Covid-19 Awareness and Covid-19 Cases in Ohio

Zheng Gu and Sheng-Lien Lee[1]

***Abstract*— This study aims to enhance the R-square score for forecasting Covid-19 cases in Ohio State by utilizing time series analysis techniques. The training and testing data sets are merged, and a hybrid imputation strategy, including a simple imputation method and MissForest, is implemented to predict missing values. An ARIMA model is subsequently employed to predict missing values that occurred in the most recent days of the data set. Ultimately, the proposed approach seeks to improve the accuracy of Covid-19 case prediction in Ohio State.**

I. Introduction

The pandemic of Covid-19 spread quickly around the world and impact people's life. The virus is not only contagious easily but also mutates rapidly which causes the effectiveness of the vaccine has been discounting. Although the virus is not deadly as other viruses such as Ebolavirus. However, the high number of infected populations still causes the number of people who die from serious Covid-19 symptoms extremely high. In addition, to against the virus and control the pandemic, the government of each country develops related measures and policies such as quarantine policy, which change people's daily life a lot.

A month after the first outbreak in China, United States also have its first case of Covid-19. Soon after, it only takes about two months from the first case to two hundred thousand cases in the US. The dramatically increasing number of cases brings a heavy burden to the healthcare system. To against the invasion of Covid-19. Governments of each state work with health departments to make related policies and develop related measures to control the pandemic. Our research target, Ohio state is one of the states that start their work before the outbreak of the pandemic in the US. Since they start some related measures earlier than other states "it is with fewer than 1/3 as many cases and "a small fraction of the deaths as other states, such as Illinois and Pennsylvania" according to the Washington Post. Although Ohio's epidemic prevention policy let the State control the pandemic better than others, it also causes some societal issues. There are some people who doubt the government seizing too much power in the nominal of pandemic prevention. While others think that their policies such as prohibiting abortion and stay-at-home orders, are too aggressive and have already exceeded the prevention need and seriously violated human rights.

Our research aimed to predict the number of cases in Ohio State by data set provided by Kaggle competition. First, we preprocess the data and do visualization to analyze the data's distribution. Second, we use simple imputation and missforest imputation to impute the missing value. Third, we smooth the data and apply the ARIMA model to predict the last few days' missing values in the data set.

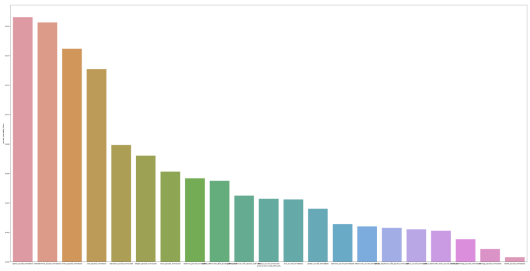
[1]Zheng Gu and Sheng-Lien Lee are both graduate students at the University of Rochester majoring in Data Science.

II. Data

The pandemic situation varies from population composition, epidemic prevention policies made by the government, and people's reactions to the related policies. Hence, the demographic of Ohio and how people react to the Ohio State government's epidemic prevention policies and increasing of the number of cases are two important factors that affect the number of infected cases. The dataset can be separated into two parts, one is the demographics data of Ohio residents another is the awareness level of different topics. We think awareness levels of different levels can be a representation of people's reactions to the pandemic development and the related policies. There are 18 demographic attributes and 126 awareness-level in each training and test data set. For the demographic attributes, the training data set contains the attribute, cases while the testing data set doesn't. The testing dataset contains the attribute, index while the training data set doesn't.

We further analyzed the awareness level part by visualization. First, we create a bar chart that shows the average normalized Jaccard similarity of different types of awareness topics. We discover that people's awareness related to entertainment activities, such as the bar of entertainment and sport, is the highest among other awareness. In addition, people care about issues related to their health since the third and fourth highest bars are related to illness and Covid-19. Moreover, we are surprised to find out that health and health technic are at the last of the bar chart, which means that people in Ohio do not really care about their health until they get sick.

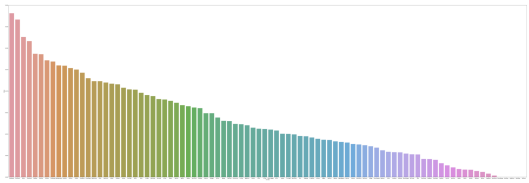
Average Values of topic awareness variables.



