Understanding Economic and Health Factors Impacting the Spread of COVID-19 Disease

Aleksandr Farseev[a,c], Yu-Yi Chu-Farseeva[c], Qi Yang[a], Daron Benjamin Loo[b]

[a] *ITMO University*
[b] *National University of Singapore*
[c] *SoMin.ai Research*

Abstract

The rapid spread of the Coronavirus 2019 disease (COVID-19) had drastically impacted life all over the world. While some economies are actively recovering from this pestilence, others are experiencing fast and consistent disease spread, compelling governments to impose social distancing measures that have put a halt on routines, especially in densely-populated areas.

Aiming at bringing more light on key economic and public health factors affecting the disease spread, this initial study utilizes a quantitative statistical analysis based on the most recent publicly-available COVID-19 datasets.

The study had shown and explained multiple significant relationships between the COVID-19 data and other country-level statistics. We have also identified and statistically profiled four major country-level clusters with relation to different aspects of COVID-19 development and country-level economic and health indicators.

Specifically, this study has identified potential COVID-19 under-reporting traits as well as various economic factors that impact COVID-19 Diagnosis, Reporting, and Treatment. Based on the country clusters, we have also described the four disease development scenarios, which are tightly knit to country-level economic and public health factors. Finally, we have highlighted the potential

Email addresses: farseev@itmo.ru (Aleksandr Farseev), bolero168@gmail.com (Yu-Yi Chu-Farseeva), diuibyang2008@gmail.com (Qi Yang), elcdbl@nus.edu.sg (Daron Benjamin Loo)

limitation of reporting and measuring COVID-19 and provided recommendations on further in-depth quantitative research.

Keywords: COVID-19, Economic Factors, Public Health

2020 MSC: 00-01, 99-00

Introduction

The rapid spread of COVID-19 has drastically impacted economies around the world. On 11 March 2020 the disease was officially classified as a pandemic and, as reported on 24 March 2020, it has infected $440,093$, and causing $19,748$ deaths worldwide, with the highest new case intensities in the USA, Spain, Germany, France, Switzerland, South Korea, United Kingdom (UK), and Hubei Province in China.

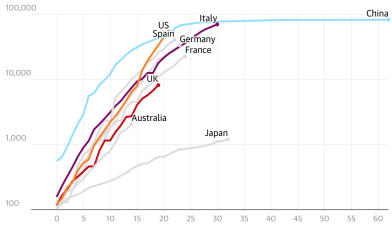


Figure 1: COVID-19 Confirmed Cases as on 24 March 2020

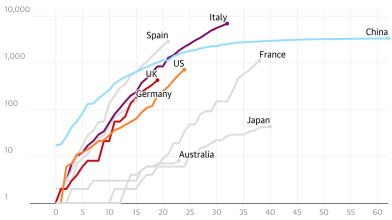


Figure 2: COVID-19 Registered Deaths as on 24 March 2020

In response to such a volatile situation in the world, governments and the scientific communities have been actively studying the underlying principles and possible reasons for the disease spread and progression. For example, Bai et.al. [1] have first discovered that COVID-19 could have been possibly transmitted by asymptomatic carriers, while Wu et.al. [2] conducted a large-scale study based on $72,314$ confirmed cases listing important actionable lessons for other societies to apply.

Furthermore, the Computer Science community has analyzed the disease spread from a statistical point of view. Specifically, in [3], the authors witnessed a potential association between COVID-19 mortality rates and health-care re-

source availability, while Chen et.al. [4] discovered a strong statistical relationship between initial emigration from Wuhan City and the infection spread to other cities in China. Finally, Chinazzi et.al. [5] suggested that travel restrictions to COVID-19 affected areas could be not as effective, as many infected individuals "...have been traveling internationally without being detected..." and as such, sharper restrictive measures are necessary to take control of the outbreak.

Even though significant efforts have been made so far towards a proper understanding of the COVID-19 outbreak from multiple perspectives, due to the constantly evolving pandemic, emerging new information and data sets, and inaccessibility of public large-scale data, literature based on quantitative research on the outbreak is still relatively sparse. To the best of our knowledge, this study is one of the first attempts to build a more holistic view on the COVID-19 pandemic, which hopes to explain relationships between the disease spread and various economic and public health factors through quantitative analysis.

Dataset

In this study, we have incorporated the "COVID19 Global Forecasting (Week 2)" dataset [6] that was released by the Kaggle[1] platform. The dataset includes daily updates of the COVID-19 confirmed cases and mortality rates for 173 countries reported by WHO between 22 January 2020 and 28 March 2020. To study the relationships between COVID-19 spread and various economic factors, we merged the the original dataset with "Country Statistics - UNData" dataset [7], "Pollution by Country for COVID19 Analysis" dataset [8], and "The World Bank (Demographics)" dataset by cross-matching country names across data sets. We also merged the original dataset with the dataset obtained by parsing the "World Life Expectancy" database [9] website for obtaining in-

[1] `https://www.kaggle.com/`

Table 1: Detailed Statistics of the COVID-19 Combined Dataset

Data Indicator Group	Number
Countries	165
Regions	286
Chronic Disease Death Rate Statistics	32
Age Demographic Groups	4
Pollution Indicators	3
Other Economic Factor Statistics	50
COVID-19 Related Indicators	2
COVID-19 Confirmed Case Speed Daily Reports	67
COVID-19 Fatalities Speed Daily Reports	67

formation on death rates from different chronic diseases across the world.

