Understanding the Variances in COVID-19 Pandemic Outcome–Excess Mortality–with Social, Cultural, and Environmental Factors

Jiner Zheng

Faculty Sponsor: Professor Maryam Gooyabadi

jzheng@wesleyan.edu

1. Introduction

With the increasing accessibility to all-cause mortality data, the excess mortality during the pandemic of COVID-19 has risen as the most reliable metric to many researchers as a way to measure the extent of its seriousness [1]. However, current research on estimating COVID-19 excess deaths is still somewhat limited. Relevant previous studies mainly used relatively simple statistical methods like regression and exponential smoothing, which did not take into account enough features in mortality time series, such as temporal autocorrelations and conditional heteroskedasticity. Moreover, due to the insufficient aggregated mortality data in many countries, studies on cross-country differences in excess deaths are still deficient.

To date, the considerable variance in the severity of the COVID-19 outbreak across countries has also drawn significant attention to researchers in various fields. Researchers have striven to understand such variability by modeling COVID-19 transmission rates with covariates such as the stringency of governmental restrictions, the duration of the infectious period, and travel restrictions [2, 3, 4]. In reality, a wide range of factors can influence the spread of COVID-19 and its outcome measures through various dimensions. More and more researchers have begun to formulate a comprehensive understanding of such factors, including urbanization levels, obesity rate, government effectiveness and testing capacity, and national cultures, to make sense of the cross-national differences in COVID-19 outcomes [5, 6, 7].

Therefore, the main goals of this presented work are: 1) to produce estimates of COVID-19 outcome measure: excess deaths, using a hybrid method of ARIMA/GARCH models; 2) to create clustered cross-country profiles using indicators within social, economic, cultural, and environmental dimensions as covariates and explore main characteristics of countries in different clusters.

1.1 The importance of analyzing COVID-19 excess mortality

Being declared as a global pandemic by World Health Organization (WHO) on March 11th, 2020, COVID-19 has infected billions of people across 213 countries and become a worldwide threat. Consequently, more statistical analyses and forecasting methods are needed to limit its spread and monitor mortality patterns. Among different measurements related to the effects of COVID-19, excess mortality has been found most reliable and objective, because other statistics such as the number of confirmed cases and disease-related deaths may be undercounted from misdiagnosis and variance in governments' testing and reporting capacity [1].

Excess mortality is defined as the increase in deaths from all causes during a crisis relative to expected deaths under "normal" conditions, mortality in the absence of exceptional events such as pandemics:

$$\text{Excess Deaths} = \text{Reported Deaths} - \text{Expected Deaths}. \tag{1}$$

Even though existing literature focusing on excess deaths during COVID-19 is relatively limited, some researchers have proposed practical methods to estimate excess mortality. For instance, Ritchie et al. [8] used a regression model to predict the expected number of deaths in 2020 under standard conditions using historical mortality data from 2015 to 2019 respectively for each country. Then, they computed excess deaths by taking the difference between reported deaths and expected deaths. They also presented a measure of excess deaths for cross-country comparisons, the P-score of excess deaths:

$$\text{P-score} = \frac{\text{Reported Deaths} - \text{Expected Deaths}}{\text{Expected Deaths}} * 100. \tag{2}$$

To produce estimates of excess mortality as the COVID-19 outcome measure for further analyses, this presented work built upon their method of computing excess deaths as well as its comparable metric of *p-score* but utilized a hybrid of ARIMA and GARCH models instead to project the number of expected deaths in each country (see Section 1.3 for detailed information about each model).

1.2 Socio-economic, environmental and cultural impacts on COVID-19

As mentioned earlier, factors within multiple dimensions can influence the COVID-19 outcome. For example, the older population has been considered to be more susceptible to COVID-related deaths [9]. Studies have also found a positive relationship between median age and the share of population aged 65 years or more with COVID-19 mortality [5, 6, 7]. Besides, since the spread of this pandemic relies heavily on physical contacts, areas with the denser population can result in more face-to-face interactions, thus exacerbating the outcome measure of such infectious disease [10]. Despite age and population density, other socio-economic factors like Gross Domestic Product (GDP) per capita can also be a potential determinant of the coronavirus diffusion. More specifically, countries that are more financially prosperous are where people have more opportunities to attend a variety of social events and large gatherings [11]. In addition, the scientific community has also explored the effect of environmental indicators, especially temperature, on the COVID-19 transmission. One cross-national study has found that as temperatures get warmer, the daily number of confirmed cases increases in temperate climate zones but decreases in tropical climate zones [12].

Apart from socio-economic and environmental impacts on the pandemic, national or societal cultures could also play an significant role in the COVID-19 morbidity and mortality given their influence on people's behaviors and relationships with the government, other individuals, and society. Studies have found that collectivist cultures or cultures with more strict norms (i.e. cultural tightness) could help mitigate the morbidity and mortality of COVID-19 [5, 13, 14]. In contrast, individualism and cultural looseness were found positively associated with measures of COVID-19 outcome such as the case fatality rate [5, 14].

As the second purpose of this study, indicators within the socio-economic, environmental, and cultural dimensions were included as covarietes in the clustering analysis to investigate characteristics of countries with similar features and explore the underlying reasons behind the cross-country differences in COVID-19 excess mortality.

1.3 Time series modeling and forecasting

Modeling and forecasting time series data have been essential in many fields like social sciences and epidemiology. Time series data usually involves repeated measurement of a univariate object at certain time intervals (every day, week, month, or year) over many observations. Many statistical and computational methodologies have been widely used in time series analysis to help researchers understand the underlying pattern and naturalistic process over time and predict future trends and values. Among a wide range of traditional statistical approaches in time series analysis, the most prevalent models are exponential smoothing and Autoregressive Integrated Moving Average (ARIMA) models, including the Seasonal ARIMA, an extension of ARIMA that particularly supports time series data with seasonality [15].

ARIMA models have been found most effective in time series forecasting in that they can capture not only basic features like seasonality and trend but also serial autocorrelations in time series. An ARIMA model can be understood by the following three major components [16]:

- **Autoregression (AR)**: incorporates the dependency between the current observation and its lagged values (i.e. values in prior time periods).
- **Integrated (I)**: refers to the degree of differencing between values at current time and previous values to make time series data become stationary.
- **Moving average (MA)**: refers to the moving average process applied to lagged values when current observations also depend on lagged residual errors.

