Thesis Manuscript Jan 8 - Current Progress and Next Steps

alexe314

January 2021

Contents

1 Methods

In this section, the methodology of the research will be defined. First, the data sources will be described. Then the data preparation and data smoothing techniques will be explained. The data interpolation algorithms will be explained. Finally, the metric generation process and corresponding metrics will be defined.

1.1 Data Sources

In this section, the data sources used in this research will be described.

1.1.1 COVID-19 Incident Data

The COVID-19 incident data and be broken down into three different categories: cases, deaths, and tests. The COVID-19 case and death data is obtained from the *Johns Hopkins Whiting School of Engineering's Center for Systems Science and Engineering's* Github page [Citation Needed]. The data is formatted as a cumulative time series, contains data on a country and state/provincial level, and is updated on a daily basis. The COVID-19 test data was obtained from the *Our World in Data's* Github page [Citation Needed]. The data is formatted as both a daily and cumulative time series, contains data on a country level, and is updated twice weekly.

1.1.2 Global Economic Data

The economic data is obtained from the *International Monetary Fund's 2020 World Economic Outlook Database* [Citation Needed]. From this data, three key metrics are selected to be used in this research: general government net lending/borrowing (percent of GDP), GDP (current prices in purchasing power parity), unemployment rate (percent of total labor force). The unemployment rate is not available for every country.

1.1.3 Government Action Data

The government action data is obtained from the *The University of Oxford's COVID-19 Stringency Index* Github page [Citation Needed]. For this data "the data from 19 indicators is aggregated into a set of four common indices, reporting a number between 1 and 100 to reflect the level of government action" [Citation Needed]. The four common indices are:

- **Overall Government Response**: "records how the response of governments has varied over all indicators in the database, becoming stronger or weaker over the course of the outbreak"

- **Containment and Health**: “combines ’lockdown’ restrictions and closures with measures such as testing policy and contact tracing, short term investment in healthcare, as well investments in vaccine”
- **Economic Support**: “records measures such as income support and debt relief”
- **Stringency**: “records the strictness of ‘lockdown style’ policies that primarily restrict people’s behaviour”

1.1.4 Healthcare Performance Data

The healthcare performance data is obtained from a *World Health Organization* report titled *Measuring Overall health System Performance for 191 Countries* [Citation Needed]. This report generated an index score for the quality of every country’s healthcare system in terms of the quality of level, equity of distribution, efficiency of health, responsiveness, and fairness in financing.

1.1.5 Pandemic Preparedness Data

The pandemic preparedness data is obtained from *Johns Hopkin’s 2019 Global Health Security Rankings* [Citation Needed]. It ranks 195 country’s pandemic preparedness across six different dimensions as well as providing an overall score. In addition to the quantitative ranking, a categorical score is given to every country across each dimension and in the overall. The six dimensions are:

- **Prevention**: “prevention of the emergence or release of pathogens”
- **Detection and Reporting**: “early detection and reporting for epidemics of potential international concern”
- **Rapid Response**: “rapid response to and mitigation of the spread of an epidemic”
- **Health System**: “sufficient and robust health system to treat the sick and protect health workers”
- **Compliance with International Norms**: “commitments to improving national capacity, financing plans to address gaps, and adhering to global norms”
- **Risk Environment**: “overall risk environment and country vulnerability to biological threats”

1.1.6 Population Data

The population data is obtained from the *United Nations Department of Economic and Social Affairs World Population Prospects 2019 Report* [Citation Needed]. Three key metrics are collected for later analyses: total population (in thousands), population density (per square kilometer), and average median age (of a country’s population).

1.2 Data Preparation

In this section, the steps taken to prepare the data for the research will be explained.

1.2.1 Labeling

To ensure data accuracy, consistency, and to make later analysis steps easier, all countries were labeled according to their *ISO 3166-1 alpha-3* code as published by the *International Organization for Standardization* [Citation Needed]. For data where countries were not originally labeled with the *ISO 3166-1 alpha-3* code, the codes needed to be applied manually.

1.2.2 Combining

All the time series data needed to be formatted into a traditional time series structure (e.g. countries vs. dates). Some of the data contains multiple instances of a single country. For example, because Canada's COVID-19 case and death data is given provincially, it is listed in the COVID-19 case and death time-series data more than once. Additionally, although the COVID-19 testing data is given as a time-series, every country and every day is treated as a new data point, so this data must be combined into the more traditional time series structure.

