# Coronavirus spread analysis in the first pandemic year

*Evgeniy* Pitukhin[1*], *Petr* Pitukhin[2], and *Mileta* Gubaeva[2]

[1]Petrozavodsk State University, 33, Lenina Avenue, Petrozavodsk, 185035, Russia
[2]Dubna State University, branch Protvino, 9, Severnii Proezd, Protvino, 142281, Moscow Region, Russia

**Abstract.** The paper examines the characteristics of coronavirus spread in different countries around the world at the beginning of the pandemic, when effective vaccines have not yet been developed. The time interval analyzed is a year and a half from the beginning of 2020 to the summer of 2021. During this period, the spread of the disease was not yet significantly affected by the uneven vaccination process of the global population, and the external environment at that time was roughly the same throughout the world in terms of the lack of effective means to counteract the spread of the coronavirus. Based on open-source data on pandemic spread statistics by country (incidence, cure, mortality), applied statistics and data mining techniques identified groups of countries with different spread of the disease. Relative values of indicators, scaled to population size, and the dynamics of their change were analyzed. Estimates were made of the relationships between country-specific pandemic indicators and key demographic and socioeconomic indicators for these countries. These results may be useful for understanding the peculiarities of viral infections spread in different countries and regions of the world in the absence of effective countermeasures.

## 1 Introduction

COVID-19 (an acronym for CoronaVirusDisease 2019), formerly Coronavirus 2019-nCoV, Coronavirus 2019 is a potentially severe acute respiratory infection caused by the SARS-CoV-2 coronavirus (2019-nCoV) [1].

COVID-19 is a very dangerous disease, occurring in both severe and milder forms of acute respiratory viral infection. This virus can affect various organs through the body's immune response or through direct infection [1].

* Corresponding author: eugene@petrsu.ru

### 1.1 COVID-19 Pandemic

On March 11, 2020, the WHO declared the spread of coronavirus worldwide a pandemic. But how are pandemics and epidemics different and what are the main dangers?

Consider the concept of an epidemic. It is the progressive spread of infection among people, and it is much higher than usual in an area. It can also be a source of emergency [1].

There is a universal epidemiological threshold of an epidemic, which is the illness of 5% of the inhabitants of a territory or 5% of any social group. However, medical departments use their own calculation for different diseases. Sometimes the epidemic threshold is as low as 1%.

Pandemic is the spread of an infection on a worldwide scale. If a new virus overwhelms people in different countries, the population has no immunity to it, and the health care system has no vaccine, then it is a pandemic. Thus, it is a disease that is widespread [1].

COVID-19 is a beta-coronavirus of the same subgenus as the Severe Acute Respiratory Syndrome-Related (SARS) coronavirus. The virus taxonomy research group, suggested that this virus be called SARS-CoV-2, because of its similarity to several bat coronavirus infections [1].

Exactly how COVID-19 was transmitted to humans from bats, via an intermediate host or directly, is now considered to be unknown.

After modeling in a study of over 1,000 patients with a confirmed diagnosis, it was found that 3% of cases developed symptoms within two days and the remaining 97% developed symptoms within 12 days. From this we can conclude that the average incubation period of COVID-19 is 5 - 6 days [1].

The main mode of transmission of COVID-19 is airborne, in which the pathogens are localized in the mucous membrane of the respiratory tract and are transferred to the new organism through the air [1].

In this route of transmission, the pathogen enters the external environment when sneezing and coughing with droplets of liquid and is introduced into the human body when inhaling air containing infected particles. If the particles are small, they remain in the air for some time, but if they are larger, they are deposited on various surfaces up to two meters around the sick person.

The likelihood of infection can also be affected by the amount of time a healthy person is in contact with a sick person: the shorter the contact, the lower the likelihood of infection.

Coronavirus spreads about ten times faster than regular flu, and the death rate from it is much higher. While the death rate for influenza outside an epidemic is less than one hundredth of one percent, it can be as high as one-tenth during a mass infection. At the same time, the death rate from COVID-19 is about three to fifteen percent, depending on the population category and country. The COVID-19 coronavirus pandemic, which affected more than 200 countries and territories of the world with a population of 7.7 billion people in a short period of time, dramatically changed the way people were used to living.

