Tackling the COVID-19 Data Crisis in NY State: Application of the Poisson Process with Considerations of Alternate Data Sources

Ronald Cheng

Abstract

As of November 2020, more than 33 million cases of COVID-19 have been recorded in the United States. In the midst of a public health crisis, the US faces another challenge: a COVID-19 data crisis. As a result of decentralized data collection and patchy state datasets, standard epidemiological models, which are used to analyze historical case data and guide public policy, such as the Susceptible-Infected-Recovered (SIR) model are difficult to produce. This paper reviewed several touted alternate data sources, finding that mobility data released by Apple and google was most useful, resulting in a machine learning model that was used to predict the percent change in cases in the next two week period with an average variance score of 0.77 produced by a K-Fold Cross Validation. To add randomness to the model, this paper considers the use of the poisson process, a stochastic modeling technique that simulates interarrival times of events, to gain a confident understanding of the bounds of the spread of the virus. Simulations were produced using the Poisson Process and the output of the machine learning model, resulting in a two week case prediction (outside the dataset the model was trained on) with 25.3% ± 3.74% error in line with the 0.77 explained variance score. This suggests that mobility data is worth investigating for use in COVID-19 modeling, and that further improvements to the model may provide an effective way to guide policymakers and estimate the short term effects of mass travel events.

Introduction

The novel coronavirus (COVID-19), an RNA virus part of a larger family of respiratory illnesses, has prompted widespread crisis. Believed to originate from Huanan, Wuhan, the virus rose to global attention after the World Health Organization (WHO) declared the illness as a pandemic on January 30th, 2020. Spread through close contact and airborne droplets, the virus was given the opportunity to propagate during China's massive Spring Festival, a celebration that attracts both tourists and Chinese citizens across the nation. As a result, cases outside of China would begin to grow exponentially[1]. Although the disease is less severe than its more deadly family member SARS, its high accumulation rate and long incubation period translate to incredible rates of infectivity (R_0 = 2-2.5)[2]. Moreover, high levels of asymptomatic cases in younger (<30), more mobile juveniles combine with high mortality rates for the elderly (30-79) exacerbate the case and death counts.[1] As of November 2020, over 33 million cases have been documented with just over a million deaths.[11] Thus, the crisis presented by the virus is both dire and in need of guidance and careful action.

Since WHO's declaration, many datasets have emerged to track the propagation of the virus, but the pathology of COVID-19 makes it difficult to glean accurate data; A high asymptomatic rate (18%), wide incubation range (0-24 days), and low RT-PCR (Reverse Transcriptase Polymerase Chain Reaction) test sensitivity (only about 64% of positives detected) contribute to this problem.[2] Moreover, experimental studies in which negative subjects are purposefully exposed to those who tested positive demonstrate that the spread of the virus is highly unpredictable.[2] Despite this, Wuhan and several other heavily affected regions such as Brazil, Italy, and India have been able to collect and provide extensive datasets that include case counts, hospitalizations, quarantined individuals, and so forth. Using this data, various epidemiologists have developed models that seek to provide predictive or analytical value for the propagation of the virus. The most common include the SIR (Susceptible-Infected-Recovered) and SEIR (Susceptible-Exposed-Infected-Recovered) models. For example, epidemiologists have developed a SEIR model integrated with intercity data, a time dependent SIR model using Wuhan data; an SEIR compartmental model [5] using data from India; and a comprehensive eight stage SIDARTHE model using data from Italy [6]. These models seek to describe the dynamics of COVID-19 in their respective regions and analyze the effects of management strategies for COVID-19 such as social distancing, travel bans, or service interruptions. As a result, these

models are helpful in guiding policymakers and developing targeted strategies to mitigate the spread of the virus. [5]

The United States, however, faces greater challenges in building an accurate model. Attributed to weak interjurisdictional coordination and a lack of federal response, data validity and consistency across states are constantly in flux. Furthermore, misleading early statements from federal officials have cultivated antagonistic public sentiment towards closures and shutdowns. Thus, the opportunity to contain COVID-19 was missed and the country's overall response lagged. As a consequence, possibly crucial data is missing and data may have been misrepresented or inaccurate, especially towards the start of datasets provided by major COVID-19 data collection organizations, whether state or independently run.[3]

