CS-595 Optimization Algorithms with ML Apps

Tushar Nitave, Naunidh Singh

Modeling the spread of COVID-19 propagation to analyze the efficacy of non-pharmaceutical measures

Abstract

In this project we aim to develop an SIR based model to estimate the spread of the coronavirus disease and evaluate the effects of government interventions in the United States. Government policies such as mask mandate, social distancing, mobility of the community are taken into consideration. We plan to show how the enforcement of these policies by government and acceptability by people has reduced the growth rate of infection. We study the particular case of the state of Illinois and plan to extend this study to other states.

Introduction

The novel coronavirus disease 2019 was announced as a pandemic by the World Health Organization (WHO) on March 11, 2020. In response to this pandemic several governmental interventions have taken place for reducing the growth rate of infection.

Even when policies to control the spread of COVID-19 have been enforced there are uncertainties regarding the efficacy of these policy measures because of hesitancy towards adhering to the rules and other behavioral responses. For example, a mask mandate may not be effective in practice, if its acceptability among the community is very low. And on the other hand, if the preventive measure is not sufficient, it may lead to increased mobility due to a false sense of security.

Mathematical modeling has proved to be an important decisive tool for controlling the spread of human as well as animal diseases [1]. Various mathematical models have described the dynamics of the growth of this infection. Three such models are presented in [2]. Our objective is to assess the direct benefits of non-pharmaceutical policies by estimating how much these policies have slowed the growth rate of infections by optimizing the parameters of our epidemiological model.

Literature Review

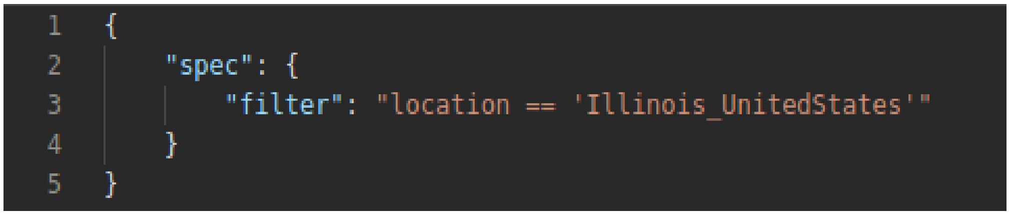
In order to stop/reduce the grow rate of infection in the given pandemic due to COVID-19 several measures have been enforced by the government of United States. Roughly 18 policies are enforced which include school closure, travel restrictions, social distancing etc.[4]. An SIR model was proposed in [5] to measure the spread of COVID-19 in Malaysia before and after Movement Control Order (MCO) and concluded that MCO significantly reduced the count of susceptible and infected. In [6] direct health benefits of these policies were measured by comparing the growth rate of infections using reduced-form econometric technique. Various policies such as school closures, work from home, religious closures were incorporated.

Authors in [7] showed that social distancing is a necessary measure to minimize the spread of infections and that combination of centralized and decentralized government policy shows optimal results. Effects of population-wide social distancing, social distancing of specific subsets of population and impact of relaxing these policies was evaluated using stochastic, individual-based transmission model of COVID-19 [8]. The effect of human mobility along with government measures on the growth rate of infection was analyzed. Three different Generalized Linear Models (GLM) were used to evaluate this hypotheses [9]. The proportion of transmission that occurs via asymptomatic transmission is also a useful statistic to summarize the feasibility of intervention measures [10]. An age-structured susceptible-exposed-infectious-removed was used to measure the effectiveness of social distancing and results showed that early government interventions delayed the epidemic curve [11].