### 1.2 Combating the pandemic

To combat the coronavirus, the authorities are taking the following measures [1]:

- many countries have introduced quarantine measures aimed at preventing the spread of infection, and there are restrictions on the accumulation and movement of people;
- new hospitals are being set up and research on treatment methods and means for coronavirus is being actively pursued;
- Sports events, including the Olympics, World Championships, and European Championships, have been cancelled or indefinitely postponed;

- Cultural and entertainment events have been banned; theaters, museums, and concert halls have been closed;
- Many businesses were shut down, except for those that are vitally important;
- educational institutions have been put on a distance mode;
- Air and railroad connections between cities and countries have been discontinued or severely curtailed;
- tourist activity has been virtually reduced to zero.

However, COVID-19 continues to spread, claiming new lives every day. Countries' GDP is declining. Therefore, knowledge about the pandemic spread in different countries of the world can help to prepare for possible changes in both people's lives and the economic situation of the countries of the world.

In this context, it is of interest to study the characteristics of the spread of coronavirus in different countries of the world at the beginning of the pandemic, when effective vaccines have not yet been developed, and thus coronavirus has not yet been affected by medical decisions. As it turned out, the ability to provide mass access to effective vaccines and drugs against the virus was not available to all countries of the world. Therefore, the time interval analyzed is a year and a half from the beginning of 2020 to the summer of 2021. During this period, the spread of the disease has not yet had time to be significantly affected by the uneven vaccination process of the world's population, and the external environment at that time was conventionally about the same throughout the world in terms of the lack of effective means of counteracting the spread of the coronavirus.

Therefore, the task arises, based on data from open sources on statistical indicators of the pandemic spread by country (infection, cure, mortality), using methods of applied statistics and data mining to identify the pandemic spread in different countries and regions of the world, as well as to plot their dependence on demographic and socio-economic indicators of these countries.

## 2 Literature review

The authors are familiar with both foreign and domestic studies on the spread of COVID-19 coronavirus.

A comparative analysis of COVID-19 spread in Russia and countries of the world was undertaken in [1].

An analysis of the relationships between the spatial distribution of pandemic coronavirus (COVID-19) in the world using self-organizing maps was performed in [2]. In [3], mathematical modeling of COVID-19 transmission dynamics with a case study of Wuhan was performed. Classification of countries and regions according to the degree of coronavirus prevalence based on statistical criteria was carried out in a study [4].

The impact of government intervention on infectious disease control was assessed in [5]. The effect of reduced population mobility on the spread of coronavirus was analyzed.

In [6], factors influencing the epidemiological characteristics of the COVID-19 pandemic were determined using the TISM approach.

Also of note are studies [7-12] devoted to modeling the spread of coronavirus; works [13-18] consider prognostic aspects of its spread, using mainly time series models.

The analysis of morbidity and mortality from coronavirus is covered in [19-23].

Despite numerous studies of coronavirus proliferation processes, the identification and analysis of factors that could influence the pandemic in different countries of the world, at a time when vaccines were not yet in use, have not been carried out.

## 3 Materials and methods

The data that would be needed for the analysis were collected and prepared from open sources. Because the coronavirus pandemic is global in scope, official indicators for countries around the world are freely available. Most of what is needed was taken from https://www.worldometers, which provides real-time statistical data on countries around the world [24].

The main indicators for analyzing the spread of coronavirus are considered to be [24]:

- Country.
- Total Cases (Total number of reported cases during all time of tracking)
- New Cases reported per day
- Total Deaths (fatal cases in all time of tracking)
- New Deaths per day
- Population

The following indicators were also calculated based on the above data:

- % Cases (Proportion of the total population who became ill, regardless of outcome)
- % Cases (Proportion of the total population who were ill, regardless of outcome)
- % Deaths (Proportion of the population with a fatal outcome as a proportion of the total population)
- Growth D (Increase in the number of deaths)

To allow comparison of countries with different populations, the data were converted to a relative form, either as a percentage or per capita.

