2020 1st International Conference on Big Data Analytics and Practices (IBDAP)

The Relations Between Implementation Date of Policies and The Spreading of COVID-19

Watcharin Sirinaovakul, Thanason Eiamyingsakul, Narawit Tubtimtoe, Santitham Prom-on, Unchalisa Taetragool
Department of Computer Engineering, Faculty of Engineering
King Mongkut's University of Technology Thonburi, Bangkok, Thailand
wchr.aun@gmail.com, titlethanason@gmail.com, narawit.tub@outlook.co.th, santitham@cpe.kmutt.ac.th, unchalisa.tae@mail.kmutt.ac.th

***Abstract*— COVID-19 is undeniably one of the worst incidents in the 21st century. There are a wide variety of factors that impact the spreading of COVID-19. This paper presents the study of relations of how public policy implementation might affect the onset and the spread of COVID-19 cases. Cluster analysis was employed to identify data patterns associating with the policy implementation profiles. The results suggest that the effectiveness of policy adoption relates to the onset spreading of COVID-19. This also indicates that the decision of public administrators was critical in the latter stage of the pandemic situation management.**

Keywords—Big Data, Clustering, Correlation, Pandemic, Policies, Feature Engineering

I. INTRODUCTION

Since the start of the COVID-19 outbreak, thousands of people have died worldwide. Each country has implemented different types of policy responses. Some policies such as limit public gathering, isolation, quarantine, and school closure have been widely adopted in many countries [5]. However, a few countries have carried out unusual policies. For example, Kenya, Panama, and Chad have followed humanitarian exemption, amendments to funeral and burial regulations, or lockdown of refugee or other minorities' camps.

Even though a number of people have explored and analyzed previously implemented policies, most of those works were done in a very limited scope. Cruz-Aponte and Caraballo-Cueto [1] developed and simulated several scenarios of the COVID-19 pandemic to find the best policies. They found that the first-best option that would optimize both fiscal and mortality from the simulations is to identify, treat, and isolate incoming infected individuals, but this option does not seem possible. The second-best option is to use strict policies, e.g. physical distancing combined with massive testing, within 3 weeks after the pandemic is then a better fit with the outbreak situation. This, however, primarily focused on the results of the simulations.

Piguillem and Shi [2] tried to find the quarantine and testing policies that optimize fatalities and cost using the SEIR epidemiology model with the COVID-19 situation in Italy. It emphasized two fundamental issues: the capacity of the health system to deal with a large inflow of patients and the ability of policymaker to distinguish the asymptomatic infected. The results show that the suppression policies, which tried to eliminate the virus, gave the best result in terms of fatality, followed by the mitigation policies, and no intervention, respectively. Moreover, the results between implementing only quarantine policies and both quarantine and testing policies are not very significant. Therefore, the observed mandatory quarantines around the world seem to be close to what can be considered optimal if the governments have no intention to identify each and every carrier of the virus.

Hale et al. [3] provided a systematic way to track government responses to COVID-19 across 150 countries and time. They combined the data into series of new indices for cross-national comparisons of government interventions. Each index is composed of different individual policy response indicators. It focused on the description of variation in the intensity of government responses and their effect on the rate of infection.

Unlike the previous works, our study is interested in the relations between the various policies' implemented date and the spread of COVID-19 cases. The real-world statistics and the impact of policies are mainly focused. Several datasets are gathered from different sources, preprocessed, explored, and analyzed by multiple techniques. In the hope of understanding the pandemic, this work provides not only crucial information of the current situation, but also the data that would increase the pool of cleaned data in this field.

II. METHODOLOGY

This section discusses the overall processes used in this paper: aggregating data, preprocessing data, clustering data, and analyzing data by studying correlations between the policies and the defined clusters.

