**COVID-19 Forecaster: Analysis of Weather Conditions and Prediction of Covid-19 via Time Series Forecasting Technique**

**CMPT 459 Data Mining – Jian Pei**

**Caijie Zhao (301354256)**

**Department of Computer Science**

**Simon Fraser University,**

**Burnaby BC V5A 1S6, Canada**

**Motivation and Background**

Due to the spread of COVID-19, many countries have locked down, and people have quarantined for a long time. Recently, people start back to their workplace, and students plan to go traveling in summer break. However, the daily confirmed cases of COVID-19 keep increasing, and scientists need more time to develop the vaccine. Most people want to forecast the spread of COVID-19, so they have basic ideas about its future trend and can take some preventive actions.

Data scientists have analyzed the effect of weather on COVID-19 and concluded that temperature changes affect the COVID-19 infection rate, influencing it at the rate of 13-16 cases a day per 1°C rise in temperature [1]. Some studies also stated that -6.28oC and +14.51oC is the most favorable temperature range for the growth of COVID-19, and regions receive low precipitation (less than 400mm) and high wind speed (10-12 m/s) were more vulnerable to the virus [1].

My project goal is to analyze the relation between the weather and the spread of COVID-19 and predict COVID-19 in a country. This project can provide suggestions about future trend of the virus and help users make the decision, like selecting their traveling regions with low-risk of the growth of COVID-19.

**Problem statement**

As COVID-19 has declared as a pandemic, people feel interested in what factors may affect the growth of COVID-19. Thus, I analyzed weather factors to determine whether they have influences on the spread of COVID-19.

**Datasets**

There are three datasets used in my project. I downloaded the first dataset from Kaggle [2], and it stores the data about accumulative confirmed, death, and recovered cases of most cities in the world from January to August. I used the OpenCage Geocoder API to achieve the latitude and longitude of each city. I applied the BigQuery API to achieve another two datasets from the NOAA\_GSOD database. The second dataset contains station information [3]. It records station number (saf), weather bureau army number (wban), country and its latitude and longitude. The third one is a weather table [4]. This dataset provides saf and wban which are same as the second table. Besides, it contains information about temperature, dew point, sea level pressure, wind speed precipitation, and humidity.

Firstly, I merged two datasets retrieved from the NOAA\_GSOD database to a weather table based on the same columns, saf, and wban. I also added a new column called ‘day\_from Jan\_first’ to concatenate the weather table and the first table to a new dataset called COVID-weather. This table contains all the necessary information for my project. For each record in the dataset, it includes attributes like the date, city/state, country, and six weather features. For cleaning the dataset, I set zero to all empty cells in the tables and removed the records without information in the state column.

**Architecture/Pipeline**

The running environment can be Windows, iOS, and Linux. Users need to run my code in Kaggle Notebook. There are two files, one is for achieving data, and another one is for analysis and prediction. Users should link the Google Cloud Service in the add-on section at Kaggle Notebook before gaining data. My program will automatically install all necessary packages when users run it in Kaggle Notebook.

When users enter a country, my program will select related data. It will replace all idle numbers and strings to zero and ‘not reported’ separately. After cleaning data, the program will start to calculate the correlation between six weather factors and the spread of COVID-19. It also finds out the most important features among six features. All the results are presented by heatmaps and scatter charts. I utilized the time series forecasting method to predict the COVID-19 situation for next week and evaluated my model by using five-fold cross-validation.

Fig1. Data-Mining Pipeline

**Methodology**

I performed online analytical processing (OLAP) to demonstrate the relation in state wise or county wise based on users’ selection. When users enter queries, roll-up, or drill-down operation from OLAP is applied to select the corresponding data. I also implemented the Pearson correlation coefficient to measure linear correlation between each two variables selected from weather factors and the set containing the number of infected, deaths, and recovered cases. Unrelated two variables will be filtered out. The reason why I applied the Pearson method is to ensure there is a relation between some weather factors and COVID-19. Since the variance of weather is related to time, I can use time series forecasting techniques to predict future trend of COVID-19 for next week.

I used an autoregressive integrated moving average (ARIMA) model to predict the

selected country or city. Since there are only eight months of data in my dataset, I chose the ARIMA model to predict the trend of next week instead of using a seasonal autoregressive integrated moving average (SARIMA). SRIMA needs more data to build the model.

**Evaluation**

There are two methods used to evaluate the predicted model. The ARIMA model uses selected data to train the model. The data don’t have information on some dates. Use this model, one-step forecasting for every day, to predict the range of day that is the same as the train data. Used China as an example in figure1, the trained model can avoid overfitting. The model predicts the date in the train data, but there are some errors. My program calculates some statistics numbers, including root mean square error (rmse), mean absolute error (mae), and median absolute error (meae). The program does the grid search for eight combinations in ARIMA to improve accuracy. It selects the parameter with the lowest median absolute error. The median absolute error decreases by 33% compared to the original model.

The second evaluation is using five-fold cross-validation. For each iteration, the median absolute error will be calculated and average median absolute error will be printed after filtering out the maximum and minimum values. The average median absolute error is around 17%.

From the figure3, we can see the predicted trend in China is decreasing, but it starts to increase at the end of next week. The result is reasonable based on the relationship found between temperature and the spread of COVID-19 in my project. The overall temperature in China for next week still high, but at the end of next week, temperatures will start to decrease since the summer season is last from May to August in China.