Descriptive statistics, data aggregation and consolidation, and Data Mining methods were used in the research process [25]. Hierarchical method and K-means method were used for clustering. The results of these methods are similar, so only the results of the hierarchical method will be shown [26].

The hierarchical method requires a measure of dissimilarity between sets of observations to separate clusters. In most hierarchical clustering methods, this is achieved by using an appropriate metric (a measure of distance between pairs of observations) and a relationship criterion, which defines the dissimilarity of the sets as a function of pairwise distances of observations in the sets.

The Euclidean distance metric was used in this work. A metric such as the unweighted average relationship clustering [26] was used as a criterion of connectivity.

The data for clustering were supplemented with the following demographic and socioeconomic parameters, among which it was assumed to find significant influencing factors on pandemic spread rates [27]:

- proportion of agricultural land to the total area of the country (Lnd Agri);
- share of agricultural population from the total population of the country (Rur Totl);
- share of population growth (Pop Grow);
- share of urban population from the total (Urb Totl);
- GDP per capita, in $US (Gdp Pcap);
- Domestic government spending on health care, %GDP (Xpd Ghed);
- Domestic public health expenditures, per capita, in $USD (Xpd Ghed Pc);
- Population density (Pop Dnst).

# 4 Results and discussion

## 4.1 Data on the number of cases worldwide and in different parts of the world

From the beginning of 2020 until the summer of 2021, the coronovirus pandemic spread around the world in the absence of mass vaccination and restrained only by hygiene and restrictive isolation methods.

Now let's analyze the data for all the countries of the world with COVID-19 disease statistics.

To begin with, let's look at which countries lead in the number of reported cases of infection. The three leading countries USA, Brazil and India have a big gap compared to the rest (See Figure 1):

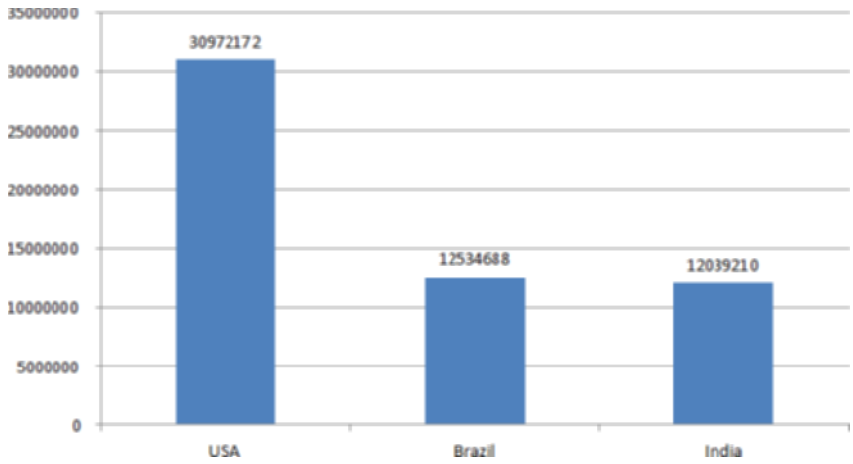


**Fig. 1.** Top 30 countries by the number of infected from above.

If we exclude them, then after the leaders we can distinguish 30 countries with high Figure 2 indicators, and similarly with low indicators Figure 3.