After the merging process, the resulting dataset consists of COVID-19 Case and Mortality Rates, Economic Statistics, and Public Health Statistics for 165 countries[2] reported between 22 January 2020 and 28 March 2020. A more detailed statistics of the resulting dataset are provided in the table 1. We have also released the dataset for public use [10].

Experimental Setup

As mentioned, the primary objective of this research is to study the relationship between the **speed** of the disease spread and various economic and health factors. Considering the uneven pace of the disease's geographical spread

[2]The actual number of the data records in the dataset is 286 as for some countries the COVID-19 statistics in the original data set was given for different regions within the same country: 54 regions in the United States of America, 33 regions in China, 10 regions in Canada, 10 regions in France, 8 regions in Australia, 7 regions in the United Kingdom, 4 regions in the Netherlands, and 3 regions in Denmark.

due to COVID-19's long incubation period [11], natural migration laws [5] and various government-imposed travel policies [5], it is not feasible to draw the analysis based on actual daily registered case and fatality rates available in the original Kaggle dataset [6], but rather necessary to perform an additional data pre-processing aiming at establishing holistic data characteristics reflecting the general worldwide COVID-19 spread tendencies. Keeping this in mind, we have performed the following data pre-processing steps:

- **Dataset Combination:** Original Dataset [6] was joined by performing Country matching to four auxiliary data sets [7, 8, 12, 9] as described in the next sections. Fifteen country names have been replaced with the naming notation used in the original dataset to perform the successful matching.

- **Normalized Daily Spread Speed Estimation:** In this study, we analyzed the last two weeks of reported data from 14 March 2020 till 28 March 2020. To estimate the disease spread speed of each day in the two-week interval, we subtracted the reported number of new cases and fatalities on the previous day from the number of the current day and then divided this number to the Median reported number during the past two weeks.

  $$Speed(day_k) = \frac{Reported(day_k) - Reported(day_{k-1})}{Mean(Reported(day_1)...Reported(day_{14}))}, k = 1...14$$

  , where $Reported()$ is the number of Confirmed Cases or Fatalities reported in the dataset, and $Mean()$ is the Arithmetic Mean of its arguments. The above normalization procedure mitigates the problem of uneven speed of spread of COVID-19 in different geographical regions as it treats each country according its outbreak "stage" and makes country statistics comparable to each other.

- **Sparse Data Indicator Filtering:** As some of the data indicators in the merged dataset were found to contain a large number of missing values, which might affect further analysis, we excluded data indicators that con-

tained more than 35% of missing values. After the sparse data indicator exclusion, the resulting dataset contained 116 data indicators.

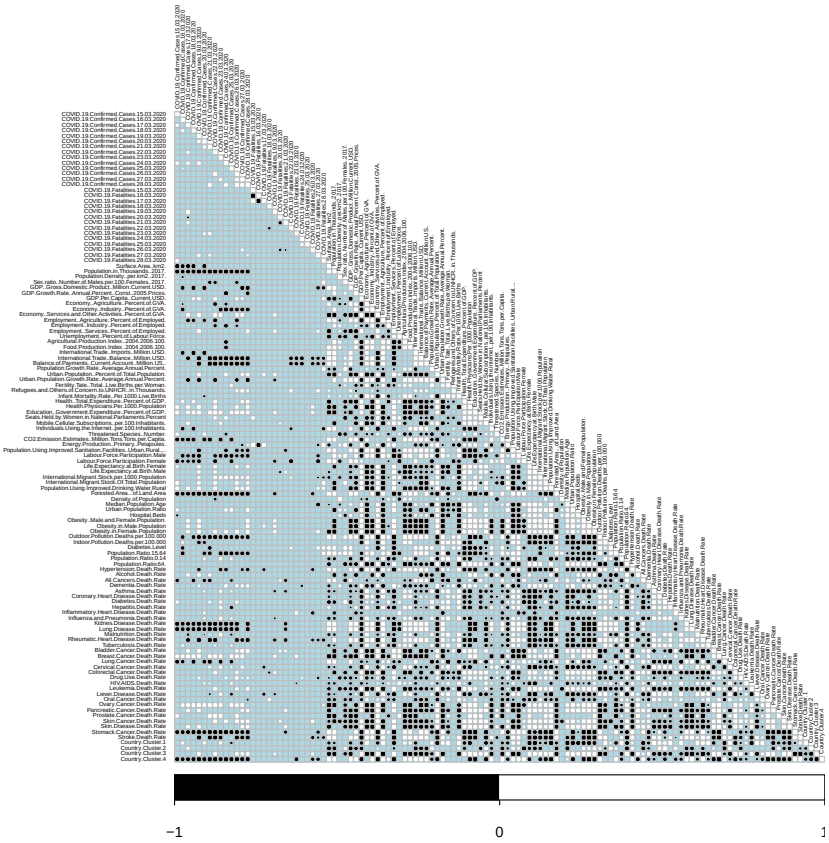
Statistical Data Analysis

To determine the relationship between the COVID-19 Spread Speed and other indicators, we applied Pearson Product-Moment Correlation [13] to 286 data samples (the countries and regions in the combined dataset) and 116 data indicators (whose data indicators that have remained after the Sparse Data Indicator Filtering step). We then filtered out all non-significant correlation values ($\alpha = 0.05$) and presented the obtained results in a form of a correlation semi-matrix (see Figure 3). On the Figure, white circles denote a positive correlation, while the black circles mean that the correlation is negative. The size of the circle is proportional to the correlation strength (the larger circle - the stronger the correlation) and the absence of a circle in a cell means that there was no correlation found or the correlation is not significant.

