TIME SERIES PREDICTION OF COVID-19 CASES BASED ON PEOPLE MOBILITY

By

Yihan Mo

Yin Xu

Instructor:

Professor Dr. Anasse Bari

New York University

Fall 2020

**Abstract -** Covid-19, declared as a global pandemic by the World Health Organization, has caused millions of infections and deaths all over the world. How to control the spread of Covid-19 is considered as a prioritized public health problem that is urged to be solved. Considering the spread features of Covid-19, people’s mobility is tracked by previous research to figure out possible actionable prevention. Therefore, this study aimed to examine the relationship and occurrence time lag between people mobility and the number of Covid-19 infected people in Los Angeles, Florida, and New York City. People mobility is measured by the number of car crashes and restaurant reservations recorded from February to September, and the dataset covers complete data offered by CDC, OpenTable, Google Trends, and the official public transportation department from Los Angeles, Florida, and New York City. Granger causality is adopted to determine whether car crashes and restaurant reservations can be used as predictors of change in Covid-19 cases. The polynomial model is further utilized to make a numerical prediction on given variables. The finding suggests that both car crashes and restaurant reservations can serve as reliable predictors of Covid-19 cases with different lags on prediction, and there is a high fitting between predicted results and real-world data. This research might bring an insight into the prevention of the Covid-19 pandemic or other similar infectious diseases by tracking people's mobility. This article's findings will serve as data references for conducting an epidemic prevention plan ahead based on people's daily movements. Further research is suggested to cover a larger scale of real-time data from different locations and periods. Moreover, advanced models are supposed to use for increasing the accuracy and effectiveness of prediction.

Keywords: Covid-19, People mobility, Time series analysis, Granger Causality, Polynomial model

# 0. INTRODUCTION

## Background

The rapid growth of the transportation network and urbanization accelerated people’s connection. The intensive and unprecedented population movement leaves people susceptible to various diseases, providing hotbeds of viruses to spread on a global scale. Coronavirus disease 2019, known as Covid-19, is a contagious respiratory and vascular disease derived from severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Covid-19 can be easily spread through several approaches, mainly through saliva and other bodily fluids and excretions during people’s daily activities, such as breaths, coughs, and talks, sneezes, and other activities that involve excrement. With the feature of being easily and quickly propagated among people, Covid-19 has caused millions of infections and death worldwide, along with global panic and lockout for nearly a year. Indisputably, public health problems and concerns aroused by the contagious disease are prioritized topics nowadays. People's mobility and communication play significant roles in controlling the speed and reach of viruses transmitted globally. Even though Covid-19 can be temporarily controlled by several self-preventions, such as wearing masks in public areas, self-quarantine for at least 14 days after traveling, social distancing, etc., people's social habits could not be changed suddenly or manipulated uniformly, especially in population-dense areas. Therefore, this research aims to reveal and predict how people’s mobility has a relationship with increasing infected cases, contributing alternative data support to further research and policy enact regarding pandemic prevention and medical resource preparation.

## Literature Review

As stated above, how to control and prevent Covid-19 is one of the crucial and urgent public health problems that needed to be solved. Plenty of researchers attach great importance to investigating Covid-19 cases based on different countries and relationships between people’s mobility and Covid-19 spread. Oztig and Askin (2020) utilize the negative binomial regression (NBR) model based on the Poisson-gamma mixture distribution, with airline passengers carried across the 144 countries as an independent variable of measuring people’s mobility and the number of Covid-19 infected case as a dependent variable. The result shows a relationship between the massive scale of human mobility among intensive airline travel and the Covid-19 infected case amount in corresponding countries. However, measuring people’s mobility using the number of the passenger during air travel might not be representative to provide the comprehensive result, since many countries locked nearly all airlines within or cross countries after Covid-19 outbreak intensively. Rather than focusing on air travel, Cartenì, Di Francesco, and Martino (2020) investigate people’ mobility by tracking around 1200 car traffic count automatic sensor data from the Italian Transport Ministry from January 2020 to May 2020, and consider other alternative sources such as PM pollution and Temperature to estimate people’s mobility habits. The result of a multiple linear regression model shows that the correlations between people's frequent mobility and increasing contagious cases with an around 21-day threshold between peaks of people’s movement and Covid-19 positive cases occurred in Italy. Kapoor, Ben, Liu, Perozzi, Barnes, Blais, and O'Banion demonstrate one of the prevalent approaches adopted by researchers for investigating epidemiological modeling is the time-series learning approach, including but not limited to the curve-fitting model, Autorepression (AR), or deep learning on time series data. Zhang (2003) suggests that time series is regarded as a vital forecasting method that relies on using past observations of the same variable to develop a model to explore the underlying relationship among variables.

## Hypotheses

Considering the uncertainty of people' mobility and various uncontrollable variables that might impact the number of infected cases, this research aims to adopt time series analysis to figure out the lag between the number of infected cases and change of people's daily movement, such as frequency of outdoor dining and driving. Therefore, the null hypothesis is using outdoor dining and driving as references to people's mobility is not a suitable predictor to the Covid-19 case. On the other hand, the alternative hypothesis is that people's mobility is a suitable predictor of the Covid-19 case.

