Team DSO 2020 COVID-19 Computational Challenge Technical Report

Gregory Faletto Mohammad Mehrabi

June 19, 2020

1 Introduction

COVID-19 has quickly upended daily life in Los Angeles. Using data science, we may be able to predict when and where the risk of contracting COVID-19 is highest. This would allow citizens to go about their lives normally when it is safe to do so and shelter-in-place less often only when needed.

Unfortunately, this is a difficult prediction task. The ideal data set would contain information on exactly where and when every person with COVID-19 contracted it. However, the available data is far from ideal:

- Most people who contract COVID-19 do not ever test positive for it [RKP+20].
- We have no way of knowing exactly where anyone got infected with COVID-19.
- Similarly, even among citizens who do test positive for COVID-19, it is hard to know with confidence what day they contracted it.
- Even "recovering" from COVID-19 is not necessarily well-defined; we do not yet fully understand the precise window of time in which someone who contracts COVID-19 is contagious.

As we will show in Section 4.1, the number of new test cases in a given neighborhood on a given day is relatively easy to predict with reasonable accuracy, given data on previous tests. However, this does not give a satisfying answer for how risky a given neighborhood will be on that day. We would prefer to know something about how many new *infections* will occur in the neighborhood on that day. Usually there will be a delay of at least one day between infection and testing, so tests conducted on a given day reflect infections from earlier days. For example, there is typically a lag of 5 to 6 days between infection with COVID-19 and the onset of symptoms [Bru20]. Many patients may not think to get tested before they show symptoms.

For this reason, the basic idea behind our risk scores was to attempt to predict the number of new infections in a given neighborhood on a given day. If predicted new infections per capita are high, then we deem the risk of contracting COVID-19 on that day in that neighborhood high.

The main data set we drew on was neighborhood-level test results from RMDS[1], which draws on the Los Angeles city data set [2]. Typically the data are updated at night or early in the morning. We designed our model so that when these new data come in, the model can generate predicted risk scores for the following day in each neighborhood. Further, our code refits the model every day as new data arrive, so the model will become more and more accurate as time goes on. Our code also makes it easy to tweak the parameters of the model if desired. Finally, our model is sophisticated, but it is also easily interpretable, unlike some machine learning models like deep neural networks or random forests. In these ways, our code and model is designed to be as useful as possible for making real-life decisions.

1.1 Outline of Report

In Section 2, we describe the data we use for our analysis. We walk through the methods we use in Section 3, and in Section 4 we describe our results, including a description of what features may increase or decrease risk of COVID-19 infection. In Section 5, we outline our implementation proposal, including our proposed methodology to implement risk score assessment. We walk through our risk mitigation recommendations in

[1] https://github.com/GRMDS/2020Covid19_challenge/blob/master/LA_County_Public_Health/LA_county_cases_city%26community.csv

[2] http://publichealth.lacounty.gov/media/Coronavirus/locations.html

Section 6, including actionable steps for risk mitigation and to improve risk scores. Since this analysis was done on a tight timeframe, in Section 7 we discuss next steps that could improve our risk scores further. Our acknowledgements are in Section 8, and our references are available at the end of our report.

2 Data

As our final goal is to predict risk scores across all neighborhoods, we need information about each area more than the history of the number of cases/infections. For example, because certain subpopulations are at a greater risk than others, it's important to know the percentage of population with age higher than 65, since the elderly are at higher risks of dying from COVID-19 (For more information about groups of people that are at higher risk, see the CDC report [fDCP20]). We used the following covariates (with sources provided):

1. Demographic information of each neighborhood (e.g. age distribution, racial makeup, etc.). [3]
2. Number of grocery stores in each neighborhood. [4]
3. Number of businesses in each neighborhood. [5]
4. Percentage of people having diabetes[6], heart disease [7], or asthma [8].
5. Annual income level of people at each neighborhood [9]
6. Average temperature on each day in Los Angeles (as measured at Bob Hope Airport Station in Burbank, CA)[10]
7. An indicator variable for whether each day was during Stage 1 or Stage 2 of Governor Gavin Newsom's roadmap for reopening[11]
8. An indicator variable for whether each day had major protests in the wake of the killing of George Floyd (the first major day of protests is defined to be May 30 2020, since this was the first day that there was a curfew in Los Angeles in response to the protests. The protests are considered by our model to be ongoing as of June 8 2020.)

A note on the shelter-in-place orders: shelter-in-place was declared in Los Angeles and statewide on March 19[12]. However, it appears that the number of test cases did not level off into linear growth until around March 30. This may be because of a lag in full compliance with the shelter-in-place orders, plus positive test results lag infections for the reasons already outlined. For this reason, we defined Stage 1 of the shelter-in-place orders as starting on March 30 2020 for the purposes of our analysis. See Figures 1 and 2.

As mentioned earlier, the main data set we drew on for neighborhood-level test results was provided by RMDS using Los Angeles city data at https://github.com/GRMDS/2020Covid19_challenge/blob/master/LA_

[3] http://geohub.lacity.org/datasets/demographics-of-neighborhood-councils/data?page=8&selectedAttribute=density
[4] https://data.lacity.org/A-Prosperous-City/Grocery-Stores/g986-7yf9
[5] https://data.lacity.org/A-Prosperous-City/Listing-of-Active-Businesses/6rrh-rzua
[6] https://geohub.lacity.org/datasets/ladot::prevalence-of-adult-diabetes-2013-2014
[7] https://geohub.lacity.org/datasets/ladot::prevalence-of-adult-heart-disease-2013-2014
[8] https://geohub.lacity.org/datasets/ladot::prevalence-of-adult-asthma-2013-2014
[9] http://www.laalmanac.com/employment/em12c.php
[10] https://www.wunderground.com/history/monthly/us/ca/burbank/KBUR/date/2020-3
[11] https://www.latimes.com/projects/california-coronavirus-cases-tracking-outbreak/reopening-across-counties/
[12] https://www.latimes.com/projects/california-coronavirus-cases-tracking-outbreak/reopening-across-counties/

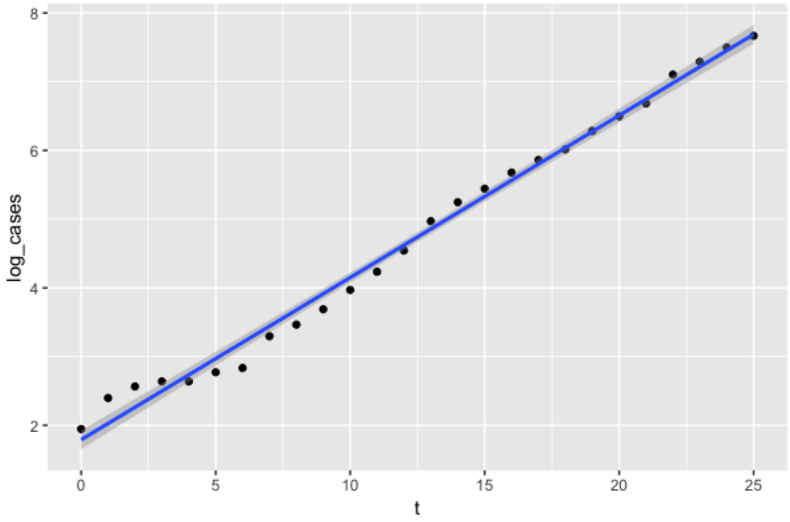

Figure 1: Plot of natural logarithm of total cases in Los Angeles against time from March 4 2020 through March 29 2020, along with linear trendline. We see that a linear model fits the data well, suggesting exponential growth in number of confirmed cases during this time.