In addition, some researchers have utilized a hybrid methodology of ARIMA and Generalized Autoregressive Conditional Heteroskedasticity (GARCH) models to analyze time series data that contains serial autocorrelations in the variance [17, 18, 19, 20]. Such an approach is particularly useful in modeling changes in the variance of time series or volatility that are time-dependent.

1.4 Clustering Time Series Data and Mixed-Type Data

This presented work also consists of clustering on time series objects and data with a mixed-type of variables, including both numerical and categorical features, respectively.

Past literature has proposed several approaches to cluster time series data such as using probability-based distance to account for the seasonality of time series and using Hidden Markov model based distance to model more complex time series [21, 22]. In particular, the Euclidean distance and the Dynamic Time Warping (DTW) distance have been found as the most popular methods to measure distances in time series. Furthermore, DTW has been found as the most accurate approach in that it allows an one-to-many match across all datapoints and does not require two time series to have the same length as compared to the Euclidean measure[23].

Clustering data where quantitative and qualitative variables coexist has also been a common issue in real life situations. One of the most useful metrics to measure similarities across observations is the Gower distance, which averages partial dissimilarities between two observations using different approaches given different types of features being evaluated [24] (see Section 2.2.3 for detailed computations). Furthermore, the Gower distance is often paired with the K-Medoids algorithm, a clustering technique that appears to outperform the traditional K-Means clustering as it is more robust to outliers and reduces noise better [25].

2. Methods

2.1 Sample and Measures

There were four primary dimensions of variables of interest in this study: 1) COVID-19 outcome measures – excess mortality and its corresponding p-score of excess deaths; 2)

Socio-economic factors (e.g. median age, GDP per capita, population density, etc.); 3) Cultural factors (e.g. degree of personal freedom and degree of civic and social participation); 4) Environmental factor – Climate class (e.g. Tropical, Dry, or Temperate climates). After excluding countries with significant missing data entries in all variables of interest, a final sample of 70 countries and regions was included in this project, with 50% in Europe ($n = 36$), 23% in Asia ($n = 16$), 11% in South America ($n = 8$), 6% in Africa ($n = 4$), 6% in North America ($n = 4$), and 3% in Oceania ($n = 2$).

2.1.1 Outcome Measure: Excess Mortality

The outcome measure of COVID-19 – excess mortality – was computed as the difference between reported deaths and expected deaths during the pandemic. Data on reported deaths came from the the World Mortality Dataset [27], which contained a collection of country-level historical and recent data on mortality from all causes since 2015, retrieved from a wide range of data sources such as the Short-Term Mortality Fluctuations (SMTF) dataset [28] and the Rosstat [29]. For the consistency of data availability across all selected countries, updated time periods were restricted to be from January 2020 to December 2021, and each mortality time series was aggregated into monthly values respectively for each country or region. Then, excess deaths were calculated using a hybrid method of ARIMA and GARCH models based on historical data on deaths from 2015 to 2019 (see Section 2.2.1 for more information).

2.1.2 Covariates

Seven social and economic factors, including policy stringency index, population, population density, median age of the population, share of population aged 65 years and older, life expectancy, and GDP per capita, were retrieved from Our World in Data [30], which provides a comprehensive collection of daily measures within various dimensions during the COVID-19 pandemic. Particularly, the policy stringency index was a compound statistic of governmental responses to COVID-19, with a scale from 0 to 100 (100=strictest) based on nine governmental response indicators such as public gathering restrictions, travel bans, and workplace closures.

Cultural factors reflecting social norms and cultural values were retrieved from Legatum Institute's 2019 Prosperity Index [31]. The following seven prosperity indicators were selected for the purpose of focusing on individuals' social relationships, involvements and freedom, as well as government effectiveness:

- **Personal and Family Relationships**: the strength of individuals' family ties and other personal relationships
- **Social Network**: the degree of individuals' opportunities and strengths to interact with their wider social networks
- **Civic and Social Participation**: the extent of individuals' involvements and contributions within the society
- **Personal Freedom**: the extent to which people are free from restrictions
- **Freedom of Assembly and Association**: the degree of individuals' freedom to assemble in public spaces with others
- **Health Prevention**: the degree of the health system's capacity and effectiveness to prevent diseases from occurring
- **Government Effectiveness**: the quality of the bureaucracy and public health provision and the competence of officials

Environmental factor–climate class–was retrieved from Köppen-Geiger updated climate zones by Kottek et al. [32] using the downscaling algorithms of Rubel et al. [33]. There were four climate classes: Tropical climates, Dry (Arid and Semiarid) climates, Temperate (Mesothermal) climates, and Continental (Microthermal) climates. Countries with missing climate class entries, including Switzerland, Mongolia, Iceland, Guatemala, Kyrgyzstan, Mauritius, and Paraguay, were manually adjusted with their most dominant climate zones, respectively.

2.2 Analytical Methods

To analyze the complete data with measures above, three primary analytical steps were taken: 1) Time series data modeling using a hybrid methodology of ARIMA and GARCH models to compute excess deaths, 2) Clustering excess deaths and policy stringency time series alone based on Dynamic Time Warping (DTW) distances and hierarchical clustering, and 3) Clustering all data with mix-typed variables of interest based on Gower distances and K-Medoids clustering (see Section 1.4 for more information on each technique).

2.2.1 Computing Excess Deaths

The estimated number and P-score of excess deaths for each country were calculated as:

$$\hat{E} = D - \hat{D}, \text{ and } \hat{P} = \frac{\hat{E}}{\hat{D}}, \tag{3}$$

where $\hat{E}$ denotes the number of estimated excess deaths during the pandemic of COVID-19, D denotes the number of reported deaths from all causes, $\hat{D}$ denotes the number of expected deaths, and $\hat{P}$ denotes the estimated P-score of excess deaths (see Section 1.1 for more information about P-score) respectively in each country.