# 1. ANALYTICS FRAMEWORK

The data we used in this study is based on public data from local governments and federal health agencies, and Google Trends scores based on personal data search records. Our project aims to investigate the incidence and relationship of people's mobility and the Covid-19 spread. Based on this target, we build our analytics framework as follows:

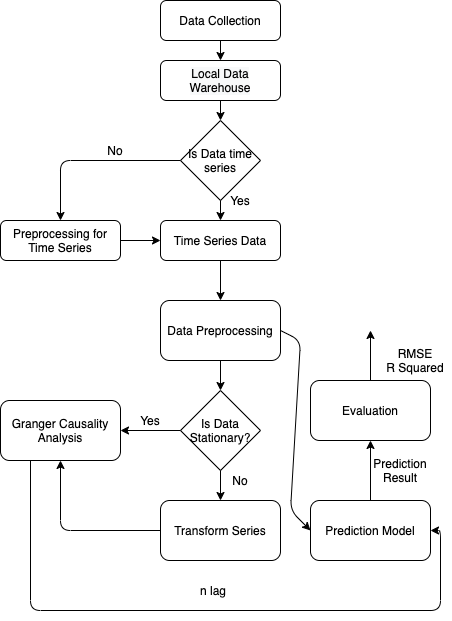


Figure 1: Analytics Process

The framework collects public data and transforms these data into time series. Based on these time-series data, we conduct further analysis. Granger causality analysis can provide evidence of whether people's movement actually affects the spread of Covid-19 and the time lag between them. Based on people's mobility and willingness to wear personal protective equipment, we established a predictive model for the number of Covid-19 cases. We compared the models with and without time lag to determine the time lag's effectiveness from the granger causality analysis. The details will be discussed in the method section.

# 2.DATA

## 2.1 Data Understanding

Considering people’s preferences for transport methods and the overall development of public transportation, this research adopts car traveling as a benchmark to evaluate the frequency of people’s daily movement. Car cash, to some extent, indicates people’s mobility under strict curfew order. Under ordinary circumstances, the more motor vehicle crashes occurred, the more likely that people move outside. Therefore, this research takes motor vehicle records as references to estimate people's mobility. Data is collected from the Traffic Crash Report provided by Florida Highway Safety, Motor Vehicle Collision Crashes offered by NYC Police Department Open Data, and Traffic Collision Data recorded by LA Police Department. In addition to car crashes recorded by different official departments, Google Trends score of the searching keyword 'mask' is used as an indicator of people's willingness to wear face masks from February to September. Beyond that, some places reopened when the number of infected cases decreased, and people began to dine outside, especially during holidays. OpenTable, a real-time online reservation network, records reservations of restaurants across the U.S. The reservation and actual dining data in the restaurants reveal people’s movement among communities. Therefore, this research refers to the restaurant reservation as an alternative data source to track people’s mobility. Regarding the number of infected cases, this research grasps official and real-time cases and deaths data issued by the Center for Disease Control and Prevention (CDC).

## 2.2 Data Preparation

This research's raw data are structured data, which could be directly extracted from various official websites and open data portals. Since car crashes data of LA, FL, and NYC are collected and displayed from different dimensions in the initial website, we converted and fused all car crashes data into time series based on uniform feature selections. Among data cleaning, we noticed that car crashes daily data of FL is decomposed by detailed data recorded by more than three departments from different countries. Therefore, we parsed the number of car crashes from all vehicle or non-vehicle crashes and combined data provided by subsidiaries and sources into the total vehicle crashes. After filtering car crash data of three representative cities, we converted the daily car crashes into time-series format. The restaurant reservation data from OpenTable are shown as the ratios of reservation reduction on a year-on-year basis. Therefore, we converted the reservation reduction rate to the ratio of reservation remaining in 2020 compared with the ratio on the same day in 2019 by adding 100% to all reservation reduction ratios. After initial data cleaning, we converted all data regarding car crashes, restaurant reservation, and positive cases into time series and visualized the comparison of the following combinations: 1) Covid-19 cases versus the number of car crashes; 2) Covid-19 cases versus restaurant reservation; 3) restaurant reservation versus the number of car crashes, of LA, FL, and NYC, respectively.

# 3.MODELING

## 3.1 Data and methods overview

After completing the data conversion and cleaning, we obtained Covid-19 cases, OpenTable restaurant reservation, the number of car crashes, and Google Trends score of 'mask' keyword time series data in three areas.