Data modeling and preprocessing

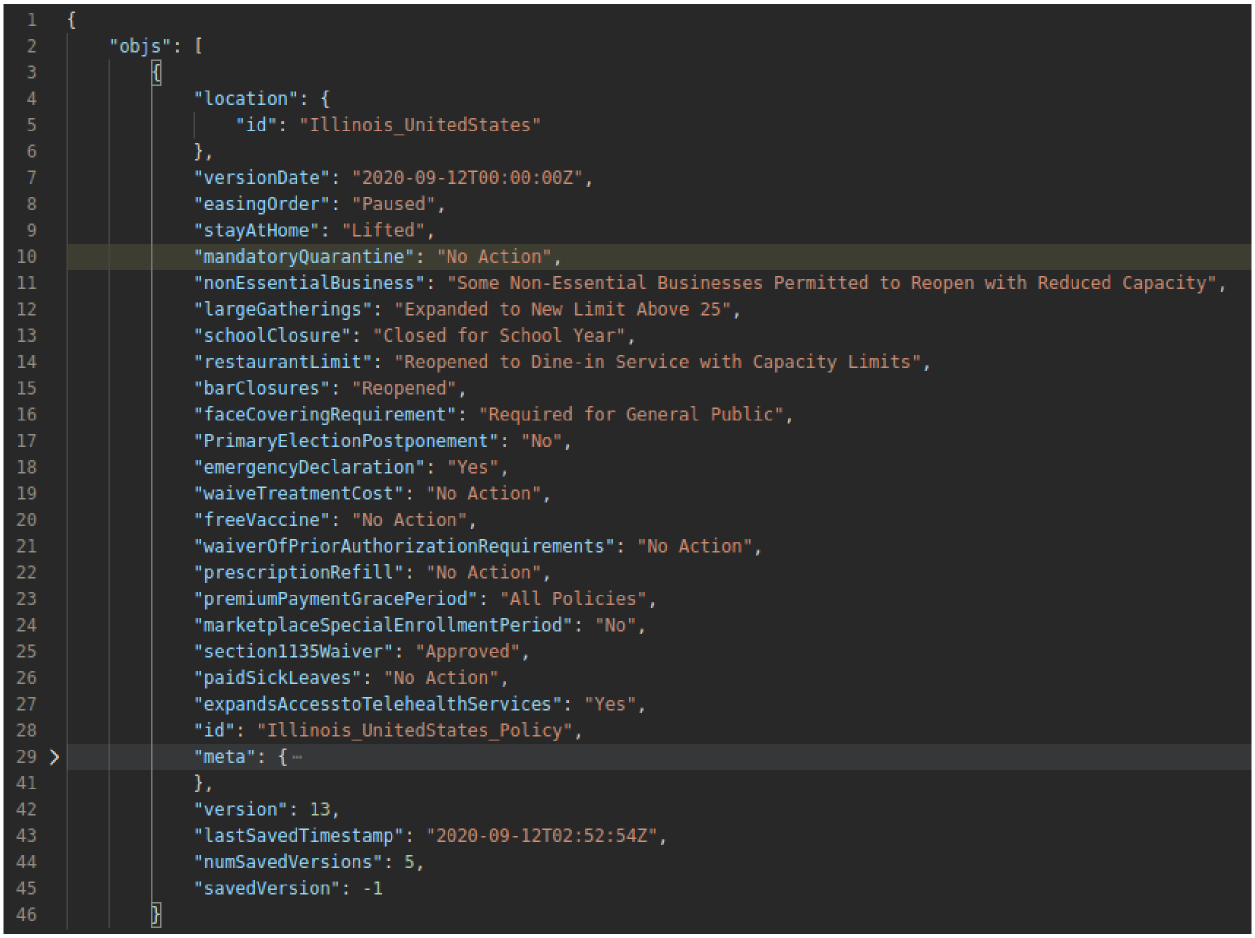
The model we have proposed relies on lot of data from different sources. In this section we provide a detailed explanation about our collection of data from various sources and preprocessing it according to our needs.

In order to get policy summary of the desired location we are using the API provided by C3.ai [12]. This API provides description about various policies such as stayAtHome, largeGathering, restaurantLimit etc via JSON response. We are also using the SurveyData API from C3.ai to know people’s acceptability of wearing a mask [13]. The sample responses for both the APIs are show in Fig. 1 b and Fig. 2 b respectively.



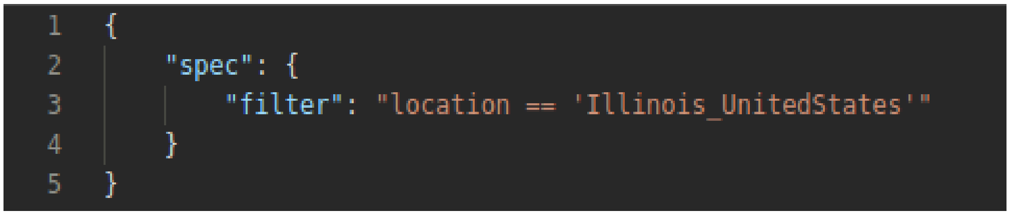
```
{
    "spec": {
        "filter": "location == 'Illinois_UnitedStates'"
    }
}
```

Fig. 1 a Sample **request** data for *LocationPolicySummary* API for the state of Illinois.



