COVID-19 Infection Forecasting

Motivation

As of mid-May 2020, various policies have been adopted in the US for roughly 2 months to reduce infection rates of COVID-19. These policies vary widely from state to state, as well as city to city, and have become sources of social tension and political debate.

An evidence-based approach needs to be a leading part of the conversation to shape policy. Specifically, based on the infection data, **when will a region's infection rate be low enough to justify relaxing policies and measures around lockdowns / stay-at-home and even social distancing?**

Background

The CDC has defined 3 metrics around relaxing shelter-at-home / social distancing policies:

1. **Infection Rates** dropping below a threshold (see below)
2. **Increased testing** on the general population (positive tests should be on the order of 2% - currently in the 15%-20% range)
3. A **contact tracing** system in place to thoroughly isolate all parties related to each newly reported infection.

Infection rate data is widely available and updated daily, so provides a good opportunity to understand how regions may / may not be approaching the first metric.

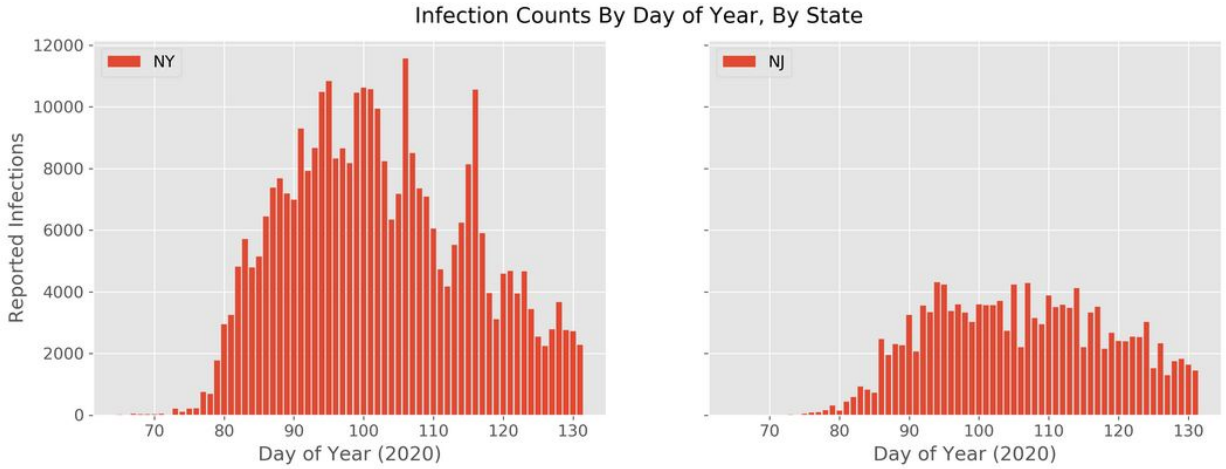
Incidence Threshold

The CDC defines low incidence as 10 or fewer new cases per 100,000 people over a period of 14 days. This rate is equivalent to **0.71 new cases per 100,000 people per day**, or about 2,300 new cases per day in the United States.