To get the number of expected deaths for 2020 and 2021, an ARIMA model was first fitted on the historical data on mortality in each country from 2015 to 2019, followed by an addition of GARCH model if there existed an ARCH effect in the mortality time series. In addition, the Akaike Information Criterion (AIC) was used for model selection. Following is the generalized equation of model separately for each country:

$$
\begin{aligned}
D_{t,Y} = c &+ \sum_{i=1}^{p} \alpha_i D_{t-i,Y} + D_{t,Y} - D_{t-d,Y} + Z_t + \sum_{j=1}^{q} \lambda_j Z_{t-j,Y} \\
&+ \sum_{k=1}^{P} \alpha_k D_{t-k\cdot m,Y} + D_{t,Y} - D_{t-D\cdot m,Y} + \sum_{l=1}^{Q} \lambda_l Z_{t-Q\cdot m,Y} \\
&+ \delta^2_{t-q_2,Y}\ \epsilon_{t-p_2,Y}.
\end{aligned} \tag{4}
$$

Here, $D_{t,Y}$ is the number of deaths at time t in year Y (t is a given month). p is the AR order; d is the degree of regular differencing; q is the MA order; P is the seasonal AR order; D is the degree of seasonal differencing; Q is the seasonal MA order; and m is the frequency; q_2 is the GARCH order; p_2 is the ARCH order; Z_t is the error at time t; $\delta^2_{t-q_2,Y}$ is the variance at time $t - q_2$ in year Y; $\epsilon_{t-p_2,Y}$ is the error of variance at time $t - p_2$ in year Y. These models can capture seasonal variations in deaths, yearly trends over recent years due to population growths and other social, economic factors, short-term serial correlations, and the potential conditional heteroscedasticity in time series.

Then, based on the above hybrid model, the predicted values for 2020 and 2021 were used as expected deaths in 2020-2021 to calculate excess mortality. To get the final estimate for excess mortality, excess deaths in each month were first computed by taking the difference between observed deaths and expected deaths, and then monthly excess deaths were summed up across all time periods from the beginning of pandemic (i.e. January 2020) as following:

$$
\hat{E} = \sum_{t \geq t_1} (D_{t,2020} - \hat{D}_t) + \sum_{t} (D_{t,2021} - \hat{D}_t). \tag{5}
$$

where t_1 denotes the start time period of summation in 2020.

Figure 1 and Figure 2 are time series plots of the estimated excess mortality and its corresponding p-scores in all countries with highlights on the four most severe countries.

Noticeably, the overall COVID-19 outcome appeared to be most severe in Russia, United States, Brazil, and Mexico when just looking at the raw number of excess deaths, but in Bolivia, Azerbaijian, Peru, and Ecuador when looking at p-scores instead. As mentioned earlier, the p-score of excess deaths is a more useful and meaningful metric for cross-national comparison in that it compares the ratio between excess deaths and expected deaths across countries, thus less sensitive to the variance in yearly trends of deaths resulted from population growth or other known factors in previous years.

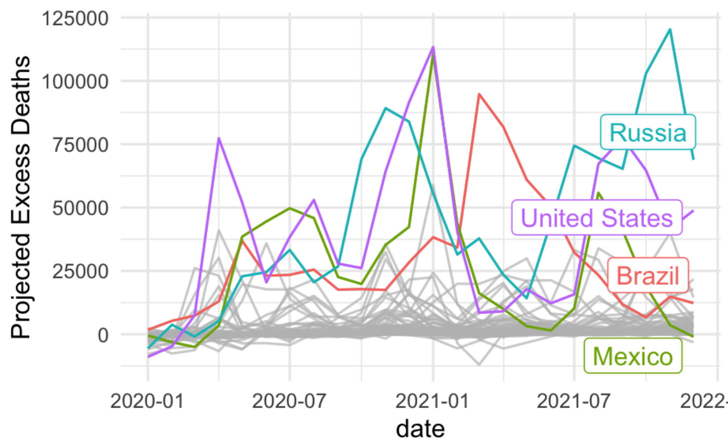


Figure 1: Number of excess deaths

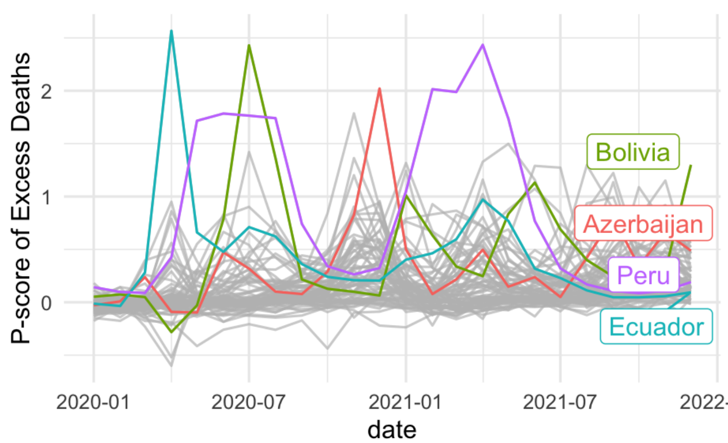


Figure 2: P-score of excess deaths

2.2.2 Clustering Excess Deaths and Policy Stringency Time Series Objects

For the purpose of clustering on all variables of interest, hierarchical clustering analyses based on Dynamic Time Warping (DTW) distances were conducted to produce a class membership variable for each of the two time series objects, excess mortality and policy stringency index, respectively.

DTW distances were calculated using the dtw package [34] after normalizing time series data. Then, hierarchical clustering using Ward's Minimum Variance method was performed. Additionally, the Silhouette coefficient was used to determine the optimal number of clusters ranging from two to eight maximum. The highest silhouette value for clustering excess deaths time series suggested a number of six clusters, but for the purpose of the simpler and better interpretability, a number of four clusters was selected. As for the policy stringency time series, a number of four clusters was selected based on its corresponding highest silhouette value.

2.2.3 Clustering Mixed-Typed Data to Extract Country Profiles

As the final step of analyses, a K-Medoids clustering analysis using Gower distances was conducted on the complete dataset with all variables within the socio-economic, cultural, and environmental dimensions as mentioned above (see Section 2.1 for more information). Time series variables including excess deaths and policy stringency index were replaced with their corresponding class membership variables respectively for clustering purposes. In addition, a cumulative p-score of excess deaths was also computed and included in the clustering analysis as a measure of the severity of the overall COVID-19 outcome.