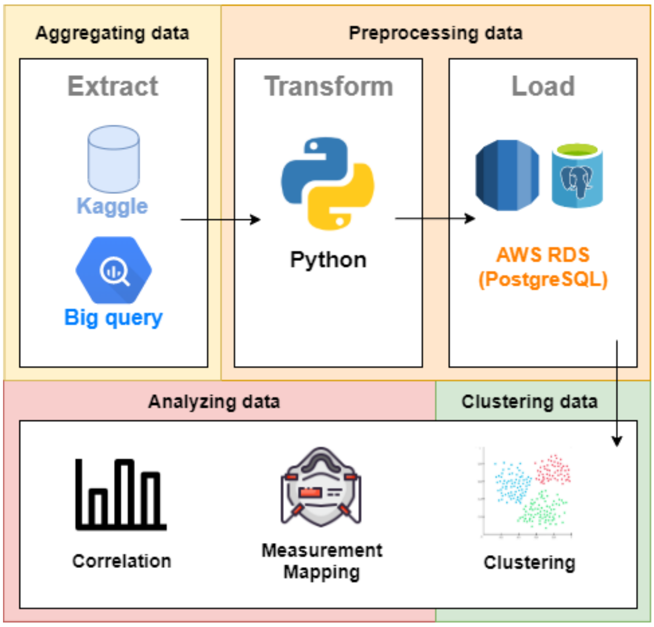

Fig. 1. The overall methodology framework

A. Aggregating data

Four datasets, namely JHU Coronavirus COVID-19 Global Cases, by country [4] from Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE), COVID-19 Government Measures Dataset [5] from Assessment Capacities Project (ACAPS), COVID-19 country lockdown days [6] and COVID-19 US Lockdown Dates

Dataset [7], were gathered for the most analysis advantages. The last two datasets [6-7] were gathered and provided as CSV files on Kaggle [8-9], which we downloaded from there.

- JHU Coronavirus COVID-19 Global Cases is a summary of a number of infected people, deaths, and recovered cases each day in each country to the level of provinces and states of major countries, e.g. United States, China, Canada, United Kingdom, and Australia.
- COVID-19 Government Measures Dataset contains policy responses of each country to the COVID-19 outbreak situation.
- COVID-19 Country Lockdown Days and COVID-19 US Lockdown Dates Dataset collect the date of a country starting the lockdown. The first dataset contains the starting date of the lockdown of every country to province and state levels, except for the US, while the second dataset contains lockdown dates of every state in the US.

For the convenience of data analysis, Numpy and Pandas, which are open-source python libraries, were used to preprocess the datasets from unformatted to well-structured formatted. Later, the data were stored in an SQL database on AWS's PostgreSQL after the preprocessing.

B. Preprocessing data

To prepare data for analyses, four aggregated datasets were divided into two groups: COVID-19 data and policy data.

First, the COVID-19 data were extracted from the JHU Coronavirus COVID-19 Global Cases dataset. The dataset was then transformed from the original horizontal data to the vertical data. The data contained country, province, date, confirmed cases, deaths, and recovered cases, in a total of 6 columns. Unfortunately, the recovered cases must be dropped since they contained a lot of missing values for many countries. Also because of the onset of the spreading was very slow in most cases, all the data before the number of confirmed cases exceeding 100 cases are also removed. The sample of COVID-19 data are shown in Table I.

TABLE I. EXAMPLE OF COVID-19 DATA

Country	Province	Date	Confirmed cases	Deaths
China	Hubei	23/01/20	444	17
China	Hubei	24/01/20	549	24
…				

Second, the policy data are composed of the rest three datasets: COVID-19 Government Measures dataset [5], COVID-19 Country Lockdown Days [8], and COVID-19 US Lockdown Dates Dataset [9]. COVID-19 Government Measures dataset [5] was combined with the lockdown datasets and cleansed to ensure data format correctness. Policy adoption distributions were also explored. For consistency, policies with adoption rates lower than 25% were dropped from this analysis. Table II presents the sample of policy data.

TABLE II. EXAMPLE OF POLICY DATA

Country	Policy	Type	Level	Implementation date
Thailand	Schools closure	Introduction	Whole	18/03/2020
Thailand	Schools closure	Phase-out	Whole	26/05/2020
…				

C. Clustering Data

In this process, two experiments were performed for two different purposes. The first experiment focused on the overall situation from the past to the present of each country, while the second experiment studied the changes and developments of each country throughout each week.