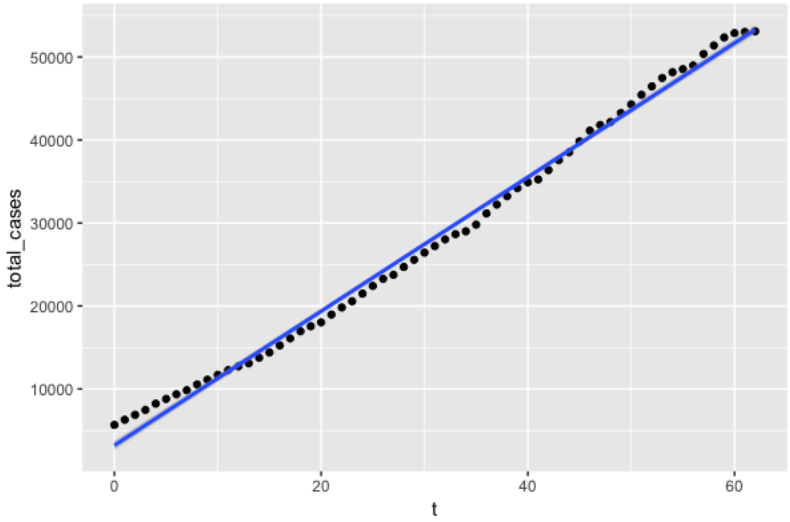

Figure 2: Plot of total cases in Los Angeles against time from March 30 2020 through the end of May, along with linear trendline. We see that a linear model fits the data well, suggesting linear growth in number of confirmed cases during this time.

County_Public_Health/LA_county_cases_city%26community.csv. Figure 3 shows a visualization of some example data from five neighborhoods in the data set.

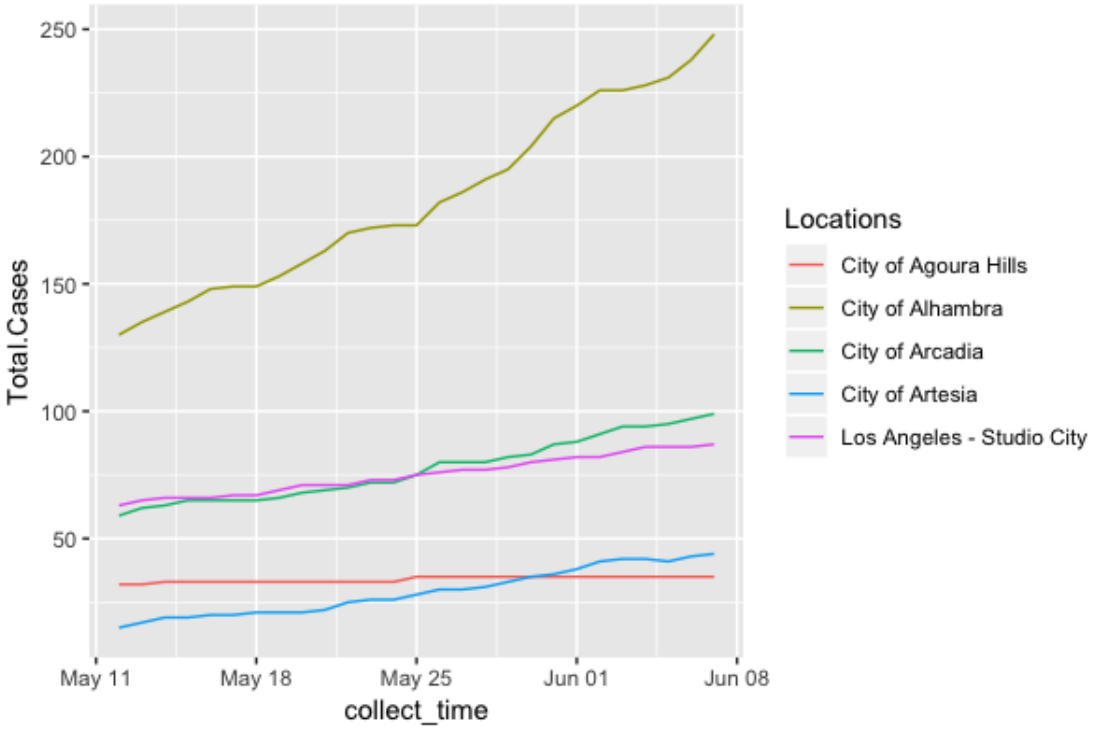

Figure 3: Plot of total cases in a few Los Angeles neighborhoods over time, from the data set at https://github.com/GRMDS/2020Covid19_challenge/blob/master/LA_County_Public_Health/LA_county_cases_city%26community.csv.

3 Methodology

With the goal of creating daily risk scores for every neighborhood in Los Angeles in mind, we decided to try to predict the number of new infections that will occur in each neighborhood tomorrow based on the most recent test results available today.

3.1 Model

One of the most fundamental models for predicting the spread of infectious disease is the SIR model [KM27]. In this model, the population through which a disease spreads is divided into three groups. The *susceptible* population is the group that has not yet contracted the disease, the *infected* group has the disease and may spread it to others, and the *recovered* group has recovered from the disease and has antibodies, so people in this group cannot contract the disease again. The SIR model is defined by three simple differential equations governing the flow of people in and out of these groups. Despite its simplicity, it has had notable success in predicting the trajectories of many diseases [BCC12], including COVID-19 [RRS19].

