PROJECT REPORT

Effect of Mobility on Covid-19

CHE398A (UGP-1)

April 30, 2021

Varun Gupta
Indian Institute of Technology, Kanpur
Project Supervisor : Dr. Himanshu Sharma & Dr. Harshwardhan H. Katkar
Department of Chemical Engineering
Indian Institute of Technology, Kanpur

Abstract

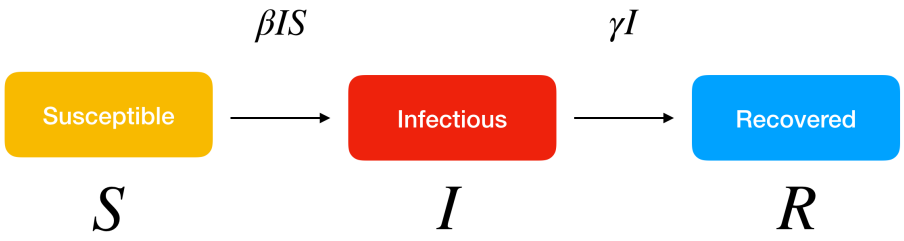
The outbreak of novel coronavirus, COVID-19, has been declared a pandemic by the WHO. Covid-19 virus is spread via contact with a carrier or infected individual. Therefore is absolutely trivial that rate at which the virus is expanding is a direct result of human interactions. Mobility is the method to associate human interactions with quantifiable numbers representing the total contacts made by the given population. Scientific modelling has been widely used on the infection data-sets available inorder to devise public policies and to help institutions. Their are certain compartmental models available for epidemiology but factors such as mobility are not captured very well. This study aims to create mobility time series and to suggest improvements into the current SIR model to make it robust.

1. Introduction

Inorder to get started with the study we first need to understand the underlying process to which we are trying to suggest improvements. The model used in the accompanying project is a simple Susceptible-Infected-Recovered Model which is one of the many epidemiology models used.

1.1 SIR Modeling

If we consider the population to be finite, people can be categorized into various compartments such as - susceptible, infected, recovered. This is a basic SIR classification and complexity of this can be increased by addition of other compartments specific to the pandemic. Currently we can think of adding a Asymptotic compartment where the person acts as a carrier and does not reflect any symptoms. Further inclusion of birth and death rates can be accounted into the model. The compartments interact with each other in first-order differential manner and can be treated as a rate equation with the concentration of the compartment indicated by the number of species present.



$$\frac{dS(t)}{dt} = -\beta S(t) I(t) \tag{1.1}$$

$$\frac{dI(t)}{dt} = \beta S(t) I(t) - \gamma I(t) \tag{1.2}$$

Beta (β) measures the effective transmission rate whereas **Gamma** (γ) represents the recovery rate for the model. Measuring these two values we define another important parameter which is the **Reproduction Number (R_0)**

$$R_0 = \frac{\beta}{\gamma} \tag{1.3}$$

This represents the average number of people infected by a single individual which helps us determine how infectious this virus is in local population.

The classical epidemic model does not capture the fact that mobility has increased multi-folds and plays an important role in determining rates for the model.

1.2 Mobility

Since the virus is mainly spread due to close contact with infected individual it is important to restrict exposure of such kind of interactions. Public policies such as **lockdown** for short or even extended period reduces the spread of virus to a great extent and prevent an outbreak. The major source of social contacts in day to day life includes -

- Workplace (Office places, Hospitals etc.)
- School
- Home
- Local markets (Grocery shops, Shopping Malls etc.)
- Recreational Regions (Local parks, Zoological park etc.)

The above list is not exhaustive but gives us the major proportion of the total contacts. We will make use of the data-set provided by Google to map this mobility to a time-series representing the total contacts made on a particular day for every district/city within our study.
We will also try to suggest certain improvement to the model inorder to incorporate the mobility time-series.

2. Mobility Data-set

Their has been enormous advancement in terms of technology in the past few decades. Global Positioning System (GPS) has been the key to the Multinationals such as Google and Apple to approximately track crowd movement locally. Both these companies have released their separate mobility data-sets specifically for research studies. We will be using Mobility data-set by Google since it covers wider categories and more number of districts specifically in India.

