Modeling the Association Between Population Mobility and New COVID-19 Diagnoses in Urban Wisconsin

Jeremy Johnson, Kevin Kristensen, Liban Mohamed

Abstract

Local stay-at-home orders and economic shut-downs assume some connection between the spread of Coronavirus infection and activity at population-dense public areas. One way of measuring such activity is through 'mobility' statistics that track the fequency with which local populations visit such areas. We collect county-specific mobility data and Coronavirus case counts for the state of Wisconsin and develop a variety of models to predict the number of new cases given mobility changes and and infection counts in the recent past. We find that regression models are prone to overfit the training data, and that therefore some strategy is needed to mitigate this effect. We also find that the convolutional and recurrent neural networks perform similarly, though both outperform all the regression models.

1. Introduction

Local governments have been issuing a variety of stay-at-home orders and economic shut-downs in an effort to reduce the spread of Coronavirus. One way to measure the efficacy of such measures is by a variety of 'mobility' metrics, such as changes in workplace or grocery store attendance. Precise quantitave correlations have been established between such mobility changes and new Coronavirus diagnoses [1], reinforcing the importance of such measures. However, a predictive model, which uses past mobility data to predict the number of new Coronavirus diagnoses on a given day, has yet to be developed, to our knowledge. We have collected mobility data from urban counties in Wisconsin along with Coronavirus case counts [see section 3] to devleop a variety of predictive models, including neural networks and regression models, which we compare and evaluate.

2. Related/Similar Work

Managing responses to the COVID-19 epidemic has been a major challenge for governments, individuals and corporations for the majority of the year 2020 [9]. To date, there have been at least 1.5 million deaths due to COVID-19 worldwide, with at least 300 thousand in the United States alone [16].

The magnitude of the public health crisis renders understanding and modelling the spread of the epidemic one of the most important challenges facing the data centered communities today. This problem is made especially difficult because of the novelty of the disease, the lack of pharmacological interventions and the apparent heterogeneity of its behavior in different countries [10].

However, unlike other epidemics over the course of history history, technological advances allow access to large amounts of data that can assist in making predictions about the near-term course of the epidemic. In this report, we exploit data describing the mobility of users of Google services in order to develop models that predict the incidence of COVID-19 cases in some urban Wisconsin counties.

The standard toolset used to make predictions about the spread of epidemics consists in a family of models related to the SIR-model, originally developed by Kermack and KcKendrick in [11]. In its most basic form, the SIR model consists in a system of differential equations in the variables S, the portion of the population that is susceptible to infection, I, the portion of the population that is actively infected, and R, the portion of the population that has been removed, whether due to recovery or death. These variables are related to ach other and time via the system:

$$\begin{aligned} \frac{dS}{dt} &= -\frac{\beta IS}{N}\,, \\ \frac{dI}{dt} &= \frac{\beta IS}{N} - \gamma I\,, \\ \frac{dR}{dt} &= \gamma I\,. \end{aligned}$$

From the parameters β and γ can be derived the *basic reproduction number*, R_0, defined by

$$R_0 = \frac{\beta}{\gamma}\,,$$

which captures the expected number of infections directly due to a single infection in an otherwise susceptible population. There are a number of elaborations on the basic SIR model that account for aspects of epidemics like dynamic population size, incubation period and loss of immunity [12].

However, there is a major limitation of the basic SIR model for the purposes of understanding the spread of COVID-19 in particular. This limitation arises from the assumption implicit in the model that individuals from the susceptible and infectious portions of the population interact with a probability uniform across individuals and uniform across time.

An approach used to account for the lack of uniformity across populations is to model populations as weighted graphs ("social networks"), where each person or region is assigned a node and the edges connecting pairs of nodes are assigned weight equal to the probability of transmission. This technique is able to produce more realistic descriptions of disease spread over large distance scales than the naïve SIR model. However, characteristics of disease spread can vary dramatically in response to small changes in the connectivity of the underlying graph, and so it relies on having an accurate model of the social graph [13].