While our approach does draw on some of the principles of the SIR model, ultimately we decided against predicting new infections using an SIR model. The first reason why is that the SIR model does not include

covariates. Using an SIR model, we would not have been able to include variables characterizing each neighborhood (like density, demographics, health characteristics, and so on) to aid our predictions. Second, we would have had to fit a different SIR model for every neighborhood, rather than "borrowing strength" across neighborhoods by fitting one model to make predictions for every neighborhood. While each neighborhood has its own characteristics, every neighborhood in Los Angeles has common characteristics as well (known as fixed effects). What happens in one neighborhood can tell us about what might happen in the next, so we preferred a model that reflected that.

Ultimately, we decided on a more modern approach. [RRS19] model the spread of an influenza outbreak in a boys boarding school using an integer-valued generalized autoregressive conditional heteroskedasticity (INGARCH) model. In particular, under the assumption that the number of infections on day t, Y_t, is a Poisson(λ_t) random variable, and using a log link function, the traditional INGARCH model has the form

$$\log(\lambda_t) = \beta_0 + \sum_{j=1}^{p} \beta_j \log\left(Y_{t-j}\right) + \sum_{\ell=1}^{q} \alpha_\ell \log\left(\lambda_{t-\ell}\right), \tag{1}$$

where $\beta_0, \beta_1, \ldots, \beta_p, \alpha_1, \ldots, \alpha_q \in \mathbb{R}$ are coefficients.

[RRS19] extend this model to include covariates, using the model

$$\log(\lambda_t) = \beta_0 + \sum_{j=1}^{p} \beta_j \log\left(Y_{t-j}\right) + \sum_{\ell=1}^{q} \alpha_\ell \log\left(\lambda_{t-\ell}\right) + \boldsymbol{\eta}_t^\top \boldsymbol{X}_t, \tag{2}$$

where $\boldsymbol{\eta}_t \in \mathbb{R}^k$ are fitted coefficients and $\boldsymbol{X}_t \in \mathbb{R}^k$ is a set of k covariates. [RRS19] rely on only one covariate: the number of newly recovered students on a given day (the number of students who transition from the "Infected" group to the "Recovered" group that day).

We modify model (2) for our problem in a couple ways. First, because we seek to model the trajectory of COVID-19 in many neighborhoods rather than just one, we use a panel data model instead of a time series model. That is, we interpret the observed case counts in each neighborhood to be independent realizations of the same time series process, rather than assuming we only observe one time series following this process. Second, because we were unable on short noticed to find software that could fit such a model, we omit the lagged values of λ_t from our regression (that is, we restrict q to equal 0). These adjustments result in the following model: let Y_{it} represent the number of new infections in neighborhood i on day t. Then we model Y_{it} as

$$\begin{aligned} Y_{it} \mid \boldsymbol{X}_{it}, Y_{i,t-1}, \ldots, Y_{i1} &\sim \text{Poisson}(\lambda_{it}), \quad \text{where} \\ \log(\lambda_{it}) &= \beta_0 + \sum_{j=1}^{p} \beta_j \log\left(Y_{i(t-j)}\right) + \boldsymbol{\eta}^\top \boldsymbol{X}_{it}. \end{aligned} \tag{3}$$

That is,

$$\mathbb{E}\left[Y_{it} \mid \boldsymbol{X}_{it}, Y_{i,t-1}, \ldots, Y_{i1}\right] = \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j \log\left(Y_{i(t-j)}\right) + \boldsymbol{\eta}^\top \boldsymbol{X}_{it}\right). \tag{4}$$

This is similar to the approach used by [Ran18] to model cyberattacks in cybersecurity risk modeling, though [Ran18] omits covariates $\boldsymbol{X}_{it}$ in her model.

The model (3) addresses the concerns we had about the SIR model. We are free to include covariates $\boldsymbol{X}_{it}$ for each neighborhood (that may be time-dependent), and by assuming that each neighborhood's disease trajectory is a realization of the same process (3), we exploit the similarities across neighborhoods in Los Angeles (and the fact that we are modeling the same disease in each neighborhood).

3.2 Predicting New Test Cases

We first use the model (3) to predict cumulative test cases on a given day in Los Angeles. That is, we fit model (3) letting Y_{it} equal the cumulative number of positive tests in neighborhood i at time t. Given data on test cases in each neighborhood for days $1, \ldots, t-1$ of the COVID-19 pandemic, we fit the model (3) and forecast the number of test cases in neighborhood i on day t using the estimator

$$\hat{\mathbb{E}}\left[Y_{it} \mid \boldsymbol{X}_{it}, Y_{i,t-1}, \ldots, Y_{i,1}\right] = \exp\left(\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j \log\left(Y_{i(t-j)}\right) + \hat{\boldsymbol{\eta}}^\top \boldsymbol{X}_{it}\right). \tag{5}$$

3.3 Predicting New Infections

As we described in the Introduction, we believe that predicting new test results is not a good indicator for the actual risk level on a given day. With those concerns in mind, we sought to estimate the number of people who will be *newly infected* in neighborhood i on day t based on test results available from days $1, \ldots, t-1$. We again used model (3) for this task, but this time the response variable is Z_{it}, the number of people newly infected in neighborhood i on day t. The tricky part was attempting the map the number of positive test results on day t to a number of new infections on some previous day.

To do this, we made a number of assumptions.

1. We borrowed from the SIR model and assumed that the population of each neighborhood could be divided into susceptible, infected, and recovered groups.
2. We assumed that infected individuals who seek testing do so 7 days after infection. This seemed like a reasonable average because some people may be tested within a day or two of infection (for example, if they are a medical professional or other member of a vulnerable group who may seek frequent testing, or if they get tested early because they have a strong reason to believe they may have been infected), and most people who get tested may be tested within a few days of showing symptoms.
3. Following [RKP+20], we assumed that only one eighth of the people who get infected test positive for COVID-19 (that is, for every positive test result there are 8 actual infections in the population).
4. Based on [VOD+20], we assumed that patients move from the infected group to the recovered group 23 days after infection (where the "recovered" group includes both patients who recovered from COVID-19 and patients who died from COVID-19).
5. Unfortunately, neighborhood-level data on test cases only goes back to May 12. However, city-wide data is available prior to May 12. We assumed that the new test cases every day prior to May 12 were split proportionally among neighborhoods in Los Angeles by population.

These are oversimplifying assumptions; for example, clearly the number of days between infection and testing varies from person-to-person. However, hopefully these assumptions are close enough to give reasonable estimates for the number of people infected (and recovered) at a given time (A more algorithmic way to estimate the true exposure to infection rather than a fixed shift is provided in [XRG+20], section 3). The numbers for all of these assumptions are easily changed in our code in order to accommodate different assumptions if this is desired.

These assumptions led to estimates of Z_{it}, the number of people who were newly infected in neighborhood i on day t. These assumptions also yield estimates for the number of newly recovered people on each day; like [RRS19], we include this as a predictor in our model.