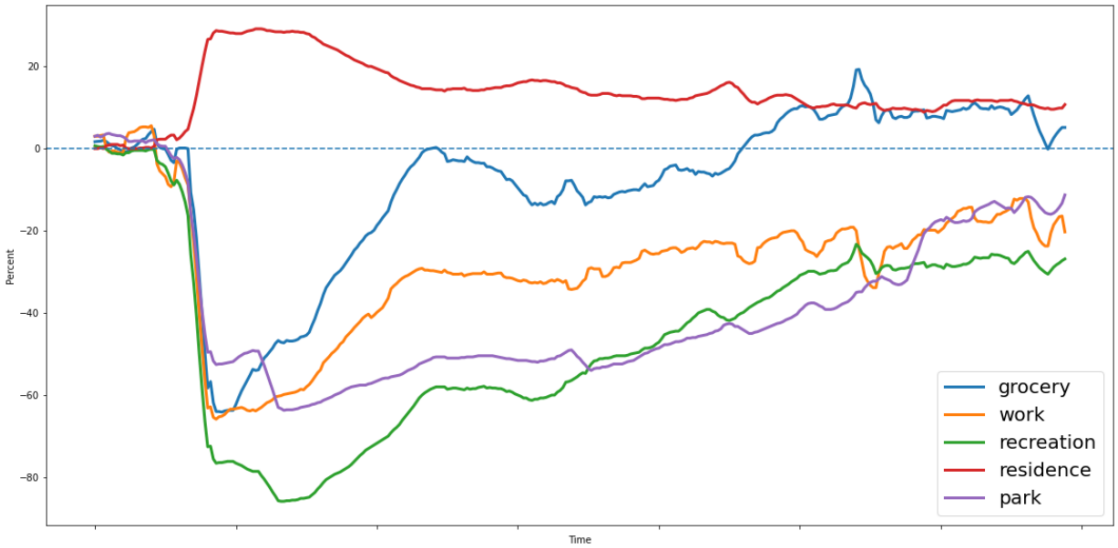
2.1 Data Overview

The data-set is generated using a similar mechanism used in Google Maps to display traffic conditions. To maintain user privacy, Google makes use of aggregated, anonymized data. It analyses the trends in visit to high level places mainly -

- Workplaces
- Retail and Recreational
- Grocery and Pharmacy
- Transit stations
- Parks
- Residential Areas

The data is provided from *15-Feb-2020* and it is updated in every 24 hours. The numbers provided in each category are the percentage change to the particular baseline values determined by Google and not released in public due to privacy constraints.
Note - Every column mentioned above except Residential areas indicates a change in the number of visitors to places belonging to that category, whereas Residential maps the difference in duration of people residing in their home.

We require absolute numbers to feed into the model and make sense of the data. We need to construct a single mobility index using the various categories provided to us which we will discuss later in the study.



The above diagram is the variation in the various categories in the given data-set. The above trend-line indicative of the various public policies implemented to reduce social contact. Certain trends observed are -

- The increase in the time spent in Residential areas shows a peak during the months of April to July indicating that India was in a lockdown period with people spending most of the time in their home.
- All the other trend-line show a steep decrease which is also as expected since these activities required movement outside our home which was prohibited during that time.
- It is also interesting to note how the trend-line start reaching about 80% - 90% of the baseline value indicating the upliftment of lockdown in various cities.

2.2 Baseline

This is one of the important factor that we need to determine in order to construct the time series. This values has not been provided by any similar data-sets available, so it is not possible to switch to any other data-set.
The baseline is the median value obtained in a 5 weeks period starting from *03-Jan-2020* to *06-Feb-2020*. Every other value is the relative change to this particular value. The baseline is different for each of the categories.
Intuition - Going through the process we can assert that baseline value is an attempt to get the total number of contacts in a particular day, when the conditions were normal.

Underlined Assumptions - Mobility is of two kinds - inter-city and intra-city. We are currently focusing on the intra-city mobility since this plays a major role in the growth of virus in local population. It is difficult to make a reasonable prediction for the intercity travel with the current model. Since we are not using inter-city mobility, the entries for *transit stations* in data-set will not be used.

