster | Euclidean distances of clusters<br>Distance below the diagonal. The square of the distance above the diagonal | | |
|---|---|---|---|
| | No 1 | No 2 | No. 3 |
| 1 | 0.000000 | 2.460733 | 1.288744 |
| 2 | 1.568672 | 0.000000 | 1.120679 |
| 3 | 1.135229 | 1.058621 | 0.000000 |

Source: own elaboration.

For the analysed data for Italy, the clusters were divided into three. When divided into three clusters, the significance of $p$ in the analysis of variance indicates the existence of significant differences between the groups (Everitt et al., 2011).

For the analysed data for the American countries, i.e. Mexico and Chile, the clusters were divided into four and two clusters, respectively. When divided into

four clusters, the significance of $p$ in the analysis of variance indicates the existence of significant differences between the groups. The existence value is very small, which means that the values differ significantly from each other. When dividing the Chilean observations into two clusters, the analysis of variance shows that there are variables with a significance $p > 5\%$ at the first observations, but were considered insignificant due to their very small number compared to observations with a significance value of less than 5%. K-means clustering indicated that the best solution is to divide the Mexico data into four clusters and the Chilean data into two clusters. Due to the use of both the hierarchical and k-means methods, the same variables were separated from a separate cluster when dividing the data into groups (Stanisz, 2006).

For the data analysed for the Asian countries, i.e. India and Israel, the clusters were divided into four and three clusters, respectively. When divided into three clusters for Israel, the significance of $p$ in the analysis of variance indicates the existence of significant differences between the groups. K-means clustering indicated that the best solution is to divide the data from India into four clusters and from Israel into three clusters. Due to the use of both the hierarchical and k-means methods, the same variables were separated from a separate cluster when dividing the data into groups.

## 4. Visualisation of the results of data analysis

The resistance to the virus achieved by the administration of the vaccines was a requirement for the achievement of herd immunity. Less than 12 months after the pandemic began, vaccines were developed to protect against SARS-CoV-2. The results of the analysed studies show that equally significant factors determining,

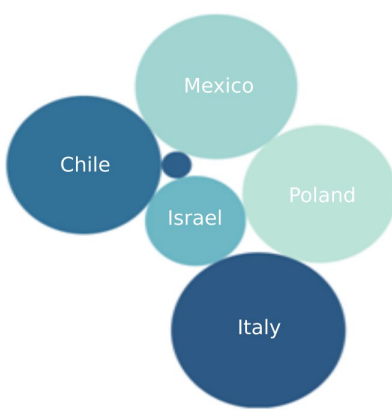


**Fig. 5.** Strictness Index

Source: own elaboration with the help of Tableau.

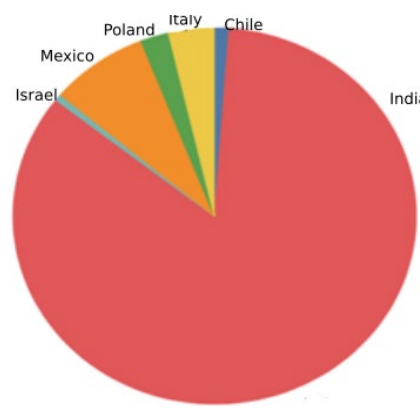


**Fig. 6.** Population

Source: own elaboration with the help of Tableau.

for example, the mortality caused by the virus, were, for example, the stringency index discussed in the previous chapters. This is an indicator of the severity of the government's response. Measurement based on nine response rates caused by the pandemic, including school closures, work place closures and travel bans, were scaled from 0 to 100. Figure 5 shows the relationship between the severity index and the total number of all recorded deaths due to the virus (Our World in Data, 2022). The size of the circles reflects the measurement of the number of deaths – the larger the size, the greater the death rate, whilst the darker the colour of the circle, the more stringent the measure. As can be seen, Chile and Italy with the highest stringency index also have the highest number of people who died from the Covid-19 virus (WHO, 2020). Nevertheless, it can be seen that India, with its very high severity index, has the lowest number of deaths than any other selected country.