Note that this prediction task is much more difficult than predicting the number of test cases. First, though we have data on tests up to day $t-1$, these tests reflect infections only up to day $t-8$. Having a lag of 8 days rather than 1 makes prediction much harder. Second, the data we use to fit our model are noisy estimates of the real data; the actual number of infections on a given day in a given neighborhood is unobservable.

Again, after determining the number of lags to use, we fit the model (3) and forecast the number of test cases in neighborhood i on day t using the estimator

$$\hat{\mathbb{E}}\left[Z_{it} \mid \boldsymbol{X}_{it}, Y_{i,t-1}, \ldots, Y_{i,1}\right] = \exp\left(\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j \log\left(Y_{i(t-j)}\right) + \hat{\boldsymbol{\eta}}^\top \boldsymbol{X}_{it}\right). \tag{6}$$

3.4 Estimating The Models

After deciding on a number of lags p and estimating coefficients $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p, \hat{\boldsymbol{\eta}}$ for each model, we estimated the models using maximum likelihood estimation using the programming language R [R C19] and the R package `pglm` [Cro17]. We note that we obtained different results on two different computers running different versions of R. We ruled out the possibility that this is due to randomization by setting seeds in the code right before fitting each model. Our best case is that this has to do with quirks in exactly how the optimization algorithm runs depending on the version of R used. The results should be replicable using the same version of R and the `pglm` package that we used (note the versions in the References section).

We also note here that the model (1) is motivated by theory and consistency results are available when the model is fit by maximum likelihood estimation. Even for the panel data extension (3) with no covariates, [Ran18] cites results from [Bol86] and [Whi82] to show that the maximum likelihood estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p, \hat{\alpha}_1, \ldots, \hat{\alpha}_q$, when appropriately scaled, are asymptotically normally distributed with unbiased mean and stable variance. However, no asymptotic consistency results are available when extending this model to include covariates, as in (2) and our model (3). The predictions from our model should still be reliable, but the p-values generated for the coefficients by the `pglm` package are not trustworthy at face value.

3.5 Generating Predictions for Neighborhoods With Missing Data

We were able to collect data for 78 neighborhoods in Los Angeles, but unfortunately there remained other neighborhoods for which we did not have full data. In order to generate some kind of predictions for the remaining neighborhoods, we adopted the technique of *synthetic controls* from causal inference for panel data. See Figure 4 to illustrate the idea (figure borrowed from [ABD+18]). Each row represents a neighborhood and each column represents a day. The check marks represent day/neighborhood pairs for which we observe

data (or make predictions). Because of missing covariates, we are unable to make predictions for some of the neighborhoods; these missing predictions are denoted in the figure by question marks.

$$\mathbf{Y}_{N\times T} = \begin{pmatrix} \checkmark & \checkmark & \checkmark & \checkmark & \dots & \checkmark \\ \checkmark & \checkmark & \checkmark & \checkmark & \dots & \checkmark \\ \checkmark & \checkmark & \checkmark & \checkmark & \dots & \checkmark \\ \checkmark & \checkmark & \checkmark & ? & \dots & ? \\ \checkmark & \checkmark & \checkmark & ? & \dots & ? \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \checkmark & \checkmark & \checkmark & ? & \dots & ? \end{pmatrix}$$

Figure 4: Figure borrowed from [ABD+18].

The idea is that we will attempt to approximate the number of infections in each neighborhood for which we cannot make predictions by a linear combination of the number of infections in each neighborhood for which we can make predictions. We can do this for each neighborhood with missing data by regressing the observed infections in the missing data neighborhoods against the observed infections in the full data neighborhoods on the past dates for which data is available for all neighborhoods (represented by the left three columns in Figure 4). These learned coefficients are then used as weights to combine the forecasts for the known neighborhoods into forecasts for the missing neighborhoods.

4 Results

4.1 Predicting New Test Cases

Our model fit to predict new test cases works yield reasonably satisfying results. We fit a model on the full data set (using data available as of the night of Sunday 6/7/20) to predict the number of new test cases that would appear in the data arriving on Monday 6/8/20. The fitted coefficients are presented in Table 1, and a sampling of predictions for 20 neighborhoods is shown in Table 2.

To validate our approach, we fit the same model using only data from the first 22 days of data available, holding out the last 5 days as a test set. The out-of-sample root mean squared error (RMSE) was 3.364532. This suggests that a typical one-day-ahead out-of-sample forecast for the cumulative number of positive test cases from this model will be within a margin of error of 7 test cases away from the actual number of positive test cases, so this model's predictions are fairly accurate.

Figure 5 shows a plot of observed positive test results against both in-sample and out-of-sample predictions from the model, using Studio City as a representative of a typical neighborhood. We see that the predictions track fairly closely with the observed data, and that the out-of-sample predictive performance is similar to the in-sample predictive performance, suggesting that our model is not overfitting.