This process consists of two main parts: feature engineering and data clustering. In the feature engineering part, the data were grouped by country and province and aggregate based on specific statistical measurements.

1) Experiment I: The features used in this experiment contains:

- *days*: Number of days from the day that confirmed cases exceed 100 cases to the latest. It was used to separate countries into stages, such as early, middle, and late.
- *after_peak*: Number of days passed by after the peak of incremental confirmed cases. It was used to separate countries that already pass the peak for a period of time.
- *c_thisweek/avg*: Ratio of the number of new confirmed cases over this week to the overall average.
- *c_thisweek/peakweek*: Ratio of the number of new confirmed cases over this week to the number of the same case over the week that contains peak date.

2) Experiment II: The features used in this experiment contains:

- *weeks*: Number of weeks from the day that confirmed cases exceed 100 cases to the latest. It was used to separate countries into stages, such as early, middle, and late.
- *after_peak*: Number of weeks passed by after the peak of incremental confirmed cases. It was used to separate countries that already pass the peak for a period of time.
- *c_thisweek/avg*: Ratio of the number of new confirmed cases of this week to the overall average.
- *c_thisweek/peakweek*: Ratio of the number of new confirmed cases over this week to the number of the same case over the peak week. Unlike this number in the first experiment, the highest *c_thisweek/peakweek* number possible was 1 as it used the number of confirmed cases of over the real peak week not the week of the peak date.

After extracting necessary features, *days,* and *after_peak* used in the first experiment were rescaled by taking the natural log to scale down the differences of each country. On the other hand, the second experiment did not need any rescaling. Then, all the features are normalized to the scale of 0 to 1.

In the data clustering part, this work uses K-means algorithm [10-11] as the clustering method for both experiments. Then, to find the number of clusters, the silhouette method was used [12-13].

D. Analyzing data

After clustering, correlations between each defined cluster from the two experiments and the implementation date of policies were studied. In this process, the COVID-19 data and the policy data prepared in the preprocessing data were merged.

1) Experiment I: The correlation of the overall clusters and policies.

After the two sets of data were consolidated, the value in each column of the policy data, which originally contained the implemented date, was changed to the number of days passing the date confirmed cases exceeding 100 cases. Then, the correlations [14] were calculated. The policies that have a fairly high correlation score compared to other policies were examined. The result of this study would indicate how the implementation date of policies affects a country's overall statistics at the current stage (7th June 2020).

2) Experiment II: The correlation of the each-week clusters and policies.

In this experiment, before joining the two sets of data, the implementation date of policies in the policy data were shifted by two weeks based on the assumption that policy would affect the spreading of COVID-19 after it has been implemented for two weeks. Then, the modified policy data were combined with the each-week clusters on the country, province/state, and week. Next, the correlations [14] between the number of weeks from implementation date and features for each policy were computed. The result of this experiment would demonstrate how the implementation date of policies affects a country two weeks after it is implemented.

III. Experimental Results and Discussion

Table III shows the abbreviations of variables and policy names. These are used all across this section.

TABLE III. THE ABBREVIATIONS OF POLICIES AND VARIABLES.

Abbre.	*Stands for*
CiP	Changes in prison-related policies
DTR	Domestic travel restrictions
Econ	Economic measures
HeScrn	Health screenings in airports and border crossing
InFS	International flights suspension
IsoQ	Isolation and quarantine policies
Psy	Psychological assistance and medical social work
RqPro	Requirement to wear protective gear in public
Test	Testing policy
Visa	Visa restrictions
cta	c_thisweek/avg
ctp	c_thisweek/peakweek
till_peak	Number of days since it passes 100 cases to its peak
c_diff_p	Ratio of confirmed cases this week to the previous week

A. Experiment I

1) Clustering results: The overall clusters

To select the number of clusters and evaluate the quality of data clustering, the silhouette method was used. The higher the average silhouette score indicates better clustering results. Fig. 2 shows the average score after performing the silhouette method ten times. 2-clusters yielded the maximum average silhouette score, while 4-clusters was the second-best option. After studying the two options, it was found that the 2-clusters were only separated by *cta* and *ctp*. On the other hand, the other features were considered when partitioning the 4-clusters. To obtain more conclusive information for the further analysis, the number of clusters using in this experiment was then 4.