0.1. COVID-19 Confirmed Cases and Fatalities

Let's take a look at the first 28 lines of the correlation semi-matrix. From the correlation plot, it can be seen that there are several strong correlations between individual COVID-19 reported statistics. For example, the negative correlation between the Fatality Speed on March 15 (Sunday) and March 16 (Monday) as well as the strong positive correlation between Fatality Speed on March 15 (Sunday) and March 17 (Tuesday) could possibly suggest a reporting time-lag on weekends. At the same time, a significant positive correlation was also found between several subsequent Confirmed Case Speed dates, such as 19 (Thursday), 20 (Friday), and 21 (Saturday) March 2020 and 26 (Thursday), 27 (Friday), and 28 (Saturday) March 2020. Such correlation sequences towards the end of the week could possibly be explained by the testing capacity of the hospitals entailing the situation when test results from the beginning of the week were received only towards the end of the week. As this study does not aim at a

Figure 3: Pearson Product-Moment Correlation Between COVID-19 Spread Speed, Economics, and Health Factors.



detailed analysis of the longitudinal properties of the COVID-19 measurement and test procedures, in this work, we would only like to highlight the importance and the influence of time-related measurement and testing arrangement aspects as well as to recommend future research along this direction.

Despite several single negative or positive correlations mentioned above, one might find many significant correlations neither between consequent measurements of the same metric (i.e. Confirmed Case Speed on Different Days) nor between different COVID-19 metrics (i.e. Confirmed Case v.s. Fatalities). These suggest that overall there is NO strong relation between Confirmed Case Speed and Fatality Speed within the 14 day-interval and, therefore, additional information sources, such as economic and health data indicators, are necessary for gaining a deeper insight into the COVID-19 Spread Speed trends.

0.2. Chronic Diseases and Health Factors

Let's now move our attention to the last 39 data indicators in the lower part of the correlation semi-matrix where the Chronic Disease Death Rates and other Public Health-Related factor correlations are displayed. From the plot, it can be seen that there are multiple significant and consistent correlations that can be found mainly for Confirmed Case Speed measurements. For example, such indicators as **Skin Cancer** (91.7% 5-year survival rate [14]), **Prostate Cancer** (98.6% 5-year survival rate [14]), **Ovary Cancer** (46.5% 5-year survival rate [14]), **Breast Cancer** (89.7% 5-year survival rate [14]), and **Bladder Cancer** (77.3% 5-year survival rate [14]) Death Rates were found to be **significantly positively correlated** with the COVID-19 Spread Speed.

In order to explain the above relationship, it is necessary to consider the factors related to country-level chronic disease data indicators. First, taking into consideration the corresponding 5-year survival rates (indicated in the brackets), one can notice that most of the diseases in the group are the cancers with a high general chance of patient survival (let's call them "high-survival cancers"). Generally, it is reasonable to assume that the countries exhibiting higher death rates for such "high-survival cancers" might experience overall difficulties in

proper and timely patient treatment. When facing COVID-19 pandemic, such countries might not be always well prepared for proper patient isolation and treatment as well, which is essential for COVID-19 disease spread control [15]. Correspondingly, in such countries, the COVID-19 Spread Speed could be higher entailing the above-reported significant positive correlation.

Supporting the above findings, we would also like to highlight the **strong positive correlation** of **Obesity Rates (especially for Female demographics)** to COVID-19 Spread Speed: similarly to Wells et. al. [16], our correlation semi-matrix visualisation uncovers that obesity is tightly knit to the countries' Gross Domestic Product (GDP[3]) rates. Consequently, for the countries with lower GDP, the higher COVID-19 spread rates could be attributed to the poorer readiness of these countries to the COVID-19-associated risks.

At last, let's consider an alternative explanation of the reported correlations. From the correlation semi-matrix, one can find that such variables as **Skin Disease Death Rate**, **Influenza and Pneumonia Death Rate**, **Diabetes Death Rate**, **Dementia Death Rate**, **Alcohol Death Rate** are also significantly correlated to COVID-19 Spread Speed. At the same time, it was also reported [17] that COVID-19 development and consequences might be directly related to the overall health status of the population, especially with regards to the existing pre-conditions affecting the human immune system. Such pre-conditions could further entail various fatal complications, such as Cytokine Storm, and, ultimately, affect the countries' COVID-19 Confirmed Case Speed and Fatality Speed statistics. In other words, it could be possible that the significant correlations between chronic diseases death rates and COVID-19 Spread Speed reflect the overall country population health status and, therefore, predisposition to infection and complication of COVID-19.