3. Initial Contact Rates

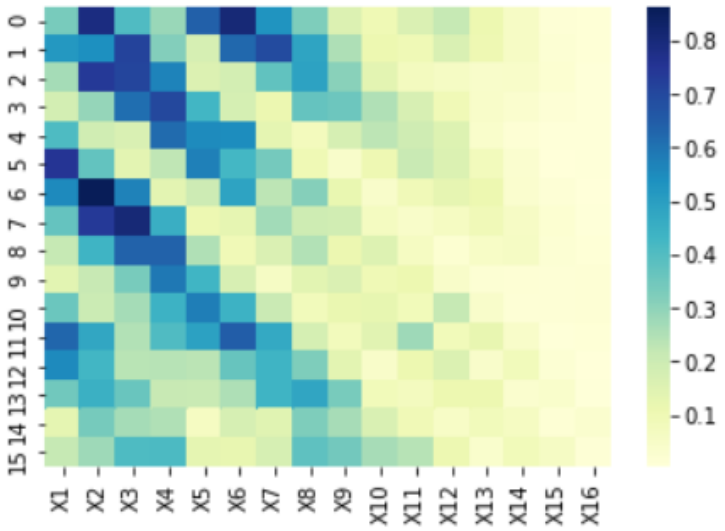
The most important part in modeling mobility is to estimate a model for the initial contacts before the pandemic. Intuitively, the major dependant variables for predicting total contacts (ϕ) are - Population (P) and Population Density (ρ)

$$\phi = f(P, \rho) \tag{3.1}$$

The above relation can be linear, polynomial or even exponential. We will be making use of separate data-set in order to create the model.

3.1 Data Overview

The data used is an outcome of a research study conducted to incorporate various factors such as age to introduce heterogeneity into the epidemic models. The data-set comprises of contact-matrix for 152 different countries. To generate the data for each and every country, they conducted surveys in certain countries and then expanded the data with help of *Markov Chain Monte Carlo simulations* and validated it differently for each and every country.



The above diagram is for the number of contacts in Residential Areas in India.

The diagram gives us the correlation for the number of contacts made by people of different ages with each other. Both the x as well as the y axis have been divided into certain age bins -

$$\begin{bmatrix} x1 \\ x2 \\ . \\ . \\ x16 \end{bmatrix} \Longrightarrow \begin{bmatrix} 0-4 \\ 5-9 \\ . \\ . \\ 75-79 \end{bmatrix} \tag{3.2}$$

3.2 Preprocessing

The data is distributed into 4 major categories - Work, School, Home and Others. We need to calculate the total number of contacts made by the population in a day.

$$\phi^k = \sum_{i=1}^{16} \left(\sum_{j=1}^{16} M_{i,j}^k N_i \right) \tag{3.3}$$

$$k \in (Home, Work, School, Others)$$

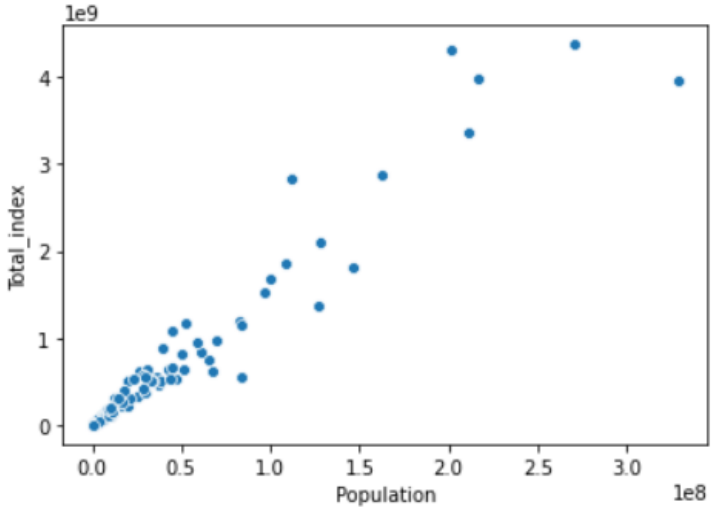
Here $M_{i,j}$ represents the value in the contact matrix and N_i represents the total population for that particular country in i^{th} age-bin.

Since, $M_{i,j}$ indicates the number of contacts made by a single individual belonging to the Age bin (i) with any individual from the Age bin (j) over a period of time. Therefore if we multiply $M_{i,j}$ with N_i we will get the total number of contacts made by X_i with X_j.

3.3 Modeling and Observations

After running the calculations we were able to generate a new data-set for each of the 4 categories mentioned previously. Since we initially assumed that total contact rate is a function of Population and Population density, we have augmented the data-set with both these values for all the countries.
The diagram below is a scatter plot of Total contact rate against population of the all the countries.

Note - The graph does not include India and China due to the enormous population as compared to the other countries in the data-set to observe the trend followed by the majority of the countries.