The difficulty is worse than simply that of identifying an accurate graph model for population interaction, because individuals have dramatically varied their social activity in response to changes in public policy, changing attitudes and beliefs about the disease, and social movements. Merely varying edge weights uniformly in response to some notion of generalized social temperature doesn't capture the behavior peculiar to this disease. For example, it has become common practice among some communities in the United States to preferentially increase social distance to elderly people, which decreases transmission probability along edges connecting elderly nodes to the rest of the graph [14]. One consequence of

this behavior has been a marked reduction in the case fatality associated to the disease, at least in part because more of the cases are found are in younger people [15].

As part of our background research, we examined the thirteen most prominent COIVD-19 forecast models to investigate the methods they use to skirt these difficulties. The majority [16, 17, 18, 19, 21, 22, 24, 25, 26, 27, 28] bite the bullet, devising elaborate adjustments to the SIR model that attempt to incorporate the dependency of epidemic dynamics on changing context. Elaborations in the UCLA model [22] include a rigorous attempt to quantify unreported cases and recoveries; in the IHME model [19] they include a detailed study of the effects of e.g. the seasonality of pneumonia sensitivity and mask compliance. Several of them require manual adjustments for policy changes: for example, the MIT DELPHI [18] model manually includes government intervention as a modulation of the infection rate, multiplying it by

$$1 - \frac{2}{\pi} \arctan(at + b) + c \exp^{d(t-f)^2},$$

where the parameters a, b, c, d, f are fit using data collected from previous lockdown cycles.

A major difficulty faced by these SIR-related models remains the dynamic behavior of the transmission parameters. The UMass model [24] takes a Bayesian approach to this problem, where the model out of the Los Alamos National Labs [26] performs a regression on the estimated parameters and extrapolates to the future.

Two of the models avoid the epidemiological SIR models altogether. The model out of the University of Arizona [20] fits a parabola to the curve of active infections plotted against cumulative infections for each peak during the course of the epidemic. The model has no internal epidemiological structure, so doesn't allow transmission parameters to vary in response to changing case rates. More impressive is the model from Georgia Tech's Aditya Lab [23], which is based on previous attempts to describe the course of influenza outbreaks. The model consists in a deep neural network trained on a bank of outbreaks of influenza-like illnesses and from which a continuous latent space of outbreak time-courses is learned. Its authors claim that it outperforms competitor models by up to 40% in standard flu-season severity prediction tasks. However, there are some concerns about its validity as a tool for modelling the COVID-19 outbreak, due to the novel nature of the disease and the unusual nature of interpersonal interactions.

Of the greatest interest to us for present purposes is the model developed by University of Texas COVID-19 Modeling Consortium [28], which is a SIR-related model, but one that estimates transmission parameters from observed social-distancing behavior. This social-distancing behavior is gathered from mobile phone GPS traces, which is correlated to SIR model transition probabilities using data collected over the course of the epidemic.

One of the deficits of the Texas model is that it only forecasts deaths, rather than positive cases. There are good reasons for this: deaths are surely counted where cases many cases are likely missed, and also varying levels of testing have changed the relationship between observed COVID-19 cases and the underlying spread of the disease. However, at this point in the progression of the epidemic, testing rates - and thus the relationship between observed cases and actual cases - has essentially stabilized, which provides an opportunity to augment the work of the Texas lab with this data.

3. Dataset

Due to the severity of the current pandemic, Google has temporarily provided access to aggregated, anonymized user mobility data [2], [3]. These data provide insight into movement trends over time across various categories of public locations. Specifically, the data describe per day, the deviation from baseline use of public places such as as retail and recreation centers, grocery stores and pharmacies, parks, transit stations, workplaces, and residential areas.

The data are county-specific and were aggregated and anonymized from mobile users who have turned on the Location History and Location Reporting settings in their Google Account, settings which are turned off by default. It is not clear to the authors whether the data are collected exclusively from Android users, though it is our assumption that they not biased by these specific users.

In order to simplify the assumptions of our model, and to make specific the question our model attempts to answer, we chose to use only data for Wisconsin. Furthermore, it was observed that data for park use and transit station occupancy were largely missing. We therefore use only data for retail and recreation centers, grocery stores and pharmacies, workplaces, and residential areas.