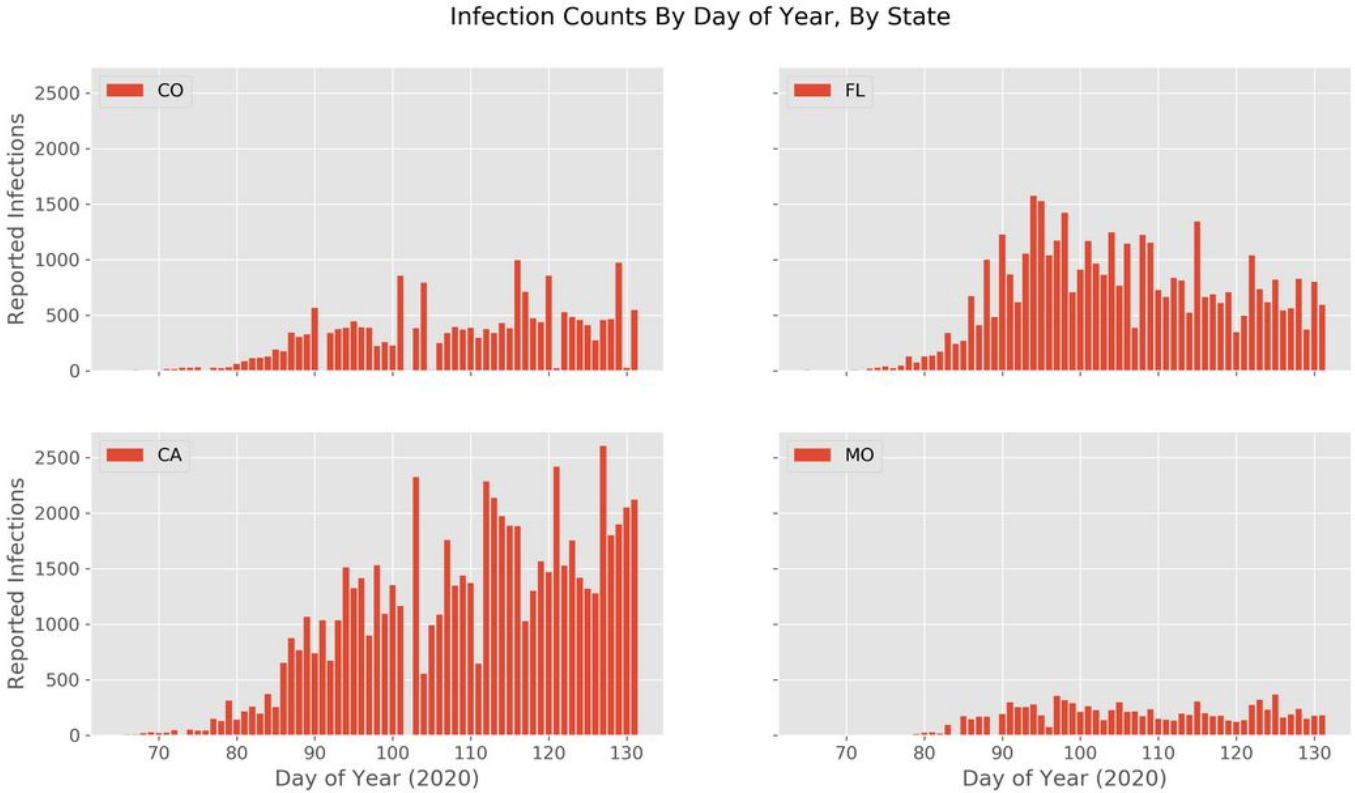
EDA - Time Series Processing

The data is simply daily and cumulative counts - bar charts are most appropriate to explore this.