New York, one of the heaviest hit by the pandemic, suffers from both weak federal response and initial overconfidence. Having faced Ebola, Zika, and H1N1, local officials were quick to downplay concerns after their first case on March 1st, citing New York's exemplary health care system and the fact that plans were in place. As a consequence, the state would have a delayed response with only 50 disease detectives compared to 9,000 workers in Wuhan and issues getting access to tests, ventilators, and masks. The state's response may have further contributed to underestimates and gaps in the data [4]. Although the NY State Health Department has released open source data on NYC stats, which has been recently used in an agent-based model [9] used to determine the effects of a quarantine, a coherent state level dataset is lacking. The Covid Tracking Project, a volunteer organization run by two journalists from the *Atlantic*, attempts to unify various datasets with varying levels of success. While most of the data is clean, several gaping holes are apparent. For instance, recovery numbers in the dataset, released only during the now discontinued Governor press conferences, are cut off abruptly, and large sections lack data.

As a consequence, SIR and similar models are difficult to generate with the given data. In this investigation, we consider the usage of the poisson process in epidemiological modeling. The poisson process is a counting process that models the arrival of events over a continuous time interval and produces a discrete time series count. It is commonly used to model arrivals of events like hospitalizations, customer flow, and phone calls over a specified time interval[12] , and, unlike the SIR model, does not require many interrelated variables, randomizing times of infection only. This paper seeks to evaluate the following hypotheses on alternate data,

COVID-19 modeling, and the Poisson Process: (1) alternative data sources may be used as part of a COVID-19 model, (2) mobility data will be most useful alternative data source in creating a model since it may mirror social contact, the main way COVID-19 propogates[2], and (3) the poisson process can provide adequate bounds and randomness to an epidemiology model measuring infections. The end goal of this project is to create a robust model that can be used to predict the

Methods

SECTION A. Data and Math Background

1. Data Collection

Data to create our model was taken from the COVID-19 tracking project. The group aims to centralize and standardize COVID-19 data collection. The open source dataset they provided for New York contains information on case, death, hospitalization, and recovery counts either cumulative or daily. It also includes extra information on ICU capacity and ventilator status, however, those numbers were not useful in our model. In our test case, we apply the poisson process to the infected category. We used the data range from 3/4/20 to 8/05/20 since it included the peak cases along with some of the aftermath.

Several other open source datasets, many of which were released by various companies to help combat the COVID-19 data crisis, we considered and reviewed. The Apple mobility dataset contains information on relative changes in routing requests for walking, driving, and transit (how often apple device users request for information or use features tied to those actions). The Google mobility dataset contains information about relative changes in search trends for various locations including parks, residential areas, grocery stores, workplaces, and transit stations. The AirBNB dataset contains information on user bookings, host locations, and activity. The NY times dataset includes data on mask usage across states and counties. The Openaq dataset contains information on the air quality of various locations across New York, including measurements of ozone (o3), particulate matter (pm10 & pm25). Appendix 1 summarizes the data they contain and other relevant information.

2. Analyzing Alternative Data Sources

The contents of various alternative data sources were reviewed to look for possibly useful data. The alternate data sets include secondary mobility reports (through routing requests or search trends), air quality data (of pm25 and o3 values), other compilations of COVID-19 data and AirBNB listings. These datasets were chosen because they may be tangentially related or affected by the spread of COVID-19

This paper summarized each data source, including the types of data included and noting whether it had relevant time series data. If the datasets contained relevant and workable time series data, correlations were taken with respect to standard case data. Then, the usefulness of each data source was assessed: whether the data could contribute to the model or if it may have opened up new interpretations of the model.

3. Stochastic Modeling

Stochastic modeling is a broad term for statistical techniques that use probability distributions to generate "simulations", or a set of values, based on information from a dataset. These simulations generate random outcomes that change each time the model is run. Stochastic modeling is an alternative to stricter deterministic models, which provide a constant output each time they are run and cannot take uncertainty into account. With stochastic models, a wide range of outcomes are presented after running a large number of simulations. Those outcomes form a range of possible outcomes useful for predicting future outcomes or analyzing the worst and best outcomes of historical data. This paper used a stochastic poisson process model, which produced a simulation of events over a continuous time series with exponentially distributed interarrival times.