Gower distances were computed using the daisy() function within the cluster package [35], which first automatically standardized all numeric features before calculating distances, followed by the implementation of k-medoids clustering via the pam() function in the same package. Below is the computation of Gower distances between countries in each pair:

$$D_{Gower}(c_1, c_2) = 1 - \left(\frac{1}{p} \sum_{i=1}^{p} s_i \left(c_1, c_2\right) \right), \tag{6}$$

where $D_{Gower}(c_1, c_2)$ denotes the Gower distance between two given countries c_1 and c_2; p denotes the number of features, and $s_i\left(c_1, c_2\right)$ is the partial similarity function based on the type of the i_{th} feature. For a numerical feature such as population density, the partial similarity is computed as:

$$s_i\left(c_1, c_2\right) = 1 - \frac{|x_{1,i} - x_{2,i}|}{R_i}, \tag{7}$$

where $|x_{1,i} - x_{2,i}|$ is the absolute difference between values of the i_{th} feature in countries c_1, c_2, and R_i is the maximum range of feature i observed from all countries (i.e. the absolute difference between the largest value and the smallest value of feature i across all countries). As for a categorical feature such as climate class, the partial similarity is assigned to be 1 if two countries had the same value, and 0 otherwise. Similarly, silhouette widths were

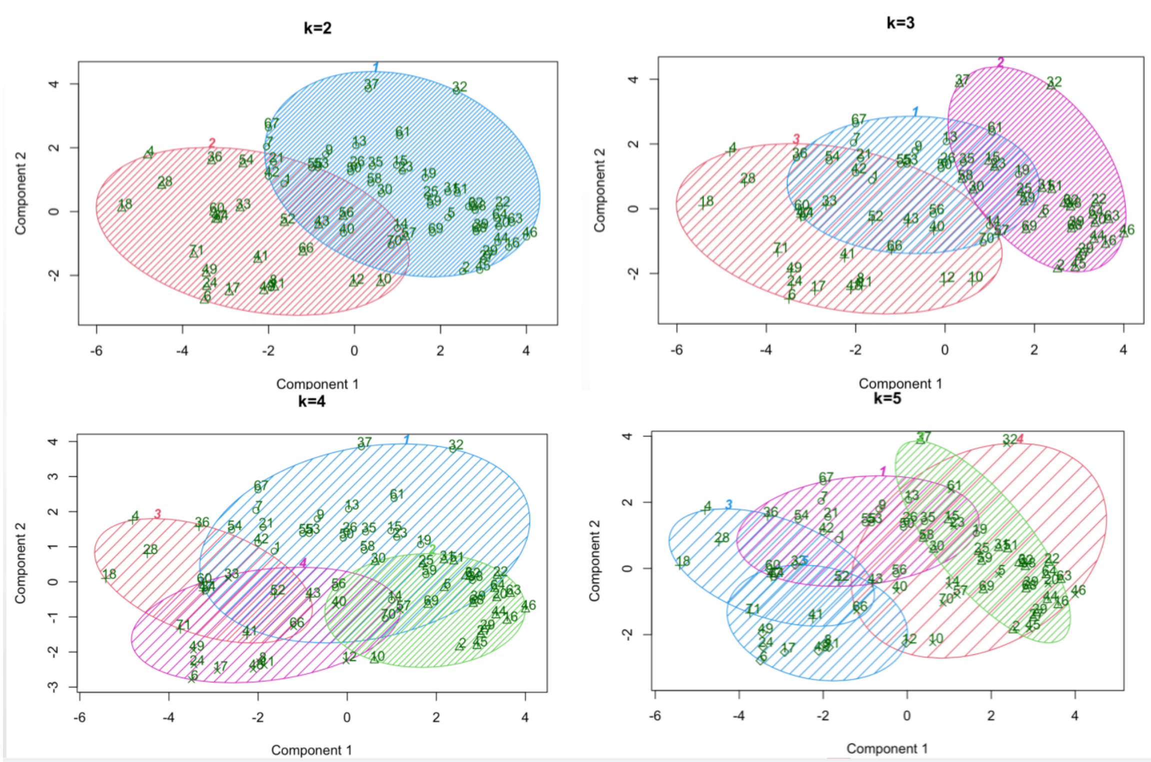
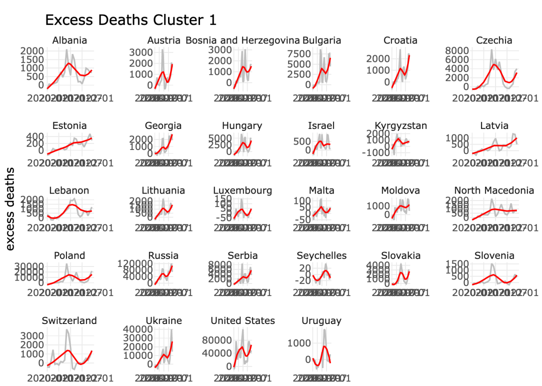


Figure 3: Clustering output with different number of clusters

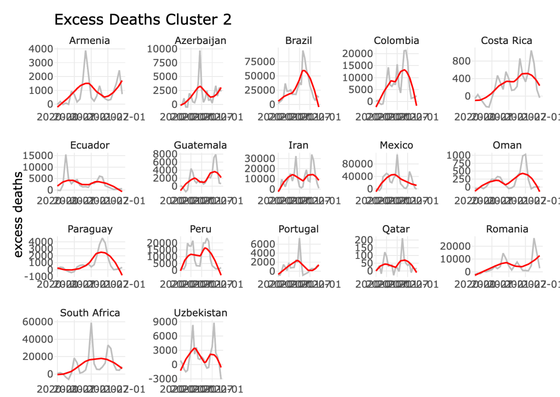
calculated to determine the optimal number of clusters. After searching for a number of clusters ranging from 2 to 8, a number of three clusters was selected based on the second highest silhouette width for the purpose of more comprehensive interpretations and coverage of clustered country profiles. Figure 3 shows a glimpse of the clustering output with different number of clusters.