Another aspect of the analysis is the population of countries (Figure 6). The country with the largest population, i.e. India, has had the fewest deaths, which in relation to the way the virus spreads, proves that the restrictions imposed by the severity index have worked for India.

India, as described in Figure 9, is characterised by both the lowest number of deaths and the virus cases, and also the number of vaccinated people. It is known from previous analyses that the number of deaths has a high correlation with the number of vaccinated persons. Figure 10 also shows how low the number of vaccinated people and deaths is in India.

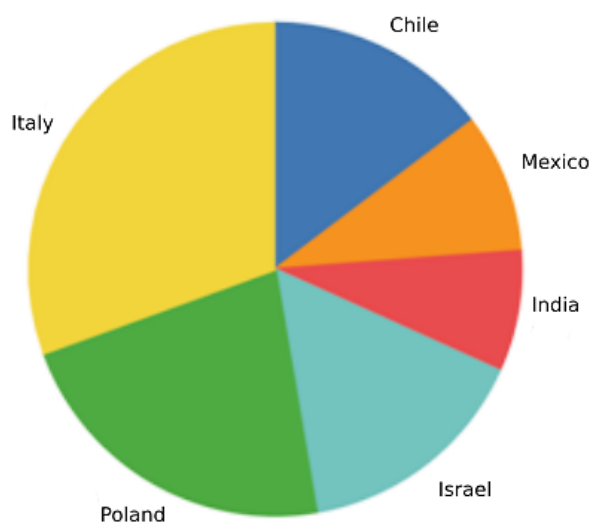


**Fig. 7.** Number of people over 65

Source: own elaboration with the help of Tableau.

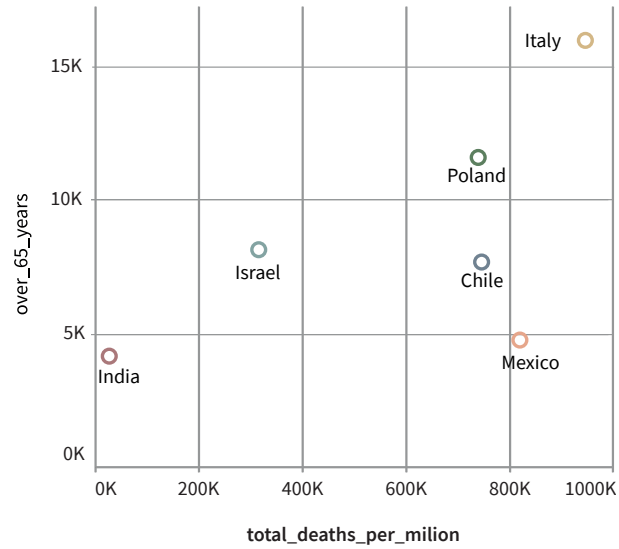


**Fig. 8.** Deaths of people over 65

Source: own elaboration with the help of Tableau.

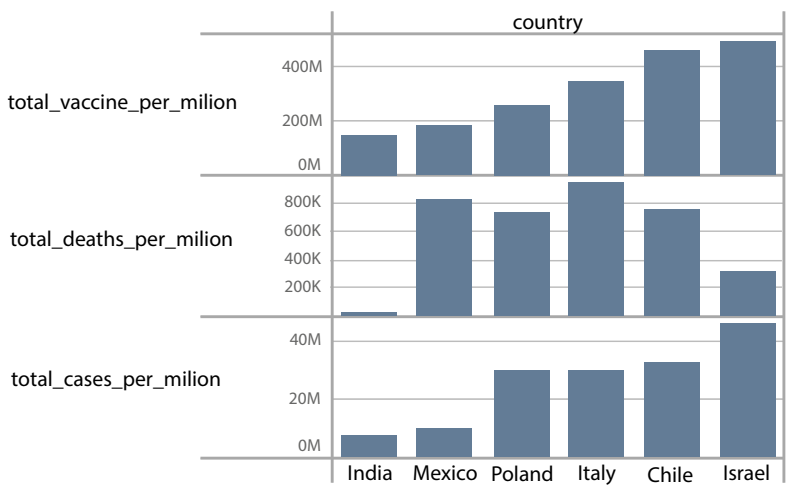


**Fig. 9.** Diagram for the selected countries

Source: own elaboration with the help of Tableau.