4. Exponential and Poisson Distributions

The model generated used the exponential and poisson distributions as assumptions. The exponential distribution is defined as a random variable given the following probability density function:

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases} \quad [1.1]$$

The value of its cumulative distribution function, generated by integrating the probability density function from 0 to *x*, and setting it equal to 1 ($lim_{x \to \infty}(-e^{-\lambda x}) + C = 1$), is given by:

$$F(x) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases} \quad [1.2]$$

The expected value of the exponential distribution, found by taking taking the improper integral of $x \bullet f(x)$ is given by:

$$E[X] = \int_{-\infty}^{\infty} \lambda x e^{-\lambda x} dx = \frac{1}{\lambda} \quad [1.3]$$

Thus in an exponential distribution, its parameter, λ, is the reciprocal of the mean of the distribution. This parameter determines the distribution of random time intervals that may occur. In our model, this distribution was used to generate randomized arrival times using a variable λ produced through curve fitting (see 5.) and an iterative *acceptance rate* (see 6).

We chose to assume an exponential distribution of interarrival times because of its memoryless property: regardless of how long a person has remained infected, the probability that they will enter a particular compartment within a certain time interval is unchanged. So for random variable *X*– the day when a person will enter the infected compartment, the probability *P* that the day a person remains the infected compartment at day *s*, given that they have been in the compartment for *t* days, is equivalent to the the probability that they the will remain there after day *s*:

$$P\{X > s + t \mid X > t\} = P\{X > s\} \text{ for all } s, t \geq 0 \quad [1.4]$$

The poisson distribution is used to approximate the number of events that will occur in some time interval. The probability mass function is given by:

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!} \quad [2]$$

With the parameter λ, the average rate at which the event occurs. This distribution underlies the poisson process, and approximates the number of events that occur by time x.

5. Poisson Process

The poisson process is a type of counting process. Counting processes are defined as a stochastic process that satisfies

$$\{N(t),\ t \geq 0\} \quad [3.1]$$

Where *N(t)* describes the number of events that occurred by time t. Counting processes thus follow that (1) $N(t) \geq 0$, (2) *N(t)* is discrete and integer valued, (3) If $s < t$, then $N(s) \leq N(t)$, and (4) for $s < t$, $N(t) - N(s)$ equals the number of events that occur in the interval $(s, t]$.

A poisson process is a special case that satisfies the following: (1) $N(0) = 0$, (2) $\{N(t), t \geq 0\}$ has independent increments (what has already occurred in a process will not affect what occurs next), and (3) $P(N(t + h) - N(t) = 1) = \lambda h$. The poisson process simulates the number of events that occur within a given time interval from 0 to T, and uses an exponential distribution with parameter λ to generate separate time intervals between 0 and T.

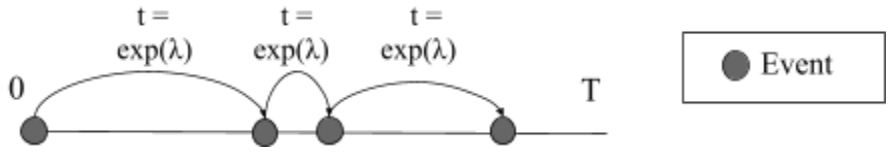

Figure 1: Visual demonstration of the poisson process on the time interval [0,T].
Total # events generated in the interval ~Poi(λT).

Since the exponential distribution is memoryless (see eq. 1.4) , the poisson process can be performed iteratively, and the time interval after t is not dependent on what has occured before t. Programmatically, this means that we can reset t each time we run a simulation using the exponential distribution.

Each time interval $t + s$, where $s \sim exp(\lambda)$, s is equally likely to be at any point in the interval 0 to T because the integral of a cumulative distribution function, in this case the exponential distribution, at any point is 0. Consequently, every discrete point on the distribution has an equivalent probability, thus giving an equal chance for any point on the interval to occur.

However, this does not mean that an arrival is equally likely to occur at any interval from 0 to ∞– the arrival is more likely to occur wherever the density function is most dense.

We can assume that this counting process will be poisson distributed because of the relationship between the poisson and exponential distributions, which are linked as such:

First we define three random variables– *T*, the interarrival time, *N(t)*, the number of events that occurred by time *t*, and S_n, the sum of the interarrival times by event *n*– with the following distributions:

$$T \sim Exponential \qquad N(t) \sim Poisson \qquad S_n \sim Gamma$$

S_n is gamma distributed because it is the sum of a series of exponentially distributed variables. We then define the probability density function that describes *s*, the time at which *n* events occur, to S_n , the probability that *n* independent events will occur by time *s*:

$$f_{S_n}(s) = \lambda e^{-\lambda s}\left(\frac{(\lambda s)^{n-1}}{(n-1)!}\right) \text{[4.1]}$$