3. Results

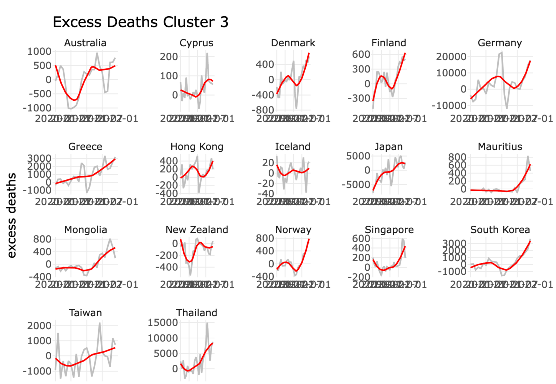
3.1 Excess Deaths Class Profiles



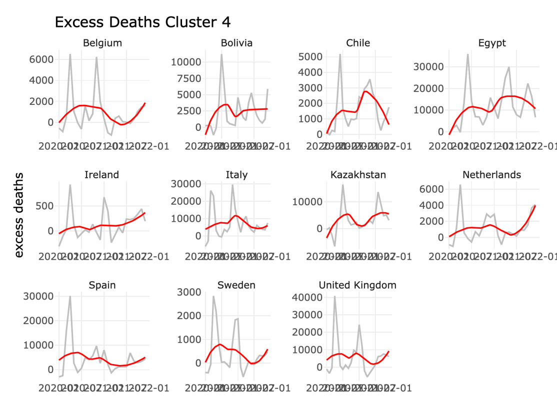
(a) Excess Deaths Cluster 1



(b) Excess Deaths Cluster 2



(c) Excess Deaths Cluster 3



(d) Excess Deaths Cluster 4

Figure 4: Excess Deaths Class Profiles

The above figure provides a visual representation of the excess mortality over time in all countries within each of the four clusters. There were 28 countries in Excess Deaths Cluster

1, including Albania, Austria, Switzerland, etc. Figure 4(a) shows two salient similar features across countries in Cluster 1: 1) Most recent trend of excess deaths in these countries seemed to be slightly going upwards; 2) Most severe outbreaks often had occurred by the end of year 2020.

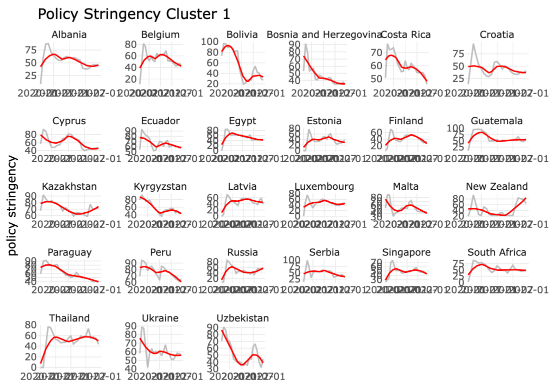
17 countries were classified as Excess Deaths Cluster 2 (e.g. Brazil, Mexico, Peru, and Costa Rica). Figure 4(b) suggests that the most recent trend of excess mortality in these countries seemed to be downward, while there was no obvious common pattern in the occurring of most severe outbreaks across countries within cluster 2.

Excess Deaths Cluster 3 had 17 countries such as Germany, Denmark, and Iceland. According to Figure 4(c), excess mortality in most recent time periods was also increasing but more dramatically than cluster 1. The most severe outbreaks appear to often have happened by the end of year 2021.

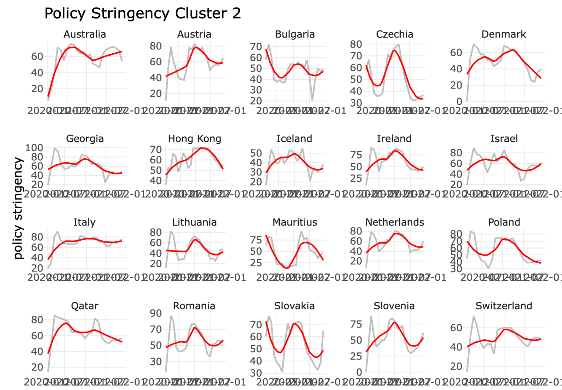
Lastly, 11 countries were grouped into Excess Deaths Cluster 4 (e.g. Spain, Italy, and United Kingdom). Figure 4(d) indicates that most recent trend of excess mortality seems to be decreasing but with some slight cyclic increases in certain countries, while most severe outbreaks often had occurred at the beginning of the pandemic (around March 2020).

3.2 Policy Stringency Class Profiles

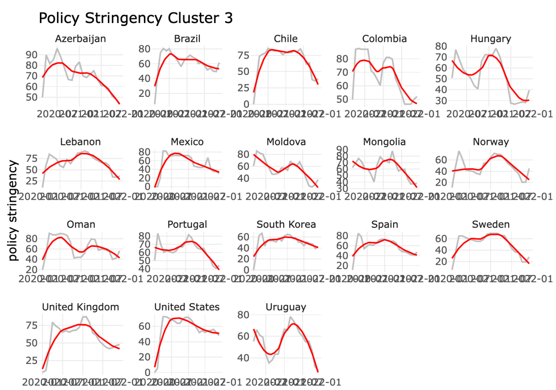
Despite the excess deaths class, countries were also assigned with a class membership based on policy stringency time series alone, and there was a final number of 4 clusters where Stringency Cluster 1 had 27 countries, Stringency Cluster 2 had 20 countries, Stringency Cluster 3 had 18 countries, and Stringency Cluster 4 had six countries. The following collection of time series plots of policy stringency index in all countries provides a snapshot of the common pattern in governmental responses stringency over time within each cluster:



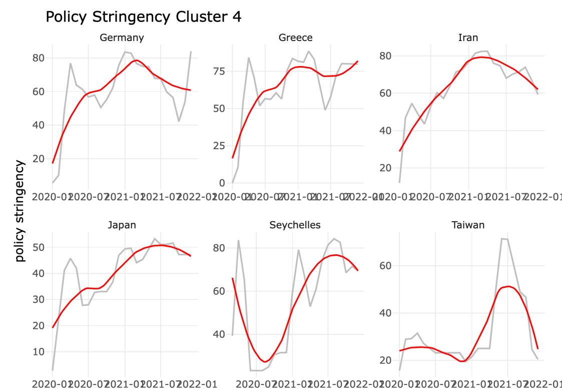
(a) Policy Stringency Cluster 1



(b) Policy Stringency Cluster 2



(c) Policy Stringency Cluster 3



(d) Policy Stringency Cluster 4

Figure 5: Policy Stringency Class Profiles

According to Figure 5(a), there appears to be a significant downward trend in most recent policy stringency level across countries in Stringency Cluster 1. Besides, countries in this cluster seem to have had their strictest policy responses to COVID at the beginning of pandemic (March 2020), but such level of stringency after the peak appears to have gone down dramatically. Similar to cluster 1, a decreasing trend in policy stringency level after its peak around March 2020 can also be detected in countries within Stringency Cluster 2 (as shown by Figure 5(b). However, such downward trend has been more stable for a longer period of time as compared to that in cluster 1.

