Austin Bellis, Ethan Chang, Reetom Gangopadhyay, Dhruv Gandhi, Gui Marques, Neeza Singh
DS 310
Prof. Seferlis
Dec 20, 2023

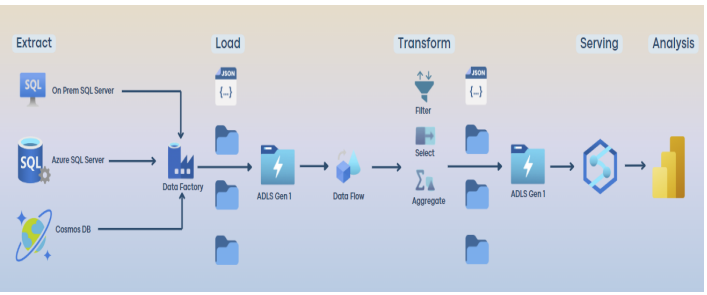
Final Project Report

Introduction

In response to the ongoing global challenges posed by the Coronavirus pandemic, our team has undertaken a critical project in collaboration with the Commonwealth of Caladan, a midsize Commonwealth with a population of 3.2 million. To successfully create a plan to mitigate the next wave of the pandemic, we used Microsoft Azure and Power BI to determine whether specific policies other countries have implemented would be effective for Caladan. The success metrics we used to assess efficacy were the growth rates of cases.

Specifically, we analyzed if countries recommending workplaces to close is enough and if governments replacing 50% or more of individuals' lost salaries could lower cases. These policies would have less of an impact on death rates than healthcare-oriented policies, like vaccination requirements or protection of the elderly, so we decided to compare with the change in cases only.

The countries we used in our analysis were Russia, the United Kingdom, France, Italy, Germany, Canada, Sweden, Japan, South Korea, and New Zealand. For each country, we collected data on which policies the country implemented and COVID-19 cases, deaths, and recoveries from each day in that country from January 2020 to March 2021. The dataset also contained identifiers for each case, tags for each country, and different formats for dates and times. These would not be useful in our analysis, but they did allow us to create a schema between our tables.

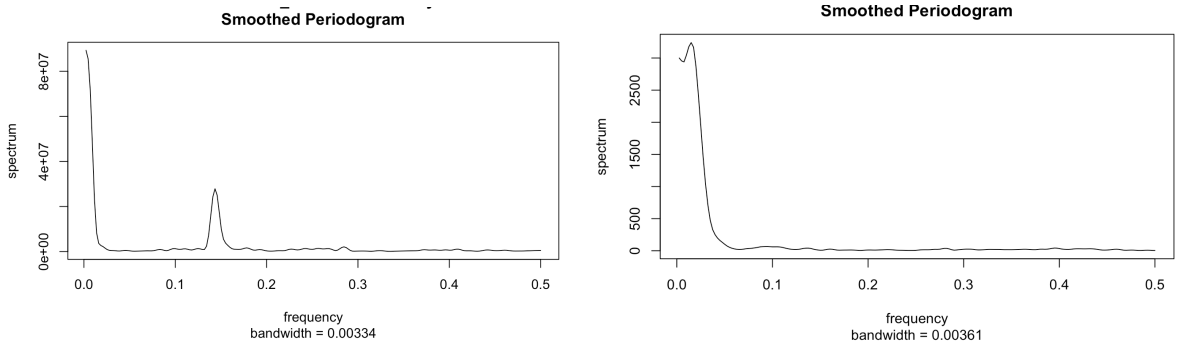


We extracted this data from three different sources using an Azure Data Factory, then loaded the data, as parquet files, into an Azure Data Lake. From there, we used data flows to filter and join the data, creating one parquet for each table. We used Synapse to create these external tables, then we connected to PowerBI to create visualizations.

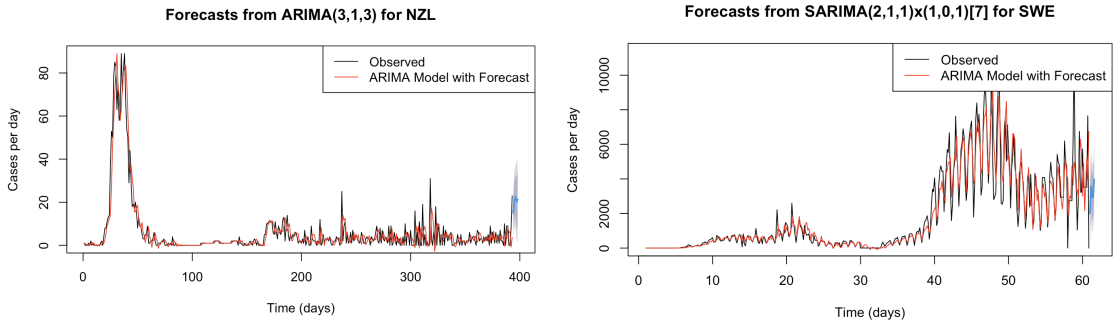
Using PowerBI, we created a line and stacked column chart for each policy, where we modeled policy usage against the monthly average change in COVID-19 cases. We utilized a slicer to change one visual to fit any country. During this step, we also decided that it would not be logical to compare larger countries to Caladan, a country of three million citizens. We opted to give extra weight to the smallest two countries: New Zealand (five million citizens) and Sweden