To prove the relationship between the two distributions, we must demonstrate that the probability that *n* events occurs at time *t* is equivalent to the poisson probability distribution:

$$P(N(t) = n) = e^{-\lambda t}\left(\frac{(\lambda t)^n}{n!}\right) \text{[4.2]}$$

We begin by stating that the probability that there are more than or equal to *n* events by time *t* is equivalent to the probability that the sum of the interarrival times up to *n* is less than or equal to *t*, or rather that the *nth* event occurs at or before time *t*:

$$P(N(t) \geq n) = P(S_n \leq t) \quad \text{[4.3]}$$

This is logical given that an event must occur before the time interval to be counted within the time interval of the process. It then follows that $P(S_n \leq t)$ is equivalent to product of the cumulative density function, or the integral of the probability density function of S_n from 0 to t, and the joint probability in which there are *n* events by time *t* given that the time of the *nth* event is *s*, which is some time at or before *t*:

$$P(N(t) = n) = P(S_n \leq t,\ N(t) = N(S_n)) \qquad \text{[4.4]}$$

Then by bayes theorem, we can expand the joint probability:

$$P(N(t) = n) = P(S_n \leq t) \bullet P(N(t) = N(S_n) \mid S_n \leq t) \quad [4.5]$$

The probability that the sum of the interarrival times up to event *n* is less than *t* is equivalent to the integral of the gamma distributed random variable from 0 to t, so $P(S_n \leq t) = \int_0^t f_{Sn}(s)$ (eq. 4.1). The probability that the number of events by time *t* is equivalent to the number of events by time S_n, when event *n* occurs, given that time S_n is less than *t* must mean that the number of events at time *t* is *n* given that S_n occurs at time *s*, the time before *t* at which event *n* occurs. Thus, we can state that $P(N(t) = N(S_n) \mid S_n \leq t)$ is equal to $P(N(t) = n \mid S_n = s)$.

$P(N(t) = n \mid Sn = s)$, which implies that the *nth* event occurring at time *s* is the last event in the interval 0 to *t*, is thus equivalent to the probability that no events occur from time *s* to *t*:

$$P(N(t) = n) = \int_0^t f_{S_n}(s)\, P(N(t - s) = 0)\, ds \quad [4.6]$$

We can then expand the probability density function of S_n and then model the probability that no events occur from time *s* to time *t* as the inverse of the cumulative density function, $e^{-\lambda(t-s)}$:

$$P(N(t) = n) = \int_0^t \lambda e^{-\lambda s} \frac{(\lambda s)^{n-1}}{(n-1)!} \bullet e^{-\lambda(t-s)}\, ds \quad [4.7]$$

From this equation we can see that the memoryless property of the exponential distribution as it relates to *N(t)*. As we iterate through the poisson process, s can be reset to zero and *t* can be iteratively increased. Finally, when we integrate with respect to *ds*, we find that *N(t)* is indeed poisson distributed:

$$P(N(t) = n) = \lambda e^{-\lambda t} \bullet \frac{(\lambda)^{n-1}}{(n-1)!} \int_0^t s^{n-1} ds$$

$$P(N(t) = n) = e^{-\lambda t} \bullet \frac{\lambda(\lambda)^{n-1}}{(n-1)!} \bullet \frac{t^n}{n}$$

$$P(N(t) = n) = e^{-\lambda t}\left(\frac{(\lambda t)^n}{n!}\right) \quad [4.8]$$

Equation [4.7] is equivalent to equation [2] with the exception that it is also in terms of *t*, giving the probability that *n* events will have occurred by time *t*. Thus, the mean number of

events that occur between 0 to t generated by the count of exponentially distributed interarrival times is poisson distributed with mean λt.

SECTION B. Generating a Predictive Simulation

1. Assumptions

After running correlations on various datasets, it was found that mobility data returned the overall highest correlations (see r^2 values in appendix 1) in comparison to the percent change in cases. It was assumed that the mobility data should instead be compared to cases two weeks in the future. This is supported by current literature that implies that two weeks is when most people start to present symptoms[2] and thus become likely to test and by the much higher correlational values that result from comparing mobility data changes with cases two weeks later (see figure 2 below).