The example of Israel, where the number of cases of illness is the highest compared to the analysed countries, and the number of deaths is one of the smallest, proves that as a result of a large number of vaccinations, as shown in Figure 10 the risk of death has been reduced.

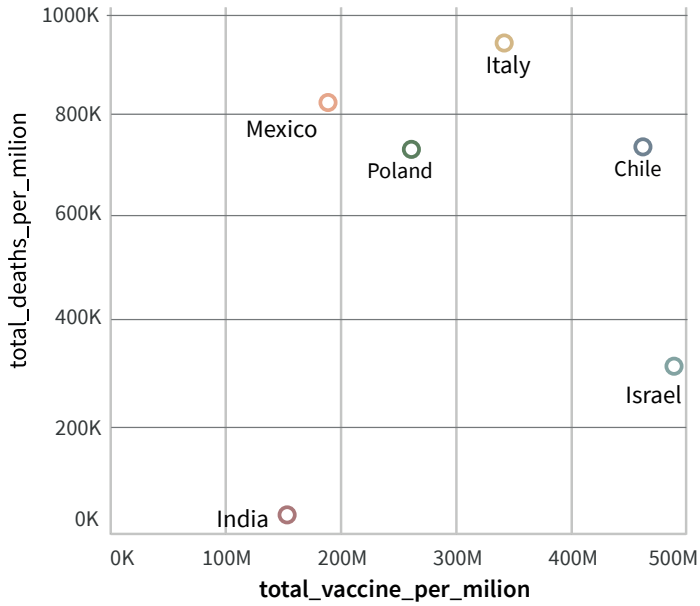


**Fig. 10.** All the vaccinated vs all the deaths relationship

Source: own elaboration with the help of Tableau.

Only Chile has a high number of vaccinated people, as well as mortality, and a high stringency index. However, Chile is one of the countries with a high number of elderly people, and the country's population is the smallest of all, hence the result of the analysis (WHO, 2021).

It is worth highlighting the differences in geographical location for all the countries considered, as they may be another variable showing dependencies on the number of cases or deaths caused by the virus (EMA, 2021). This also gives another opportunity to study the phenomenon of the pandemic in both economic and demographic terms.

## 5. Conclusion

The article describes the phenomenon of morbidity in relation to the administered vaccinations and analysed the observations describing the Covid-19 pandemic in the period from 1 March 2020 to 22 January 2022. The authors gathered information on patterns of occurrence between the implementation at the turn of 2020 and the turn of 2021 of vaccines, and mortality due to this virus and the incidence of new cases.

Thanks to the use of the Statistica package, it was possible to analyse regularities and anomalies between the examined observations and group them according to similarities. Thanks to the visualisations of the data presenting the observed

observations, it was also possible to describe the phenomena in question in more detail and highlight the relations between the data. Thus the assumed goals were achieved.

As a result of the research, the selected countries were classified based on public health indicators. Cluster analysis was used to divide the countries into groups based on selected and previously described indicators, such as the number of cases of Covid-19 infections, death rates or the Rigor index, etc. The groups created as a result of the analyses made it possible to describe them according to their characteristics, and enable a thorough comparison of the results of the research carried out for the selected countries. The countries were classified in three groups, broken down by their location by continent: Poland and Italy, Chile and Mexico, India and Israel The analysis allowed to describe the studied groups and to draw conclusions.