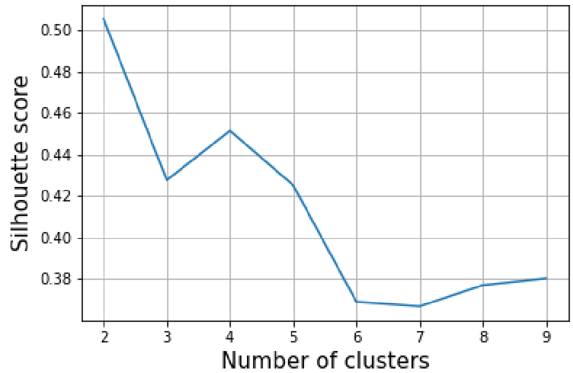

Fig. 2. The overall clusters' average silhouette score

Table IV presents the centroid, i.e., mean of all data points (countries), and the number of countries, named size, that belongs to each cluster. The sample country of each cluster is shown in Table V.

TABLE IV. THE CENTROIDS OF OVERALL CLUSTERS

Cluster	**Features**				**size**
	days	***after_peak***	***cta***	***ctp***	
0	72.50	4.82	2.34	1.12	62
1	73.87	31.18	1.15	0.86	61
2	89.43	72.67	0.25	0.11	119
3	29.50	18.89	0.58	0.29	18

TABLE V. SAMPLE COUNTRY IN EACH CLUSTER

Cluster	**Country**	**Province**	**days**	**after_peak**	**cta**	**ctp**
0	Brazil	-	87	9	3.04	1.15
1	US	Florida	85	67	1.42	1.21
2	Thailand	-	85	78	0.12	0.07
3	Zambia	-	12	12	0.72	0.18

The results can be classified into stages starting from the beginning to almost the end based on how a country performs in the latest week (7th June 2020) compared to the average (*cta*) and the peak week (*ctp*), how long a country has already passed its peak (*after_peak*), and how long the infection has started in a country (*days*).

- Cluster #0 represents the countries that are currently facing their peak of infection. It can be observed from the very low *after_peak* values and the very high *cta* and *ctp* values.
- Cluster #1 represents the countries that have already passed their peak, but the number of infections is still high. The *cta* and *ctp* values, which are around 1 and below 1, respectively, imply that the number of the

new confirmed cases in the latest week has decreased from its peak but is still around their average.

- Cluster #2 represents the countries that have already passed their peak and come close to the end of the infection because the number of *days* and *after_peak* are very high, while the ratios of *cta* and *ctp* are significantly low.
- Cluster #3 represents the countries that might be recently added to the data. It is because the number of *days* in this cluster is noticeably low compared to the other clusters. In other words, other statistical numbers cannot be very well explained since the pandemic just began.

2) Correlation results: The correlation of the overall clusters and policies

This section demonstrates the outstanding correlations between the implementation date of policies and the overall statistics of each cluster at the current stage (7th June 2020).

First, the countries in cluster #0 are in the stage of facing their peak. *IsoQ, Psy,* and *RqPro* have high levels of correlations as shown in Table VI.

TABLE VI. CORRELATION OF POLICIES AND CLUSTER#0

Features	**Policies**		
	IsoQ	***Psy***	***RqPro***
till_peak		0.71	0.66
death_rate	0.65	0.65	

Second, the countries in cluster #1 are in a steady stage. The correlations are not as relatively high as those in cluster #0, but *DTR*, *InFS*, and *IsoQ* still have a moderate level of correlation as displayed in Table VII.

TABLE VII. CORRELATION OF POLICIES AND CLUSTER#1

Features	**Policies**		
	DTR	***InFS***	***IsoQ***
till_peak	0.54		0.52
death_rate		0.44	

Third, the countries in cluster #2 are reaching the ending stage. The correlations are relatively low compared to the two previous results, but similar policies are shown up in Table VIII.