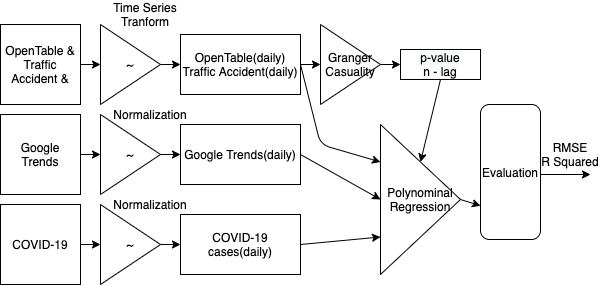


Figure 2: Methodology Diagram

As shown in Figure 2, the modeling proceeds in three phases. In the first phase, for each region we select, we collected the number of Covid-19 cases, and car crashes statistics disclosed by the government and health agencies, the restaurant reservations from OpenTable, and the Google Trends score of the 'mask' keyword in each region, recorded from February 2020. The collected data is reorganized and transformed into a time series.

In the second stage, we proposed the hypothesis that the number of traffic accidents and restaurant reservation is used to predict future Covid-19 cases in a particular area. Granger Causality analysis is applied to test if there is a correlation between Covid-19 cases to car crashes cases and restaurant reservation rate in the past n days.

In the last phase, we apply a polynomial model to determine how the input factors will drive response and the changing direction. In this case, the input factors are lags of change in people's movement and the number of Covid-19 cases. The response is predictive direction and fitting among variables in a given by the total derivative.

## 3.2 Bivariate Granger Causality Analysis of Vehicle Crashes Cases/ OpenTable Restaurant Reservation Rate

Granger causality analysis is used to test predictors of the relationship between traffic accidents and OpenTable versus Covid-19 cases. Granger causality analysis is based on the assumption that if a variable X causes Y, then changes in X will systematically occur before changes in Y. Base on these hypotheses, if the structure of the Granger causality test is significant, though correlation does not represent causation, by granger causality analysis we can answer following questions:

1)Whether OpenTable restaurant reservation or car crashes data can serve as predictors to the Covid-19 cases or not?

2)What is the time lag of prediction using OpenTable restaurant reservation or car crashes data as predictors to predict the Covid-19 cases?

We denoted our Covid-19 cases as , the value represents the Covid-19 case number from day 1 to day t. In the Granger causality test, we selected two different models for comparison, the sampled formula Eq. 1 and Eq. 2 shown:

Where Eq. 1 means only use n lagged value of (,..., ) to predict , which denotes n days lag in this case. Meanwhile, Eq. 2 introduced a new variable, which is the OpenTable or car crashes time series data in this case. We denote the new variable as . Beyond Eq.1, Eq.2 using (,..., ) and (,..., ) to predict Ct together. According to the definition of Granger Causality, if the results from model Eq.2 are statistically significantly better than the result from model Eq. 1, then we can believe that variable provides predictive information for .

Based on the result of Granger causality analysis, if the p-value is lower than 0.05, then we can reject the null hypothesis that the OpenTable or car crashes do not predict Covid-19 cases values. In this case, both OpenTable and traffic accidents provide predictive information to Covid-19 cases. We noticed that the OpenTable has a Granger causality relation with Covid-19 cases with lags ranging from 19-22 days in all three areas. The car crashes has a Granger causality relation with Covid-19 cases with lags in 16 days in all three areas.

Table 1:

Lag of OpenTable vs Covid-19 Cases

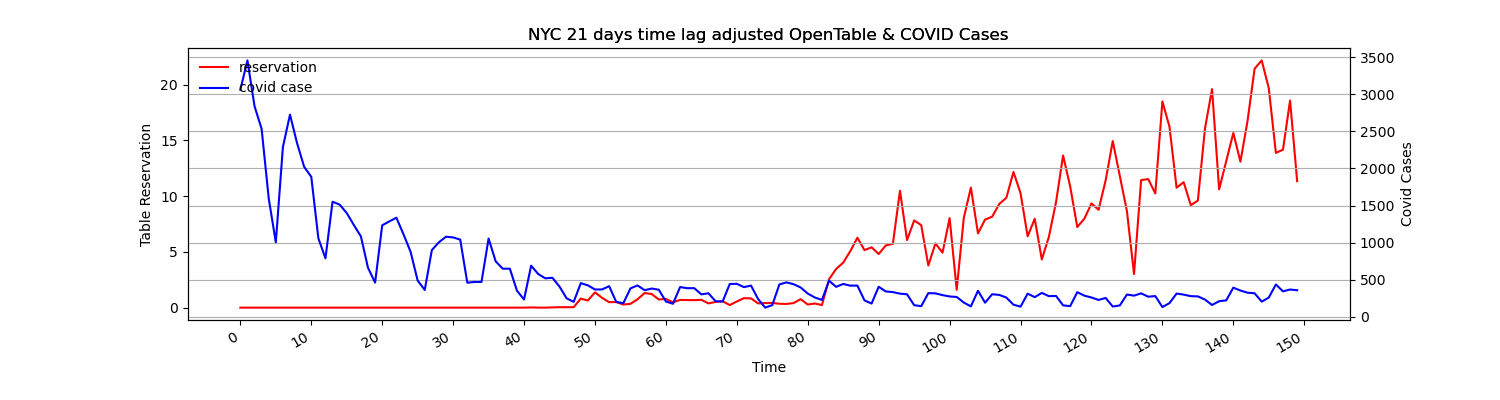
|  |  |  |  |
| --- | --- | --- | --- |
| lag/days | NYC | FL | LA |
| 18 | 0.0000\*\* | 0.0154\*\* | 0.0993\* |
| 19 | 0.0000\*\* | 0.0065\*\* | 0.0446\*\* |
| 20 | 0.0015\*\* | 0.0137\*\* | 0.0075\*\* |
| 21 | 0.0003\*\* | 0.0014\*\* | 0.0074\*\* |
| 22 | 0.0003\*\* | 0.0015\*\* | 0.0042\*\* |