1.2.3 Converting

In this section, the two different data conversions will be described.

Conversion to Incident Rate

Two types of data conversions needed to be performed on the data. The first was to convert the COVID-19 incident data from cumulative to daily incidents. This conversion allows for a more intuitive expression of the state of the COVID-19 pandemic in a country.

For the COVID-19 test data, some countries had report both cumulative and daily tests and some countries only report one or the other. If only daily or cumulative tests were reported, the data would be converted to daily tests if necessary. However, there were instances where, if a country reports both daily and cumulative tests, that the data would not be consistent (this appeared to be a result of a country's tendency to update either daily or cumulative tests over the other). In this case, the larger of the two options was chosen and converted to daily tests if necessary.

Conversion to Daily Incidents

The second conversion was to convert the COVID-19 incident data from total number of incidents (e.g. total cases) to incidents per million (e.g. cases per million). This conversion allows for a more equitable comparison between countries as it removes the confounding variable of population size in relation to the number of COVID-19 incidents in a country.

1.3 Data Smoothing

In this section, the step taken to smooth the data, so it could be analyzed, will be explained.

1.3.1 Removing Outliers

This section will deal with the three different types of outliers and the algorithm used to to remove and correct them. Note that because the COVID-19 test data was more sparse and the reporting frequency was less than the COVID-19 case and death data, outlier removal was not performed on the COVID-19 test data. A further explanation of the consequences, in terms of how the test data can appropriately be used, of this will be discussed in the metric generation section.

Types of Outliers

There are three types of oultiers in the COVID-19 data that need to be handled:

- **Known**: outliers that are explicitly tracked and noted by the by the original data sources
- **Negative**: outliers that appear as a negative value in the daily time series data
- **Manual** outliers that, although not explicitly tracked by the data sources or negative, result in visible and temporary change in a country's COVID-19 time series data

The known outliers are complied from lists found on the data source's websites or Github pages and tracked in an excel file. The negative outliers could be automatically detected and, upon their removal, would be listed in an excel file. Due to the wide variation between different country's data and within a country's data, algorithmic outlier detection proved infeasible. Rather, for reputability and consistency considerations, these outliers were manually tracked.

Outlier Removal Algorithm

The outlier removal algorithm was made up of two parts: determining the new incident value on the date of the outlier and redistributing the excess incidents to prior dates. The these two algorithms work together to remove the outlier, not change the total number of instances in the data, and keep the signal of the original data intact.

New Value Algorithm

The New Value Algorithm determines the new value of incidents to set on the date of the outlier. It functions by determining the average number of incidents on the seven dates before and after the date with the outlier. This served to keep the signal of the distribution as intact as possible after the outlier was removed. It then returned the number of incidents to remain on that date (e.g. the new value) and the number of incidents to be redistributed (e.g. the excess incidents). The full algorithm is below:

Algorithm 1: New Value Algorithm

$Outlier\_Date, First\_Date, Last\_Date, Incident\_Data$

$Avg\_Before\_Outlier = MEAN(Incident\_Data[Outlier\_Date - 7 : Outlier\_Date))$

$Avg\_After\_Outlier = MEAN(Incident\_Data(Outlier\_Date : Outlier\_Date + 7])$

$New\_Value = (Avg\_Before\_Outlier + Avg\_After\_Outlier)/2$

$Excess\_Value = Incident\_Data(Outlier\_Date) - New\_Value$

$Incident\_Data(Outlier\_Date) = New\_Value$

Redistribution Algorithm

The Redistribution Algorithm determines how to redistribute the excess incidents after the outlier has been removed. It does so on a proportional basis over the prior 30 days. This means that dates with more incidents will get slightly more incidents removed than dates with less incidents and ensures that, after removing cases, the number of incidents on each date remains positive. This helps minimize the changes in the signal of the overall data. The full algorithm is below:

Algorithm 2: Redistribution Algorithm

$New\_Value, Excess\_Value =$
$New\_Value\_Algorithm(Outlier\_Date, First\_Date, Last\_Date, Incident\_Data)$
$Incidents\_Over\_Range = SUM(Incident\_Data[Outlier\_Date - 30 : Outlier\_Date])$
$Range\_Weight = Incident\_Data[Outlier\_Date - 30 :$
$Outlier\_Date]/Incidents\_Over\_Range$
$Range\_Adjustments = Range\_Array * Excess\_Value$
$New\_Range\_Data = Incident\_Data[Outlier\_Date - 30 :$
$Outlier\_Date] - Range\_Adjustments$