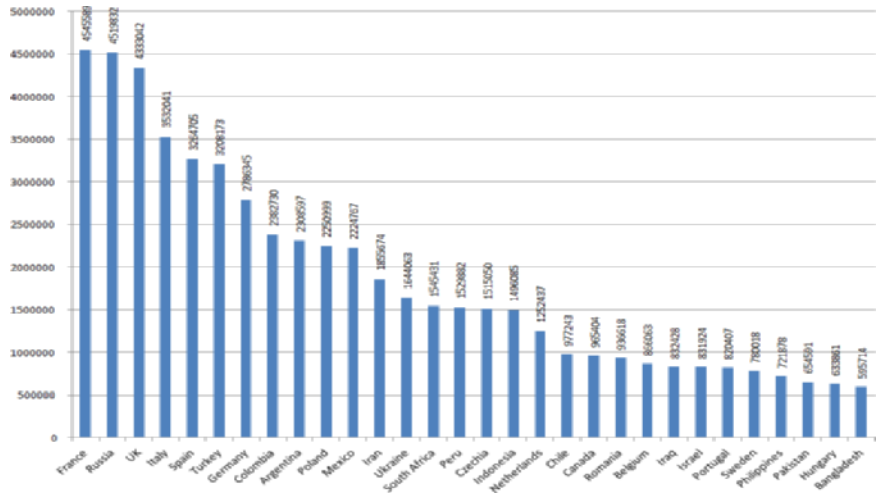


**Fig. 2.** Top 30 countries by the number of infected from above.

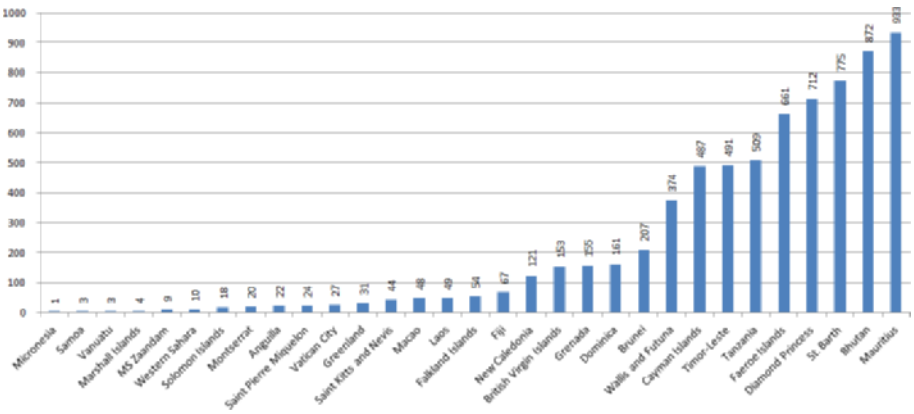


**Fig. 3.** Top 30 countries by the number of infected from below.

Then let us plot the scatter plots of the increase in the number of diseases and the proportion of the population that got sick, Figure 4. Now let us trace the obtained result by sorting the dots by parts of the world. After that it will be possible to find out whether there are regularities between the values of the results of scattering of points and their belonging to one or another part of the world.

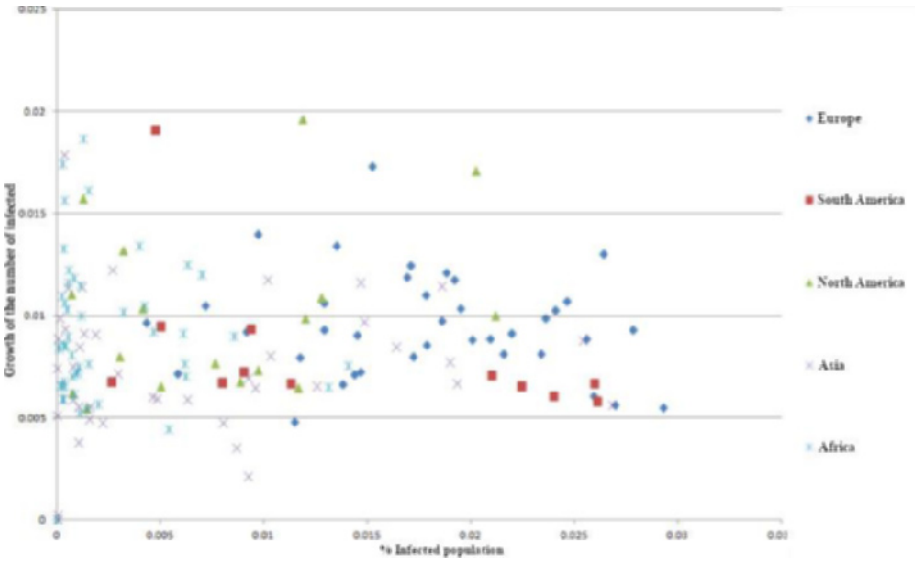


**Fig. 4.** Dependence of the proportion of cases and increase in the number of cases by parts of the world.

There are some patterns:

- African countries have high values of growth, but small values of % of diseased population this is due, in my opinion, to the fact that in these countries this virus does not progress so much. Since hot countries are not well suited for viruses like COVID-19, the percentage of those infected is low, but the increase values are alarming.