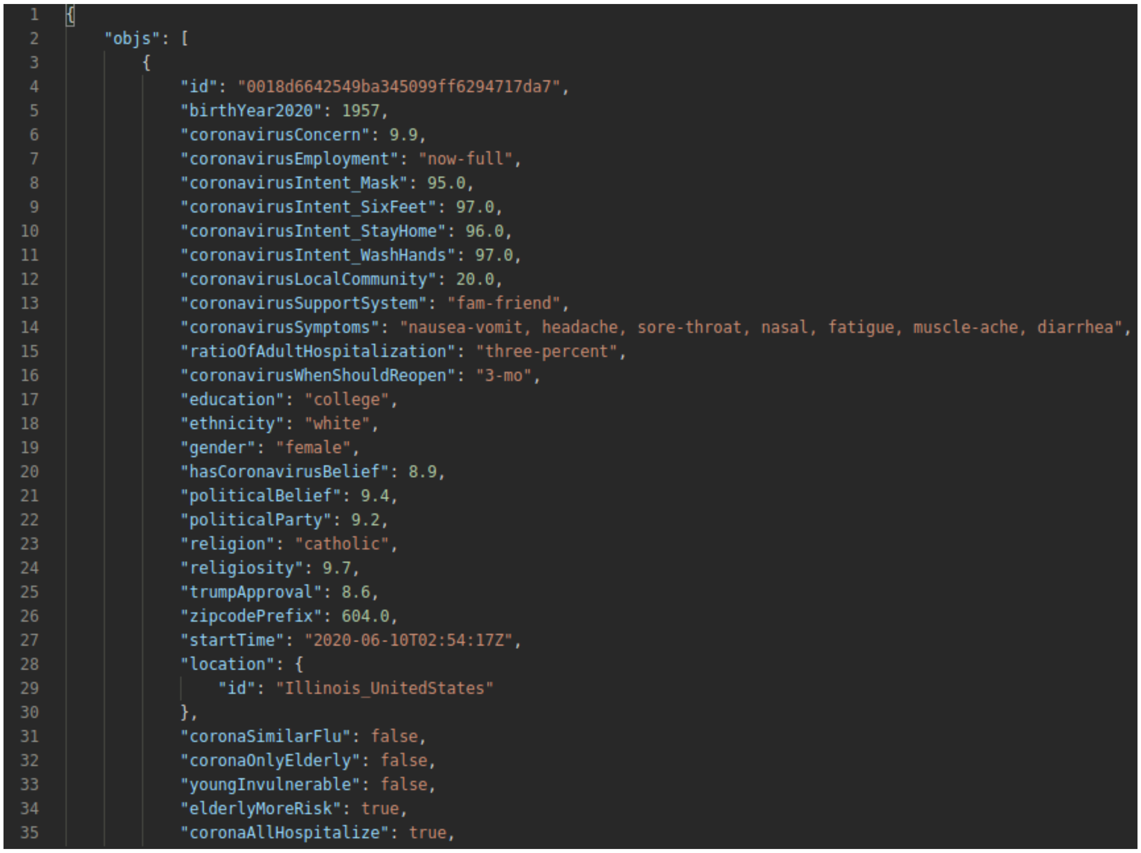
```
{
    "objs": [
        {
            "location": {
                "id": "Illinois_UnitedStates"
            },
            "versionDate": "2020-09-12T00:00:00Z",
            "easingOrder": "Paused",
            "stayAtHome": "Lifted",
            "mandatoryQuarantine": "No Action",
            "nonEssentialBusiness": "Some Non-Essential Businesses Permitted to Reopen with Reduced Capacity",
            "largeGatherings": "Expanded to New Limit Above 25",
            "schoolClosure": "Closed for School Year",
            "restaurantLimit": "Reopened to Dine-in Service with Capacity Limits",
            "barClosures": "Reopened",
            "faceCoveringRequirement": "Required for General Public",
            "PrimaryElectionPostponement": "No",
            "emergencyDeclaration": "Yes",
            "waiveTreatmentCost": "No Action",
            "freeVaccine": "No Action",
            "waiverOfPriorAuthorizationRequirements": "No Action",
            "prescriptionRefill": "No Action",
            "premiumPaymentGracePeriod": "All Policies",
            "marketplaceSpecialEnrollmentPeriod": "No",
            "section1135Waiver": "Approved",
            "paidSickLeaves": "No Action",
            "expandsAccesstoTelehealthServices": "Yes",
            "id": "Illinois_UnitedStates_Policy",
            "meta": {⋯
            },
            "version": 13,
            "lastSavedTimestamp": "2020-09-12T02:52:54Z",
            "numSavedVersions": 5,
            "savedVersion": -1
        }
```

Fig. 1 b Sample **response** data for *LocationPolicySummary* API for the state of Illinois.



```
{
    "spec": {
        "filter": "location == 'Illinois_UnitedStates'"
    }
}
```

Fig. 2 a Sample **request** data for *SurveyData* API for the state of Illinois.