Based on the cluster analysis, conclusions can be drawn about the effectiveness of different health strategies applied in different groups of countries. For the group of examined European countries characterised by a high number of infections and high mortality, the high rigor index parameter and the current vaccination doses turned out to be important. The greater the restrictions and the more people vaccinated, the lower the morbidity and mortality rates. For the remaining groups of the respondents, referring to the countries of Asia and South America, data on the number of current cases also decreased significantly as the number of people vaccinated increased. Thanks to the analysis, a feature characteristic of selected Asian countries was also distinguished, namely a high rigor index, in connection with which the morbidity and mortality rates for them were significantly lower compared to other countries. The group of American countries, was identified with their high mortality and morbidity rate, which proves that these are countries that are struggling and require additional measures.

As part of the research, it was possible to conclude that there are strong links between the variables determining the number of cases and the overall course of the pandemic and the number of people vaccinated. The introduction of vaccination was associated with a decrease in the number of cases, mortality, etc. for the analysed countries. It should be noted, however, that the key variables affecting the spread of the virus were primarily the introduction of restrictions in connection with the pandemic, information about the age of the population and the geographical location and climate in a given country, which may be the subject of further research. This also opens up a topic for discussion about vaccinations, their effectiveness and future solutions to similar pandemic threats. Analyses already carried out show that a large population does not have such a significant impact on morbidity, because in a large population the previously described variable of the stringency index was an important factor. The research results presented in the article enable further

analysis of the presented data and subjecting them to time series analysis, to characterise the observations in more detail, and may also initiate forecasting of the discussed phenomenon.

Summing up, the goal of detecting data patterns in selected observations and describing the phenomena accompanying them as part of opening the discussion on vaccination against the Covid-19 virus and dealing with the reality of the pandemic was achieved on the basis of the analyses carried out.

## References


Aczel, A. (2000). *Statystyka w zarządzaniu.* Warszawa: Wydawnictwo Naukowe PWN.

CEFARM24. (2021). Retrieved from https://www.cefarm24.pl/czytelnia/odpornosc/rodzaje-szczepionek-przeciw-covid-19-jak-dziala-szczepionka-mrna-i-wektorowa

Centres for Disease Control and Prevention. (2022). Retrieved from https://www.cdc.gov/coronavirus/2019-ncov/vaccines/effectiveness/index.html

Ciesek-Ślizowska, B. D. (2021). Sceptycyzm wobec szczepień przeciwko COVID-19. Raport z badań. *RE-BUŚ*.

*Covid19 track vaccines.* (2023). Retrieved from https://covid19.trackvaccines.org/vaccines/

EMA. (2021). Retrieved from https://www.ema.europa.eu/en/human-regulatory/overview/public-health-threats/coronavirus-disease-covid-19/treatments-vaccines-covid-19

Everitt, B. S. Landau, S., Leese, M. & Stahl, D. (2011). *Cluster analysis*. Wiley.

Gan, G., Ma, Ch. & Wu, J. (2007). *Data clustering: Theory, algorithms and applications.*

Haworth, R. (2021). *Szczepionki przeciwko covid-19.* Retrieved from https://ourworldindata.org/covid--vaccinations

Jach, Ł. L. (2021). Psychologiczne korelaty postaw wobec szczepionek na COVID-19 wśród polskich respondentów – migawkowe badanie przed rozpoczęciem masowej kampanii szczepień. *Przegląd Psychologiczny*.

Johns Hopkins Medicine. (2022). Retrieved from https://www.hopkinsmedicine.org/health/conditions--and-diseases/coronavirus/is-the-covid19-vaccine-safe

Kamińska, E. (2021). *Pandemia COVID-19. Wszystko co warto wiedzieć o koronawirusie, testach i szczepionkach*. Retrieved from https://www.zwrotnikraka.pl/szczepionka-na-covid-19/

Kreps, S. P. (2020). Factors associated with US adults' likelihood of accepting COVID-19 vaccination. *JAMA Network Open*.

Lund, B., & Ma, J. (2021). Retrieved from https://www.emerald.com/insight/content/doi/10.1108/PMM-05-2021-0026/full/html?skipTracking=true

Luszniewicz, A. (2001). *Statystyka z pakietem komputerowym Statistica.* Warszawa: Wydawnictwo C.H. Beck.

Mathieu, H. R.-O.-G. (2021). A global database of COVID-19 vaccinations. *Nature Human Behaviour*.