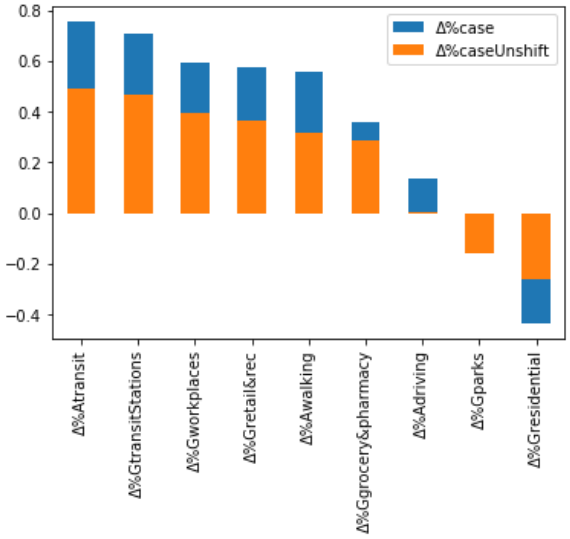

Figure 2. Graphical comparison of correlations between percent change in cases shifted two weeks (blue) and unshifted (orange) with mobility data

In addition, it had to be assumed that the mobility data, released from Google[B] and Apple[C], could be considered a representative sample of the real mobility of the population even though the mobility data measures only changes in mobility by those who use Google and Apple services[B C].

2. Multiple Linear Regression

Multiple linear regression is a machine learning regression technique that takes in a set of *n* features to be trained through a linear fitting algorithm in relation to some output *y*. The result of training returns an equation in the form:

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + ... + \theta_n x_n \quad [5.1]$$

This equation can be used to predict a continuous set of outputs, *ŷ,* given all *n* features that the model has been trained on. Each feature is represented by an *x* value that maps to *y*, each paired with a coefficient θ and a bias constant θ_0.

3. Choosing Features

Features were chosen based on a variance threshold of 0.5. As long as the candidate feature (see figure 3) had a correlation, or variance, greater than 0.5, it was included as a feature to be trained on the multiple linear regression model.

	Δ%case
Δ%Atransit	0.755002
Δ%GtransitStations	0.707311
Δ%Gworkplaces	0.596760
Δ%Gretail&rec	0.575315
Δ%Awalking	0.558817
Δ%Ggrocery&pharmacy	0.362187
Δ%Adriving	0.134277
Δ%Gparks	-0.105286
Δ%Gresidential	-0.434393

Figure 5. Chart of the correlations between percent change in cases and various mobility data factors

After running correlations, it was found that the percent change in transit from both Apple and Google, percent change in workplace, retail and recreation from Google, and the percent change in walking from Apple met this variance threshold, and was thus chosen to be features of the multiple linear regression model.

4. Root Mean Squared Error and Variance with K-Cross Fold Validation

To evaluate the model, we used root mean square error, which measures the average error of the model by backtesting with historical data, or when attempting to predict the test portion of a train test split. The equation is given below, which states that it is the square root of the average squared difference between the actual value y and the predicted value $\hat{y}$ of each output given the values of the features of the model:

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2} \quad [5.2]$$

In addition, the explained variance was taken, which measures how much of the model (and its features together) can account for the change in the output and how well the model performs with 1 being the highest score. The value is determined by normalized the mean squared error and subtracting that value from 1:

$$Var(\hat{y}, y, \bar{y}) = 1 - \frac{\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}{\sum_{j=1}^{n}(y_j - \bar{y})^2} \quad [5.3]$$

K-cross fold validation is the process of creating k train test splits on the data with testing sets that are mutually exclusive. This process is used to generate more accurate evaluation metrics by averaging them over k iterations. In this project, a four fold cross validation was used with four iterations of 75-25 train test splits.

5. Homogeneous Poisson Process

To generate randomized simulations we used a homogeneous poisson process, we follow the following pseudocode, which takes in parameters T, the length of the time interval being simulated, and λ, the average number of cases each discrete time step of the interval:

```
1 function poisson_p(λ, T):
2    t = 0
3    repeat until t >= T:
4       x = randn_uniform 0-1
5       Δt = F_exp^-1(λ,x)
6       t := t + Δt
```

```
7       store t in array
```

Figure 4. General pseudocode for the poisson process implementation

First we set *t*, our initial variable representing time, equal to 0 (l.2). Randomized, exponentially distributed intervals were generated by mapping a pseudo randomly generated uniformly distributed number to the inverse of the cumulative density function (CDF) of the exponential distribution.