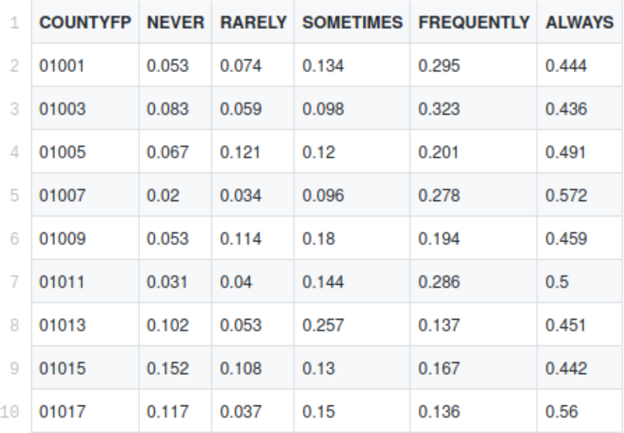


```
{
    "objs": [
        {
            "id": "0018d6642549ba345099ff6294717da7",
            "birthYear2020": 1957,
            "coronavirusConcern": 9.9,
            "coronavirusEmployment": "now-full",
            "coronavirusIntent_Mask": 95.0,
            "coronavirusIntent_SixFeet": 97.0,
            "coronavirusIntent_StayHome": 96.0,
            "coronavirusIntent_WashHands": 97.0,
            "coronavirusLocalCommunity": 20.0,
            "coronavirusSupportSystem": "fam-friend",
            "coronavirusSymptoms": "nausea-vomit, headache, sore-throat, nasal, fatigue, muscle-ache, diarrhea",
            "ratioOfAdultHospitalization": "three-percent",
            "coronavirusWhenShouldReopen": "3-mo",
            "education": "college",
            "ethnicity": "white",
            "gender": "female",
            "hasCoronavirusBelief": 8.9,
            "politicalBelief": 9.4,
            "politicalParty": 9.2,
            "religion": "catholic",
            "religiosity": 9.7,
            "trumpApproval": 8.6,
            "zipcodePrefix": 604.0,
            "startTime": "2020-06-10T02:54:17Z",
            "location": {
                "id": "Illinois_UnitedStates"
            },
            "coronaSimilarFlu": false,
            "coronaOnlyElderly": false,
            "youngInvulnerable": false,
            "elderlyMoreRisk": true,
            "coronaAllHospitalize": true,
```

Fig. 2 b Sample response data for *SurveyData* API for the state of Illinois.

Another source of data is Global Policy Lab [14] which provides enforcement of various government policies on daily basis and it is encoded in binary format.

Our third source of data is New York Times which provides the estimates of mask usage at county level in the United States [15]. It is a static survey data obtained from 250,000 responses between July 2 and July 14. The fig. 3 shows the sample data for mask usage.



COUNTYFP	NEVER	RARELY	SOMETIMES	FREQUENTLY	ALWAYS
01001	0.053	0.074	0.134	0.295	0.444
01003	0.083	0.059	0.098	0.323	0.436
01005	0.067	0.121	0.12	0.201	0.491
01007	0.02	0.034	0.096	0.278	0.572
01009	0.053	0.114	0.18	0.194	0.459
01011	0.031	0.04	0.144	0.286	0.5
01013	0.102	0.053	0.257	0.137	0.451
01015	0.152	0.108	0.13	0.167	0.442
01017	0.117	0.037	0.15	0.136	0.56

Fig. 3 New York Times mask usage estimates in percentage.

Our final source of data is Google which provides the mobility data of the community during COVID-19 according to country, state and county for United States as well as other countries [13]. It tracks movements of community at different public places like grocery stores, pharmacy, parks etc. The baseline is the median value which corresponds to day of the week during the time period of January 3 to February 6, 2020.