Regardless of the actual reason behind the discovered relationships, it is likely that more developed economies are able to respond to COVID-19 out-

[3]Gross domestic product (GDP) is a monetary measure of the market value of all the final goods and services produced in a specific time period.

break better as compared to the less developed ones. At the same time, such developed world populations might be more affected by unhealthy life habits and various other urban-living factors [18, 19], which, in turn, would potentially entail higher chronic disease rates and, correspondingly, COVID-19 health predispositions. While in this study we are only witnessing existing relationships, further research is necessary in order to make more conclusive observations regarding the cause for these relationships.

0.3. Economic and Other Factors

In the previous Section, we discovered that multiple economic factors exhibit a strong relationship to the chronic diseases across the globe, and can be used to characterise the profiles of these countries with relation to the economic development stage, and ultimately, COVID-19 Spread Speed.

To gain further insight into the relationship between such economic factors and the COVID-19 disease spread, let's now discuss the actual correlations between the two groups of variables. Precisely, from the correlation semi-matrix, it can be seen that six economic attributes are strongly positively correlated to the COVID-19 Spread Speed. Below, we will discuss the possible reasons for the discovered relationships.

Number of Health Physicians Per 1000 Population, **Health Total Expense (% of GDP)**, and **GDP Per Capita (in USD)** strong **positive correlations** could be attributed to the higher ability of the countries with stronger health systems in performing timely patient assessment, diagnosis, and reporting of the disease. In contrast, countries with weakly-subsidised health systems, many, especially asymptomatic [20], **COVID-19 cases could remain unreported** bringing the COVID-19 confirmed case statistics down and entailing the inverse correlation traits that we have discovered from the dataset.

In the cases of **International Migrant Stock Per 1000 Population** and **International Migrant Stock % of Total Population** variables, the discovered strong positive correlations to COVID-19 Spread Speed could be possibly explained by the higher rates of imported cases in countries with larger

proportions of the migrant population who often travel abroad for business and personal purposes. Interestingly, both variables exhibit a strong positive correlation during the second week of the observed period (22 March 2020 - 28 March 2020), which could be potentially explained by the travel restrictions imposed by the governments during that week, resulting in the situation when many migrants were rushing to return back to their countries of residence prior to border closures [21, 22].

Finally, the strong positive correlation of the **Services and Other Activities % of Gross Value Added (GVA[4])** measurement can be attributed to the more intense human interaction rates in countries with larger population involved in the service sector of economics, making the risk of COVID-19 infection higher [23].

0.4. Negative and Insignificant Correlations, Mortality Rate Speed

Last but not least, let's discuss several other relationships that can be observed from the correlation plot.

For example, an interesting observation is the strong **negative correlation** of the **Population in Thousands 2017** variable to the COVID-19 Spread Speed. At the same time, one can notice that the **Population Density Per km2 in 2017** does not exhibit any significant correlation with COVID-19. A possible conclusion from these two findings could be that the population size and density criteria might not be the key factors impacting the COVID-19 Spread Speed. As it was previously reported in the literature [24], and was also observed in this current study, the cultural and behavioural factors, such as human interaction habits, or government-regulated factors, such as social distancing rules, could be of a much higher influence on the ability of the country government to manage the COVID-19 disease outbreak.

At the same time, the reader could observe that such variables as **CO2**

[4]In economics, gross value added (GVA) is the measure of the value of goods and services produced in an area, industry or sector of an economy.

Emission Estimates (Million Tonnes Per Capita) and **Forested Area Ratio** also exhibit a strong **negative correlation** to the COVID-19 Spread Speed. While the latter can be easily hypothesized by the natural geographical sparsity of population introduced by the forested landscape and entailing a limited inter-human interaction, the former two cases require additional clarifications. For example, one can further observe that, evidently, **CO2 Emission Estimates** metric is strongly **positively correlated** to the **Lung Cancer Death Rate**, which, in turn, is also **negatively correlated** to the **COVID-19 Spread Speed**. Taking into consideration that the two metrics might not be related directly to the disease spread speed (as COVID-19 disease gets "...transmitted between people through close contact and droplets" [25] and, thus, more depends on the inter-human close contacts), it is then reasonable to assume that, correspondingly, the two variables could also be positively correlated to the COVID-19 Fatality Speed as we could expect more patients with lung preconditions in the countries with more polluted environments. However, such a relationship could not be observed from the data we have, except for the **only one negative correlation** of the **Lung Disease Death Rate** to the **COVID-19 Fatality Speed on 23 March 2020**. The high sparsity of the correlation semi-matrix in the area of the COVID-19 Fatality Speed might possibly suggest that, at the moment, the currently-available COVID-19 data on Fatality Rate might not be sufficient for making conclusive observations regarding the relationships with all economic and health factors. Therefore, we recommend further research in this direction based on extended data collections.