$$F(x) = 1 - e^{-\lambda x}, \text{ D: } [0,\infty), \text{ R: } [0,1] \quad [6.1]$$

$$F^{-1}(x) = -\frac{1}{\lambda} ln(1 - x), \text{ D: } [0,1], \text{ R: } [0,\infty), \text{ X} \sim Unif(0,1) \quad [6.2]$$

As written in pseudocode (l.5), $F_{exp}^{-1}(\lambda,x)$, returns an interarrival time given a parameter lambda, λ, and a pseudo randomly generated number uniformly distributed number *X*, where $0 \leq X \leq 1$ to fit to the domain of the inverse function so that it maps properly to the inverse CDF. Because the domain of *1-X* and *X* are equivalent, it is sufficient to use *ln(X)* instead of *ln(1-X)* in the function. The updated *t* value is then stored into an array (l.7). By the end of the simulation, the total expected cases can be summed from the array.

The equation used to calculate percent change in cases was rewritten as follows so that λ is mapped to day *n*, where Ω represents the output of the regression model mapped to mobility data features *n-14* or 14 days before day *n*, and I_n is the number of cases on day *n*:

$$\lambda(n) = \Omega(x_{1_{n-14}}, x_{2_{n-14}}, x_{3_{n-14}}, x_{4_{n-14}}, x_{5_{n-14}})I_n + I_n \quad [7.1]$$

A weekly average was used in place of the equation 7.1 because it occasionally produced negative values. This was likely due to the nature of shifting the cases back two weeks. The new equation is given below:

$$\lambda = \frac{1}{14}\sum_{n=0}^{14} \Omega(x_{1_{n-14}}, x_{2_{n-14}}, x_{3_{n-14}}, x_{4_{n-14}}, x_{5_{n-14}})I_n + I_n \quad [7.2]$$

Equation 7.2 was used as the parameter λ for the simulations with a *T* value of 14 days.

Results

1. Regression Model

The regression model, after a four fold cross validation, resulted in an average explained variance score of 0.77, meaning that the five features of mobility data– apple transit, walking and google workplace, retail and rec, and transit–seemed to account for about 77% of the percent change in cases. Although the average root mean square error was 4.91%, this was likely due to large amounts of error at the start of the model where cases were high (see figure 5 below).

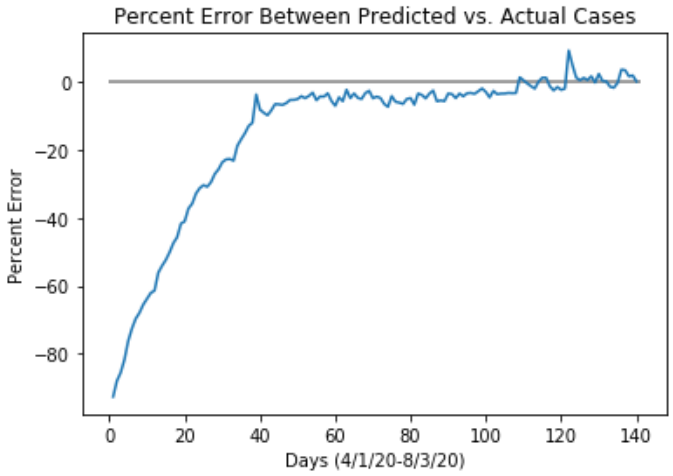

Figure 5. Historical data testing of the model trained on the full dataset over the course of a four month period.

The actual error seemed to decrease significantly as cases increased over time, but demonstrates that the model performs poorly when there is initially a very low case count but tends to improve as time goes on.

2. Simulations

The simulations run by the poisson process were attempted on the two weeks immediately outside of the trained dataset, 8/4/20-8/18/20, resulting in a distribution of cases between 11147-11832 (see figure 6 below).

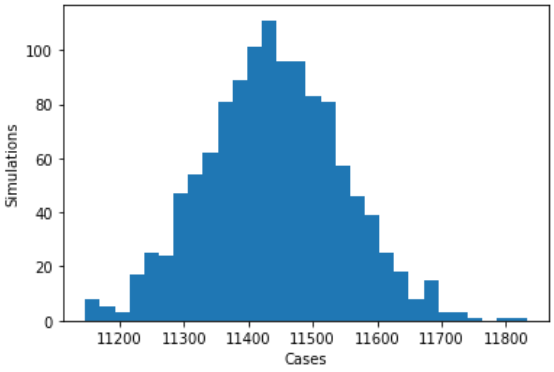

Figure 6. Distribution of 2000 simulations produced by the Poisson Process

That actual number of cases that week was 9165. Compared to the actual cases that week, there was about a 25.3% ± 3.74% percent error in the prediction.