1.3.2 Creating Moving Averages

The vast majority of countries' COVID-19 incident data exhibits weekly periodicity. For example, for any given week the USA's COVID-19 cases tend to be higher on Friday than Monday. To remove this periodicity and further smooth the COVID-19 data, a seven-day moving average was applied to the COVID-19 data. A seven-day moving average was chosen because there were seven days in a week.

1.3.3 Savitzky-Golay Filtering

Even after the outliers are removed and the seven-day moving average is applied, the data for some countries may still contain sudden spikes that result in a time series that is not smooth. These spikes can be considered as noise as, although they do not affect the overall signal of the data, they could cause issues with later data analysis. To correct this, a Savitzky-Golay filter [Citation Needed] was applied to the COVID-19 data. This allows for the noise in the COVID-19 data to be removed while preserving the overall signal. Every country's COVID-19 case and death data is smoothed using a Savitzky-Golay filter with a window length of 15 (days) and a third-order polynomial.

1.4 Data Interpolation

In this section, the steps taken to interpolate the data will be explained. The overall goal of the data interpolation is to take the smoothed COVID-19 data and fit a continuous, derivable function to it so that advanced metrics (e.g. rate of growth) can be extracted from it later on. The crux of the problem is that, even though the COVID-19 data has been smoothed, the data between outbreaks is not always strictly increasing or decreasing.

1.4.1 Peak Finding Algorithm

In order to solve this problem, a Peak Finding Algorithm was created to systematically determine the peaks and valleys (i.e. the key-points in the data for function fitting) of the COVID-19 data. The resulting algorithms work for both the COVID-19 case and death data for every country. The Peak Finding Algorithm is made up of four sub-algorithms (Initial Peaks Algorithm, Peak Pruning Algorithm, Final Pruning Algorithm, and Valley Finding Algorithm).

Initial Peaks Algorithm

The Initial Peaks Algorithm works in two main steps. First, points are analyzed to determine if they dominate over a specific window of points. Note that this is different than determining whether points are relative maximums. For example, it is possible for a point to be a relative maximum while not dominating over a range of points (of sufficient size). Thus, by searching for dominating points instead of relative maxes, one solves the problem of the data between peaks and valleys not being strictly increasing or decreasing.

The second step is to determine if a peak is a plateau and only to keep the middle point if it is. For example, especially for country's with less COVID-19 incidences, a peak will be at a certain level for multiple days (e.g. a plateau). Keeping every point in a plateau is redundant, so only the middle point on a plateau should be kept. The algorithm also accounts for the instance (usually only present in smaller data sets) where two small peaks of equal size appear next to each other. The full algorithm is below:

Algorithm 3: Initial Peaks Algorithm

```
Incident_Data, Window_Length, Peaks
for i ∈ Incident_Data do
    max_value = 0
    max_value = MAX(Incident_Data[i − Window_Length/2 : i + Window_Length/2])
    if Incident_Value(i) = max_value AND max_value > 0 then
        Peaks.append(i)
end
for j ∈ Peaks do
    if Peaks(j).index = Peaks(j + 1).index − 1 then
        if j + 2 = Peaks(last).index then
            n = 1
            plateau_points.append(Peaks(j : j + n).index)
        else
            n = 0
            while
             k+n < Peaks(last).index AND Peaks(j+n).index = Peaks(j+n+1).index−1
             do
                plateau_points.append(Peaks(j + n).index)
                n = n + 1
            end
        end
    else if
     Peaks(j+1).index−Peaks(j).index < 15 AND Peaks(j+1).index = Peaks(j).index
     then
        plateau_points.append(Peaks(j + 1).index)
end
Peaks = Set_Diff(Peaks, plateau_points)
```

Valley Finding Algorithm

The Valley Finding Algorithm is run after every step of the overall peak finding algorithm (e.g. after the Initial Peaks Algorithm, Peak Pruning Algorithm, and Final Pruning Algorithm). It works by looking for the absolute minimums between peaks and endpoints of the COVID-19 incident data. It does so peak by peak, which allows for instances where ranges of dates contain zero incidents to be preserved in the data. Duplicate valleys are removed at the end. The algorithm is below:

Algorithm 4: Valley Finding Algorithm

```
Peaks, Incident_Data
for i ∈ Peaks do
    min_points.append(MIN(Incident_Data[i : i + 1]).index)
    min_points.append(MIN(Incident_Data[i − 1 : i]).index)
end
min_peaks.unique_values
```

Peak Pruning Algorithm

The Peak Pruning Algorithm prunes peaks based on the peak's width, relative height (height above neighboring valleys), width percentage, height percentage, and relative height percentage (where percentage refers to the percent of that value of the summation of all values, e.g. Peak i Width / All Peak Widths. The combination of these metrics allows the conditional removal of peaks that are not large enough to be considered true peaks in the COVID-19 data. Furthermore, the inclusion of the percentage metrics allows for the function to be universally applied to every country's COVID-19 incident data. An additional check on the first peak in the data to ensure it is appropriate that it is included. The algorithm is below:

Algorithm 5: Relative Maximum Pruning Algorithm

```
Peaks, Incident_Data
for i ∈ Peaks do
    if i's Width is Too Thin AND i's Width % is Too Small then
        if i's Height % is Too Low then
            Prune(i)
    if i's Height % is Too Small then
        if i's Width % is Too Small then
            Prune(i)
    if i's Relative Height is Too Short then
        if i's Relative Height % is Too Small then
            Prune(i)
    if i is the first peak AND i's Height % is Too Small then
        Prune(i)
end
```

Final Pruning Algorithm

The Final Pruning Algorithm works by pruning peaks, and their corresponding valleys, if they do not dominate over their range (e.g. if a point larger than the peak appears between it and a neighboring valley). This algorithm is used in the rare instance that a country's COVID-19 incidence data had a significant drop in cases over a very short time period (e.g. a couple of days). When this happens, it can cause valleys to be improperly identified and thus a final correction is needed. The algorithm is below:

Algorithm 6: Final Pruning Algorithm

```
Peaks, Valleys, Incident_Data
for i ∈ Peaks do
    max_point = MAX(Incident_Data(i's Left Valley : i's Right Valley)
    if max_point > Incident_Data(i) then
        Removed_Peaks.append(i)
        Prune(i)
end
for j ∈ Removed_Peaks do
    if i < i + 1 then
        if Next Two Valleys = 0 then
            Remove Next Two Valleys
        else
            Remove Next Valley
        end
    else if i > i + 1 then
        if Last Two Valleys = 0 then
            Remove Last Two Valleys
        else
            Remove Last Valley
        end
end
```

1.4.2 Interpolation

Now that the the peaks and valleys from the COVID-19 incident data has been determined, they must be fit to a continuous and derivable function so that advanced metrics can be derived later on. To do so, interpolation rather than regression will be used. This is because interpolated functions can be fitted to data that does not follow a standard function (i.e. polynomial, exponential, etc.), which nearly every country's COVID-19 incidence data does not. Thus, an interpolated function will results in a much more accurate fitting of a country's COVID-19 data. However, using exclusively the peaks and valleys does not guarantee that the interpolated function will be a good fit. For example, if the convexity of a function is misjudged between a peak-valley pair, the interpolation will be entirely incorrect. Therefore, before the interpolation algorithm is run, and algorithm to add additional key points for the interpolation to utilize must be run.

Key Point Addition Algorithm

The Key Point Addition Algorithm is used to add additional points to the peak and valley points to increase the accuracy of the interpolation. It works by fitting a cubic polynomial between every single peak-valley pair. It then searches for all the points in the peak-valley range where the COVID-19 incident data and the cubic polynomial intersect. Then, based on whether the data is increasing or deceasing, points are pruned based on whether their neighboring points increase or decrease. This process ensures that the data between the peaks and valleys will always be purely increasing or decreasing while greatly improving the accuracy of the final interpolation. The algorithm is below:

Algorithm 7: Key Point Addition Algorithm

```
Peaks_and_Valleys, Incident_Data
for i ∈ Peaks_and_Valleys do
    data_direction = 0
    if Incident_Data(i) < Incident_Data(i + 1) then
        data_direction = −1
    else if Incident_Data(i) > Incident_Data(i + 1) then
        data_direction = 1
    Create Cubic Polynomial using Incident_Data(i : i + 1)
     Identify Intersection Point between Cubic Polynomial Data and Incident_Data
    for j ∈ Intersection_Points do
        if direction = 1ANDj > j + 1 then
            Removej + 1
        else if direction = −1ANDj < j + 1 then
            Removej + 1
    end
end
min_peaks.unique
```

Interpolation

Piecewise Cubic Hermite Interpolating Polynomial (PCHIP) [Citation Needed] was chosen as the interpolation over other interpolation methods (like splines). This is because PCHIP interpolation does not overshoot that data and it includes less oscillation, which is important due to the wide variety of data distributions that are performed. Additionally, PCHIP interpolation is able to handle instances of flat data, which is important as it would not be unexpected for flat data ranges (e.g. consecutive days with zero new COVID-19 cases) to occur in the data set. Indeed, when testing PCHIP interpolation against splines and Akima interpolation, PCHIP interpolation produced the best results.

1.5 Metric Generation

In this section, the metrics and their generation process will be explained.

1.5.1 Summation Metrics

A *summation* metric is a metric where all values over the entire range of a time series dataset are summed together. It is performed on a per country basis. The three summation metrics calculated were as follows:

- Total Cases per Million
- Total Deaths per Million

- Total Tests per Million

1.5.2 Ratio Metrics

A *ratio* metric is a metric where two summation metrics are compared against one another as a ratio. It is performed on a per country basis. The two ratio metrics calculated were as follows:

- Cases per Million to Death per Million Ratio
- Tests per Million to Cases per Million Ratio

1.5.3 Maximum Value Metrics

A *maximum value* metric is a metric where the maximum value over the entire range of a given time series data is determined. It is performed on a per country basis. The four maximum value metrics calculated were as follows:

- Maximum Stringency
- Maximum Government Response
- Maximum Containment Health
- Maximum Economic Support

1.5.4 Number of Peak Metrics

A *number of peaks* metric is a metric that indicates how many peaks occurred in a country's COVID-19 data. The two number of peaks metrics calculated were as follows:

- Number of Peaks in Cases
- Number of Peaks in Deaths

The number of peaks for tests was not calculated. This was because of how differently country's reported their COVID-19 testing number, if they reported them at all. Whereas every country reported the number of COVID-19 cases and deaths on a daily basis, some countries would report their test data much less frequently (regardless of the fact that the source data was updated twice weekly). Reporting patterns ranged all the way from China's (who reported all of their testing data in bulk on two separate days) to the USA's (who reported it daily). Thus, the highest level metric it would be appropriate to calculate from the test data would be a summation metric.

1.5.5 Rate of Growth Metric

A *rate of growth* metric is a metric where the rate of growth of COVID-19 data between a peak and its preceding valley is calculated. It is calculated both in term of the average rate of growth rate and the maximum rate of growth over that range. It is calculated for every single peak in a country's COVID-19 data and on a per country basis. The four rate of growth metrics calculated are as follows:

- Average Rate of Growth of Cases

- Maximum Rate of Growth in Cases
- Average Rate of Growth in Deaths
- Maximum Rate of Growth in Deaths

1.5.6 Rate of Submission Metric

A *rate of submission* metric is a metric where the rate of growth (which in this case is negative) of COVID-19 data between a peak and its succeeding valley is calculated. It is calculated both in terms of the average rate of submission and the maximum rate of submission over that range. It was calculated for every single peak in a country's COVID-19 data ans on a per country basis. The four rate of submission metrics calculated are as follows:

- Average Rate of Submission of Cases
- Maximum Rate of Submission in Cases
- Average Rate of Submission in Deaths
- Maximum Rate of Submission in Deaths

1.5.7 Length of Time Metrics

A *length of time* metrics is a metric where the length of a specific observation relating to the COVID-19 data is recorded. Three length of time metrics were calculated for every peak in the COVID-19 data on a per country basis: length of growth (the amount of time between a peak and it's preceding valley), length of submission (the amount of time between a peak and its succeeding valley), and total peak length (a peak's length of growth and length of submission). The six length of time metrics calculated are as follows:

- Length of Case Growth
- Length of Case Submission
- Length of Case Peak
- Length of Death Growth
- Length of Death Submission
- Length of Death Peak