- European countries mostly have average values for both indicators. This means that even with a small population, a certain percentage, from 1 to 3, will be infected with the virus.

- South American countries have an average of 0.5% to 1%, regardless of the percentage of the population that is infected. And this means that South America is in some state of stability.

- North American countries are scattered, but their values are mostly within certain limits. Usually, these countries have disease rates between 0 and 1.5%, and increases between 0.5% and 1.5%.

- The countries of Asia, on the other hand, have half of their cases with a very low proportion of cases and a constant increase of 0.5% to 1.5%, which may indicate the emergence of the disease in new areas of Asia.

We do the same with the mortality dependence Figure 5.

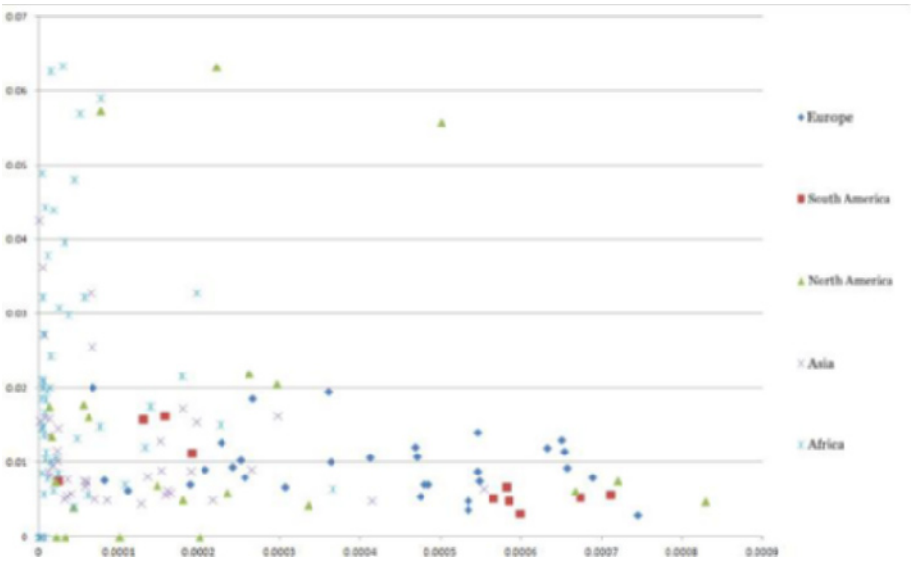


**Fig. 5.** Relationship between the proportion of fatalities and the increase in the number of deaths by parts of the world.

The following patterns can be seen:

- African countries have very high values of increase, but extremely low values of %fatal cases this is due, in my opinion, to the fact that in these countries the virus began to spread recently. And since African countries are known for their low standard of living, it is unfortunate that the number of deaths can start to increase there.
- European countries are mostly average on both counts. This means that even with a small population, a certain percentage, between 1 and 3, died from the disease. Which is surprising with the measures taken in these countries.
- South American countries have fairly high rates of deaths compared to others. But the increase in such cases is low, suggesting a gradual decline in the number of deaths.
- North American countries are scattered, but the values for both indicators are not high. We can assume that these countries are in some state of stability.
- Countries in Asia, on the other hand, mostly have the lowest values on both parameters, but there are exceptions with increases above 2%. It is possible that the new regions are beginning to catch the disease.

## 4.2 Clustering

For the convenience of clustering, the indicators were divided into two subgroups [27].

The first [27]:

- The share of agricultural land in the total area of the country;
- share of agricultural population from the total population of the country;
- share of population growth;
- share of urban population from the total;
- population density.

The second [27]:

- GDP per capita, in $US$;
- Domestic government spending on health care, %GDP;
- Domestic government spending on health care, per capita, in $US.

For each subgroup, clustering was done by adding either the proportion of cases and increase in cases, or the proportion of deaths and increase in deaths.

Cluster names were chosen by the visible features of the groups of countries or by the name of the country that is the individual cluster.