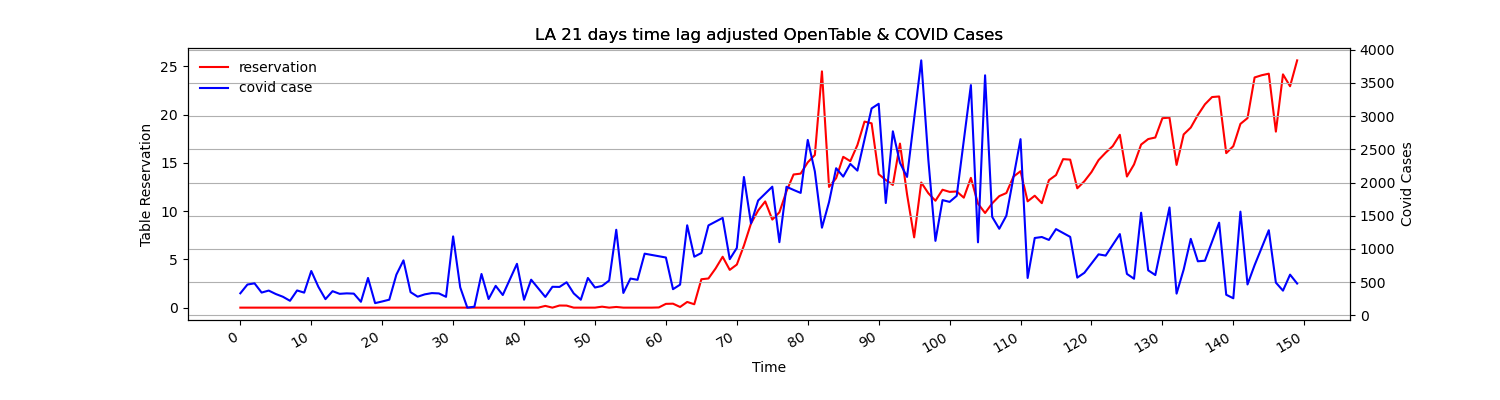
Table 2:

Lag of Car Crashes vs Covid-19 Cases

|  |  |  |  |
| --- | --- | --- | --- |
| lag/days | NYC | FL | LA |
| 15 | 0.0001\*\* | 0.1661 | 0.1661 |
| 16 | 0.0020\*\* | 0.0262\*\* | 0.0384\*\* |
| 17 | 0.0027\*\* | 0.0414\*\* | 0.1096 |
| 18 | 0.0023\*\* | 0.0618\* | 0.1372 |
| 19 | 0.0081\*\* | 0.0852\* | 0.1611 |

To draw a direct observation of the time lag between Xt and Ct, we replot the series of Covid-19 cases vs OpenTable and car crashes. In Figure 3, the is replaced by , where n is 21 days for OpenTable, and Covid-19 cases and n is 16 for car crashes and Covid-19.