TABLE VIII. CORRELATION OF POLICIES AND CLUSTER#2

Features	**Policies**			
	CiP	***DTR***	***IsoQ***	***RqPro***
till_peak	0.49	0.57		
death_rate			0.45	0.39

The results of these three sets of correlations demonstrate that the implementation date of policies has relations to *death_rate* and *till_peak*. For instance, *IsoQ* (isolation and quarantine policies) has high correlations with all three clusters. Moreover, after taking a closer look, some policies are also related to each other. For example, *CiP* (changes in prison-related policies) and *DTR* (domestic travel restrictions) have a decent correlation with *till_peak* in cluster #2. Furthermore, *CiP* also has a high correlation with *DTR*.

Nonetheless, the correlations of policies and cluster #3 are not mentioned in this section due to two reasons. First, the countries in this cluster currently were in the early stage of the pandemic. Second, the number of countries in this cluster was relatively low compared to other clusters. Therefore, the correlation in this cluster could be biased and hard to conclude.

B. Experiment II

1) Clustering results: The each-week clusters

Similar to the previous experiment, the optimized number of clusters was calculated using the silhouette method as shown in Fig 3.

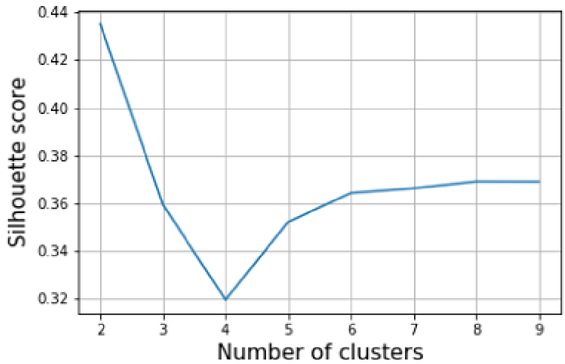

Fig. 3. The each-week clusters' average silhouette score

Eight clusters were chosen in this experiment. Table IX presents the centroids and the size of data points in each cluster. It should be noted here that, in this experiment, each data point does not represent each country. It represents the weekly outbreak situation in each country. The sample data of each cluster is shown in Table X.

TABLE IX. THE CENTROID OF EACH-WEEK CLUSTERS

Cluster	**Features**				**size**
	weeks	***after_peak***	***cta****	***ctp****	
0	9.94	7.28	0.45	0.17	499
1	5.17	-4.18	0.89	0.43	361
2	2.73	0.05	3.00	0.83	375
3	15.37	13.50	0.08	0.01	147
4	8.77	0.21	1.72	0.75	441
5	5.41	3.58	0.60	0.20	447
6	0.99	-2.64	0.42	0.16	367
7	1.30	-8.47	0.33	0.14	315

TABLE X. SAMPLE DATA IN EACH CLUSTER

Country	**Province**	**weeks**	**after_peak**	**cta**	**ctp**	**cluster**
Thailand	-	1	-1	2.10	0.61	2
Thailand	-	2	0	3.42	1.00	2
Thailand	-	3	1	3.39	0.99	2
Thailand	-	4	2	1.66	0.48	5
Thailand	-	5	3	0.93	0.27	5
Thailand	-	6	4	0.68	0.20	5
Thailand	-	7	5	0.20	0.06	5
Thailand	-	8	6	0.17	0.05	0

Again, the results of clusters can be primarily separated into stages by weeks such as before-peak, peak, and after-peak.

- Cluster #0 represents the weeks of countries that have already passed their peak. The low *cta* and *ctp* numbers indicate the decreasing number of new confirmed cases compared to the overall average and the peak week. Countries in this cluster could potentially be developed to cluster #3, which is reaching the end of the pandemic, in the near future.

- Cluster #1 represents the weeks of countries that are rising to their peak. Considering the negative number of *after_peak* and the *ctp* value, it can be remarked that the outbreak situation in this cluster is around halfway to its peak. According to a further investigation, it is found that countries in this cluster have the peak week around the 10^{th} week, which is a high number compared to the other clusters.

- Cluster #2 represents the weeks of countries that are facing its peak of the infection as stated by the highest *cta* value. This cluster is similar to cluster #4, but the *weeks* value is not as high. It means that the countries in this cluster reach their peak shortly after confirmed cases exceed 100 cases.