Finally, we would like to highlight the **International Trade Balance (Million USD)** and **Balance of Payment Current Account (Million USD)** are strongly **negatively correlated to the COVID-19 Fatality Speed** during the second week of the observed period (22 March 2020 - 28 March 2020). As the above two indicators describe countries' economics strength [26], the observed relationship can be easily explained as reflecting the countries' ability to manage the spiking of COVID-19 disease outbreak: countries with stronger economies and medical equipment reserve might be able to provide patients

with necessary care, as compared to the economies experiencing a shortage of resources.

Cluster Analysis

In the previous sections, we have determined multiple economic and public health factors that are strongly and significantly correlated to the COVID-19 disease spread. We have also witnessed various governments and population behavioural traits possibly explaining the different scenarios of COVID-19 development around the world.

Even though the above-discovered findings bring more light into the approaches that governments have adopted to mitigate the crisis, it is still unclear what the exact differences are in these approaches as well as in the country profiles affected by the COVID-19 disease spread.

Aiming at answering this second research question of the study, we have adopted a Clustering Analysis technique [27] to study the groups of countries in our dataset that could be found by separating it based on the economic and public health factors. Specifically, we have adopted the "X-Means" clustering algorithm, that have been reported to be effective in determining the number of clusters in the dataset without necessarily having a prior assumption on the number of clusters [28]. The X-Means clustering was applied to the whole dataset and have determined the four country clusters listed below:

- **Country Cluster 1:** AFGHANISTAN, ANGOLA, BANGLADESH, BENIN, BHUTAN, BURKINA FASO, CAMBODIA, CAMEROON, CENTRAL AFRICAN REPUBLIC, CHAD, COTE D'IVOIRE, DJIBOUTI, EQUATORIAL GUINEA, ERITREA, ETHIOPIA, GABON, GHANA, GUINEA, GUINEA-BISSAU, HAITI, INDIA, INDONESIA, KENYA, LIBERIA, MADAGASCAR, MALI, MAURITANIA, MOZAMBIQUE, NAMIBIA, NEPAL, NIGER, NIGERIA, PAKISTAN, PHILIPPINES, RWANDA, SENEGAL, SOMALIA, SRI LANKA, SUDAN, TANZANIA, TIMOR-LESTE, TOGO, UGANDA, ZAMBIA, ZIMBABWE.

- **Country Cluster 2:** ALBANIA, ALGERIA, ANDORRA, ANTIGUA AND BARBUDA, ARGENTINA, ARMENIA, AZERBAIJAN, BAHRAIN, BARBADOS, BELARUS, BELIZE, BOLIVIA, BOSNIA AND HERZEGOVINA, BRAZIL, BRUNEI, BULGARIA, CABO VERDE, CHILE, COLOMBIA, COSTA RICA, CUBA, CYPRUS, DOMINICA, DOMINICAN REPUBLIC, ECUADOR, EGYPT, EL SALVADOR, FIJI, GEORGIA, GRENADA, GUATEMALA, GUYANA, HOLY SEE, HONDURAS, IRAN, IRAQ, JAMAICA, JORDAN, KAZAKHSTAN, SOUTH KOREA, KUWAIT, KYRGYZSTAN, LAOS, LEBANON, LIBYA, LIECHTENSTEIN, MALAYSIA, MALDIVES, MAURITIUS, MEXICO, MOLDOVA, MONGOLIA, MONTENEGRO, MOROCCO, NICARAGUA, OMAN, PANAMA, PAPUA NEW GUINEA, PARAGUAY, PERU, QATAR, ROMANIA, SAINT KITTS AND NEVIS, SAINT LUCIA, SAINT VINCENT AND THE GRENADINES, SAN MARINO, SAUDI ARABIA, SERBIA, SEYCHELLES, SINGAPORE, SOUTH AFRICA, SURINAME, SYRIA, THAILAND, TRINIDAD AND TOBAGO, TUNISIA, TURKEY, UKRAINE, UNITED ARAB EMIRATES, UZBEKISTAN, VENEZUELA;

- **Country Cluster 3:** AUSTRALIA (8 REGIONS), AUSTRIA, BELGIUM, CANADA (10 REGIONS), CROATIA, CZECHIA, DENMARK, DENMARK, DENMARK, ESTONIA, FINLAND, FRANCE (10 REGIONS), GERMANY, GREECE, HUNGARY, ICELAND, IRELAND, ISRAEL, ITALY, JAPAN, LATVIA, LITHUANIA, LUXEMBOURG, MALTA, MONACO, NETHERLANDS (4 REGIONS), NEW ZEALAND, NORTH MACEDONIA, NORWAY, POLAND, PORTUGAL, RUSSIA, SLOVAKIA, SLOVENIA, SPAIN, SWEDEN, SWITZERLAND, UNITED KINGDOM (7 REGIONS), URUGUAY, US (54 REGIONS).

- **Country Cluster 4:** CHINA (33 REGIONS).

We then adopted correlation analysis as it was described in Section c to uncover the statistical profiles for each of the clusters. Our findings are discussed immediately next.