*Pharmaceutical technology covid-19 vaccination*. (2022). Retrieved from https://www.pharmaceutical--technology.com/covid-19-vaccination-tracker/

Pierre Vergera, E. D. (2020). *Restoring confidence in vaccines in the COVID-19 era.* Taylor&Francis.

Pogue, K., Jensen, J. L., Stancil, C. K., Ferguson, D. G., Hughes, S. J., Mello, E. J., Burgess, R., Berges, B. K., Quaye, A., & Poole, B. D. (2020). Influences on attitudes regarding potential COVID-19 vaccination in the United States. *Vaccines*, *8*(4), 582. https://doi.org/10.3390/vaccines8040582

Roozenbeek, J. (2020). *Susceptibility to misinformation about COVID-19 around the world*. The Royal Society.

Sherman, S. M. (n.d.). C*OVID-19 vaccination intention in the UK: Results from the COVID-19 vaccination acceptability study (CoVAccS), a nationally representative cross-sectional survey*. Taylor& Francis.

Stanisz, A. (2006). *Przystępny kurs ststystyki w oparciu o program Statistica PL na przykładach z medycyny.* Kraków: StatSoft Polska.

StatSoft. (2011). Retrieved from https://www.statsoft.pl/textbook/stathome_stat.html?https%3A-%2F%2Fwww.statsoft.pl%2Ftextbook%2Fstcluan.html

Trzpiot, G. (2017). *Statystyka a Data Science.* Katowice: Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach.

WHO. (2020). Retrieved from https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19—21-august-2020

WHO. (2021). Retrieved from https://covid19.who.int/

*WHO Coronavirus (COVID-19) Dashboard*. (2021). Retrieved from https://covid19.who.int

Wouters, O. J., Shadlen, K. C., Salcher-Konrad, M., Pollard, A. J., Larson, H. J., Teerawattananon, Y., & Jit, M. (2021). Challenges in ensuring global access to COVID-19 vaccines: production, affordability, allocation, and deployment. *The Lancet*, *397*, 1023-1034. https://doi.org/10.1016/S0140-6736(21)00306-8

## Analiza skupień i wizualizacja opisująca zjawisko pandemii Covid-19

**Streszczenie:** Artykuł poświęcono pandemii wywołanej wirusem SARS CoV-2. W tekście skupiono się na działaniu szczepionek przeciwko niemu. Związek między szczepionkami a rozwojem pandemii na świecie jest oczywisty – cały świat walczy bowiem z pandemią, która jest przyczyną bardzo wysokiej śmiertelności i wywołuje kryzys gospodarczy. Wykazanie wzorców oraz możliwych anomalii między danymi dotyczącymi liczby osób zaszczepionych oraz przebiegiem choroby i liczbą zgonów jest ważnym czynnikiem zwiększania świadomości dotyczącej rozprzestrzeniania się wirusa. Metody przedstawione w drugim punkcie artykułu to aglomeracja danych i metoda *k*-średnich. W badaniu porównano wyniki uzyskane w sześciu wybranych krajach z różnych regionów świata i wskazano najważniejsze czynniki wpływające na rozwój pandemii. Zaprezentowana metodologia jest podstawą do głębszej dyskusji nad czynnikami warkującymi rozprzestrzenianie się wirusa oraz może być wprowadzeniem do analizy szeregów czasowych. Równocześnie umożliwiła ona stworzenie wzorców związanych z badanym zjawiskiem (dla wybranych krajów), określających lokalne czynniki przyczyniające się do rozprzestrzeniania się choroby i decydujących o skuteczności podawanych szczepionek. Analizę empiryczną przeprowadzono na podstawie danych dostępnych w elektronicznej publikacji naukowej https://ourworldindata.org/. Wizualizacje wykonano w programie Tableau, a analizę skupień przeprowadzono z wykorzystaniem pakietu Statistica.

**Słowa kluczowe:** wirus, Covid-19, szczepionki, wskaźnik zachorowalności, analiza skupień, koronawirus.