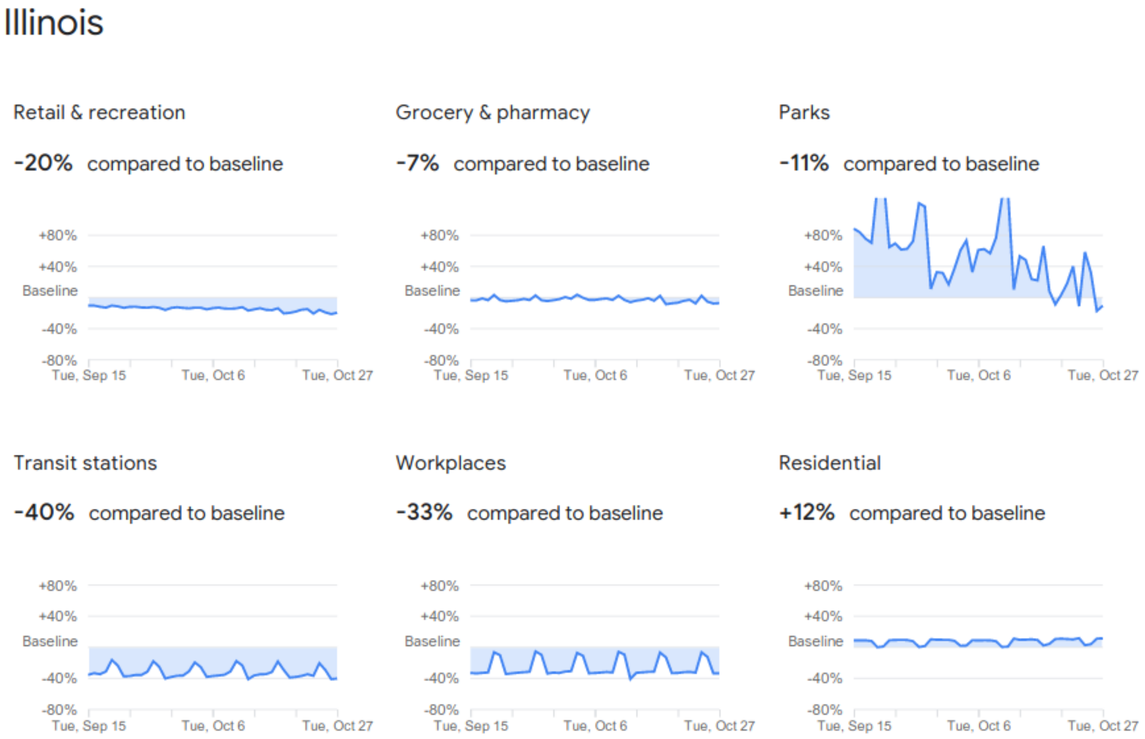


Fig. 3 Sample *mobility* data from Google for the state of Illinois.

All of the above mentioned data sources provide data in JSON and CSV formats. The data for desired county/state/country along with only certain features is aggregated and stored in a single CSV file. In order to achieve we are using a *python* script developed by us. The fig. 4 shows the overview of data collection and preprocessing framework.

The granularity of the data is kept limited to the state level for the scope of this project as all the data required is not available at county level.