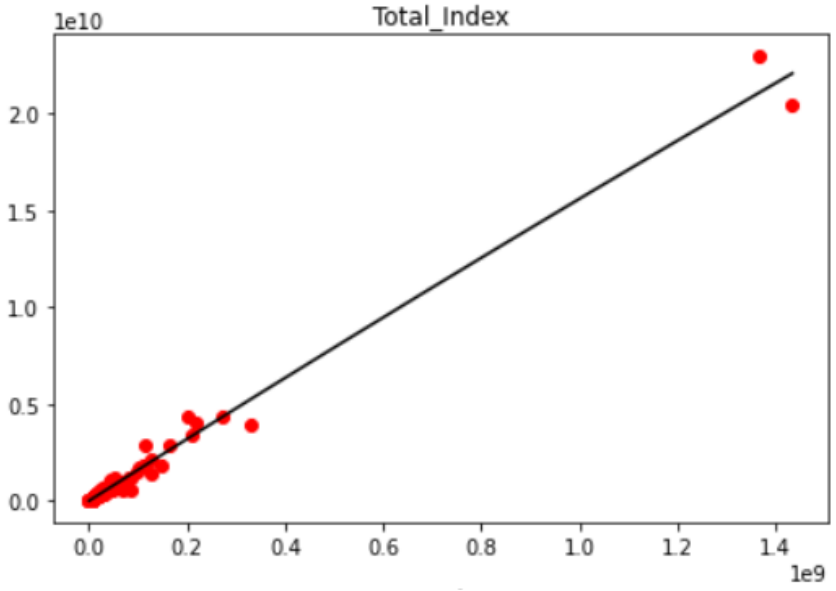
Population Density - In spite of the strong intuition this does not play an important role in our model since it does not correlate with the predictive variable. One of the major reasons is the fact that we are considering country wide data and population density is not exact. The population density must be much higher due to the presence of uninhabitable land decreasing the actual number. Also the fact that the population is not distributed homogeneously creates a problem while accounting this term. This factor can come into play when we try to design for smaller areas such as individual cities where we can assume that the local population is evenly distributed.

Since we now know that contact rate is a function of Population alone, we need to find if this correlates with its higher order terms. The data-set is subjected to various modeling techniques revolving around regression. It was observed that for higher order terms of population (≥ 3) the accuracy dropped instead of increasing, therefore the dependant variables chosen were -

- Population (P)
- Population squared (P^2)
- Population Density (ρ)

The model was trained for the various permutation on the above mentioned dependant variables. *Polynomial Features*, a technique in machine learning was used to include P^2 within the regression models used. Various regression techniques such as - Linear, Non Linear, Polynomial, Decision tree, Lasso etc. was done. Polynomial Regression reached the highest accuracy of **95.87%** on the testing data-set.

$$\phi = f(P, P^2)$$



This model predicts the baseline when we feed in the population of the particular location. This completes the baseline modeling for the mobility time series.

4. Mobility Time Series

After obtaining the model for the baseline we will be building mobility time-series for India - entire country and districts/cities that are used in the corresponding study conducted. The contact matrix values are mapped to the column in the mobility data-set as follows -

- Home baseline (ϕ^H) - It represents the Residential Area (μ^H)
- Work baseline (ϕ^W) and School baseline (ϕ^S) - It represents the Workplaces (μ^W)
- Others baseline (ϕ^O) - It represents Groceries (μ^G), Parks (μ^P) and Recreational (μ^R)

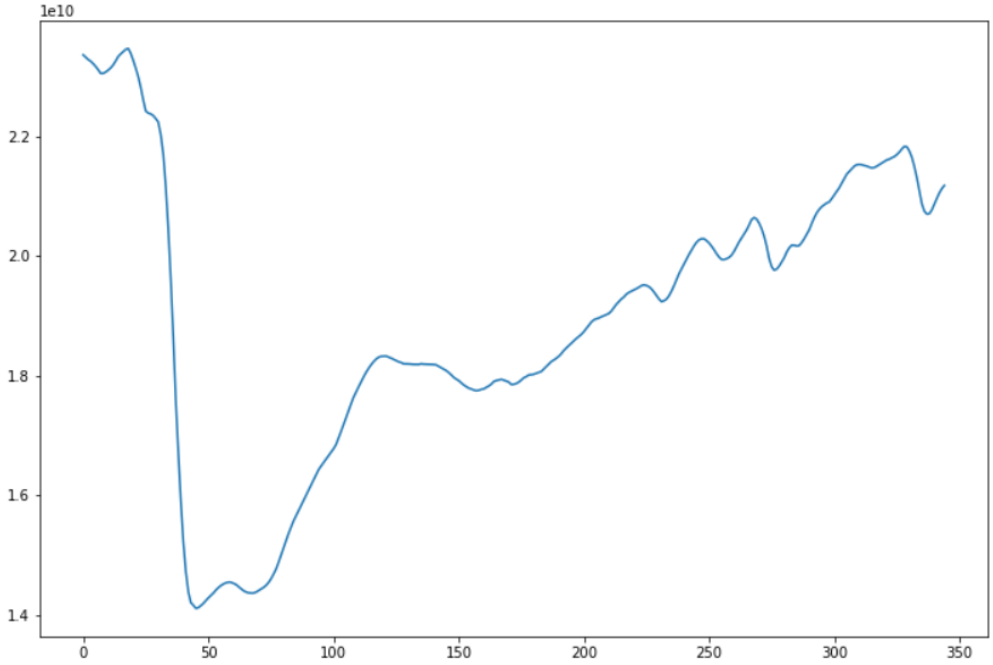
The terms mentioned in the left are obtained by contact matrix data-set and the term on the right are obtained from google mobility data-set.