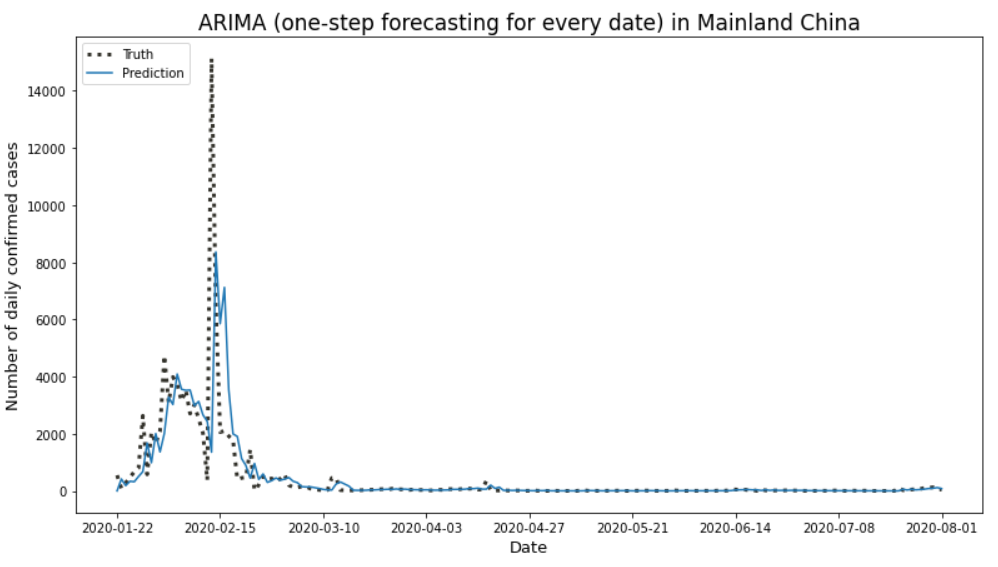


Fig1. One-step forecasting in Mainland China before Grid Search

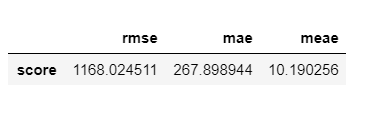


Fig2. Calculated Errors before Grid Search

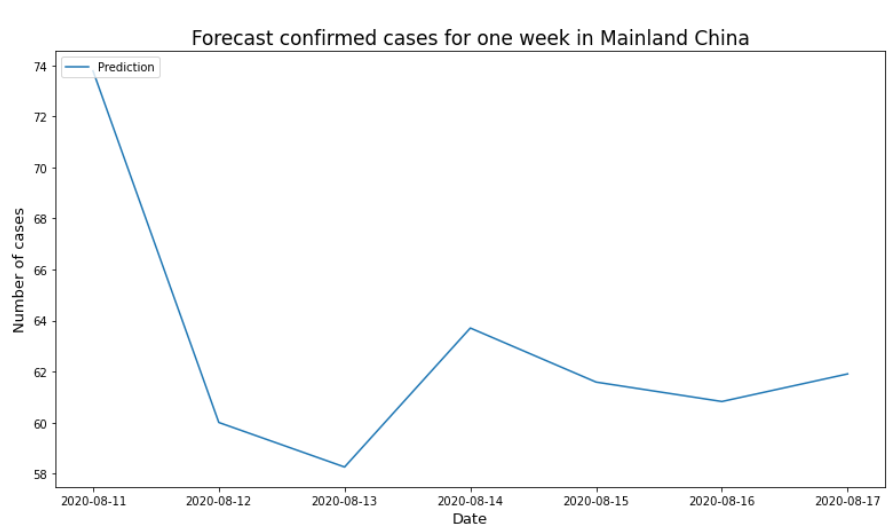


Fig3. Forecast Confirmed Cases for one week in China

**Results**

There are several outcomes shown to users. When users enter a city, my program will present the weather and COVID-19 situation about this city. The figures about the relation between the growth of COVID-19 and its most related weather factors, highest positive and negative Pearson coefficients, will be shown to the user as well.

Selecting a country, the outcome is a correlation heatmap and a bar chart of each weather feature’s importance. Users can receive the predicted result for next week. When users want to view worldwide situation, there are two interactive maps to show the growth of confirmed and death cases from January to August. Besides, another two figures are illustrating the distribution of deaths and confirmed cases in different temperatures (Fahrenheit).

From this project, I have learned how to make use of datasets like cleaning datasets, integrating different datasets to complete my goal. I have basic ideas on how to implement OLAP based on users’ queries. Besides, I have known a new type of forecasting technique, a time series forecasting. Based on the result of my project, I have basic ideas about how the weather factors affect the spread of COVID-19. People living in cold areas tend to be easier infected by Covid-19 compared to warm regions.

**Summary**

My project is to analyze the relation between weather and the spread of COVID-19 and then forecast the trend. Users can get suggestions from my project. Since people have stayed at home for a long time and students start their summer break, many families have planned about traveling. My work can help users to decide which cities or countries have a low risk of the spread of the virus when they want to go outside.

To complete my project, I made integration between the weather table and Covid-19 data. I also applied OLAP, statistic method, and time-series technique to select data, analyze the association, and make predictions for a country, respectively. I used some visualization like interactive maps, scatter, and bar figures to present the result and relation between weather and virus.

**References**

[1] M. Dr. Liji Thomas, "The effect of weather on COVID-19 infection rate", *News-Medical.net*, 2020. [Online]. Available:

<https://www.news-medical.net/news/20200506/The-effect-of-weather-on-COVID-19-infection-rate.aspx>. [Accessed: 10- Aug- 2020].

Datasets:

[2] <https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset?select=covid_19_data.csv>

[3] <https://www.kaggle.com/carinazhao/stations>

[4] <https://www.kaggle.com/carinazhao/gsod2020>