1.5.8 Peak/Valley Value Metrics

A *peak/valley value* metric is a metric that measures the number of new incidents observed at either a peak or valley in a country's COVID-19 data. The peak value metric measure the worst severity of a given peak and the valley value metrics measure how submitted the COVID-19 outbreak actually was. These metrics were calculated for every peak and every peak's succeeding valley on a per country basis. The four peak/valley value metrics calculated were as follows:

- Peak Value of Cases

- Peak Value of Deaths
- Valley Value of Cases
- Valley Value of Deaths

1.5.9 Case-Death Pair Peak Ratio

A *case-death pair peak ratio* metric is a metric that compares the ratio between the peak number of cases and the peak number of deaths of a country's corresponding COVID-19 case peak and death peak. KMeans clustering was used in order to pair peaks in cases with peaks in deaths [Citation Needed]. A KMeans clustering was used because the number of case peaks and death peaks is not guaranteed to match in quantity or location. This could be due to a number of factors. For example, one case peak could be related to two death peaks in the situation where ICU's reach capacity (i.e. a peak in deaths occurs), but a better treatment method is found (i.e. the number of daily deaths begin to fall), yet the number of daily cases continues to grow (i.e. the ICU's begin to fill up again and another peak in deaths occurs). The data used for the KMeans clustering was a peak's start date, peak date, and end date. The optimal number of clusters to use was calculated by finding the optimal silhouette value of different sized clusters. The resulting clusters are the case-death peak pairings. The algorithm is below:

Algorithm 8: Valley Finding Algorithm

$Case\_Peaks\_Data, Death\_Peaks\_Data$
$min\_clusters = MIN(num\_Case\_Peaks, num\_Death\_Peaks)$
$max\_clusters = MAX(num\_Case\_Peaks, num\_Death\_Peaks)$
if $min\_clusters > 1$ **then**
    for $i \in min\_clusters : max\_clusters$ **do**
        $Perform\ KMeans\ Clustering\ with\ i\ clusters$
        $Record\ largest\ silhouette\ score$
    end
$num\_clusters = cluster\ with\ best\ silhouette\ score$
$Perform\ Final\ KMeans\ Clustering(k = num\_clusters, Case\_Peaks\_Data, Death\_Peaks\_Data)$

Note that algorithm above does not necessarily force the pairing of case and death peaks. Thus, it is possible for a "paring" to only be a single case or death peak. The final part is an algorithm that will calculate the case-death pair peak ratios. It does so by, after ensuring there is at least one case peak and one death peak in a pairing, finding the maximum case and death peak value over the ranges.

1.5.10 Days Above Sub-index Level

A *days above sub-index level* metric is a metric that tracks the total number of days that the level of one of *Oxford's Stringency* indexes [Citation Needed] are at or above a specified level. Each level in a sub-index corresponds to a certain stringency that can be tracked. For example, one of the sub-indexes that is used to calculate a country's containment policy index is workplace closings, of which a level 2 corresponds to some sectors being closed. Some of these sub-indexes have a great deal of direct importance to this research, thus the necessity of these metrics. The eight days above sub-index level metrics calculated are as follows:

- Number of Days Some Sectors Closed
- Number of Days All Non-Essential Sectors Closed
- Number of Days Public Transport Closed
- Number of Days of Stay-at-Home Lock-downs except for Essential Trips
- Number of Days of Stay-at-Home Total Lock-downs
- Number of Days Internal Movement Restricted
- Number of Days of International Bans for Some Countries
- Number of Days of Total International Border Closure

1.5.11 Total and Rate Before Sub-Index Level

A *total and rate before sub-index level* metric is a metric that describes the total number of incidents, and the rate of growth of instance, before the first time a specified sub-index level was reached. The rate of growth is captured by both an average rate and a maximum rate. In total 48 metrics were calculated on a per country basis, represented in the following list:

- (Total Instances, Average Rate of Growth, Maximum Rate of Growth) of (Cases, Deaths) Before Some Sectors Closed
- (Total Instances, Average Rate of Growth, Maximum Rate of Growth) of (Cases, Deaths) Before All Non-Essential Sectors Closed
- (Total Instances, Average Rate of Growth, Maximum Rate of Growth) of (Cases, Deaths) Before Public Transport Closed
- (Total Instances, Average Rate of Growth, Maximum Rate of Growth) of (Cases, Deaths) Before Stay-at-Home Lock-downs except for Essential Trips
- (Total Instances, Average Rate of Growth, Maximum Rate of Growth) of (Cases, Deaths) Before Stay-at-Home Total Lock-downs
- (Total Instances, Average Rate of Growth, Maximum Rate of Growth) of (Cases, Deaths) Before Internal Movement Restricted
- (Total Instances, Average Rate of Growth, Maximum Rate of Growth) of (Cases, Deaths) Before International Bans for Some Countries
- (Total Instances, Average Rate of Growth, Maximum Rate of Growth) of (Cases, Deaths) Before Total International Border Closure

1.5.12 Total, Rate, and Incident-to-Index Ratio Before First Peak in Index

A *total and rate before first peak in index* metric is a metric describes that the total number of instances and the rate of growth in COVID-19 incidents before the first peak in one of Oxford's four main indexes was reached. The rates are calculated by both an average rate and a maximum rate. The ratio is made between the number of incidents and the index level at its first peak. In total, 32 metrics were calculated on a per country basis, which are described in the list below:

- (Total Incidents, Average Growth Rate, Maximum Growth Rate, Incident-to-Index Ratio) of (Cases, Deaths) of Stringency Index
- (Total Incidents, Average Growth Rate, Maximum Growth Rate, Incident-to-Index Ratio) of (Cases, Deaths) of Containment Health Index
- (Total Incidents, Average Growth Rate, Maximum Growth Rate, Incident-to-Index Ratio) of (Cases, Deaths) of Economic Support Index
- (Total Incidents, Average Growth Rate, Maximum Growth Rate, Incident-to-Index Ratio) of (Cases, Deaths) of Government Response Index

1.5.13 Index-to-Index Ratio Metrics

A *index-to-index ratio* metric is a metric that creates a ratio between the values in one of Oxford's indexes to another. The ratio is calculated by finding the average level of both indexes over the recording time frame, then creating the ratio between the resulting averages. The three index-to-index ratio metrics calculated on a per country basis are as follows:

- Economic Support to Containment Health Ratio
- Economic Support to Stringency Ratio
- Economic Support to Government Response Ratio

1.6 Statistical Analysis (YET TO BE COMPLETED)

This section contains some details of the statistical analyses that will be performed to uncover relationships in and evaluate the COVID-19 data between countries and other datasets. Once these statistical tests are completed, the research itself will be finished. All that will be left is to properly summarize the conclusions of the from the research, apply it to a case (either theoretical or practical), and potentially create a ranking system and country profiles detailing their suitability as geographical locations for supply chain diversification to hedge the black swan risk of a pandemic outbreak.

1.6.1 Confounding Variable Exploration

Statistical testing between the some of the COVID-19 metrics and outside data to determine if possible confounding variables can be uncovered. The datasets that will be compared against the COVID-19 data are:

- Population Data

- Economic Data
- Healthcare Performance Data
- Press Freedom Data

1.6.2 Pandemic Preparedness Rankings to COVID-19 Data Accuracy

The *Global Health Security Rankings* will be compared to the actual COVID-19 metrics in order to determine how accurately the can be used as predictors as the the effectiveness of a country's COVID-19 response. The analyses will be performed on the overall index, qualitative sub-indices, and the categorical sub-indices.

1.6.3 Pandemic Preparedness to Oxford Stringency Analysis

The *Global Health Security Rankings* will be compared to the *Oxford Stringency Index* in order to determine how accurately the global health security rankings were at predicting the magnitude of a country's COVID-19 response. The overall and sub-indices of the global health security rankings will be compared to the four main Oxford indices.

1.6.4 COVID-19 Data to Oxford Stringency Index Analysis

The COVID-19 data will be compared with the *Oxford Stringency Index* to quantity an compare the thresholds required for a country to increase the severity of their action to combat COVID-19. This will help will help determine every country's observed risk tolerance, which would be directly related to their desirability as being a new geographical location for supply chain diversification in a future pandemic scenario.

1.6.5 Post Pandemic Rankings

A ranking system can be designed to systematically rank a country's attractiveness as a geographical location for future supply chain diversification in terms of combating the black swan risk of a pandemic outbreak. The COVID-19 data, the Oxford Stringency Index, and their relationship will be used for this analysis as well as any confounding variables discovered.