- Cluster #3 represents the weeks of countries that approach the end of the infection as the ratios of confirmed cases to the average (*cta*) and the peak week (*ctp*) are extremely low. Furthermore, the numbers of *weeks* and *after_peak* are extremely high.

- Cluster #4 represents the weeks of countries that have been facing its peak of the infection since the number of weeks is comparatively high. However, the *after_peak* is very low while the *cta* and *ctp* ratios are very high. It is noticed that almost all the countries in this cluster are now facing the stage of high infection numbers at the latest dates in the data.

- Cluster #5 represents the weeks of countries that have already passed their peak, but not as far as cluster #0. The circumstances of countries in this cluster could potentially be developed to cluster #0 in the near future.

- Cluster #6 represents the weeks of countries that are in the early stage of COVID-19 since the *weeks* value in the cluster is incredibly low and the *after_peak* value is negative.

- Cluster #7 represents the weeks of countries that are in the early stage of COVID-19 similar to cluster#6. However, the much lower *after_peak* values in this cluster when compared to cluster #6 suggested that the countries in this cluster take a longer time to reach their peak.

Additionally, the path of changes to different clusters throughout the time can be separated into two groups. In the first group, the infection rises to its peak in a short time after the first week (*weeks* 0), such as Thailand or China. It starts at cluster #6, then #2, #5, #0, and #3, respectively. On the other hand, the countries in the second group take a longer time to reach their peak, and almost all countries have not yet passed the stage of high infection numbers at the moment (7^{th} June 2020). It starts at cluster #7 and moves to #1 and #4, respectively.

2) Correlation results: The correlation of the each-week clusters and policies

To study the correlation between implementation date and each stage of the virus spreading, the defined clusters from the previous process were grouped into three stages: before-peak, peak, and after-peak. The after-peak group, however, had a small amount of implemented policies during the periods compared to other groups. Therefore, the correlations of the after-peak group were not analyzed.

Firstly, the correlations of the before-peak group, including clusters #1, #6, and #7, were calculated. From the results shown in Table XI, *HeScrn* and *IsoQ* are the policies that have a high correlation compared to the others.

TABLE XI. CORRELATION OF POLICIES AND CLUSTER#1 #6 AND #7

Features	**Policies**	
	HeScrn	***IsoQ***
death_rate	0.48	0.60

Next, the correlations of the peak group, including clusters #2 and #4, were computed. The results in Table XII display three policies, namely *IsoQ*, *Econ*, and *Visa,* that have high correlations to the death rate. Furthermore, the testing policies (*Test*) had a moderate negative correlation to the ratio of confirmed cases this week to the previous week (*c_diff_p*). This result suggests that the faster *Test* is implemented, the higher number of the new confirmed cases is.

TABLE XII. CORRELATION OF POLICIES AND CLUSTER#2 AND #4

Features	**Policies**			
	Test	***IsoQ***	***Econ***	***Visa***
death_rate		0.51	0.69	0.71
c_diff_p	-0.55			

C. Discussion of both experiments

In this work, the two experiments, which are policies implementation on the overall situation and the impromptu situations, are performed. The results from the first experiment demonstrate that the early implementation of movement restrictions (*DRT*, *IsoQ,* and *InFS*) and public health policies (*RqPro*) seems to help reduce the overall death rate and expedite the peak. In the second experiment, the early implementation of restriction policies (*Visa* and *IsoQ*) seems to help reduce the death rate after two weeks of implementation. Moreover, *Test* seems to help identify infected people faster.

Despite the differences in policies between the two experiments, the similarity is worth mentioning. Not only does the isolation and quarantine policies (*IsoQ*) stand out in both experiments, but it also has an outstanding correlation with every studied cluster. It could then be concluded that *IsoQ* might help with the overall situation and impromptu situation if it is implemented early on. The presence of *IsoQ* also agrees with the results from the balancing paper [1] and the optimization paper [2]. The results of the balancing paper [1] indicated that the early implementation of physical distancing policies is the best possible option. Meanwhile, the results of the optimization paper [2] demonstrated that quarantine policy helps reduce the death rate significantly.