Even within these features, data are missing. We accounted for this in two ways. First, we excluded counties which are missing more than 10% of the entries from the residential feature. The residential feature was chosen for this purpose because it had the most missing data. There remained 23 of the original 71 counties. It was our loose assumption that this effectively restricted our scope to more 'urban' counties, helping to remove bias across counties. Second, we imputed the remaining missing entries as the mean of the existing data across the counties for each particular day. This resulted in 275 days of sequential data for 23 different counties in Wisconsin.

Time series data describing confirmed COVID-19 case counts at the county level were taken from the Johns Hopkins University COVID-19 database [4]. These data were collated with the mobility data collected above to produce a single dataset $X \in \mathbb{R}^{6325\times 5}$ with 5 features and 6325 samples, consisting, for a given county and day, of the percent deviations from baseline use of retail and recreation centers, grocery stores and pharmacies, workplaces, and residential areas, in order, and the total number of confirmed cases through that day in that county.

4. Approach

For the neural network models, we chose convolutional and recurrent neural networks to take advantage of the time-series nature of the data. In either method, the data are packaged into windows of specified input length, a potential offset, and label width. Each sample fed into the neural network therefore consists of some number of input days for which the model considers all five features, a number of offset days to ignore, then finally a label day for which the model predicts the number of new infections. The model is then applied to sliding windows of inputs. Refer to Figure 1 for an example window.

The neural network models were created using the TensorFlow Keras library [6]. The models were run using Mean Square Error loss functions and ADAM-Optimizer. The models

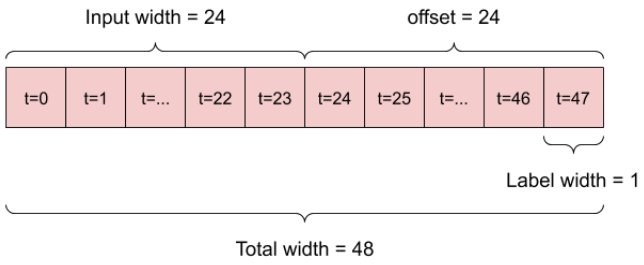

Figure 1: An example window which uses data from days 0-23 to make a prediction about day 47.

are compared using Mean Absolute Error. We chose to use 16 counties chosen at random as test data, 4 counties as validation data, and 3 counties as test data. It would likely be preferred to split the training, validation, and testing data randomly but was foregone for ease of processing. These windows are created for each county, placed into batches of 32, then shuffled. The models are run over all batches for 20 epochs.

Either model runs over the samples in a slightly different way. For the convolution neural network, the model uses the previous 7 days of data for each prediction. It then slides over for a total of 24 days worth of predictions. We used a recurrent neural network called Long Short Term Memory (LSTM) which stacks layers, making a prediction for each input and maintains memory of each model used before. Refer to Figure 2 for schema of the neural networks as well as sample runs.

For the regression models, to account for the incubation period of Coronavirus, mobility data and case count over the previous 14 days are used as predictor variables for the number of new cases on a given day. Thus, for a given county and day, a feature vector of the design matrix $\widetilde{X}$ is a vectorized submatrix of X. Explicitly, a row of $\widetilde{X}$ corresponding to a day (> 14) and county has the following structure:

$$[M_{14}^{RR}, \ldots, M_1^{RR}, M_{14}^{GP}, \ldots, M_1^{GP}, M_{14}^{W}, \ldots, M_1^{W}, M_{14}^{Res}, \ldots, M_1^{Res}, C_{14}, \ldots, C_1]$$

where M_j^{RR} is the retail and recreation mobility percent change from baseline j days prior to the given day, M_j^{GP}, M_j^{W}, and M_j^{Res} are the same for grocery and pharmacy, workplace, and residential mobility, respectively, and C_j is the case count j days prior to the given day. This produces the full design matrix $\widetilde{X} \in \mathbb{R}^{6003\times 70}$, which we randomly split into training and test matrices $\widetilde{X}_{train}$ and $\widetilde{X}_{test}$, with $\widetilde{X}_{test}$ consisting of 30% of the data. We used the train_test_split function from scikit-learn's model_selection library for this purpose [5]