0.5. Country Cluster 1 Correlations

From the correlation semi-matrix, it can be seen that the countries from the Cluster 1 are positively correlated to the COVID-19 Confirmed Case Speed on 16, 20, and 23 March 2020, while negatively correlated on 25 March 2020. It can also be noticed that these economies heavily rely on agriculture[5], have high fertility and infant death rates[6], skewed towards younger population[7], and have higher death rates from indoor pollution as well as such well-treatable chronic diseases and cancers[8]. All the above correlations could characterize the countries from Cluster 1 as belonging to the category of developing countries, which could also be observed from the country participant list provided in the previous section. Therefore, the non-consistent correlations with the COVID-19 Confirmed Case Speed (three significant positive correlations and one significant negative correlation), could then be explained by possible testing and reporting issues that frequently occur when facing world-scale disease outbreaks [29]. Given the limited available data in our COVID-19 dataset regarding COVID-19 reporting procedures in different countries, in this work, we would like to highlight a

[5]See significant **positive correlations** of **Economy Agriculture Percent of GVA**, **Employment Agriculture Percent of Employed**, **Agricultural Production Index 2004-2006** and **Food Production Index 2004-2006** variables to countries in Cluster 1

[6]See significant **positive correlations** of **Fertility Rate % of Total Live Births Per Woman** and **Infant Mortality Rate Per 1000 Live Births** variables to countries in Cluster 1

[7]See significant **positive correlation** of **Population Ratio 0 -14 years old** variable to countries in Cluster 1

[8]See significant **positive correlations** of **Indoor Pollution Deaths Per 100,000**, **Asthma Death Rate**, **Hepatitis Death Rate**, **Influenza.and.Pneumonia.Death.Rate**, **Malnutrition Death Rate**, **Tuberculosis Death Rate**, **Oral Cancer Death Rate**, and **Ovary Cancer Death Rate** variables to countries in Cluster 1

possible under-reporting issue for the developing world and, consequently, suggest further in-depth research towards COVID-19 spread characteristics in the developing countries.

0.6. Country Cluster 2 Correlations

When looking at the correlation profile of the Country Cluster 2, a reader could immediately notice that the cluster is not associated with any significant COVID-19 correlations except for one positive correlation with COVID-19 Spread Speed on 17 March 2020. Furthermore, one can also find that other positive correlations of the cluster are arguably weak, having its spikes in population growth[9], Obesity[10] and Diabetes[11], various heart-related diseases[12], and reproduction system cancers[13].

From the observed relationships, we can clearly acknowledge the existence of a cluster consisting of countries with population overweight, and correspondingly, heart [30] and reproductive cancer problems [31]. As the countries in the cluster do not exhibit significant correlation to COVID-19-related data indicators and, therefore, are out of focus in this article, in this work, we would not be further elaborating on such an interesting finding. Having said that, we would like to bring the readers' attention to such an interesting relationship, which could be guide future research direction.

[9] See significant **positive correlation** of **Population Growth Rate Average Annual Percent** variable to countries in Cluster 2

[10] See significant **positive correlations** of **Obesity in Female Population** and **Obesity in Male Population** variables to countries in Cluster 2

[11] See significant **positive correlations** of **Diabetes Level** and **Diabetes Death Rate** variables to countries in Cluster 2

[12] See significant **positive correlations** of **Coronary Heart Disease Death Rate** and **Inflammatory Heart Disease Death Rate** variables to countries in Cluster 2

[13] See significant **positive correlations** of **Cervical Cancer Death Rate** and **Prostate Cancer Death Rate** variables to countries in Cluster 2

0.7. Country Cluster 3 Correlations

Country Cluster 3 is the largest and also the most diverse cluster that we have discovered as it includes most of the European Countries and all states of US that have experienced spikes in COVID-19 cases over the past several weeks[14]. So what are the factors uniting all these countries except for the high speed of COVID-19 disease spread?

From the correlation semi-matrix it can be seen that the countries from the cluster are significantly positively correlated to multiple factors associated with typical modern developed economies, such as higher GDP Rates[15], the involvement of the population in Services Industry[16], high ratio of urban population[17], high health system maturity[18], and solid international migrant stocks[19].

At the same time, it can also be seen that countries from Cluster 3 exhibit a strong positive correlation with population aging and its-associated diseases[20], various types of cancers[21] and, correspondingly, urban population-linked chronic

[14] See **multiple significant positive correlations with COVID-19 spread speed and fatalities** on the correlation plot

[15] See significant **positive correlation** of **GDP Per Capita in USD** variable to countries in Cluster 3

[16] See significant **positive correlations** of **Economy Services and Other Activities Percent of GVA** and **Employment Services Percent of Employed** variables to countries in Cluster 3

[17] See significant **positive correlation** of **Urban Population Percent of Total Population** variable to countries in Cluster 3

[18] See significant **positive correlations** of **Health Total Expenditure Percent of GDP**, **Health Physicians Per 1000 Population**, **Life Expectancy at Birth Female**, and **Life Expectancy at Birth Male** variables to countries in Cluster 3

[19] See significant **positive correlations** of **International Migrant Stock Per 1000 Population** and **International Migrant Stock Of Total Population** variables to countries in Cluster 3

[20] See significant **positive correlations** of **Median Population Age**, **Population Ratio 64+ Years Old**, and **Dementia Death Rate** variables to countries in Cluster 3

[21] See significant **positive correlations** of **Bladder Cancer Death Rate**, **Breast Cancer Death Rate**, **Colo-rectal Cancer Death Rate**, **Leukemia Death Rate**, **Ovary Cancer Death Rate**, **Pancreatic Cancer Death Rate**, and **Skin Cancer Death Rate** variables to countries in Cluster 3

diseases[22] .