### *4.2.1 Case 1. Morbidity rates and demographics*

Data used: proportion of population diseased, increase in diseased, proportion of agricultural area, proportion of agricultural population, proportion of urban population, proportion of population growth, population density.

The clustering results in eight clusters. Of these, five are large, and three consist of 1-2 countries.

Let's start with Monaco. It is treated as a separate cluster because it does not have the data for which the clustering was done. Mongolia and Uruguay are cluster 1. They are combined because they have the highest rate of increase in new cases, not including Monaco. They also have a high % of agricultural land. Bahrain and Qatar have high % of new cases and % of urban population. The countries in the Urbanized group have a higher density and % urban population, but not a high incidence rate. These are mainly Asian, American, and European countries. Middle-sized countries have low densities, low prevalence rates, and low rates of new infections. They are made up of countries in all parts of the world, mostly Europe. Agricultural countries have low densities, very low % incidence rates, and high % farming population. Asian and African countries are included. High % urbanized countries that have very high % of disease with low population densities. European and North American countries.

High density c/o countries that have high densities and disease incidence rates. Mostly North American countries.

### *4.2.2 Case 2. Mortality rates and demographics*

Data used: share of deaths, increase in deaths, share of agricultural area, share of agricultural population, share of urban population, share of population growth, population density. There are 9 clusters. Five of them are one country. The following conclusions can be drawn from the results of the clustering. Monaco is again separate due to lack of data. Brunei has an extremely high mortality rate with a very low % sickness rate and population density. Mongolia stands out as well as Brunei, but density is the lowest. Moldova has low % deaths and growth rates, with an average % sickness rate, and a negative population growth rate. Belgium high % of deaths, growth, and population density. Group Urbanized high population density with high increase in deaths, but low % of deaths. Mainly Asian and North American countries.

Predominantly agricultural more than half the population, high growth rate at low population density. Asian and African countries are included. Few cases high increase in deaths at medium population density. Mostly North American countries. Half urbanized

low density at high % of urban population and increase in deaths. Mostly European countries.

#### *4.2.3 Case 3. Morbidity and economic indicators*

Data used: Proportion of population sickened, increase in sick people, GDP per capita, in $US, Domestic government spending on health care, share of GDP, Domestic government spending on health care, per capita, in $US. Six clusters are highlighted.

Luxembourg has high health care spending, high GDP, but also high % sick. Qatar has average spending and GDP, but high % sick. Bermuda does not have enough data.

Group High growth rate GDP is low, also spending is low, but the increase in illnesses is quite high. North America and Asia are included.

Group Low increase with low GDP and expenditures have a small increase. Countries in all parts of the world, most of all African countries.

Group High spending with high spending on health care have a high % of people getting sick. Almost all countries are European.

### 4.2.4 Case 4. Mortality rates and economic indicators

Data used: proportion of deaths, increase in deaths, GDP per capita, in $US, Domestic government spending on health, %GDP, Domestic government spending on health, per capita, in $US. Eight clusters were obtained. Of these, Cuba and Bermuda have incomplete data. Group Large spending and percent with high health care spending have high % fatalities. Almost all countries in Europe. The group Large GDP and average % are also European countries, but already with an average % of cases. Rich group mostly European countries, with high GDP and spending, but high increase in deaths.

Low spending countries in Asia and North America, with low health care spending, but a fall high increase in deaths.

Low GDP and medium spending group in most European countries, with low GDP, medium spending, but high increase in deaths. Poor group, consisting of countries in all parts of the world, with average growth and low health care spending. Mostly African and Asian countries.

### 4.2.5 Clustering Results

In preparing the data, most of the countries in Oceania were screened out for one reason or another, so it does not appear among the data for the parts of the world.

Of the 220 countries obtained in the data preparation phase, only 168 were used in the clustering for various reasons.

Several conclusions can also be drawn from the clustering:

- European countries have the highest readings for GDP and health care expenditures, but the readings for disease and death rates and increases are also quite high;
- In contrast, African countries have low GDP, expenditures, but incidence rates and death rates are quite low, although the increases are high;
- Not in all cases the higher share of the agricultural population compared to the share of the urban population will play a significant role;
- Population density directly affects the proportion of the population that gets sick.