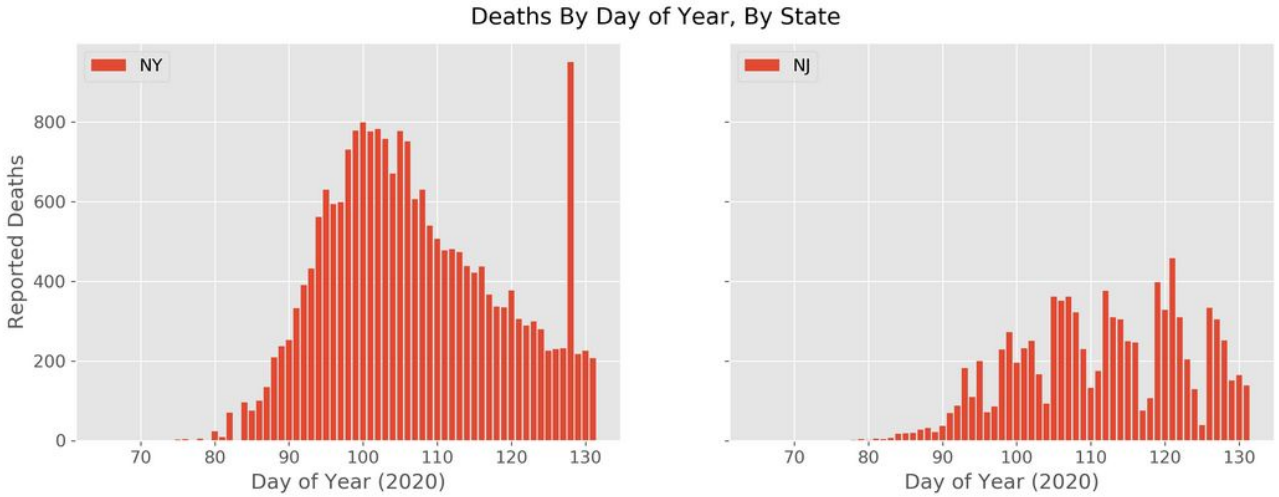
Infections



Hard to see all states on one y-axis, and a log scale loses some vertical perspective.


Infection Counts By Day of Year, By State
CO
FL
CA
MO
Reported Infections
Reported Infections
0
500
1000
1500
2000
2500
70
80
90
100
110
120
130
Day of Year (2020)
Day of Year (2020)

Deaths


Deaths By Day of Year, By State
NY
NJ
Reported Deaths
0
200
400
600
800
70
80
90
100
110
120
130
Day of Year (2020)
Day of Year (2020)

Weekly Trends

A 7-day cyclic behavior can be seen in infection rates and deaths for most states (as well as world countries). Infections and Deaths are lowest on Mondays and Tuesdays; highest on Thursdays and Friday.

It's unclear what drives this trend; i.e., if it's an artifact of patient's behavior with medical care, a lag in reporting data on certain days, etc.

This periodicity needs to be smoothed. A 7-day rolling average was applied to the data for fitting with against models.

Additionally, the target line of "0.71 infections / 100,000 population / day" (= 111 people for NY) has been added.

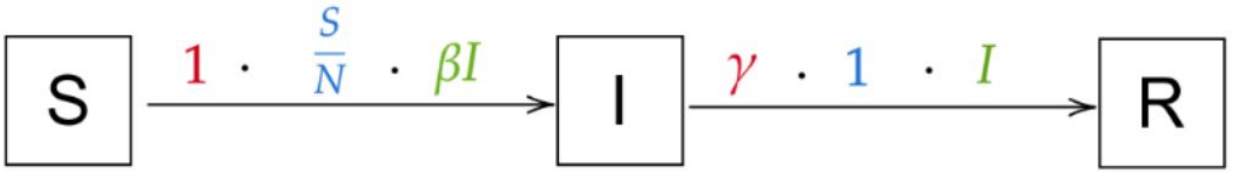
SIR Model

A compartmental model used to model infectious diseases; where every individual in a population is assigned to a compartment based on their condition. The most basic of these models is an SIR model, where all indiviuals are in one of 3 states at any given time:

- (S) usceptible
- (I) nfected
- (R) ecovered

and individuals transition from one state to another following a system of differential equations.

Generic SIR Model Flow:



Where the state transitions are given by:

$$\xrightarrow{rate \cdot probability \cdot population}$$

The sum of the 3 compartments is a constant (the population, N), and the sum of the derivatives of all compartments must be 0.

In an epidemic, initially the entire population (less some initial infected number, I_0) is healthy and in the **S**useptble compartment. Individuals can transition from **S**usceptble to **I**nfected at a rate proportional to some constant, and transition from **I**nfected to **R**ecovered at a rate proportinal to some other constant:

- **β** - Average number of people an infected person infects each day
- **γ** - the proportion of infected people recovering each day ($1/\gamma$ = duration a person is infected)
- **R_0** - Total number of people an infected person infects ($R_0 = \beta / \gamma$)

SEIR Model

Generic SEIR Model Flow:

This adds in an **(E)** xposed compartment, where an individual has contracted the virus but can't yet infect others (incubation period). This brings a new rate variable, δ.