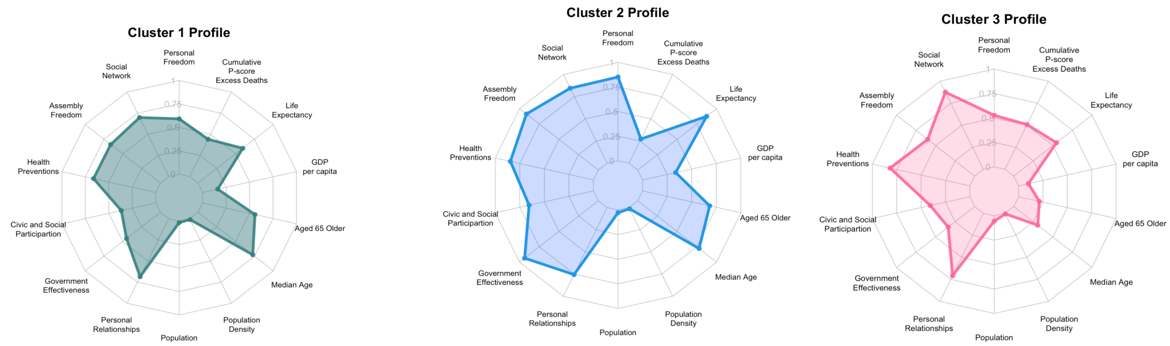
Then, Figure 5(c) shows that Stringency Cluster 3 has a similar pattern of a stable and decreasing trend in policy responses level after the highest value at the beginning of the pandemic for a long time, but such trend is slightly more dramatic than that in cluster 2. Lastly, in Figure 5(d), countries in Stringency Cluster 4 seem to have had their strictest response to COVID much later compared to other three clusters (mostly in the year of 2021).

3.3 Final Clustered Country Profiles

Taking all variables of interest into considerations, three clusters – Low, Moderate, and High Risks – were identified through K-Medoids clustering analysis:

Table 1: Summary statistics of clustered country profiles (median values)

	Cluster 1 *Moderate*	**Cluster 2** *Low*	**Cluster 3** *High*
P-score of Excess Deaths	16.36%	7.47%	19.83%
Median Age	41.30	42.60	30.65
Aged 65 Older	15.66%	18.89%	6.78%
GDP per capita	21,611	41,622	14,814
Population Density	78.84	112.75	58.14
Life Expectancy	76.47	82.27	75.47
Continent	70% in Europe (n=14)	79% in Europe (n=22)	73% in Asia or South America (n=16)
Climate Class	80% Temperate (n=16)	79% Temperate (n=22)	86% in Tropical or Dry (n=19)
Stringency Class	50% in Class 1 (n=10)	79% in Class 2 (n=14)	55% in Class 1 (n=12)
Excess Deaths Class	85% in Class 1 (n=17)	71% in Class 1 or 3 (n=20)	85% in Class 2 (n=14)



(a) Cluster 1: Moderate (b) Cluster 2: Low (c) Cluster 3: High

Figure 6: Radarcharts of clustered country profiles (median values)

Table 1 and Figure 6 together show the main characteristics of countries within each cluster, where median values of each feature were calculated for descriptive purposes in both the summary statistics table and radarcharts.

With the lowest cumulative p-score of excess deaths, cluster 2 could be considered as the low risk cluster, which contains 28 countries such as the United States, United Kingdom, and Germany. 79% of these countries are in Europe and Temperate climate zones. This low risk cluster 2 has highest median values across all socio-economic factors, especially with a much larger share of the older population and a significantly higher GDP per capita compared to the other two clusters. Besides, 50% of countries being categorized with Stringency Class 2 indicates that half of the low risk countries seem to have a stable and slightly decreasing trend in policy responses to COVID. As for the cultural dimension, Figure 6(b) demonstrates that Cluster 2 countries also scored the highest across all indicators of social norms and cultural values, suggesting that some dominant cultural features include the highest degree of freedom, the strongest social relationships or interactions with others and involvements in the society, and the most desirable government effectiveness and health prevention.

Cluster 3 could be considered as the high risk cluster based on its highest cumulative p-score of excess deaths, including a total number of 22 countries where the majority of them are in Asia or South America with Tropical or Dry climates (e.g. Brazil, Mexico, and Bolivia). In contrast to cluster 2, this high risk cluster 3 scores the highest across all measures within the socio-economic dimension. A 55% of countries in Stringency Class 1 indicates that more than half of these high risk countries have been reducing their policy responses to COVID dramatically after having taken the strictest response at the beginning of the pandemic. What's more, the radarchart in Figure 6(c) shows that societal cultures in Cluster 3 countries could be characterized with relatively lower personal freedom, government effectiveness, or civic and social participation, but significantly strong social interactions with individuals and wider social network as well as high quality of health prevention.

Lastly, there were 20 countries in Cluster 1–the moderate cluster–such as South Korea or Russia. The majority of them are in Europe and Temperate climates. Despite the moderate median scores in all socio-economic measures, Cluster 1 countries also have the moderate values across almost all cultural indicators as compared to the other two clusters, except that degrees of social network and health prevention in Cluster 1 countries are both noticeably lower than Cluster 3.

4. Discussion

4.1 Conclusions and Implications

This presented work produced estimates of COVID-19 excess mortality and its corresponding p-scores across 70 countries via a hybrid methodology of ARIMA/GARCH models and created three clustered cross-national profiles – high, low, and moderate risk clusters – under considerations of socio-economic, cultural, and environmental factors to formulate a fairly comprehensive understanding of the variance in COVID-19 outcome across the globe.

