Research report

**Introduction**

This research aims to examine how the pandemic behave across different counties. Four counties are selected based on their demographic similarity and dissimilarity. They are Los Angeles County(CA), Maricopa County(AZ), New York County(NY) and Burlington County(NJ). In step one, I used Gompertz functions and SIR model predict the spread of COVID-19. In step two, I run regression to identify any influential supply chain factor to the spread of the pandemic. Both R and python are used to build models and visualize the results.

**Gompertz Model**

Let’s first take Maricopa County as an example.

A screenshot of a cell phone

Description automatically generated

This simple line chart shows that the number of confirmed cases in Maricopa is still increasing rapidly and it hasn’t reached the upper flat line yet. I implemented the Gompertz function in Rstudio to get the best predicting parameters (see below).

A screenshot of text

Description automatically generated

Gompertz model10 = "y~a\*exp(-b\*exp(-c\*x)

Gompertz function is involved with exponential distribution. So, when my data includes some large values, I think some "infinity" values have been produced, which can be a problem. Then I standardized the data (with scale), the model provides some results. It indicates that starting from time 3.5, the number of confirmed cases stays at 3.25. But standardizing the data brings some difficulties in real-life prediction. I’m not sure if I should standardize my data as my aim is to make predictions.

Then I switched to scipy package to make prediction in python. Based on this model, the number of confirmed cases in Maricopa is estimated to increase to 3571 by June 1st.

A close up of a map

Description automatically generated

**SIR Model**

The SIR model indicates that someday after the 100th day, the number of confirmed cases will stay stable as the majority of population have been infected. With such a high transmission rate, keeping social distance is extremely important to decrease contact ratio( the fraction of population that comes into contact with infective individuals within their period of infectiousness).

A close up of a device

Description automatically generated

**Regression**

I run regression with miles of freight railroad, the number of major airports, the number of healthcare workers, the hourly mean wage of healthcare workers, and population density as independent variables and with the number of confirmed cases as dependent variables. Unfortunately, the results show that none of the independent variables are significant in influencing the number of confirmed cases (see below). I'm going to explore and include more supply chain factors in the regression to see if any of them would influence the spread of the pandemic.

A screenshot of a cell phone

Description automatically generated