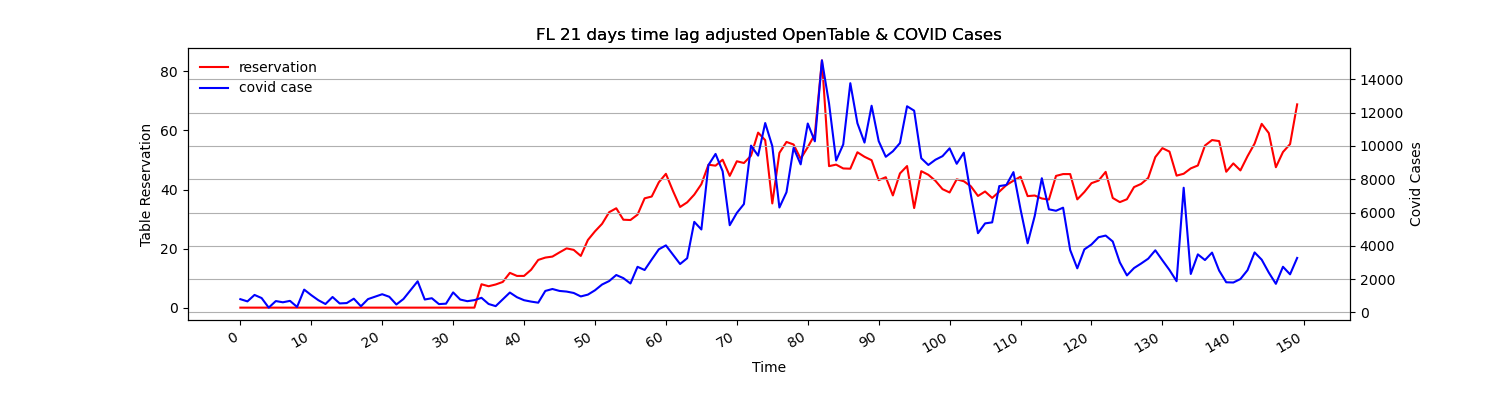
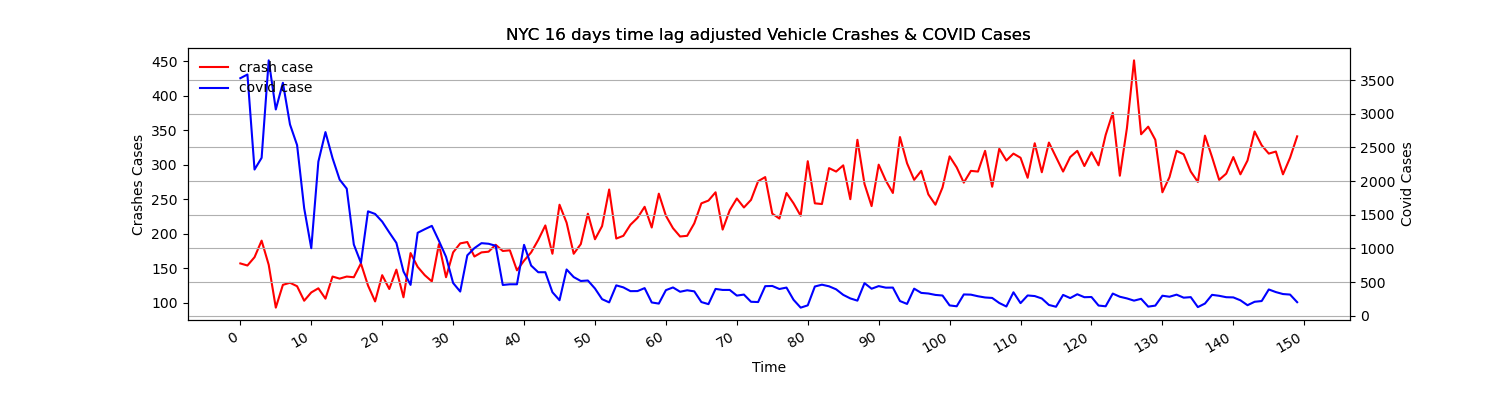


Figure : OpenTable vs COVID-19 Cases (lag = 21)



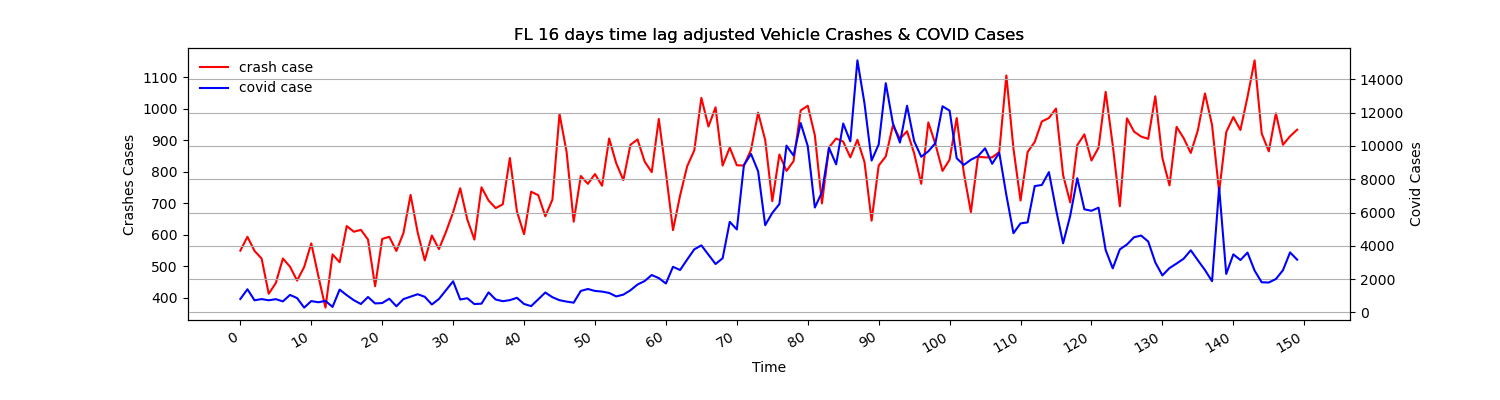
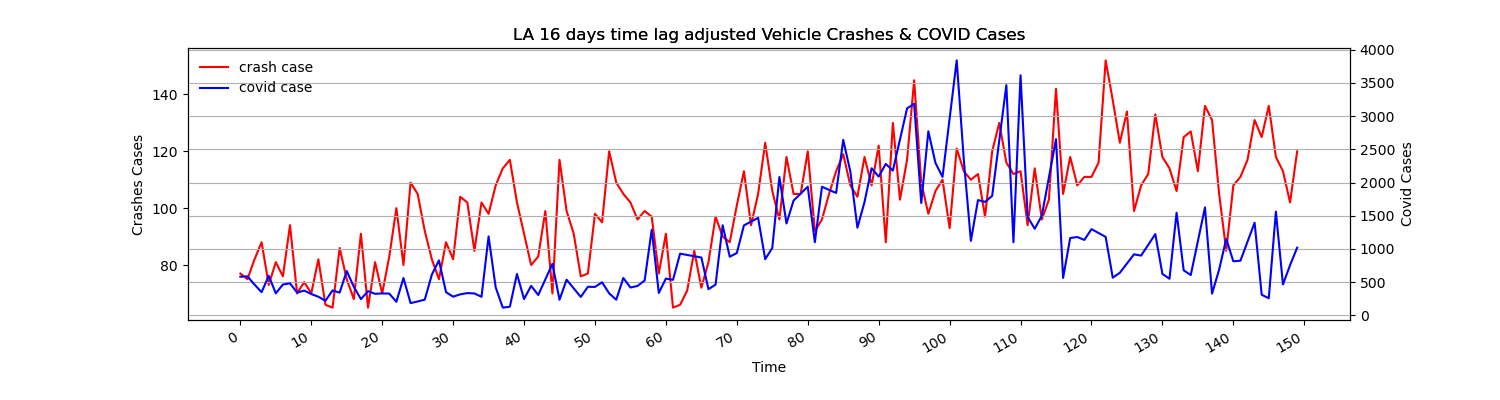