Second, we create another bar chart which shows the aggregation means awareness value of each county. We compare the chart with the Coronavirus cases map per 100,000 residents on December 29, 2022, from Wikipedia. We find out that the other five counties in the six last counties

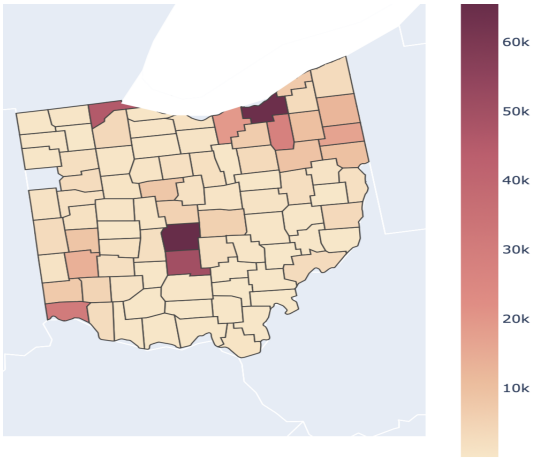
aren't counties that have fewer cases. Hence, we think the awareness is not only driven by the number of Covid-19 cases but also by other factors.

Aggregated Mean Awareness Value for each County

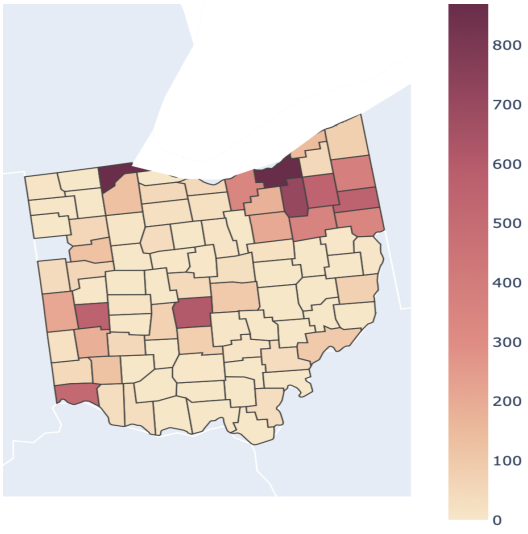


Third, we create number of cases' county-level map and number of deaths county-level map. We find out that there is a positive correlation between cases and deaths. Moreover, the pandemic is more severe in the northeast of Ohio.

Number of Cases

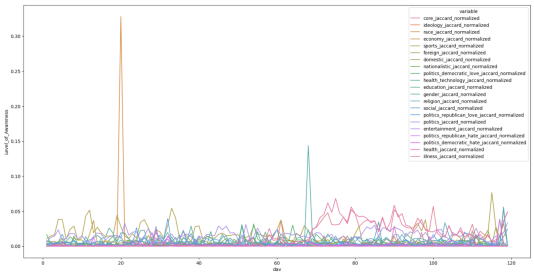


Number of Deaths



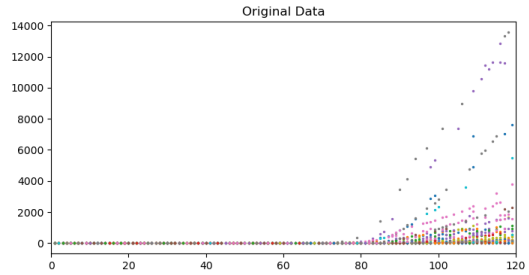
Fourth, we plot a line chart that shows the average normalized Jaccard awareness scores for each day. The chart shows that the awareness level for core and illness has a positive correlation but the awareness of illness is higher than awareness of core most of the time. The reason maybe what mention in the Wikipedia article that there are lots of cases don't be tested out.

Average Awareness Scores of each Day



For the preprocessing step, first, we merge the testing and training data set together. Second, we extract the number part of the attribute, date_index_converted. Third, we convert the data type of attributes, cases, index, and date_index_converted, from string to integer. At last, we delete normalized attributes in the awareness level part because the method, missforest, we are going to use is not sensitive to whether the data is normalized or not.

After preprocessing the data set, we plot a graph to see the distribution of the daily number of cases in each county. The graph shows that the number of cases increases dramatically after day 80.

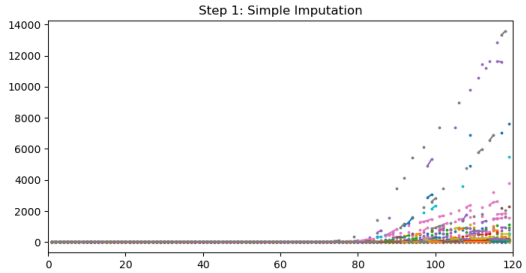


III. Methods

Due to the time continuity of the cases for each county, we combine the training and test data to create 88 time series, one for each county, with a total of 119 timestamps. Rather than relying on traditional machine learning methods to predict values for the test dataset, we approach the problem as a time series problem and utilized multiple imputation methods to fill out the test data. This method is more accurate than simply building machine learning models using numerous features because it considers the reality and trend of viral transmission, as well as taking into account the continuity of the time series.