Conclusions

1. Evaluation of the Model

The results imply that changes in mobility data may be very useful in the construction of COVID-19 models. Considering the high correlations and variance score, the model suggests that measures of mobility can account for a very large portion of the change in cases. In addition, the model produced in this paper becomes more accurate in predicting cases as total case numbers grow larger, and the Poisson Process alone doesn't produce bounds sufficient enough when tracking case spreads, likely due to the exponential nature of virus propagation[2]. The model produced in this paper is again limited by data, but seems to provide decent predictions given just mobility data. As stated previously, the percent error of the prediction was 25.3% ± 3.74%. While this may seem initially high, the percent error is not unexpected given its explained variance score of 0.77; since the model is only able to account for about 77% of the change in cases, an error between 20-30% is expected. It is likely that other factors besides mobility may result in overestimation or underestimation of cases such as safety protocols and mask usage which the model does not account for as it only checks the variance of percent change in cases with mobility data over time. Despite this error, it may be possible to use the model to estimate the effects of mass travel events like holidays or vacation times– helping guide policy makers in making decisions.

2. Limitations

The limitations of the model stem from (1) the available data and (2) the methodology. In terms of available data, there may have been issues with data sources that may have decreased inaccuracy. For instance, there was no indication of what the "baseline" meant when looking at the mobility data, nor how many users were represented in the same. In addition, varied data quality and possible underreporting of case data may have been possible due to the nature of the COVID-19 virus and US response. These factors have made it hard to determine how legitimate the predictions of the model are, and mean that some of the data the model was trained on may have varied in the conditions in which they were collected and were thus affected by outside factors unaccounted for by the model.

In terms of methodology, the model was simplified to predict cases using minimal data and made several assumptions based on those few factors that may have been inaccurate. For instance the assumptions that interarrival times for infections were exponentially distributed and that there was a linear relationship between cases and mobility data may have been naive. Furthermore, the model only takes mobility and case data into account– nothing more. Thus the model cannot predict cases perfectly, and does not take other factors such as safety protocol implements or superspreader events that do not have a great effect on mobility data, but do have a great effect on case data. Lastly, given that the model was a regression, there is the possibility that it may have been overfit.

3. Future Research

Further research could address many of the limitations of the model and affirm the validity of mobility data usage in epidemiological models. This can be done through (1) improved methodology, and (2) expansion of data sources and information. Other machine learning techniques such as the use of parametric, nonlinear, or piecewise regressions may result in better models. The use of other distributions and counting processes, or the use of parameter estimation may result in better simulations. Furthermore, this study seems to support the integration of mobility data with other COVID-19 models, whether in locations facing data crises or not, given the high correlations and variance scores.

The use of other data may improve the model or quantify the effects of different protocols and mask usage. Cross state comparisons with non temporal data, such as using mask usage surveys from the NY Times[F], may allow us to determine whether certain safety protocols or public adherence to those protocols affect how cases vary with changes in mobility, resulting in a better model that accounts for those factors. Better mobility data, such as phone usage data, may provide a more representative sample of people as a proxy metric for mobility. In addition, more information on the mobility datasets released by Apple and Google such as the volume of people or meaning of the “change in baseline” may help to draw more accurate conclusions from the model.

Appendix 1

Datasets	Description	Included Data	Analysis
AirBNB	AirBNB is a house sharing service; Releases data related to its service and listings.	Contained a dated series of listings and whether they were filled or not filled.	The dataset released contained data within a limited time frame (only two weeks). Statistical analysis could not be performed.
NY Times	US case data compiled by the NY Times (news company) split by County	(1) Surveyed Mask Usage by County, (2) Compiled Case Data by County	Had no new time series data, mask usage had only one entry; Statistical analysis could not be performed.
Apple Mobility Report	Data released by Apple on routing requests for various mobility changes from 1/29-7/29	Percent change in routing requests of (1) driving, (2) transit, and (3) walking by day	The data had decent correlations (see figure 3)
Google Mobility Report	Data released by Google on search engine requests and other data trends in various locations.	Percent change in movement trends for (1) Retail and Rec, (2) Grocery and Pharmacy, (3) Parks, (4) Transit Stations, (5) Workplaces, (6) Residential Areas	The data had decent correlations (see figure 3)
OpenAQ	Data released by Open Air Quality by borough in NY.	Values of o3 and pm25 by ppm on an hourly basis.	The data had weak correlations (r=0.003 for o3, and r=0.002 for pm25) with infected cases.