Figure : Vehicle Crashes vs COVID-19 Cases (lag = 16)

According to Figure 4 above, especially in Florida and Los Angeles, there is a correlation between OpenTable () and traffic accidents () vs Covid-19 case (), with similar tendencies in the movement. They have predictive value regarding Covid-19 cases.

Compared with Florida and Los Angeles, OpenTable () and traffic accidents () are less similar to the trends of Covid-19 cases () in New York City. The possible explanation for this situation lies in the strength of implementing epidemic prevention and control measures in different regions and people's attention to the epidemic. New York City experienced a severe pandemic in the early stage of the epidemic and therefore implemented more stringent prevention and regulations to prevent the further spread of Covid-19. As evidence, compared with the other two regions, New York's restaurant reservations and traffic accident recovery are the slowest. Thus, these measures may have significantly suppressed the spread of Covid-19.

## 3.3 Polynomial model for Covid-19 Cases prediction

Based on the granular causality analysis results, OpenTable and traffic accidents can be used to predict Covid-19. However, this conclusion has shortcomings in two aspects. First of all, Granger causality analysis is based on linear regression, and from the results of data visualization, we can easily find that the trends of OpenTable, traffic accidents, and Covid-19 are all non-linear. On the other hand, although granger causality analysis can verify the correlation between variables, it cannot make numerical predictions.

In order to further improve our predictive analysis process, we used non-linear models to make numerical predictions on Covid-19 data. Multiple models would make numerical predictions; however, considering the tendency of data is non-linear, and the volume of data is relatively small, we choose the polynomial regression in this case. The input of modeling will be OpenTable, traffic accidents, Google Trends score of keyword 'mask' and Covid-19 cases, and the time lag we find out in the last section. For polynomial regression, higher degrees will bring higher R squares in the training set, but it will also lead to overfitting. Setting the model to a relatively low degree can avoid overfitting. To balance the performance of the model on the training and testing set, we select that degree = 3 as a suitable choice.

For the data of three different cities, we used the following input to verify the effectiveness of the model：

Covid-19() represents the Covid-19 cases of day , OpenTable () represents the OpenTable reservation rate value of day , Traffic Accident () represents the traffic accident case value of day and Google Trends () represent the google trends score of day .

# 

Figure : Prediction with 21-day lag on OpenTable vs 16-day lag on Vehicle Crashes

From our polynomial regression model's visualization result, we noticed a significant correlation between our prediction result and the real-world case number. In the next section, we will discuss the performance of our model and its improvement compared to the model that does not consider the time lag.

# 4.Evaluation

Our evaluation of the accuracy of the model is measured based on the root mean squared error (RMSE) and R squared of the model. RMSE is used to evaluate the deviation between the model's predicted value and the actual value. On the other hand, the R squared is a measurement to evaluate the proportion of the variance in the dependent variable that is predictable from the independent variable. Based on our results, the model that considers time lag obtains better results than the model that does not consider time lag in the two measurements.

We evaluated the result of the model by split validation, and the proportion of train/test is 90/10.