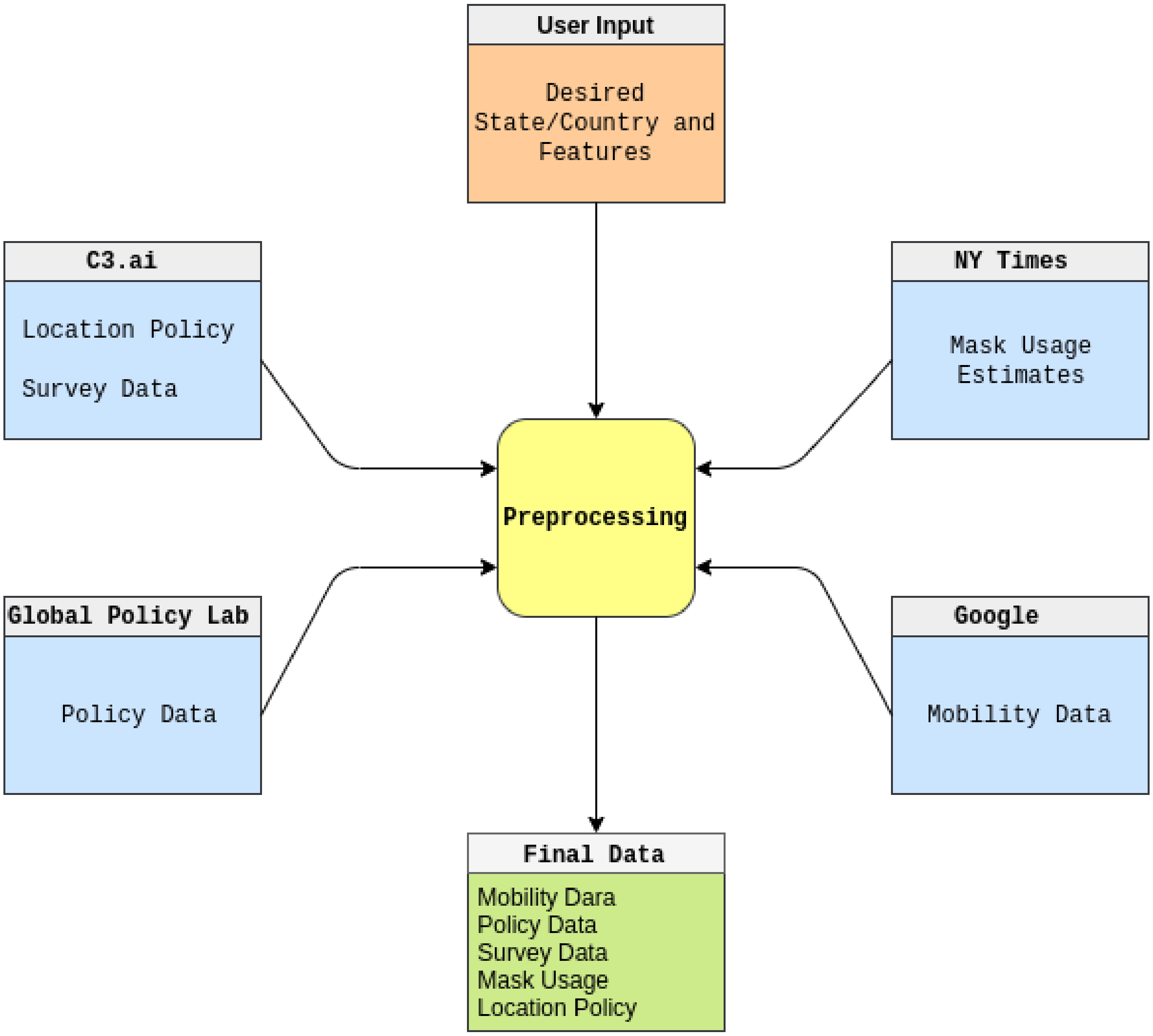


Fig. 4 Data collection and preprocessing overview

Model

Compartmental models in epidemiology - SIR:

The **SIR model** is one of the primary compartmental models, and there are various models which are derivatives of this basic form. For our study we shall begin with this basic form and we have mentioned other variants which can be used in the future to further evaluate the scenario.

Susceptible, *Infected* and *Recovered* (or alternatively *Removed*) - signifies majorly three possible categories to divide the population affected by a contagious disease.

Ordinary differential equations in SIR:
To make the calculations easier we could use the normalized values of the variables instead. Following set of variables represents the *fraction* of the total population in each of the three categories:

S(t) = s(t)/N, susceptible fraction of the population.
I(t) = i(t)/N, infected fraction of the population.
R(t) = r(t)/N, recovered fraction of the population.

$$\frac{\partial S}{\partial t} = -\beta \cdot S(t) \cdot I(t) \tag{1}$$

$$\frac{\partial I}{\partial t} = \beta \cdot S(t) \cdot I(t) - \gamma R(t) \tag{2}$$

$$\frac{\partial R}{\partial t} = \gamma R(t) \tag{3}$$

Initial Model Attempt

Using the SIR model as the base, we attempted to configure a model to establish relationship between transmission coefficient and impact due to policies in reducing the spread of the disease COVID-19.

We wanted to group regions (such as cities like NY City and Chicago) with similar features such climatic conditions, medical facilities, population, diversity etc.

We took the assumptions:

- The transmission rate for the spread of the disease, in the absence of any influence from external measure or control policy, is same for regions belonging to the same group.

- Also, we assumed the extent of impact due to an external factor depends on current transmission potential of the disease in a particular region.

Using this we get the equations:

$$\beta_c = \beta_0 + \sum \sigma_k \tag{4}$$

$$\sigma_k \propto \beta_c \tag{5}$$

$$\propto policy_k, acc_k \tag{6}$$

If we take a particular region such as a city,

$$\sigma_k = \delta_k * policy_k * acc_k * \beta_c \tag{7}$$

$$\beta_c = \frac{\beta_0}{1 + \sum \delta_k * policy_k * acc_k} \tag{8}$$

σ_k → impact due to kth policy intervention.
β_c → transmission coefficient with the current policies which are active.
β_0 → transmission coefficient in the absence of any policy.
δ_k → coefficient of influence factor.
$policy_k$ → degree of enforcement (binary) by govt.
acc_k → degree of acceptance by people.

Drawbacks:

- In this model, we were trying to estimate β_c by analyzing the current data on number of infected and recovered people in a particular region (Illinois) by fitting the trend to the SIR model. And we assumed transmission coefficient as constant for the time period taken and did not accommodate change in β_c with changes in policies between that time period.
- We compared and grouped different cities which have similar climatic condition over the year, diversity of population, medical facilities. However, this assumption did not account for difference in population density, demography, etc. parameters other than policies which may cause difference in β_c.
- Impact of a kth policy is not justified by $policy_k * acc_k$. $policy_k$ is a binary variable describing if a certain Policy k is enforced by Government or not. If the government did not enforce the policy but its still being followed by the people then its impact is not accounted for in the model.