We trained a variety of regression models on the training data and compared their performance on the test data. Specifically, we trained a linear regression model (LinReg), a RANSAC model (RANSAC), a LASSO model (LASSO), a ridge regression model (Ridge), an elastic net model (ElNet), a random forest model (RF), a quadratic polynomial regression model (Quad), and a cubic polynomial regression model (Cubic). We used the LinearRegression, RANSACRegressor, LassoCV, RidgeCV, and ElasticNetCV classes from the scikit-learn linear_model library, and the RandomForestRegressor class from the scikit-learn ensemble library. For the polynomial regression models, the PolynomialFeatures class was used from scikit-learn's preprocessing library to augment the design matrix to include higher-order monomials of the original features. The hyperparameters of RANSAC were chosen by

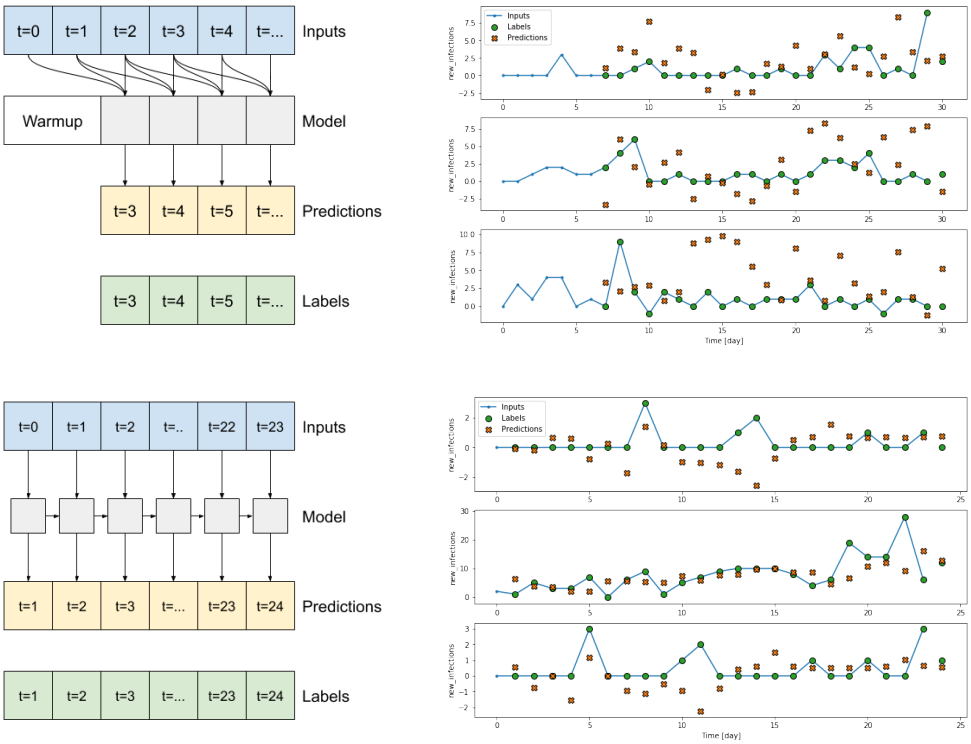

Figure 2: The left: Convolution and recurrent neural network schema respectively. The right: an example window of data with predictions. The top correspond to the CNN and the bottom correspond to the RNN.

experimentation to maximize on the average the R^2 score for the test data. The penalty parameters for Ridge, LASSO, and ElNet were chosen by cross-validation. The number of trees to average in RF was chosen, as with RANSAC, by experimentation to maximize on the average the R^2 score on the test data, though the results appeared to be somewhat independent of this choice. The R^2 scores and MAEs were calculated using the mean_absolute_error and r2_score functions from the scikit-learn metrics library [5].