## 5 Conclusion

In the course of the work the following results were achieved: COVID-19 incidence data were collected by countries; the data were arranged, put in a convenient form for analysis; dependence of the share of cases and increase in the number of deaths was analyzed; dependence of the share of deaths and increase in the number of deaths was analyzed; factors possibly influencing the spread of the virus were identified.

In conclusion, finding the presence of dependencies, in our case the influence of internal factors of a country on the proportion of registered cases and the proportion of deaths, can help predict disease behavior in other countries with a similar structure. Finding key factors helps to understand what to look for in order to change the situation in the country for the better.

## References

1. V. Tikhonov, S. Gushchina, *Analiz zabolevaemosti COVID-19 (2020g.)*, In Proceedings of the XIII International Student Scientific Conference "Student Scientific Forum", Moscow, Russia (2021). https://scienceforum.ru/2021/article/2018027186
2. P. Melin, J. C. Monica, D. Sanchez, O. Castillo, Int. J. of Chaos, Solitons & Fractals **138** (2020)
3. F. Ndaïrou, I. Area, J. J. Nieto, D. F.M. Torres, Int. J. Chaos, Solitons & Fractals **135** (2020)
4. A. Wilinski, E. Szwarc, Int. J. Expert Systems with Applications **172** (2021)
5. A. Giffin, W. Gong, S. Majumder, A. G. Rappold, B. J. Reich, S. Yang, Int. J. Spatial Statistics **52** (2022)
6. P. Lakshmi, M. Suresh, Int. J. Healthc Manag. (2020)
7. A. Babaei, H. Jafari, S. Banihashemi, M. Ahmadi, Int. J. Chaos, Solitons & Fractals **145** (2021)
8. F. A. Rihan, H. J. Alsakaji, Int. J. Results in Physics **28** (2021)
9. Y. N. Kyrychko, K. B. Blyuss, I. Brovchenko, Sci. Rep. **10**, 19662 (2020)
10. F. A. Rihan, G. Velmurugan, Prog. Fract. Differ. **7** (2021)
11. I. Ahmed, G. Modu, A. Yusuf, P. Kumam, I. Yusuf, Results Phys. (2021)
12. F. A. Rihan, H. J. Alsakaji, C. Rajivganthi, Adv. Difference Equ. **1** (2020)
13. A. Babaei, M. Ahmadi, H. Jafari, A. Liya, Int. J. Chaos, Solitons & Fractals **142** (2021)
14. D. Gomes, G. Serra, Int. J. ISA Transactions **124** (2022)
15. M. Maleki, M. Mahmoudi, M. Heydari, K. Pho, Int. J. Chaos, Solitons & Fractals **140** (2020)
16. L. Ismail, H. Materwala, T. Znati, Sh. Turaev, M. A. B. Khan, Computational and Structural Biotechnology J. **18** (2020)
17. J. A. Doornik, J. L. Castle, D. F. Hendry, Int. J. of Forecasting **38** (2022)
18. Ch. Satrio, W. Darmawan, B. Nadia, N. Hanafiah, Int. J. Procedia Computer Science **179** (2021)
19. D. D. Atsa'am, R. Wario, Scientific African **18** (2022)
20. J. Jung, J. Manley, V. Shrestha, J. of Economic Behavior & Organization **182** (2021)
21. A. Canatay, T. J. Emegwa, M. F. H. Talukder. Int. J. of Disaster Risk Reduction **64** (2021)

22. H.-J. Kremer, J. Public Health **190** (2021)
23. N. Ayoobi, D. Sharifrazi, R. Alizadehsani et al., Results in Physics **27** (2021)
24. Covid-19 coronavirus pandemic (2022). https://www.worldometers.info/coronavirus/
25. What is data mining? (2022). https://www.sap.com/insights/what-is-data-mining.html
26. Hierarchical clustering (2022). https://ru.xcv.wiki/wiki/Hierarchical_clustering
27. The worldbank. Data. Indicators (2022). https://data.worldbank.org/indicator?tab=all