Reference Model Study[6]

Model Basis: Econometric Analysis

To describe the behavior of aggregate outcomes, infection rates (denoted as y), the model attempts to identify the causal effects induced by external changes in independent policies (denoted by z). The underlying mechanisms that links the effect of a policy intervention to change in y is denoted by x_n. Variables w_m denote the determinants of y that are unrelated to the policies, for example, population density, demography, diversity, etc.

$$y = f(x_1(z_1, z_2 \dots z_k), x_2(z_1, z_2 \dots z_k) \dots x_n(z_1, z_2 \dots z_k), w_1, w_2 \dots w_M) \tag{11}$$

This process does not attempt to evaluate the structure of f, but how changes in a particular policy may affect the control on spread of the disease. For a particular region where we can assume the population is fixed over time, we can empirically estimate the dependence of changes in y to changes in the policy, which we can see, in the following equation, is independent of $\mathrm{w_m}$.

$$\frac{\partial y}{\partial z_k} = \sum_{n}^{N} \frac{\partial f}{\partial x_n} * \frac{\partial x_n}{\partial z_k} \tag{12}$$

Model:

Considering the initial phase, from the SIR model, we can assume $S(t) \rightarrow 1$. Which gives us

$$\frac{\partial I_t}{\partial t} = \beta * S(t) * I(t) - \gamma * I(t) = (\beta * S(t) - \gamma) * I(t) \tag{13}$$

$$\xRightarrow[S(t)\rightarrow 1]{} (\beta - \gamma) * I(t) = g * I(t) \tag{14}$$

Solution to this ordinary differential equation:

$$\frac{I(t_2)}{I(t_1)} = e^{g*(t_2 - t_1)} \tag{15}$$

Taking the natural log on both side

$$log(I(t_2)) - \log(I(t_1)) = g * (t_2 - t_1) \tag{16}$$

Where $g = (\beta - \gamma)$. Policy interventions are designed to alter g, by reducing β, by lowering the contact between susceptible and infected individuals.

If we take the time step between observations fixed at one day ($t_2 - t_1 = 1$), we can attempt to model g as a time-varying outcome that is a linear function of a time-varying policy

$$g_t = \log(I(t)) - log(I(t-1)) = \theta_0 + \boldsymbol{\theta} \cdot policy_t + \varepsilon_t \tag{17}$$

where θ_0 is the average growth rate without any policy intervention. $policy_t$ is a binary variable, 0 indicating a policy was not deployed at time t, and 1 indicating otherwise.

θ is the impact of the policy on growth rate g. We assume it to be average effect g over all periods subsequent to the introduction of the policy, thereby accommodating any lagged effects of policies.

ε_t is the noise to t capture inter-period changes not described by $policy_t$.

Taking the equation for a particular region c

$$g_t^c = \log(I^c(t)) - log(I^c(t-1)) = \theta_0^c + \sum_{p=1}^{P_c} \theta_p^c \cdot policy_{p,t}^c + \varepsilon_t^c \tag{18}$$

The number of infected cases is often underreported, it can be shown that the model is robust to this particular issue. Let $\tilde{I}$ be the count of reported cases which is only a fraction (ξ) of the actual cases then, If we take this fraction as constant, then we could infer that model is robust to systematic underreporting of infections.

$$log(\tilde{I}(t)) - log(\tilde{I}(t-1)) = log\big(\xi * I(t)\big) - log\big(\xi * I(t-1)\big) \tag{19}$$

$$= \log(\xi) + log\big(I(t)\big) - log(\xi) - log\big(I(t-1)\big) \tag{20}$$

$$= log\big(I(t)\big) - log\big(I(t-1)\big) = g_t \tag{21}$$

Our Modifications or Enhancements on the Model:

Using the dataset we acquired from survey analysis to gather the acceptability status from the general public

$$\begin{aligned} g_t^c &= \log(I^c(t)) - log(I^c(t-1)) \\ &= \theta_0^c + \delta_d^c + \sum_{p=1}^{P_c} \theta_p^c \cdot policy_{p,t}^c + \sum_{p=1}^{P_c} \phi_p^c \cdot acc_{p,t}^c + \varepsilon_t^c \end{aligned} \tag{22}$$

ϕ_p^c is the effect of the policy attributed with its acceptance among people. This term signifies if the people are following a policy such as wearing mask. Even if it is not a mandate, the policy would still have an effect.

We also include a variable δ_d^c to account for the day-of-week effect in the growth rate of infections.

Policies Considered:

For the scope of our project we are considering the effect of the following policies with the availability of coherent data.