Even though clustering of country profiles was only for descriptive purposes, there still existed some salient common characteristics in each cluster that are useful to help explore the underlying reasons behind such cross-national variability of the overall burden on mortality from the pandemic of COVID-19. For instance, the Low Risk Cluster 2 (with the lowest cumulative p-score of excess mortality) was distinguished by more liberal and assertive cultures, high government effectiveness and health prevention, and a stable, slightly downward trend in the stringency level of governmental responses to COVID, regardless of its considerably high scores on all socio-economic indicators included in the analyses like high GDP per capita, a larger share of older population, and high population density. In contrast, countries in High Risk Cluster 3 were found to have the lowest socio-economic values including median age, share of population aged 65 years or more, and GDP per capita, low government effectiveness, and less freedom from restrictions but high social interactions with others and stronger social relationships and family ties.

It is interesting that such findings contradict some of the previous literature to some extent. For example, higher levels of socio-economic indicators like GDP per capita or the older population were expected to have associated with an increase in COVID-19 outcomes such as morbidity and mortality [5, 6, 7, 10, 11], and the tightness of cultural norms and less personal freedom were expected to help reduce such outcome [5, 13, 14]. However, such deviation can be explained by the potential inter-dependency between various factors and the different levels of their impacts on COVID-19. For example, although countries in the Low Risk Cluster 2 such as the United States or the United Kingdom have a larger number of more vulnerable older population and more opportunities for large gatherings in overcrowded public spaces, which can result in much faster transmission rates, the high quality of their health infrastructure and better preventive interventions could have overcome these issues and effectively control for the burden on mortality. In addition, cultural tightness could also play a role in people's compliance with governmental restrictions on COVID-19. Thus, more strict social or cultural norms might not be able to mitigate COVID-19 outcome as much in the high risk cluster countries where policy stringency level has been decreasing dramatically over time.

4.2 Limitations and Future Directions

One of the biggest limitations in this presented study is that data sample was very small and not representative enough. As mentioned in the earlier section, the study only included 70 countries after excluding missing values, and the sample was also Western-centered as it consisted of 51% European countries and 6% North American countries. Another limitation is that the study only incorporated a relatively limited number of covariates within each of the socio-economic, cultural, and environmental domains of interest. Furthermore, as mentioned in the above section, most analyses were conducted only for descriptive purposes, and thus no causal inferences could be drawn from presented analytical results.

To better understand the variance in COVID-19 outcome measures across the globe, future researchers should consider conducting a larger-scale of cross-national study with a more representative sample and a more comprehensive set of covariates. Besides, it will also be interesting and meaningful to investigate which of these factors are statistically significant predictors of the COVID-19 outcome measure.

Lastly, the clustering analyses of time series objects in this presented work could also be improved by better algorithm such as the model-based clustering, which uses a probability-based approach and tries to recover the underlying probability distribution from the data [36].

References

[1] National Academics of Sciences, Engineering, and Medicine, Evaluating data types: a guide for decision makers using data to understand the extent and spread of COVID-19, The National Academics Press. (2020).

[2] Y. Fang, Y. Nie, M. Penny, Transmission dynamics of the COVID-19 outbreak and effectiveness of government interventions: a data-driven analysis. J Med Virol. 92 (6) (2020) 645-659.

[3] A. J. Kucharski, T. W. Russell, C. Diamond, Y. Liu, J. Edmunds, S. Funk, R. M. Eggo, Early dynamics of transmission and control of COVID-19: a mathematical modeling study, The Lancet Infectious Diseases. 20 (5) (2020) 553-558.

[4] R. M. Anderson, H. Heesterbeek, D. Klinkenberg, T. D. Hollingsworth, How will country-based mitigation measures influence the course of the COVID-19 epidemic? The Lancet. 395 (10228) (2020) 931-934.

[5] C. Cao, N. Li, L. Liu, Do national cultures matter in the containment of COVID-19? International Journal of Sociology and Social Policy. 40 (9/10) (2020) 939-961.

[6] L. L. Liang, C. H. Tseng, H. J. Ho, C. Y. Wu, Covid-19 mortality is negatively associated with test number and government effectiveness, Scientific Reports. 10 (1) (2020) 1-7.

[7] A. Upadhyaya, S. Koirala, R. Ressler, K. Upadhyaya, Factors effecting COVID-19 mortality: an exploratory study, Journal of Health Research. Advanced online publication (2020).

[8] H. Ritchie, E. Mathieu, L. Rodés-Guirao, C. Appel, C. Giattino, E. Ortiz-Ospina, J. Hasell, B. Macdonald, D. Beltekian, M. Roser, Excess mortality during the coronavirus pandemic (COVID-19) - Statistics and Research, Our World in Data. (2020).

[9] W. B. Applegate, J. G. Ouslander, COVID-19 presents high risk to older persons, Journal of the American Geriatrics Society. 68 (4) (2020) 681-681.

[10] S. Hamidi, S. Sabouri, R. Ewing, Does density aggravate the COVID-19 Pandemic? Journal of the American Planning Association. 86 (4) (2020) 495-509.

[11] S. Weyers, N. Dragano, S. Möbus, E. M. Beck, A. Stang, S. Möhlenkamp, K. H. Jöckel, R. Erbel, J. Siegrist, Low socio-economic position is associated with poor social networks and social support: results from the Heinz Nixdorf Recall Study, International Journal of Equity in Health. 7 (13) (2008).

[12] M. Y. Liu, Z. W. Li, Y. X. Zhu, Y. Liu, M. William, N. Kuetche, J. P. Wang, X. N. Wang, X. T. Liu, X. Li, W. Wang, X. H. Guo, L. X. Tao, Association between temperature and COVID-19 transmission in 153 countries, Environmental Science and Pollution Research. 29 (2022) 16017-16027.

[13] F. Huang, H. Ding, Z. Liu, P. Wu, M. Zhu, A. Li, T. Zhu, How fear and collectivism influence public's preventive intention towards COVID-19 infection: A study based on big data from the social media, BMC Public Health. 20 (1) (2020) 1-9.