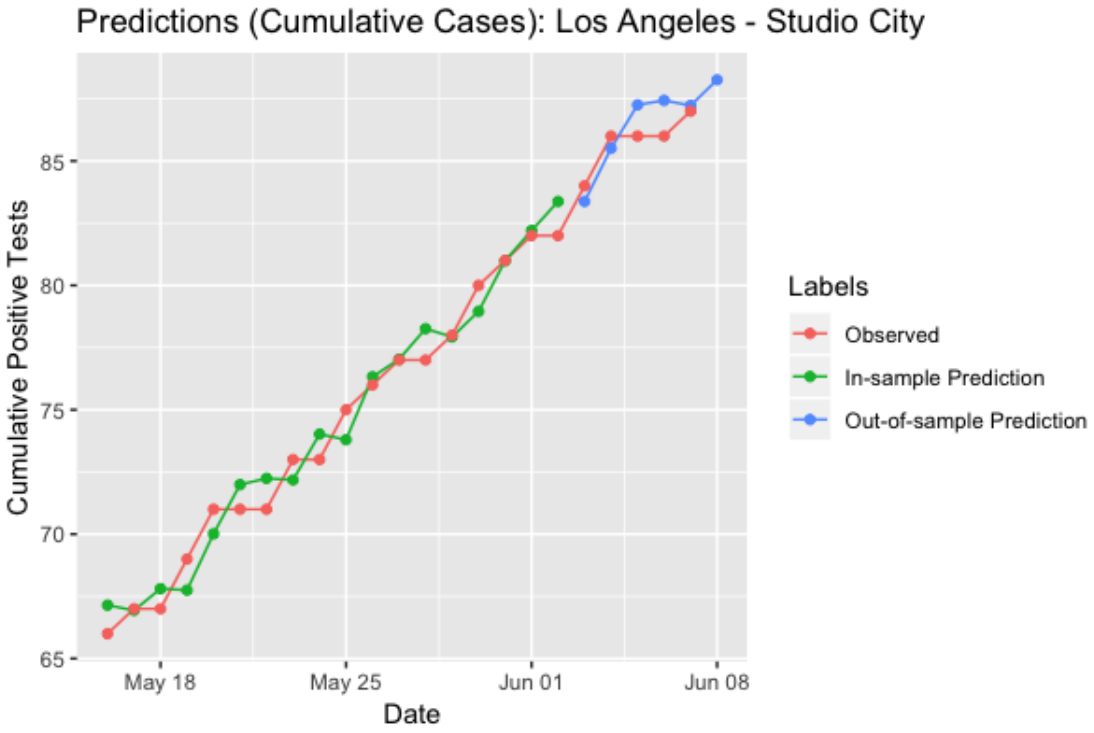

Figure 5: Predicted cumulative positive test results for COVID-19 in Studio City based on a model trained on the first 22 days of data available. The predictions are using (4) from the coefficients fit via maximum likelihood using the R `pglm` *package. The red data points represent observed positive test cases, the green data points represent one-day-ahead forecasts for days in the training set, and the blue data points represent one-day-ahead forecasts for dates in the held-out test set. (There is no observed number of positive tests for June 8, since data for June 8 was not available as of the training of this model.)*

Table 1: Regression output for model fit on the full data set to predict new test cases on 6/8/20. As we warned in Section 3.4, the asymptotic distribution of the coefficients in this model including covariates is not currently known, so we omit the p-values reported by `pglm` *since they are not valid for our model.*

	Dependent variable:
	Total.Cases
(Intercept)	2.5298016
$\log(Y_{i,t-1})$	1.1324749
$\log(Y_{i,t-2})$	-0.0954060
$\log(Y_{i,t-3})$	0.0389095
$\log(Y_{i,t-4})$	-0.0751085
black_pct	0.0108065
pct_65_plus	-0.0279372
grocery_count	-0.0001381
diabetes_perc	0.0002299
Asthma_perc	-0.0009108
Avg.Temp	-0.0001457
shelter1	0.0000000
shelter2	-2.4956366
protests	-0.0012117
Observations	2,106
Log-Likelihood	−5635.623

4.2 Predicting New Infections

We fit the same model to predict new infections, except that following [RRS19], we include as a predictor the number of newly recovered people on day $t - 1$.[13] Again, we fit a model on the full data set (using data available as of the night of Sunday 6/7/20) to predict the number of new infections that would occur on Monday 6/8/20. The fitted coefficients are presented in Table 3, and a sampling of predictions for 20 neighborhoods is shown in Table 4.

We again validated our approach by fitting the same model using only data from the first 22 days of data available, holding out the last 5 days as a test set. The out-of-sample RMSE in this case was significantly higher, at 44.45756. This suggests that this model provides some indication of what infections might look like tomorrow, but this model and the assumptions behind it may be a bit too simple for reliable predictions in this form.

Figure 6 shows a plot of estimated new infections against both in-sample and out-of-sample predictions from the model, using Studio City as a representative of a typical neighborhood. We see that the predictions track decently well with the observed infections, though the fit is not nearly as good as in the model for positive test results. This suggests that the lag in infection data—as well as the fact that infections have to be inferred rather than directly observed—makes predictions difficult.

[13] Recall that the most recent infection data available are infections occurring 8 days prior to the day of a desired forecast due to the assumed seven-day lag between the date of infection and the date of testing. However, under the assumption that patients enter the recovered population 23 days after infection, we can know how many newly recovered patients there will be on only a one-day lag—it will be the number of patients who were newly infected 24 days earlier.

Table 2: A sampling of predicted new test cases for 20 neighborhoods in Los Angeles from data arriving on Monday 6/8/20 based on data available the night of Sunday 6/7/20.

Locations	Date	Predicted New Positive Test Cases
City of Beverly Hills	2020-06-08	149
City of Commerce	2020-06-08	93
City of Culver City	2020-06-08	167
City of West Hollywood	2020-06-08	191
Los Angeles - Alsace	2020-06-08	61
Los Angeles - Angelino Heights	2020-06-08	22
Los Angeles - Arleta	2020-06-08	311
Los Angeles - Atwater Village	2020-06-08	53
Los Angeles - Beverly Crest	2020-06-08	37
Los Angeles - Century Palms/Cove	2020-06-08	372
Los Angeles - Country Club Park	2020-06-08	111
Los Angeles - Eagle Rock	2020-06-08	204
Los Angeles - East Hollywood	2020-06-08	249
Los Angeles - Echo Park	2020-06-08	55
Los Angeles - Elysian Park	2020-06-08	16
Los Angeles - Elysian Valley	2020-06-08	67
Los Angeles - Encino	2020-06-08	146
Los Angeles - Exposition Park	2020-06-08	314
Los Angeles - Figueroa Park Square	2020-06-08	46
Los Angeles - Glassell Park	2020-06-08	201

Table 3: Regression output for model fit on the full data set to predict new infections on 6/8/20. (Recall that because of the assumed seven-day lag between the date of infection and the date of testing, the most recent infection data available are infections occurring 8 days prior to the day of a desired forecast.) As we warned in Section 3.4, the asymptotic distribution of the coefficients in this model including covariates is not currently known, so we omit the p-values reported by `pglm` *since they are not valid for our model.*

	Dependent variable:
	New_Infected
(Intercept)	-1.0124960
$\log(Z_{i,t-8})$	0.1117348
$\log(Z_{i,t-9})$	0.0290261
$\log(Z_{i,t-10})$	0.0534825
$\log(Z_{i,t-11})$	0.0438092
black_pct	-0.3326574
pct_65_plus	-8.6746600
grocery_count	0.0000172
diabetes_perc	-0.0054386
Asthma_perc	0.0537494
Avg.Temp	0.0946105
shelter1	-1.9647309
shelter2	-2.3384509
protests	0.1980564
$\log(\text{New\_Recovered}_{t-1})$	0.0710361
Observations	5,382
Log-Likelihood	-194993.7

Table 4: A sampling of predicted new infections for 20 neighborhoods in Los Angeles on Monday 6/8/20 based on data available the night of Sunday 6/7/20.