Table 3: Evaluation Result for model with and Without Lag

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | With Time Lag | | | Without Time Lag | |
| Cities |  | REMS | R2 | | REMS | R2 |
| NYC | Training Set | 417.87 | 0.925 | | 681.40 | 0.803 |
|  | Testing Set | 590.89 | 0.897 | | 727.21 | 0.620 |
|  |  |  |  |  | |  |
| LA | Training Set | 457.320 | 0.663 | | 516.961 | 0.579 |
|  | Testing Set | 549.687 | 0.702 | | 484.90 | 0.456 |
|  |  |  |  | |  |  |
| FL | Training Set | 1384.488 | 0.861 | | 2237.962 | 0.596 |
|  | Testing Set | 1183.600 | 0.701 | | 2758.685 | 0.353 |

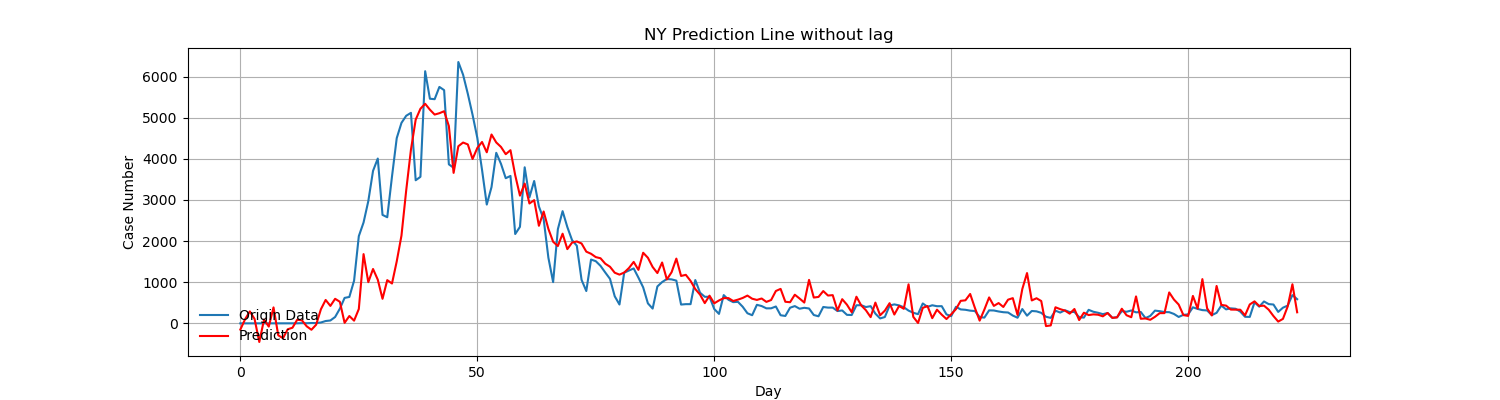


Figure : NYC Prediction vs Original Data without Lag

For the model with time lag for NYC, we get REMS = 417.87 and R squared = 0.925 in the training set, REMS = 590.89 and R squared = 0.897 in the testing set. Against the model without time lag, get REMS = 681.40 and R squared = 0.803 in the training set, REMS = 727.21 and R squared = 0.620 in the testing set. Both RMSE and R squared in the model considering time lag is better than the model not considering time lag. In Los Angeles and Florida, we get a similar result.

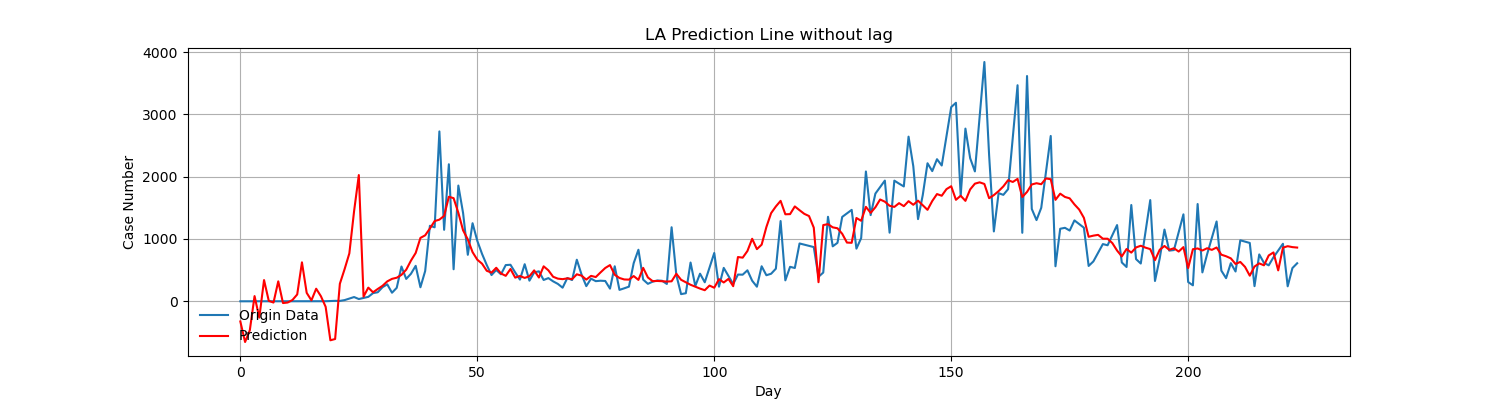


Figure : LA Prediction vs Original Data without Lag

For the model with time lag for Los Angeles, we get REMS = 457.320 and R squared = 0.663 in the training set, REMS = 549.687 and R squared = 0.702 in the testing set. Against the model without time lag, get REMS = 516.961 and R squared = 0.579 in the training set, REMS = 484.90 and R squared = 0.456 in the testing set.