Policy	Enforcement Data	Acceptability Data
Mask Mandate	Data available	Survey
School closure	Data available	Same as Enforcement
Work from home	Data available	Same as Enforcement
Groceries and pharm	(Never Closed)	Data available
Transit	Data available	Data available
Parks	Data Not Available	Data available
Retail and recreation	Aggregated from multiple sources.	Data available
Social Distance (6 feet)	Aggregated from multiple sources.	Data available

Challenges and Next Steps:

Recovery Data

Recovery data is not publicly available at sufficiently granular level such as city or county. In order to incorporate this lack of data we need to slightly change the model and take into consideration the cumulative number of infected cases registered each day.

We need to further analyze this modification in the model to exactly evaluate and account for the errors due to assumptions made at this step.

Currently we are focusing on data which is available state level. If we have partial data available at granular level, we aggregate it into the same format to use it coherently.

Considering only the Initial Scenario

Currently the model is evolved with the assumption that our focus is on the initial stages of the infections in a particular region. Where we assumed the $S(t) \to 1$, or entire population can be considered susceptible.

We need to modify the model development to incorporate the scenarios when the disease has been spreading for a while and the infected number of cases are high in a particular region.

Next Steps

- We plan research and overcome the challenges mentioned above.

- Study the optimization techniques best suited to optimize the parameters of our finalized model.

Future scope

The basis of the model discussed above is motivated by the SIR model. There can be many variants SIR model such as SEIR, which does not assume zero latent period between exposure to the disease and getting infected. Other compartments such Death, Hospitalization, Infected – Detected, Infected – Not Detected and many more can be added as per the extent of availability and granularity of the data.

In our model development we assumed that growth rate g_t is linearly dependent on the effect of the policy interventions. However, in future with the availability of more data, important nonlinearities or interactions between policies can be identified.

References

1) May, R. M., & Anderson, R. M. (1979). Population biology of infectious diseases: Part II. Nature, 280(5722), 455-461.

2) PANG, L., LIU, S., ZHANG, X., TIAN, T., & ZHAO, Z. (2020). TRANSMISSION DYNAMICS AND CONTROL STRATEGIES OF COVID-19 IN WUHAN, CHINA. Journal of Biological Systems, 1-18.

3) Wang, Y., Wang, Y., Chen, Y., & Qin, Q. (2020). Unique epidemiological and clinical features of the emerging 2019 novel coronavirus pneumonia (COVID-19) implicate special control measures. Journal of medical virology, 92(6), 568-576.

4) https://www.bsg.ox.ac.uk/sites/default/files/2020-08/BSG-WP-2020-034.pdf

5) Arifin, W. N., Chan, W. H., Amaran, S., & Musa, K. I. (2020). A Susceptible- Infected-Removed (SIR) model of COVID-19 epidemic trend in Malaysia under Movement Control Order (MCO) using a data fitting approach. medRxiv.

6) Hsiang, S., Allen, D., Annan-Phan, S., Bell, K., Bolliger, I., Chong, T., ... & Lau, P. (2020). The effect of large-scale anti-contagion policies on the COVID-19 pandemic. Nature, 584(7820), 262-267.

7) Topirceanu, A., Udrescu, M., & Marculescu, R. (2020). Centralized and decentralized isolation strategies and their impact on the COVID-19 pandemic dynamics. arXiv preprint arXiv:2004.04222.

8) Gasparek, M., Racko, M., & Dubovsky, M. (2020). A stochastic, individual-based model for the evaluation of the impact of non-pharmacological interventions on COVID-19 transmission in Slovakia. MedRxiv.

9) Kraemer, M. U., Yang, C. H., Gutierrez, B., Wu, C. H., Klein, B., Pigott, D. M., ... & Brownstein, J. S. (2020). The effect of human mobility and control measures on the COVID-19 epidemic in China. Science, 368(6490), 493-497.

10) Fraser, C., Riley, S., Anderson, R. M., & Ferguson, N. M. (2004). Factors that make an infectious disease outbreak controllable. Proceedings of the National Academy of Sciences, 101(16), 6146-6151.

11) Matrajt, L., & Leung, T. (2020). Evaluating the effectiveness of social distancing interventions to delay or flatten the epidemic curve of coronavirus disease. Emerging infectious diseases, 26(8), 1740.

12) https://c3.ai/covid-19-api-documentation/#tag/LocationPolicySummary

13) https://c3.ai/covid-19-api-documentation/#tag/SurveyData

14) http://www.globalpolicy.science/covid19

15) https://github.com/nytimes/covid-19-data

16) https://www.google.com/covid19/mobility/