Locations	Date	Predicted New Infections
City of Beverly Hills	2020-06-08	15
City of Commerce	2020-06-08	34
City of Culver City	2020-06-08	25
City of West Hollywood	2020-06-08	36
Los Angeles - Alsace	2020-06-08	19
Los Angeles - Angelino Heights	2020-06-08	25
Los Angeles - Arleta	2020-06-08	52
Los Angeles - Atwater Village	2020-06-08	14
Los Angeles - Beverly Crest	2020-06-08	8
Los Angeles - Century Palms/Cove	2020-06-08	55
Los Angeles - Country Club Park	2020-06-08	31
Los Angeles - Eagle Rock	2020-06-08	30
Los Angeles - East Hollywood	2020-06-08	33
Los Angeles - Echo Park	2020-06-08	31
Los Angeles - Elysian Park	2020-06-08	17
Los Angeles - Elysian Valley	2020-06-08	24
Los Angeles - Encino	2020-06-08	16
Los Angeles - Exposition Park	2020-06-08	50
Los Angeles - Figueroa Park Square	2020-06-08	12
Los Angeles - Glassell Park	2020-06-08	41

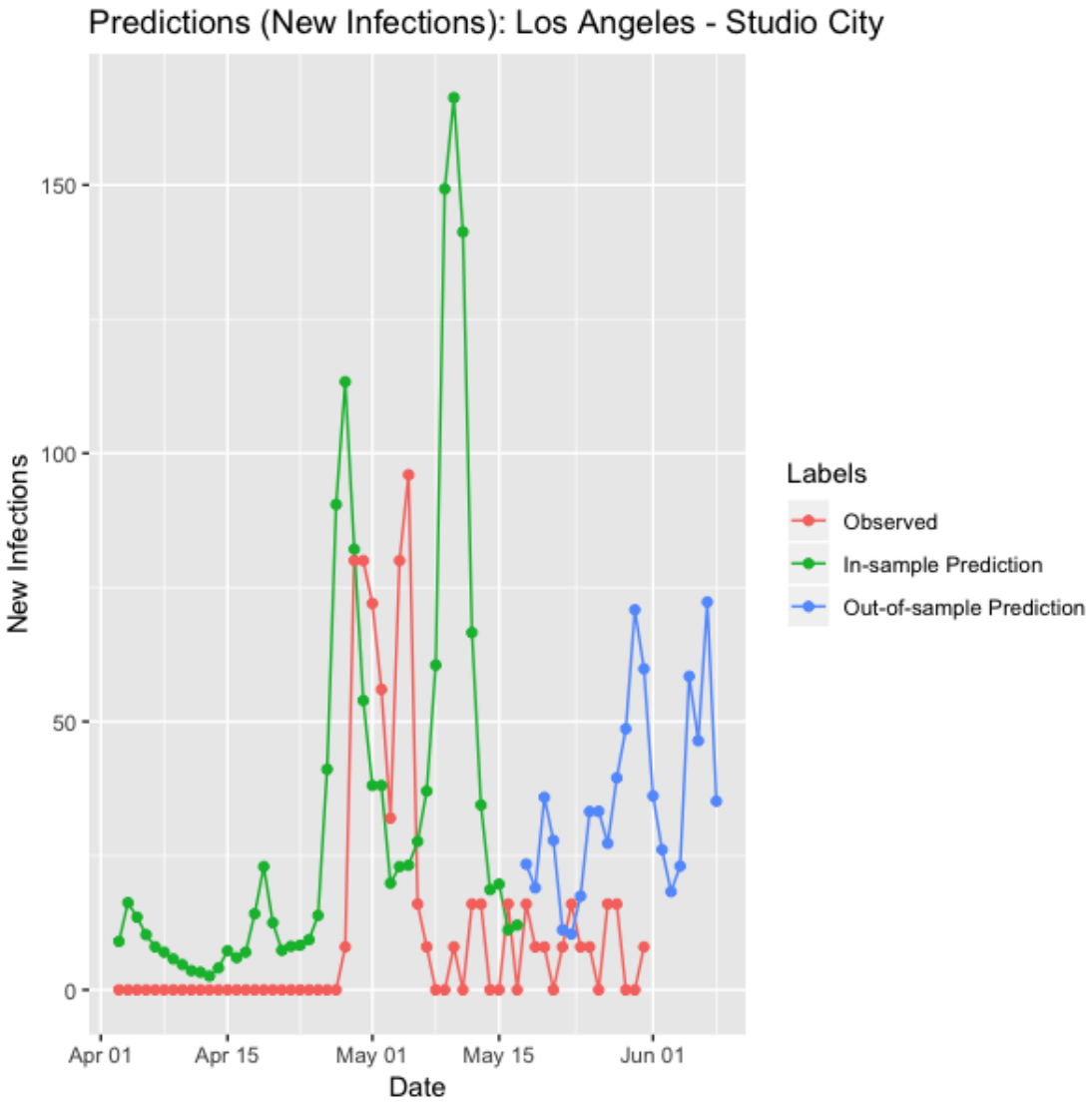

Figure 6: Predicted new COVID-19 infections in Studio City on each date based on a model trained on the first 22 days of data available. The predictions are using (5) from the coefficients fit via maximum likelihood using the R `pglm` *package. The red data points represent infections estimated from observed data, the green data points represent one-day-ahead forecasts for new infections for dates in the training set, and the blue data points represent one-day-ahead forecasts for new infections for dates in the held-out test set. (There is no observed number of new infections for the most recent 7 days, since only testing data through June 7 was available as of the training of this model, representing new infections through May 31 under our assumptions.)*

Note that the coefficients for the neighborhood and time covariates are generally much larger in this model than in the model for predicting positive test results. This is because the large lag in infections is a much weaker signal for prediction than the short lags in test results.

Again, the prediction task of predicting new infections is more difficult than predicting cumulative test cases, for the reasons outlined in Section 3.3. Nevertheless, the results of our model form for predicting test cases suggests that this model is promising for predicting the trajectory of COVID-19. These results show a promising initial attempt at predicting new infections on a given day using test data available the day before. In Section 7, we outline some next steps for this kind of analysis that could lead to improved predictions.

4.3 Features that may increase or decrease COVID-19 exposure risks

We can interpret the signs of the coefficients in Table 3 to get a sense of what factors lead to risk of COVID-19 exposure. Because the model is trained on observational data, we cannot make valid causal claims about the coefficients in this model (i.e., we cannot interpret a positive coefficient as implying that that factor causes an increase in new infections). However, we can start to draw some inferences about associations between variables and risk of COVID-19 infection.

- The shelter-in-place orders both seem to be associated with a decrease in new infections of COVID-19. This suggests that the shelter-in-place orders have been quite effective.
- Neighborhoods with higher prevalence of asthma have higher incidence of infections from COVID-19. This could suggest that having pre-existing respiratory conditions makes contracting COVID-19 more likely.
- The protests are associated with an increase in COVID-19 infection rates. This makes sense, since the protests contain large crowds of people gathering, not always with masks.