5. Results

MODEL	MAE (train)	MAE (test)	R^2 (train)	R^2 (test)
LinReg	22.65	20.63	0.6308	0.7415
RANSAC	18.92	16.59	0.5836	0.7546
LASSO	19.98	18.04	0.6118	0.7494
Ridge	22.65	20.63	0.6308	0.7415
ElNet	19.99	18.04	0.6115	0.7497
RF	7.73	18.11	0.9445	0.7286
Quad	11.23	55.69	0.9797	-1.2881
Cubic	0	435.56	1	-2901.8128

Table 1: Summary Statistics: Regression

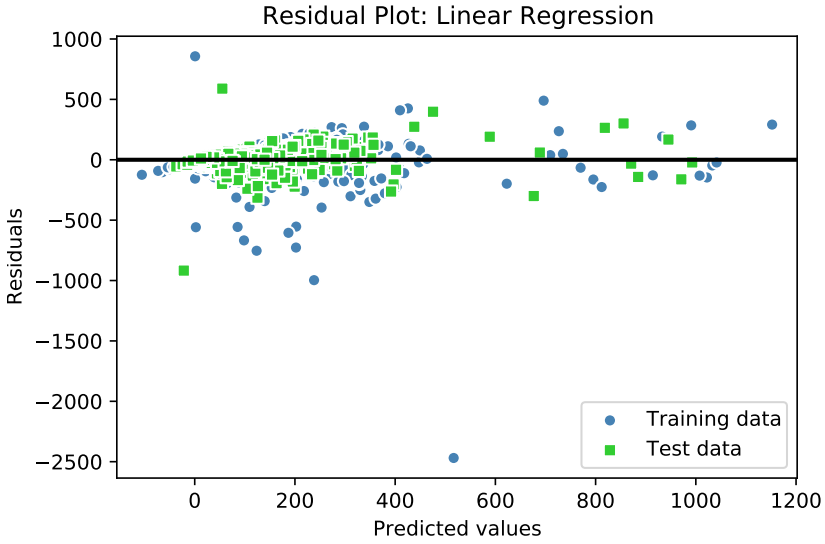

Figure 3: Linear Regression Residuals

NN	MSE (val)	MSE (test)	MAE (val)	MAE (test)
CNN	424.32	417.96	9.98	10.64
RNN	859.22	611.26	11.50	10.19

Table 2: Summary Statistics: Neural Networks

In Table 1 are reported the MAEs and R^2 scores on training and test data for the regression models listed above. Note that, for the linear models, the test MSE is lower than the training MSE. This is due to the presence of an outlier, as can be seen in the residual plot in Figure 3 corresponding to a predicted value of approximately 600 for LinReg. As can be seen from the table, the Quad and Cubic models drastically overfit the training data. The RF model also overfits the training data, but much less dramatically. It appears that the RANSAC and ElNet models have the best generalization performance, though the improvement over LinReg is not dramatic. In Figure 4 are plots of true and predicted numbers of new cases vs. day for Dane County for the RANSAC and ElNet models.

Similar statistics for the convolutional and recurrent neural networks are displayed in Table 2. Their generalization performance is very similar, and demonstrates a significant improvement over the regression models.

6. Conclusions and Future Work

We have developed two classes of models which use user mobility data to predict the number of new Coronavirus cases in urban Wisconsin counties. We note that different random splits of the data into test and training sets occasionally produced significantly different results. Though most of the linear models were robust to the choice of split, the random forest model sometimes strongly overfit the training data (as is illustrated in the results above), though sometimes it demonstrated the best generalization capability of all the regression