Taking into consideration both the above-described traits, we could further hypothesise that the populations in countries from Cluster 3 might be also initially predisposed to COVID-19 infection. The latter assumption raises from the two known COVID-19 risk factors that are also to be found related to the countries from Cluster 3, namely older population demographics [32] and existing pre-conditions that could lead to, for example, Cytokine Storm [17] or other highly-lethal COVID-19 complications.

At last, we would like to mention that, based on our findings in the previous section, it is also reasonable to assume that more developed countries might be able to diagnose and report COVID-19 cases timely and at a necessary scale as compared to some developing economies, that, ultimately, may lead to the strong correlations of such countries to COVID-19 Spread Speed in our dataset.

0.8. Country Cluster 4 Correlations

As China is the only country in the Cluster 4 and its economic, population, and, for example, pollution statistics are commonly known, we will omit some strongly positively-correlated indicators in the data commentary.

At the same time, we would like to bring the readers' attention to the possible bias in some of the conclusions that we have drawn from our data. For example, it is easy to notice that such variable as **Lung Disease Death Rate**, **Stomach Cancer Death Rate**, **Malnutrition Death Rate Rheumatic**, and **Heart Disease Death Rate**, are strongly **positively correlated to Country Cluster 4**. At the same time, these variables are also strongly **negatively correlated to the COVID-19 Spread Speed**. By drawing a parallel between these two observations, readers then could conclude that data from China might affect the overall analysis results in relation to cases of China-

[22] See significant **positive correlations** of **Obesity Male and Female Population**, **Obesity in Female Population**, **Obesity in Male Population**, **Drug Use Death Rate**, **Skin Disease Death Rate**, and **Alcohol Death Rate** variables to countries in Cluster 3

specific population chronic diseases and the shift of the disease development timeline between China and other countries.

Last but not least, we would also like to highlight the importance of the proper alignment and synchronization of the data that comes from the regions with large territories and specific disease development timelines.

Conclusions

In this work, we have performed a preliminary statistical analysis aiming at understanding the relationship between various economic and public health factors and COVID-19 disease spread cadence. The study had shown and explained multiple significant relationships between the COVID-19 data and other country-level statistics. We have also identified and statistically profiled four major country-level clusters with relation to different aspects of COVID-19 development and country-level economic and health statistics. Finally, we have highlighted the limitations of our adopted data and approach, encouraging further larger-scale research along the direction.

In future works, we are aiming at establishing automotive Machine Learning and Statistical frameworks, that would be attempting to predict the future development of COVID-19 disease based on our COVID-19 dataset. We will be also extending the dataset with more dynamic and comprehensive data, such as medical resource availability, government-imposed control measure, and culture-related aspects.

Declaration of Interests

We declare no competing interests.

Declaration of Author Contributions

Aleksandr Farseev and Yu-Yi Chu-Farseeva conceived of the presented idea. Aleksandr Farseev and Qi Yang developed the theory and performed the data

analysis. Yu-Yi Chu-Farseeva and Daron Benjamin Loo verified the analytical methods. Yu-Yi Chu-Farseeva encouraged Aleksandr Farseev to investigate public health-related aspects of the COVID-19 Disease Spread and Fatalities Speed and supervised the findings of this work. All authors discussed the results and contributed to the final manuscript.

References

[1] Y. Bai, L. Yao, T. Wei, F. Tian, D.-Y. Jin, L. Chen, M. Wang, Presumed asymptomatic carrier transmission of covid-19, Jama.

[2] Z. Wu, J. M. McGoogan, Characteristics of and important lessons from the coronavirus disease 2019 (covid-19) outbreak in china: summary of a report of 72 314 cases from the chinese center for disease control and prevention, Jama.

[3] Y. Ji, Z. Ma, M. P. Peppelenbosch, Q. Pan, Potential association between covid-19 mortality and health-care resource availability, The Lancet Global Health.

[4] Z. Chen, Q. Zhang, Y. Lu, Z. Guo, X. Zhang, W. Zhang, C. Guo, C. Liao, Q. Li, X. Han, et al., Distribution of the covid-19 epidemic and correlation with population emigration from wuhan, china., Chinese medical journal.

[5] M. Chinazzi, J. T. Davis, M. Ajelli, C. Gioannini, M. Litvinova, S. Merler, A. P. y Piontti, K. Mu, L. Rossi, K. Sun, et al., The effect of travel restrictions on the spread of the 2019 novel coronavirus (covid-19) outbreak, Science.

[6] Covid19 global forecasting (week 2): orecast daily covid-19 spread in regions around world, `https://www.kaggle.com/c/covid19-global-forecasting-week-2/data`, accessed: 2020-03-29.

[7] Country statistics - undata dataset, `https://www.kaggle.com/sudalairajkumar/undata-country-profiles`, accessed: 2020-03-29.