Some of the results from this regression are surprising or contradict otherwise known information about COVID-19.

- The percentage of black residents of a given neighborhood is negatively associated with risk of contracting COVID-19 on a given day, even though most studies so far suggest the opposite is true. (Indeed, even our own model for predicting test cases suggests a positive association between the percentage of black residents of a neighborhood and likelihood of contracting COVID-19.)
- The percentage of residents over 65 in a given neighborhood is sharply negatively associated with risk of contracting COVID-19. This seems counterintuitive since elderly individuals are at higher risk of dying from COVID-19 if they contract it. However, this could make sense if it means that elderly Angelenos are taking special care to avoid risks that could lead to contracting COVID-19.
- Increasing average temperatures seem to be associated with an increase in risk of contracting COVID-19, even though existing evidence indicates the opposite might be true.

These counterintuitive results could have arisen because of the inclusion of unneeded covariates in our model, leading to biased estimates of the coefficients of the model. More exploration of this model and data would be needed to make stronger conclusions.

Some other results are worth mentioning: we initially included density and income as variables in our model, but they did not seem to predict infections or positive tests well. This matches some analysis that suggests that city-wide density is not a strong predictor for infection, though crowdedness within a home is[14].

5 Implementation Proposal

Table 5: Risk scores for a sample of 20 neighborhoods in Los Angeles for June 8, 2020.

Locations	Predicted New Infections	New Infections Per Capita	Risk Level
City of Beverly Hills	15	0.001	Medium Risk
City of Commerce	34	0.003	High Risk
City of Culver City	25	0.002	High Risk
City of West Hollywood	36	0.005	High Risk
Los Angeles - Alsace	19	0.001	High Risk
Los Angeles - Angelino Heights	25	0.001	Medium Risk
Los Angeles - Arleta	52	0.006	High Risk
Los Angeles - Atwater Village	14	0.004	High Risk
Los Angeles - Beverly Crest	8	0.001	Medium Risk
Los Angeles - Century Palms/Cove	55	0.002	High Risk
Los Angeles - Country Club Park	31	0.002	High Risk
Los Angeles - Eagle Rock	30	0.001	Medium Risk
Los Angeles - East Hollywood	33	0.001	Medium Risk
Los Angeles - Echo Park	31	0.002	High Risk
Los Angeles - Elysian Park	17	0.001	Medium Risk
Los Angeles - Elysian Valley	24	0.006	High Risk
Los Angeles - Encino	16	0.0003	Medium Risk
Los Angeles - Exposition Park	50	0.002	High Risk
Los Angeles - Figueroa Park Square	12	0.001	Medium Risk
Los Angeles - Glassell Park	41	0.003	High Risk

We used these predictions to generate risk scores. We looked at the past data of new infections on each day in each neighborhood and divided these new infections by population to get per capita new infections each day in each neighborhood. Then after predicting new infections in each neighborhood, we divided by population to get per capita predicted new infections. If these per capita new infections were in the top third of per capita new infections from the earlier data, we deemed this day a "high risk" day in that neighborhood. Predicted per capita infection rates in the middle third were deemed "medium risk" days, and predicted per capita infection rates in the bottom third were labeled "low risk" days.

Figure 7 contains a visualization of our risk scores across Los Angeles County. Table 5 contains our predicted risk scores for June 8, 2020 based on data available the night of June 7, 2020. Note that all of the risks are "medium" or "high;" this matches perceptions that we are near a relative high in the pandemic's outbreak so far.

Evaluating our models can be done by retroactively estimating how many infections occurred on a given day in a given neighborhood and comparing them to our model's predictions.

[14] https://www.wsj.com/articles/covid-19-households-spread-coronavirus-families-navajo-california-second-wave-11591553896

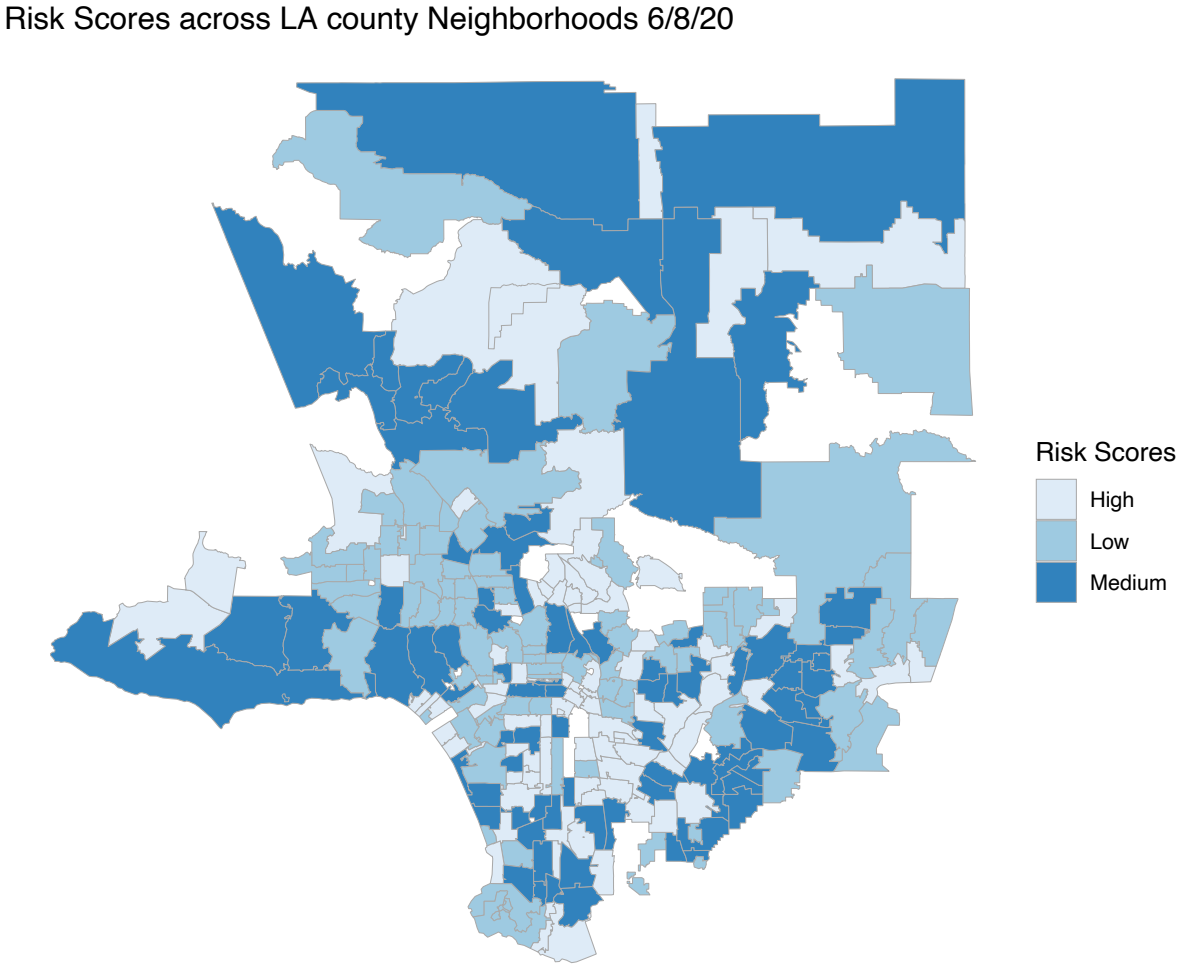

Figure 7: Visualization of predicted risk scores across Los Angles on June 8, 2020 using data available the night of June 7, 2020.