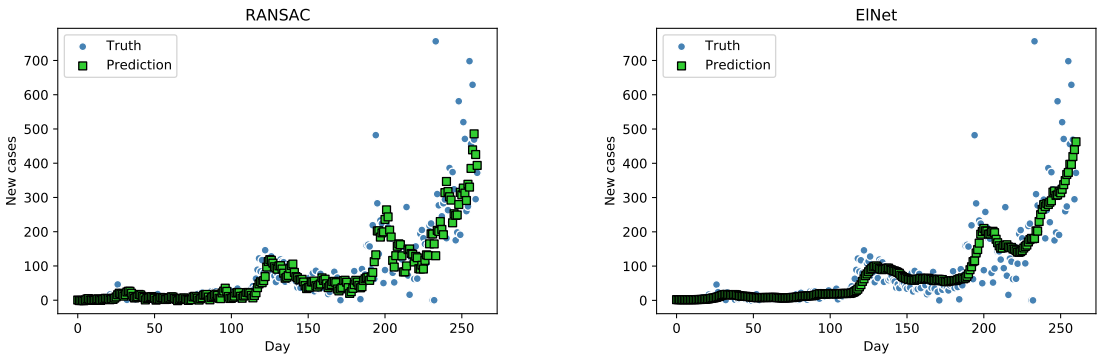

Figure 4: Left: RANSAC (Dane County). Right: ElNet (Dane County).

models. We decided for this reason that the random forest model was unreliable and preferred instead the linear regression models which were more consistent across the choice of split. Of these regression models, the RANSAC model had the best generalization performance, which is to be expected given that RANSAC is designed to reduced overfitting. Both neural networks performed similarly on a consistent basis, with either performing slightly better depending on the random split of the data. The neural nets demonstrate considerably better performance than the regression models, and we therefore recommend them as the preferred predictive model. We note that, given that the neural networks exploit the time-series nature of the data, we expected them to outperform the regression models, which ignore this structure entirely.

In future work, we will use more precise scientific estimates of the incubation period of Coronavirus to choose the time-profiles for prediction. It is unlikely that mobility changes from a few days before a given day are strongly correlated with the number of new cases on that day. Since the range of such an estimate is likely wide, a more narrow profile might be obtained by cross-validation accross a variety of choices for this hyperparameter.

7. References

[1] Abdul Hannan, Katie House, US Army Engineer Research and Development Center. GitHub repository, https://github.com/reichlab/covid19-forecast-hub/blob/master/data-processed/USACE-ERDC_SEIR/metadata-USACE-ERDC_SEIR.txt

[2] Aktay A, Bavadekar S, Cossoul G, Davis J, Desfontaines D, Fabrikant A, et al. Google COVID-19 Community Mobility Reports: Anonymization Process Description (version 1.0). https://github.com/nytimes/COVID-19-data

[3] Alessandro Vespignani et al. Modeling of COVID-19 epidemic in the United States. https://uploads-ssl.webflow.com/58e6558acc00ee8e4536c1f5/5e8bab44f5baae4c1c2a75d2_GLEAM_web.pdf

[4] Bijaya Adhikari, Xinfeng Xu, Naren Ramakrishnan, and B. Aditya Prakash. 2019. EpiDeep: Exploiting Embeddings for Epidemic Forecasting. In Proceedings of the

25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19). Association for Computing Machinery, New York, NY, USA, 577586. DOI:https://doi.org/10.1145/3292500.3330917

[5] Castex, G., Dechter, E. & Lorca, M. COVID-19: The impact of social distancing policies, cross-country analysis. EconDisCliCha (2020). https://doi.org/10.1007/s41885-020-00076-x

[6] COFFEE: COVID-19 Forecasts using Fast Evaluations and Estimation. Lauren Castro, Geoffrey Fairchild, Isaac Michaud, and Dave Osthus. LANL reference LA-UR-20-28630. https://covid-19.bsvgateway.org/static/COFFEE-methodology.pdf

[7] Dan Sheldon, Casey Gibson, Nick Reich. Bayesian compartmental models for COVID-19, (2020), GitHub repository, https://github.com/dsheldon/covid

[8] Difan Zou, Lingxiao Wang, Pan Xu, Jinghui Chen, Weitong Zhang, Quanquan Gu. Epidemic Model Guided Machine Learning for COVID-19 Forecasts in the United States. medRxiv 2020.05.24.20111989; doi: https://doi.org/10.1101/2020.05.24.20111989

[9] Estrada E 2020 COVID-19 and SARS-CoV-2. Modeling the present, looking at the future Phys. Rep. 869 151

[10] Li X, Rudolph AE, Mennis J. Association Between Population Mobility Reductions and New COVID-19 Diagnoses in the United States Along the Urban-Rural Gradient, February-April, 2020. Prev Chronic Dis 2020;17:200241. DOI: http://dx.doi.org/10.5888/pcd17.200241.