The optimization paper [2] also suggested that a complete lockdown could reduce the total fatalities more than a quarantine. However, in the current situation, there are a little number of countries implementing lockdown compared to the quarantine policy. If the number of countries implementing lockdown policy increases, it would be interesting to inspect the correlation of it.

IV. Conclusion

The objective of this work is to study whether or not the implementation date of policies relates to the spreading of COVID-19. This work then begins with gathering data from a variety of sources and preprocessing the raw data to create well-structured data for further analysis. Then, the clustering technique is adopted to separate the stages of each country. The experiments are performed in two ways: the overall clusters and the each-week clusters. After analyzing the characteristics of each cluster, the clusters data are merged with policies to examine their correlations.

In the first experiment, correlations of the overall clusters are calculated to find out how the implementation date of policies affects a country's overall statistics. In the second experiment, correlations of each-week clusters are computed to find how the implementation date of policies affects a country two weeks after it is executed.

The results show that faster high-correlation policies are implemented, the better the situation would be. Even though different policies correlate to different clusters, the isolation and quarantine policies appear on all the studies of correlations with an interestingly high correlation coefficient. The appearance of the policy also corresponds to *Balancing Fiscal and Mortality Impact of SARS-CoV-2 Mitigation Measurements* [1] and *The optimal COVID-19 quarantine and testing policies* [2] in terms of the implementation date of isolation-like policies affects the death rate and how fast a country would pass its peak.

For more detailed data and the source codes of this work, please refer to https://github.com/nonna4822/policies-and-the-spreading-of-covid19

References

[1] M. Cruz-Aponte, and J. Caraballo-Cueto, "Balancing Fiscal and Mortality Impact of SARS-CoV-2 Mitigation Measurements," arXiv preprint arXiv:2006.01363, 2020.

[2] F. Piguillem, and L. Shi, "The optimal COVID-19 quarantine and testing policies," Einaudi Institute for Economics and Finance, 2020.

[3] T. Hale, N. Angrist, B. Kira, A. Petherick, T. Phillips, and S. Webster, "Variation in government responses to COVID-19," Blavatnik school of government working paper, 31, 2020.

[4] Johns Hopkins University Center for Systems Science and Engineering, "JHU Coronavirus COVID-19 Global Cases, by country," [Online]. Available: https://console.cloud.google.com/marketplace/details/johnshopkins/covid19_jhu_global_cases [Accessed Jun 7, 2020].

[5] Assessment Capacities Project, "Covid19 Government Measures Dataset," [Online]. Available: https://www.acaps.org/covid19-government-measures-dataset [Accessed Jun 7, 2020].

[6] S. Mervosh, D. Lu and V. Swales, "See Which States and Cities Have Told Residents to Stay at Home," New York Times, 2020.

[7] "National responses to the COVID-19 pandemic," Wikipedia. [Online]. Available: https://en.wikipedia.org/wiki/National_responses_to_the_COVID-19_pandemic [Accessed Jun 7, 2020].

[8] Jinghua Li, "Covid-19 country lockdown days," [Online]. https://www.kaggle.com/letili0417/covid19-country-lockdown-days [Accessed Jun 7, 2020]

[9] Sloth, "Covid19 US Lockdown Dates Dataset," [Online]. https://www.kaggle.com/lin0li/us-lockdown-dates-dataset [Accessed Jun 7, 2020]

[10] "API Reference – scikit-learn document," [Online]. Available: https://scikit-learn.org/stable/modules/classes.html [Accessed Jun 7, 2020]

[11] S. Lloyd, "Least squares quantization in PCM," IEEE Transactions on Information Theory, vol. 28, no. 2, pp. 129–137, 1982.

[12] P. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis", Journal of Computational and Applied Mathematics, vol. 20, pp. 53-65, 1987.

[13] R. Thorndike, "Who belongs in the family?", Psychometrika, vol. 18, no. 4, pp. 267-276, 1953.

[14] D. Freedman, R. Pisani, and R. Purves, Statistics. New York: W.W. Norton & Co., 2007