We have already shown how our code can be updated with new data the night before (or, at latest, the morning of) a new day using data from `https://github.com/GRMDS/2020Covid19_challenge/blob/master/LA_County_Public_Health/LA_county_cases_city%26community.csv` in order to generate risk scores for a new day. Further, if data arrive late for whatever reason, the code is easily tweaked to allow for a longer delay in reporting data in order to generate predictions anyway, though the predictions will likely be of worse quality. Our code and model are designed for simple implementation and use.

6 Risk Mitigation Recommendations

Neighborhoods and businesses in Los Angeles want actionable steps for risk mitigation and to improve their risk scores. Unfortunately, the coefficients in our model suggest that there are no quick fixes, and the best way to avoid contracting COVID-19 is to stay home. That said, our data also indicate that individuals with pre-existing respiratory conditions like asthma should be particularly careful. Our model also suggests that increasing temperatures may not help with avoiding COVID-19 as much as hoped.

Still, this is a preliminary analysis. More analysis of the data could lead to more robust conclusions and more specific recommendations for neighborhoods and businesses in Los Angeles.

7 Next Steps

More can be done to improve our model's predictions (and consequently, the risk scores). As mentioned earlier, the assumptions made in Section 3.3 are oversimplifying. More sophisticated assumptions would lead to better data to fit the model, which would likely lead to a big improvement in predictions. For example, rather than assuming everyone gets tested seven days after infection, we could assume there are two separate populations of people—at-risk populations like essential workers who get tested regularly, and typical citizens who get tested only after showing symptoms. We could also assume that the amount of time between infection and testing (as well as the amount of time between infection and recovery) follow probability distributions rather than being fixed numbers. With more time, we could have even replicated the analysis done by [RKP+20] to estimate how many Angelenos have COVID-19 for each positive test case, rather than assuming the 8x multiplier they found applies to Los Angeles as well.

Data cleaning would also help; in particular, the number of test results that are available the first day the results come out is lower than the number of test results available for that day a few days later, when all the tests are fully assessed. We did not account for this in our model, so accounting for this could help our predictions.

We also assumed that each neighborhood is a more or less "closed system" and that infections in a neighborhood result from other positive cases in that neighborhood. Accounting for infections between neighborhoods could improve our model's performance.

Another source of improvement would be more careful feature selection. We put some thought into what covariates to include in the model, and we omitted some variables that we originally included because they didn't seem to affect predictions much. However, in an analysis with more time, feature selection could be done in a more deliberate and effective way. In particular, the number of lags for the model could be done more systematically, perhaps using cross-validated prediction error or a criterion like AIC or BIC.

Lastly, more covariates could be included in the model to aid prediction. In particular, we did not use any

mobility data to predict infections. This could be very helpful for predicting new infections, and it could tell us something about how mobility affects risk of contracting COVID-19.

We could also try machine learning methods like random forests and deep neural networks to predict infections, as done by [RRS19]. However, this would make interpreting our models harder and make it more difficult to make recommendations to reduce risk scores.

8 Acknowledgement

We thank RMDS, the City of Los Angeles, the Los Angeles County Department of Public Health, and the Chamber of Commerce for hosting this challenge and making data available.

References

[ABD+18] Susan Athey, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. Matrix Completion Methods for Causal Panel Data Models. NBER Working Papers 25132, National Bureau of Economic Research, Inc, October 2018.

[BCC12] F. Brauer and C. Castillo-Chavez. *Mathematical models in population biology and epidemiology*. Springer, New York, vol. 40 edition, 2012.

[Bol86] Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327, April 1986.

[Bru20] Bruce Aylward (WHO); Wannian Liang (PRC). Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19). *The WHO-China Joint Mission on Coronavirus Disease 2019*, 1(February):40, 2020.

[Cro17] Yves Croissant. *pglm: Panel Generalized Linear Models*, 2017. R package version 0.2-1.

[fDCP20] Centers for Disease Control and Prevention. *People Who Are at Higher Risk for Severe Illness*, 2020.

[KM27] W.O. Kermack and A.G. McKendrick. A Contribution to the Mathematical Theory of Epidemics. *Proceedings of the Royal Society of London, Series A*, 115(772):700–721, 1927.

[R C19] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019. R version 3.6.0 (2019-04-26) – "Planting of a Tree".

[Ran18] Yashika Ranjan. Dynamic generalized poisson panel-data models: An application to cybersecurity risk modeling. Masters thesis, Università Ca'Foscari Venezia, 2018.

[RKP+20] Lionel Roques, Etienne K Klein, Julien Papaix, Antoine Sar, and Samuel Soubeyrand. Using early data to estimate the actual infection fatality ratio from covid-19 in france. *Biology*, 9(5):97, 2020.

[RRS19] Maziar Raissi, Niloofar Ramezani, and Padmanabhan Seshaiyer. On parameter estimation approaches for predicting disease transmission through optimization, deep learning and statistical inference methods. *Letters in Biomathematics*, pages 1–26, 2019.

[VOD+20] Robert Verity, Lucy C Okell, Ilaria Dorigatti, Peter Winskill, Charles Whittaker, Natsuko Imai, Gina Cuomo-Dannenburg, Hayley Thompson, Patrick GT Walker, Han Fu, et al. Estimates of the severity of coronavirus disease 2019: a model-based analysis. *The Lancet infectious diseases*, 2020.

[Whi82] Halbert White. Maximum Likelihood Estimation of Misspecified Models. *Econometrica*, 50(1):1 – 25, 1982.

[XRG+20] Ran Xu, Hazhir Rahmandad, Marichi Gupta, Catherine DiGennaro, Navid Ghaffarzadegan, Heresh Amini, and Mohammad Jalali. Weather conditions and covid-19 transmission: Estimates and projections. *Available at SSRN 3593879*, 2020.