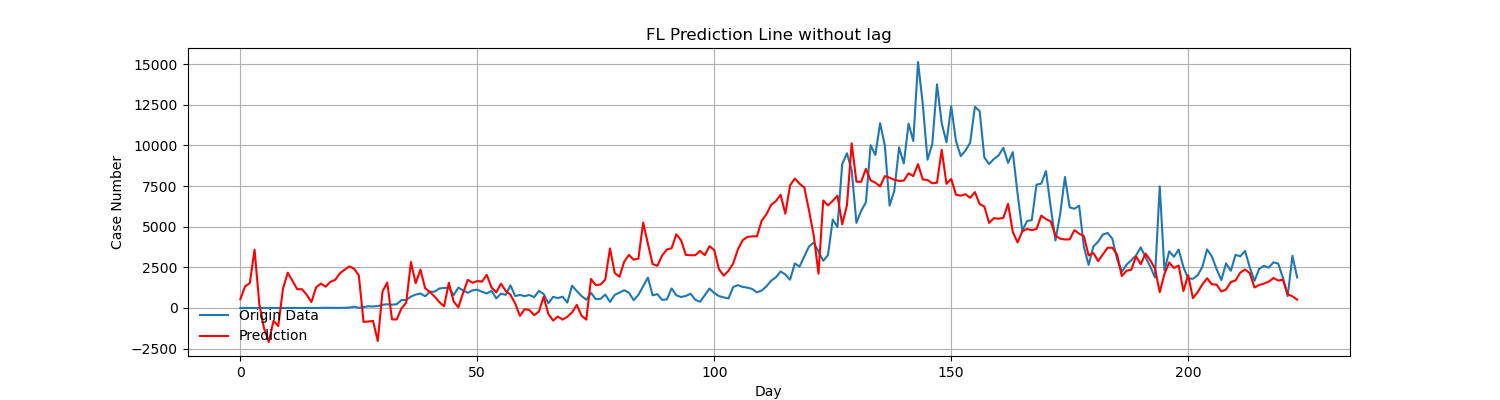


Figure : FL Prediction vs Original Data without Lag

For the model with time lag for Florida, we get REMS = 1384.488 and R squared = 0.861 in the training set, REMS = 1183.600 and R squared = 0.701 in the testing set. Against the model without time lag, get REMS = 2237.962 and R squared = 0.596 in the training set, REMS = 2758.685 and R squared = 0.353 in the testing set.

As a result of our evaluation, the time lag between Covid-19 cases and traffic accident cases and the time lag between Covid-19 cases and OpenTable reservation rate, improves the prediction accuracy of the polynomial model of predicting Covid-19 cases.

# 5.Discussion

From the result shown above, car crashes and restaurant reservations can be referenced as reliable predictors of Covid-19 cases. This research provides data support for further research on predict Covid-19 or other infectious diseases with similar spread features. By tracking people's mobility, the researcher might generate a possible trend of how diseases spread and provide actionable suggestions for the public or government to conduct precautionary measures, such as medical supply preparation before the holiday when people move frequently and broadly. Despite the result with high accuracy between prediction and real-time data, several limitations should be addressed. Firstly, personal mobility is confidential data, usually recorded and monitored by governmental departments using a Satellite GPS or public transportation system. Even though governmental departments have open data to the public, it is quite challenging to tell data integrity and effectiveness. Secondly, a limited and small amount of dataset might distort the accuracy of the prediction model. Lastly, there are other moderator variables that affect the spread of diseases, such as pollution density, prevention measures, unpredictable gathering events, integration frequency, and so forth, and different factors that vary from different locations and periods. Therefore, further study is supposed to consider and rank various moderator variables based on significance in the model's design. Moreover, the time series forecast for a given location is highly dependent on observed patterns of data. Advanced models, such as Bayesian vector auto-regression, dynamic stochastic generalized equilibrium model, deep learning, neural networks, etc., are supposed to analyze more accurate real-time data and consider temporal and spatial interactions for the spread of infectious disease.

# 6.Reference

Cartenì, A., Di Francesco, L., & Martino, M. (2020). How mobility habits influenced the spread of the COVID-19 pandemic: Results from the Italian case study. *Science of the Total Environment*, *741*, 140489.

Kapoor, A., Ben, X., Liu, L., Perozzi, B., Barnes, M., Blais, M., & O'Banion, S. (2020). Examining covid-19 forecasting using spatio-temporal graph neural networks. *arXiv preprint arXiv:2007.03113*.

Oztig, L. I., & Askin, O. E. (2020). Human mobility and coronavirus disease 2019 (COVID-19): A negative binomial regression analysis. *Public health*, *185*, 364-367.

Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, *50*, 159-175.