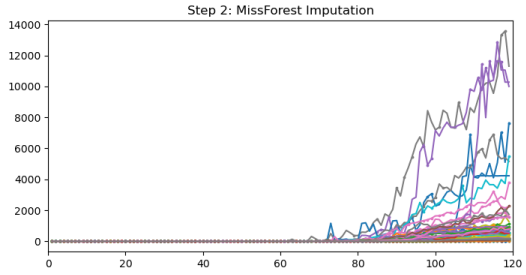
A. Simple Imputation

The initial step employs a basic imputation technique in which a missing value is replaced with the last known value either before or after it, provided that the neighboring known values are identical. This approach is particularly effective in swiftly completing numerous zero values in the initial phase of the time series.



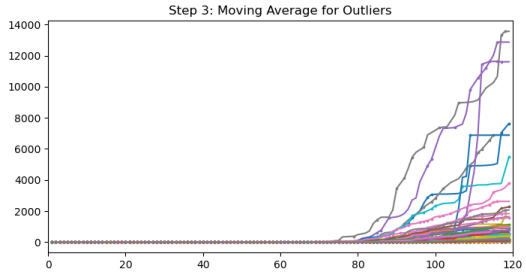
B. MissForest Imputation

The MissForest imputation method is used in the second step to impute all other missing values in the dataset. MissForest is a non-parametric imputation method that imputes missing values by building a Random Forest model on the observed values and then predicting the missing values using the model. By utilizing this technique, more accurate prediction of missing data will be obtained, which in turn improves the overall quality of our analysis.



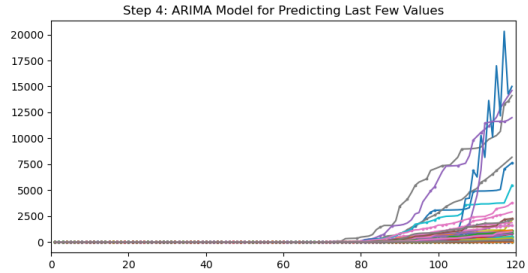
C. Moving Average for Outliers

Considering the data generated by MissForest is volatile, we apply the moving average method to replace outliers with a more representative value. Outliers are defined as the values larger or smaller than both neighboring values. The method involves calculating the moving average of a subset of neighboring data points and using the result as the replacement for the outlier.



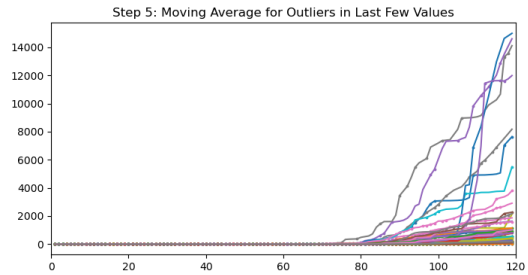
D. ARIMA Model for Predicting Last Few Values

Autoregressive Integrated Moving Average (ARIMA) is used for each time series to predict the last few values. For each county, a model is trained on the observed data from the beginning to the last timestamp with known values. And then this trained model is utilized to predict the missing values that occurred in the most recent days of the data set.



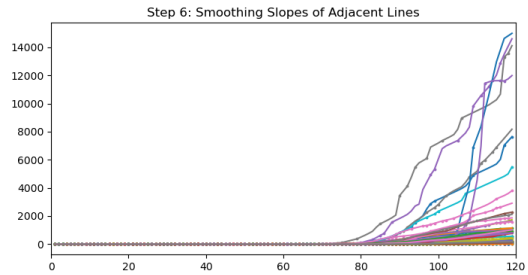
E. Moving Average for Outliers in Last Few Values

Similarly, the values generated by the ARIMA model has up and down oscillating, so moving average method is also applied to replace outliers in the predicted values generated by the ARIMA model.



F. Smoothing Slopes of Adjacent Lines

The final method used was to smooth the slopes of adjacent lines in the time series. This was achieved by calculating the average slope of adjacent lines and adjusting the slope of the lines to match the average slope.



IV. Results

In conclusion, our model is able to achieve an R-squared value of 0.98396, placing it fourth among all the teams on the 70% public test dataset. Obtaining R-squared values by ourselves is challenging for us because the training dataset represents only 30% of the total dataset, and splitting it into a training set and test set would yield meaningless results of R-squared value when it is compared with 70% of test dataset. Furthermore, in this challenge, we utilize imputation methods to predict test data instead of traditional machine learning methods, making it unnecessary to report R-squared values since we don't need to fine-tuned model parameters for a better result.

References

[1] COVID-19 pandemic in Ohio from https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Ohio.