[11] Google. COVID-19 Community Mobility Reports. https://www.google.com/covid19/mobility/.

[12] https://coronavirus.jhu.edu/

[13] IHME COVID-19 Forecasting Team., Reiner, R.C., Barber, R.M. et al. Modeling COVID-19 scenarios for the United States. Nat Med (2020). https://doi.org/10.1038/s41591-020-1132-9

[14] Joceline Lega and Heidi E.Brown. Data-driven outbreak forecasting with a simple nonlinear growth model. Epidemics, Volume 17, December 2016, Pages 19-26. https://doi.org/10.1016/j.epidem.2016.10.002

[15] Johns Hopkins University & Medicine. Coronavirus Resource Center. https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series.

[16] Joseph Chadi Lemaitre, Kyra H Grantz, Joshua Kaminsky, Hannah R Meredith, Shaun A Truelove, Stephen A Lauer, Lindsay T Keegan, Sam Shah, Josh Wills, Kathryn Kaminsky, Javier Perez-Saez, Justin Lessler, Elizabeth C Lee. A scenario modeling pipeline for COVID-19 emergency planning. medRxiv 2020.06.11.20127894; doi: https://doi.org/10.1101/2020.06.11.20127894

[17] Kermack W O and McKendrick A G 1927 A contribution to the mathematical theory of epidemics Proc. R. Soc. A 115 70021

[18] Leora I Horwitz, MD, MHS, Simon A Jones, PhD, Robert J Cerfolio, MD, Fritz Francois, MD, Joseph Greco, MD, Bret Rudy, MD, Christopher M Petrilli, MD, Trends in COVID-19 Risk-Adjusted Mortality Rates. J Hosp Med. Published Online First October 23, 2020. DOI: 10.12788/jhm.3552

[19] Li Wang, Guannan Wang, Lei Gao, Xinyi Li, Shan Yu, Myungjin Kim, Yueying Wang, Zhiling Gu. Spatiotemporal Dynamics, Nowcasting and Forecasting of COVID-19 in the United States. https://arxiv.org/abs/2004.14103

[20] Michael Lingzhi Li, Hamza Tazi Bouardi, Omar Skali Lami, Nikolaos Trichakis, Thomas Trikalinos, Mohammad Fazel Zarandi and Dimitris Bertsimas. Overview of DELPHI Model V3 - COVIDAnalytics. July 2020. https://www.covidanalytics.io/DELPHI_documentation_pdf

[21] Miller, I. F., Becker, A. D., Grenfell, B. T. & Metcalf, C. J. E. Disease and healthcare burden of COVID-19 in the United States. Nat. Med. 26, 12121217 (2020).

[22] Neil Pearce, Deborah A Lawlor, Elizabeth B Brickley, Comparisons between countries are essential for the control of COVID-19, International Journal of Epidemiology, Volume 49, Issue 4, August 2020, Pages 10591062, https://doi.org/10.1093/ije/dyaa108

[23] Opuszko, M. and Ruhland, J. Impact of the Network Structure on the SIR Model Spreading Phenomena in Online Networks. ICCGI 2013: The Eighth International Multi-Conference on Computing in the Global Information Technology.

[24] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[25] Spencer Woody. UT-Austin forecast for COVID-19 mortality in the US. GitHub repository, https://github.com/UT-Covid/USmortality

[26] TensorFlow: Large-scale Machine Learning on Heterogeneous Systems, Abadi et al., 2015

[27] Teresa Yamana, Sen Pei, Sasikiran Kandula, Jeffrey Shaman. Projection of COVID-19 Cases and Deaths in the US as Individual States Re-open May 4,2020. medRxiv 2020.05.04.20090670; doi: https://doi.org/10.1101/2020.05.04.20090670