(ten million citizens). If a policy worked for those two countries, we could check their effectiveness in the other eight. We found that our first policy, recommending workplace closing, benefitted both countries, but also that both countries did require closings when cases were higher. Our second policy, replacing over 50% of individuals' lost salaries, seemed to have little effect in two countries that utilized this policy for close to a year. We found a third policy, controlling international travel, was beneficial with more substantial results.

Using R and its packages, we generated two different modeling frameworks. The first method is the ARIMA/Seasonal ARIMA (SARIMA) model. This modeling ideology accurately captures and accurately predicts patterns in time series data. After creating a time series plot, there are two distinctly different patterns using data from our selected countries, Sweden and New Zealand. To fit the ARIMA model, we must first check for unit roots and investigate spectral analysis plots to investigate whether the model requires differencing and whether any seasonality is present in the series. Spectral analysis is used to identify any dominant frequencies in the data. The results are the following:

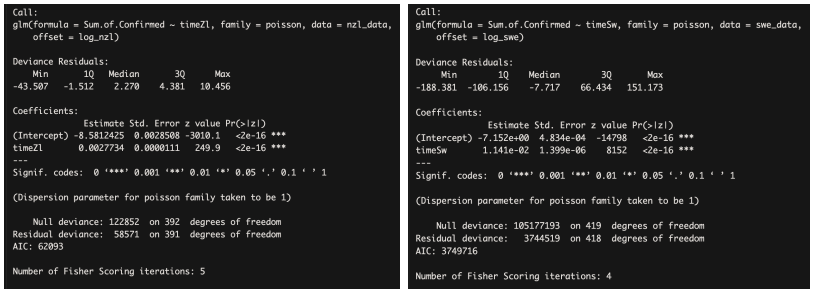


On the left is Sweden's periodogram, and on the right is New Zealand's. We need to account for the seasonality in the Swedish data. Fitting the respective models and using forecasting, we achieve the following results:



The models fit the data well and forecast five days in the future for both countries. This modeling methodology is valuable since we can predict daily cases with relative accuracy, giving

us an idea of how the case numbers will change. The second modeling method is a Poisson regression approach for count data with an “offset” term. This method utilized the sum of the cases over time. We can employ a Poisson Regression with a Generalized Linear Model (GLM) with the Poisson family. Adding an offset term gives us some power to compare data generated from different population sizes. In the context of our problem, using Poisson regression with an offset including the log(population) of a country will let us compare the rates of how quickly the number of cases grew for different countries.



```
Call:
glm(formula = Sum.of.Confirmed ~ timeZl, family = poisson, data = nzl_data,
    offset = log_nzl)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-43.507   -1.512    2.270    4.381   10.456

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.5812425  0.0028508 -3010.1   <2e-16 ***
timeZl       0.0027734  0.0000111   249.9   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 122852  on 392  degrees of freedom
Residual deviance:  58571  on 391  degrees of freedom
AIC: 62093

Number of Fisher Scoring iterations: 5
```

```
Call:
glm(formula = Sum.of.Confirmed ~ timeSw, family = poisson, data = swe_data,
    offset = log_swe)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-188.381  -106.156    -7.717    66.434   151.173

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.152e+00  4.834e-04  -14798   <2e-16 ***
timeSw       1.141e-02  1.399e-06    8152   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 105177193  on 419  degrees of freedom
Residual deviance:   3744519  on 418  degrees of freedom
AIC: 3749716

Number of Fisher Scoring iterations: 4
```

The parameter estimates for Sweden (right) and New Zealand (left) are significant, as shown by the P-Values in the readout. With slope estimates of 0.001141 for Sweden and 0.0027734 for New Zealand, the cases in New Zealand grew faster. This modeling framework suggests a rudimentary method for comparing case rates in different population sizes.

Combining exploratory and statistical data analysis gives some importance to this project. By using the ARIMA/SARIMA and Poisson regression frameworks, we can model the situation retroactively and predict future values. This gives us a holistic view of the representative sample. We have two views that offer a strong reinforcement against COVID-19. Our project could be more robust with a stronger sample. Our ten countries were one of the most extensive limitations we experienced. Additionally, surveying twenty policies, each in use at different periods by each nation, made it difficult to understand the impact truly. We have scientific proof that facial coverings limit COVID-19 spread, but policies like income support were generally implemented to keep countries running during downtime, which we could not measure. This project is functional in measuring COVID-19 cases, but there are better ways to advise a country.