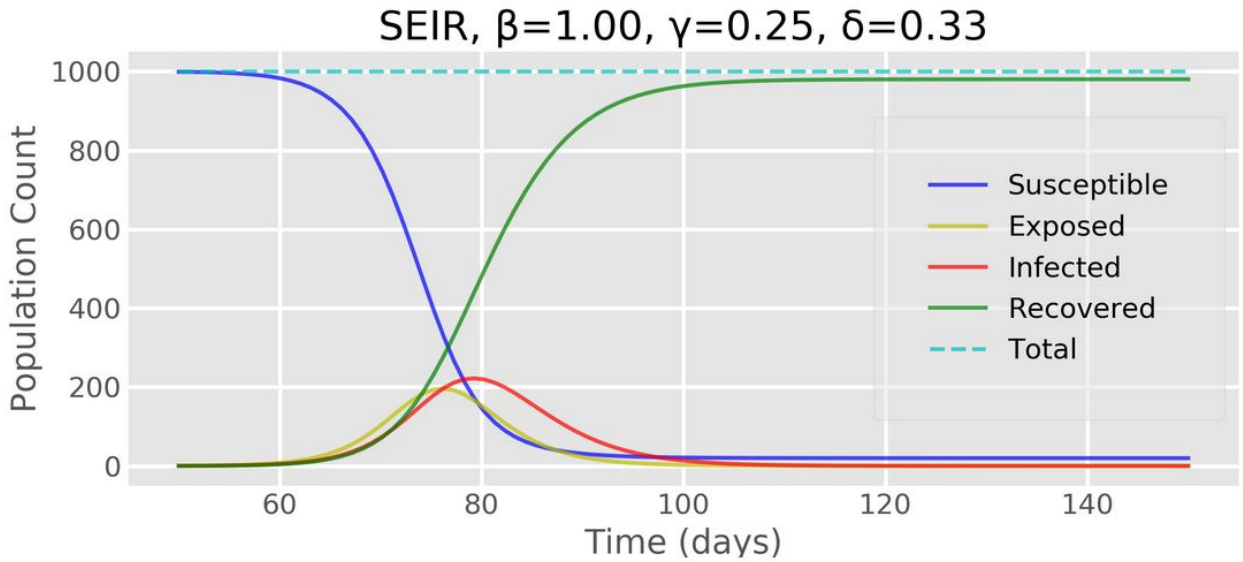
- **δ** - the rate at which exposed people become infected/infectious (1/δ = incubation period)