[8] Pollution by country for covid19 analysist, https://www.kaggle.com/brandonhoeksema/pollution-by-country-for-covid19-analysis, accessed: 2020-03-29.

[9] World life expectancy database, https://www.worldlifeexpectancy.com/, accessed: 2020-03-29.

[10] Covid-19 combined dataset (somin.ai), https://covid19.somin.ai/, accessed: 2020-03-29.

[11] S. A. Lauer, K. H. Grantz, Q. Bi, F. K. Jones, Q. Zheng, H. R. Meredith, A. S. Azman, N. G. Reich, J. Lessler, The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: estimation and application, Annals of internal medicine.

[12] The world bank data, https://data.worldbank.org/indicator/SP.POP.0014.TO.ZS, accessed: 2020-03-29.

[13] D. Freedman, R. Pisani, R. Purves, Statistics (international student edition), Pisani, R. Purves, 4th edn. WW Norton & Company, New York.

[14] Eer cancer statistics review, 1975-2013, https://seer.cancer.gov/archive/csr/1975_2013/, accessed: 2020-03-30.

[15] J. Hellewell, S. Abbott, A. Gimma, N. I. Bosse, C. I. Jarvis, T. W. Russell, J. D. Munday, A. J. Kucharski, W. J. Edmunds, F. Sun, et al., Feasibility of controlling covid-19 outbreaks by isolation of cases and contacts, The Lancet Global Health.

[16] J. C. Wells, A. A. Marphatia, T. J. Cole, D. McCoy, Associations of economic and gender inequality with global obesity prevalence: understanding the female excess, Social science & medicine 75 (3) (2012) 482–490.

[17] P. Mehta, D. F. McAuley, M. Brown, E. Sanchez, R. S. Tattersall, J. J. Manson, Covid-19: consider cytokine storm syndromes and immunosuppression, The Lancet.

[18] V. L. Feigin, C. M. Lawes, D. A. Bennett, C. S. Anderson, Stroke epidemiology: a review of population-based studies of incidence, prevalence, and case-fatality in the late 20th century, The Lancet Neurology 2 (1) (2003) 43–53.

[19] M. Pereira, N. Lunet, A. Azevedo, H. Barros, Differences in prevalence, awareness, treatment and control of hypertension between developing and developed countries, Journal of hypertension 27 (5) (2009) 963–975.

[20] Z. Hu, C. Song, C. Xu, G. Jin, Y. Chen, X. Xu, H. Ma, W. Chen, Y. Lin, Y. Zheng, et al., Clinical characteristics of 24 asymptomatic infections with covid-19 screened among close contacts in nanjing, china, Science China Life Sciences (2020) 1–6.

[21] Coronavirus: Travel restrictions, border shutdowns by country, https://www.aljazeera.com/news/2020/03/coronavirus-travel-restrictions-border-shutdowns-country-200318091505922.html, accessed: 2020-04-04.

[22] W. H. Organization, et al., Coronavirus disease 2019 (covid-19): situation report, 67.

[23] J. Hilton, M. J. Keeling, Estimation of country-level basic reproductive ratios for novel coronavirus (covid-19) using synthetic contact matrices, medRxiv.

[24] S. Merler, M. Ajelli, L. Fumanelli, M. F. Gomes, A. P. y Piontti, L. Rossi, D. L. Chao, I. M. Longini Jr, M. E. Halloran, A. Vespignani, Spatiotemporal spread of the 2014 outbreak of ebola virus disease in liberia and the effectiveness of non-pharmaceutical interventions: a computational modelling analysis, The Lancet Infectious Diseases 15 (2) (2015) 204–211.

[25] W. H. Organization, et al., Rational use of personal protective equipment for coronavirus disease (covid-19): interim guidance, 27 february 2020, Tech. rep., World Health Organization (2020).

[26] T. Beck, Financial development and international trade: Is there a link?, Journal of international Economics 57 (1) (2002) 107–131.

[27] X. Wu, S. Hu, A. B. Kwaku, Q. Li, K. Luo, Y. Zhou, H. Tan, Spatio-temporal clustering analysis and its determinants of hand, foot and mouth disease in hunan, china, 2009–2015, BMC infectious diseases 17 (1) (2017) 645.

[28] D. Pelleg, A. W. Moore, et al., X-means: Extending k-means with efficient estimation of the number of clusters., in: Icml, Vol. 1, 2000, pp. 727–734.

[29] L. Bakalikwira, J. Bananuka, T. Kaawaase Kigongo, D. Musimenta, V. Mukyala, Accountability in the public health care systems: A developing economy perspective, Cogent Business & Management 4 (1) (2017) 1334995.

[30] A. Keys, Obesity and heart disease, Journal of Chronic Diseases 1 (4) (1955) 456–461.

[31] N. M. Maruthur, S. D. Bolen, F. L. Brancati, J. M. Clark, The association of obesity and cervical cancer screening: a systematic review and meta-analysis, Obesity 17 (2) (2009) 375–381.

[32] Q. Ruan, K. Yang, W. Wang, L. Jiang, J. Song, Clinical predictors of mortality due to covid-19 based on an analysis of data of 150 patients from wuhan, china, Intensive care medicine (2020) 1–3.