[14] M. J. Gelfand, J. C. Jackson, X. Pan, D. Nau, D. Pieper, E. Denison, M. Dagher, P. A. M. Van Lange, C. Y. Chiu, M. Wang, The relationship between cultural tightness-looseness and COVID-19 cases and deaths: a global analysis, The Lancet Planetary Health. 5 (3) (2021) e135-e144.

[15] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, Time series analysis Forecasting and control, (3rd ed.), Englewood Cliffs, Prentice-Hall, New Jersey, 1994.

[16] V. Kotu, B. Deshpande, Chapter 12 - Time Series Forecasting, In: Kotu V, Deshpande B (eds) Data science, 2nd edn. Morgan Kaufmann, Burlington, 2019, pp 395–445.

[17] F. Yusof, I. L. Kane, Z. Yusuf, Hybrid of ARIMA–GARCH modeling in rainfall time series, Jurnal Teknologi. 63 (2) (2013) 27–34.

[18] C. Ding, J. Duan, Y. Zhang, X. Wu, G. Yu, Using an ARIMA-GARCH modeling approach to improve subway short-term ridership forecasting accounting for dynamic volatility, IEEE Transactions on Intelligent Transportation Systems. (2017).

[19] T. Chhorn, C. Chukiat, Modeling and forecasting tourist arrivals to Cambodia: an application of ARIMA-GARCH approach, Journal of Management, Economics and Industrial Organization. 2 (2) (2018) 1–19.

[20] M. Pahlavani, R. Roshan, The comparison among ARIMA and hybrid ARIMA-GARCH Models in forecasting the exchange rate of Iran, Intern. J. of Business and Development Studies. 7 (1) (2015) 31–50.

[21] G. Rote, Computing the minimum Hausdorff distance between two point sets on a line under translation, Information Processiing Letters. 38 (3) (1991) Elsevier: 123-27.

[22] J. P. Zeng, J. J. Duan, C. G. Wu, A new distance measure for Hidden Markov models, Expert Systems with Applications. 37 (2) (2010) Elsevier: 1550-5.

[23] X. Y. Wang, A. Mueen, H. Ding, G. Trajcevski, P. Scheuermann, E. Keogh, Experimental comparison of representation methods and distance measures for time series data, Data Mining and Knowledge Discovery. 26 (2) (2013) Springer: 275-309.

[24] J. C. Gower, A general coefficient of similarity and some of its properties, Biometrics. 27 (1971) 623-637.

[25] P. Arora, Deepali, S. Varshney, Analysis of K-Means and K-Medoids algorithm for big data, Procedia Computer Science. 78 (2016) 507-512.

[26] H. Ritchie, E. Mathieu, L. Rodés-Guirao, C. Appel, C. Giattino, E. Ortiz-Ospina, J. Hasell, B. Macdonald, D. Beltekian, M. Roser, Policy responses to the coronavirus pandemic - Statistics and Research, Our World in Data. (2020).

[27] A. Karlinsky, D. Kobak, Tracking excess mortality across countries during the COVID-19 pandemic with the World Mortality Dataset, eLife. (2021).

[28] Short-Term Mortality Fluctuation Data Series (STMF), Human Mortality Database University of California Berkeley (USA) and Max Planck Institute for Demographic Research (Germany), https://mpidr.shinyapps.io/stmortality/.

[29] Rosstat: https://rosstat.gov.ru/.

[30] H. Ritchie, E. Mathieu, L. Rodés-Guirao, C. Appel, C. Giattino, E. Ortiz-Ospina, J. Hasell, B. Macdonald, D. Beltekian, M. Roser, Coronavirus Pandemic (COVID-19), Published online at OurWorldInData.org (2020). Retrieved from https://ourworldindata.org/coronavirus.

[31] Legatum Institue, The Legatum Prosperity Index 2019–a tool for transformation. (2020).

[32] M. Kottek, J. Grieser, C. Beck, B. Rudolf, F. Rubel, World Map of the Köppen-Geiger climate classification updated, Meteorologische Zeitschrift. 15 (3) (2006) 259-263.

[33] F. Rubel, K. Brugger, K. Haslinger, I. Auer, The climate of European Alps: shift of very high resolution Köppen-Geiger climate zones 1800-2100, Meteorologische Zeitschrift. 26 (2) (2017) 115-125.

[34] T. Giorgino, Computing and visualizing Dynamic Time Warping alignments in R: the dtw package, Journal of Statistical Software. 31 (7) (2009) 1-24.

[35] M. Maechler, P. Rousseeuw, A. Struyf, M. Hubert, K. Hornik, cluster: Cluster Analysis Basics and Extensions, R package version 2.1.3 – For new features, see the 'Changelog' file (in the package source), https://CRAN.R-project.org/package=cluster, 2022.

[36] S. Fröhwirth-Schnatter, S. Kaufmann, Model-Based Clustering of Multiple Time Series, Journal of Bussiness and Economic Statistics. 26 (1) (2012) 78-89.

Acknowledgement

First and foremost, I would like to give my warmest thanks to my esteemed faculty sponsor Professor Maryam Gooyabadi. This presented work could not have been accomplished without her invaluable advice and continuous support through all stages of conducting and writing up my senior capstone project. My senior year back at Wesleyan after taking a gap year due to COVID-19 also would not have been as meaningful and enjoyable without having met Maryam from her Applied Time Series Analysis Course in Fall 2021, where the inspiration of this capstone project came from. Maryam has made my journey at Wesleyan much more fruitful and valuable. More importantly, without her genuine help, encouragement, love, and support, I definitely could not have had the opportunity to continue my graduate study at Stanford University in Fall 2022. Words alone can't not express how much Maryam means to me and how much I've learned from her sincerity, vision, and dynamism. Again, I would like to express my deepest appreciations to Maryam for her friendship, patience, understanding, and great sense of humor.

My sincere gratitude also goes to all other faculty members in the Wesleyan Quantitative Analysis Center, Professor Valerie Nazzaro, Professor Robert Kabacoff, Professor Jennifer Rose, Professor Pavel Oleinikov, and, most importantly, my dearest advisor Manolis, Professor Emmanuel Kaparakis, for their consistent support and guidance throughout my journey in the QAC since freshman year.

Lastly, I would also like to give my special thanks to my dearest friends in the QAC, Philippe Bungabong and Ori Cantwell, for motivating me to accomplish this paper by the stellar quality of work on their own capstone projects.