$$\phi_k^T(t) = \left(\frac{100 + \mu_k^H(t)}{100}\right)\phi_k^H + \left(\frac{100 + \mu_k^W(t)}{100}\right)\left(\phi_k^W + \phi_k^S\right) + \left(\frac{300 + \mu_k^G(t) + \mu_k^P(t) + \mu_k^R(t)}{300}\right)\phi_k^O \tag{4.1}$$

$\phi_k^T(t)$ represents the mobility index or the total contacts made on the particular day and k represents the city/district.
After obtaining the time-series we apply a 7 day **moving average filter** to smooth the curve and to reduce the abrupt peaks within the data. Their are different kinds of these filters we used a filter which takes into consideration both future and past values.

$$\phi_k^{Tnew}(t) = \frac{\sum_{i=-3}^{3} \phi_k^T(t+i)}{7} \tag{4.2}$$

While implementing this we took care of the fact that it does not go out of bounds for the initial or the final values and therefore tweaked this setup to ensure that.



The above graph is a visualization of the time series obtained for the whole country. We can observe the effect created due to the lockdown in India during the months of April - September. This also helps us visualize the upliftment of lockdown throughout the country after this period.
Similar Time-series have been obtained for different cities in India.

4.1 Model Improvements

$$\frac{dS(t)}{dt} = -\beta \phi_k^{Tnew}(t) S(t) I(t) \tag{4.3}$$

$$\frac{dI(t)}{dt} = \beta \phi_k^{Tnew}(t) S(t) I(t) - \gamma I(t) \tag{4.4}$$

We introduce the mobility time-series index in the infection rate itself to optimize the model. Studies have also shown that their is a certain delay in observing the effect of mobility termed as incubation period which is estimated about 14 days. Another approach could be to use $\phi_k^{Tnew}(t-14)$ instead of $\phi_k^{Tnew}(t)$.

Future Work

- One of the major portions which we were not able to cover was Intra-city mobility. This area has to be explored in detail in both perspective - Modeling as well as its effect on existing Covid-19 model.
- Adding heterogeneity to the compartments in the model such as - Age and gender is also a promising direction to increase the robustness of the model and to give better predictions.
- Another approach on which we have started working is to make use of the average of the mobility time-series inorder to extract the relative values of beta and to see if it correlates with the findings we obtain using the Covid-19 model.

References

Google COVID-19 Community Mobility Reports: Anonymization Process Description. [arXiv:2004.04145v4]

Rajesh Singh and R. Adhikari : Age-structured impact of social distancing on the COVID-19 epidemic in India. *University of Cambridge* [arXiv:2003.12055v1]

Kiesha Prem, Alex R Cook and Mark Jit : Supporting Information Projected contact matrix in 152 countries for models of contact-transmissible infectious diseases.

Karima Nigmatulina, Philip Eckhoff and Hao Hu : The scaling of contact rates with population density for the infectious disease models

Reza Sameni : Mathematical Modeling of Epidemic Diseases; A case study of the COVID-19 coronavirus [arXiv:2003.11371v4]

Eran Tock, Boaz Lerner and Eyal Ben-Zlon : Analyzing Large Scale Human Mobility; a survey of machine learning methods and application.

GitHub repository for Data-set and Python Code:
`https://github.com/varungupta333/Effect-of-Mobility-on-Covid-19`