WORKS CITED

1. Sahin AR, Erdogan A, Mutlu Agaoglu P, Dineri Y, Cakirci AY, Senel ME, et al. 2019 Novel Coronavirus (COVID-19) Outbreak: A Review of the Current Literature. *EJMO* 2020;4(1):1-7.
2. Karim, Umar. "Review of 'Epidemiology and Clinical Features of COVID-19: A Review of Current Literature.'" 2020, doi:10.14322/publons.r7789065.
3. Haffajee, Rebecca L., and Michelle M. Mello. "Thinking Globally, Acting Locally — The U.S. Response to Covid-19." *New England Journal of Medicine*, vol. 382, no. 22, 2020, doi:10.1056/nejmp2006740.
4. Goodman, J. David. "How Delays and Unheeded Warnings Hindered New York's Virus Fight." *The New York Times*, The New York Times, 8 Apr. 2020, www.nytimes.com/2020/04/08/nyregion/new-york-coronavirus-response-delays.html.
5. Mohamadou, Y., Halidou, A. & Kapen, P.T. A review of mathematical modeling, artificial intelligence and datasets used in the study, prediction and management of COVID-19. *Appl Intell 50*, 3913–3925 (2020). https://doi.org/10.1007/s10489-020-01770-9
6. Giordano, G., Blanchini, F., Bruno, R. et al. Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. *Nat Med 26*, 855–860 (2020). https://doi.org/10.1038/s41591-020-0883-7
7. Kucharski, Adam J, et al. "Early Dynamics of Transmission and Control of COVID-19: a Mathematical Modelling Study." *Centre for Mathematical Modelling of Infectious Diseases COVID-19 Working Group*, 2020, doi:10.1101/2020.01.31.20019901.
8. Zeng, Nianyi, et al. "Epidemiology Reveals Mask Wearing by the Public Is Crucial for COVID-19 Control." *Medicine in Microecology*, vol. 4, June 2020, p. 100015., doi:10.1016/j.medmic.2020.100015.
9. Hoertel, Nicolas, et al. "Facing the COVID-19 Epidemic in NYC: a Stochastic Agent-Based Model of Various Intervention Strategies." *MedRxiv*, 2020, doi:10.1101/2020.04.23.20076885.
10. He, Sha, et al. "A Discrete Stochastic Model of the COVID-19 Outbreak: Forecast and Control." *Mathematical Biosciences and Engineering*, vol. 17, no. 4, ser. 2792-2804, 16 Mar. 2020. *2792-2804*, doi:10.3934/mbe.2020153.
11. "COVID-19 Situation Update Worldwide, as of 27 September 2020." *European Centre for Disease Prevention and Control*, 27 Sept. 2020, www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases.
12. Tse, Kung-Kuen. (2014). Some Applications of the Poisson Process. Applied Mathematics. 05. 3011-3017. 10.4236/am.2014.519288.
13. Ricciulli, V. (2020, April 01). New York City finally closes playgrounds due to coronavirus pandemic. Retrieved November 11, 2020, from https://ny.curbed.com/2020/4/1/21203101/nyc-coronavirus-playgrounds-close-covid-19

BIBLIOGRAPHY

1. Ross, S. M. (2019). *Introduction to probability models*. Amsterdam: Academic Press.

DATA SOURCES

A. COVID-19 Tracking Project. (2020). New York COVID-19 Case Data CSV. Retrieved from https://covidtracking.com/data/state/new-york
B. Apple. (2020). Apple Maps Mobility Trends Reports. Retrieved from https://covid19.apple.com/mobility
C. Google. (2020). Google Mobility Trends Report. Retrieved from https://www.google.com/covid19/mobility/
D. AirBNB. (2020). New York Data. Retrieved from http://insideairbnb.com/get-the-data.html
E. OpenAQ. (2020). New York Open Air Quality Data. Retrieved from https://openaq.org/#/locations?_k=vwsfxw
F. New York Times. (2020). NY Times COVID-19 Data. Retrieved from https://github.com/nytimes/covid-19-data