Assumptions

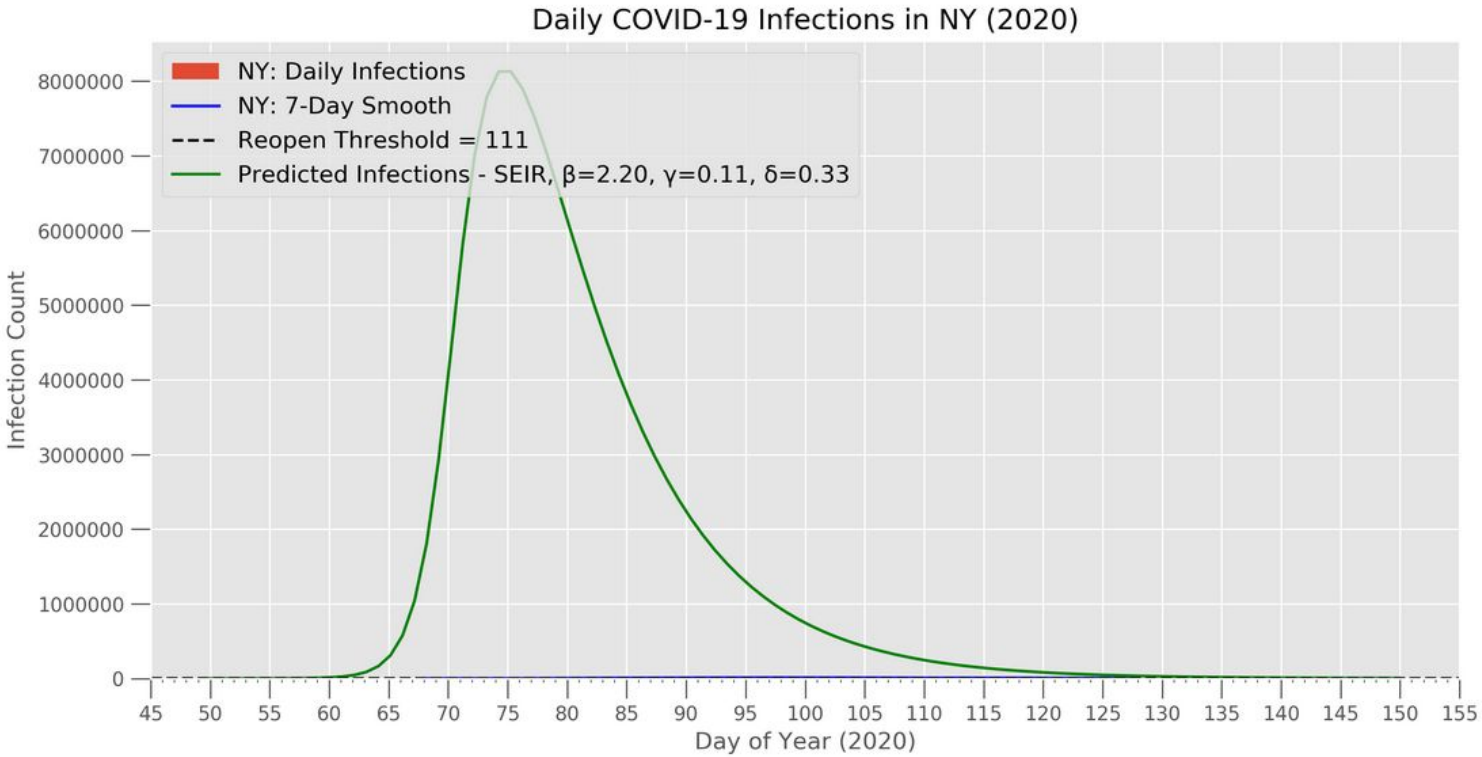
- A given population is treated as homogenous (i.e., within a region/state, there are no population density or age impacts). There is also assumed to be a single initial infection point per geographic area, and not multiple clusters.
- Each geographic region is self-contained (i.e., infections only come from existing sources within and not from travel, immigration, etc.)
- Infection data is accurate - Testing is widespread enough to catch nearly all the infections in the population.
- This SIR/SEIR model assumes immunity upon recovery - this may be true with COVID-19, or true for a short time period, but has yet to be confirmed. An SEIDS (or SEIRD, SEICDS, SEIRCDS, etc.!) may ultimately be a better approximation of the underlying situation.
- ...there are likely many others

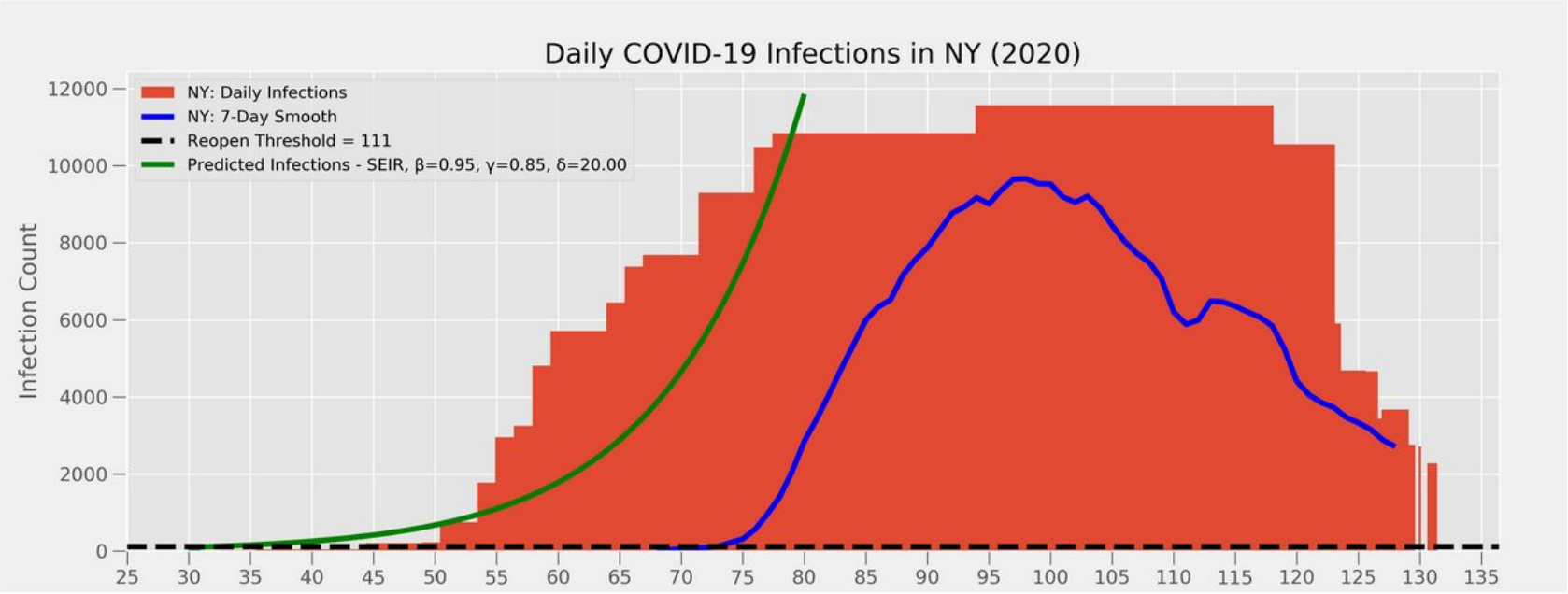
Models and Fits

The SEIR base model with population N=1000. The SciPy `odeint` function is used to integrate the differnetial equations into time series.



SEIR model - Infections and NY Infection data - "literature parameters"


Daily COVID-19 Infections in NY (2020)
NY: Daily Infections
NY: 7-Day Smooth
Reopen Threshold = 111
Predicted Infections - SEIR, β=2.20, γ=0.11, δ=0.33
Infection Count
8000000
7000000
6000000
5000000
4000000
3000000
2000000
1000000
0
45
50
55
60
65
70
75
80
85
90
95
100
105
110
115
120
125
130
135
140
145
150
155
Day of Year (2020)


Daily COVID-19 Infections in NY (2020)
NY: Daily Infections
NY: 7-Day Smooth
Reopen Threshold = 111
Predicted Infections - SEIR, β=0.95, γ=0.85, δ=20.00
Infection Count
12000
10000
8000
6000
4000
2000
0
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
105
110
115
120
125
130
135

Next Steps

- The SEIR model used to try to fit the data was overly simplistic given the expected lack of full reporting of infected persons (low test prevalence). Need to explore if additional complexity will address this issue (e.g., look at modeling β as a function of time).
- The x-axis is in day of year to allow offsetting the onset of infection in a region more easily - this needs to be enabled.
- The SEIR model also doesn't split out deaths from recovered - this would be useful to aiding in the next point.