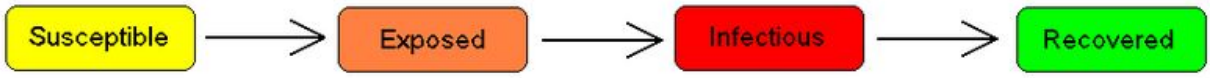
Levente Szabo
6/20/2020

US County Level COVID-19 Forecasting

Abstract: The new daily counts of COVID-19 across the most populous 1250 U.S. counties is predicted using a modified compartmental epidemiological model, denoted NNMRA-SEIR. Modifications to traditional compartmental models include localized measures of nearby outbreak severity through a nearest neighbor heuristic. The ability of NNMRA-SEIR to predict COVID-19 cases is shown to outperform a baseline SEIR model on a multi week prediction task. Further, NNMRA-SEIR indicates a tighter fit around the true infection curve compared to a baseline SEIR model.

Introduction: Beginning in early 2020 the government and healthcare resources of the United States have been responding to the pandemic of highly infectious coronavirus disease 2019 (COVID-19). Forecasting the scale of the pandemic remains a difficult task due to the highly variable rate of infection across the country and even across states. A variety of factors are at play and may influence the outbreak in a given county including but not limited to; population density, testing, socioeconomic status, health care resources, proximity to epicenters, etc. We believe that the daily change in confirmed infection counts can be accurately predicted for separate counties by fitting a compartmental epidemiological model which takes into account the geographic distribution of infections and physical proximity of the county to large outbreak centers. For the purpose of local forecasting of COVID-19 we present an SEIR model with nearest neighbor attack rates.

SEIR Models: Compartmental epidemiological models are an approach to modelling the spread of infectious diseases by separating the population into labeled compartments. In the case of an SEIR model these compartments are respectively Susceptible, Exposed, Infected and Recovered/Removed [1]. The SEIR model is just one of the various compartmental models and it builds on the simpler SIR model which makes the simplification that susceptible individuals go directly to becoming infected upon exposure to the agent.



The SEIR model has parameters alpha,beta, gamma which govern the transition from S to E, E to I and I to R respectively. We make the simplification that the population N = 1 and S+E+I+R=N so that we are interested in the dynamics of the compartment proportions for a given outbreak. Note that we also make the simplification that no new individuals are entering the population or leaving and that once recovered or removed one cannot become reinfected.

$$dS/dt = -\alpha(S * I)$$
$$dE/dt = \alpha(S * I) - (\beta * E)$$
$$dI/dt = (\beta * E) - (\gamma * I)$$

$dR/dt = \gamma * I$

$S + E + I + R = N = 1$

In order to fit an SEIR model for a given outbreak we fix some initial
$S0,\ E0,\ I0,\ R0$
And then find the parameters $\alpha,\ \beta,\ \gamma$ that give the closest fit according to RMSE

Contributions:
We were able to improve on the predictive power of the baseline SEIR model by making adjustments to the difference equations concerning localized outbreak metrics. Fitting an SEIR model for a given county results in a predictive capacity that relies only on the time series infection data and ignores the physical proximity to epicenters. A county's physical proximity to epicenters provides insight into the infection risk it faces. For this reason we devise Attack Rate denoted AR and nearest neighbor mean attack rate denoted NNMAR.

$AR = TotalInfected / Population$

$NNMAR(d) : Average(AR)_{N(d)}$

Where N(d) indicates the neighborhood of counties within d units of longitude & latitude. For simplification we set d=1 with NNMAR referring to NNMAR(1). The NNMAR is used as a measure of localized infection and it is incorporated into the difference equations concerning the Exposed and Infected compartments.

Problem Statement: For the most populous 1250 US counties predict the number of new daily infections 3 weeks ahead using a NNMAR-modified SEIR model. Compare the RMSE fit of the modified SEIR model with a baseline SEIR model. Furthermore, compare the predicted new cases over the next 3 weeks with baseline SEIR.

Methods
Exponential Smoothing: The number of new daily infections can be extracted from a counties cumulative infection time series by taking the 1st order discrete difference. Since the infected compartment in a SEIR model measures the current active number infected it is a smooth curve, however the number of new confirmed infected is not. To project the new case counts with our model we will have to smooth out the discrete difference, we do this with exponential smoothing using a weight parameter of 0.45

$Y = expsmooth(diff(cumulative)/\ countypopulation,\ 0.45)$

$expsmooth(array, weight) :$

$\quad for(int\ i = 1;\ i < len(array);\ i++) :$

$\quad\quad array[i] = weight * array[i] + (1 - weight) * array[i-1]$

$\quad return\ array$

Fitting and Integrating ODEs: Given a certain time range, x, parameters alpha,beta,gamma, set of differential equations and initial conditions we need to generate the S, E, I, R curve (particularly the I curve). This can be done by integrating a system of ordinary differential equations according to the given parameters

$$
\begin{aligned}
&fitodeint(x, alpha, beta, gamma):\\
&\quad curves = integrate(seirmodel, (S0, E0, I0, R0),\ x,\ args = (\alpha, \beta, \gamma))\\
&\quad return\ curves[infected]
\end{aligned}
$$

Finally we need to select the values of alpha,beta,gamma who generate the system of ODEs which when integrated give the minimum RMSE against our exponentially smoothed daily case counts Y.

NNMAR-modified SEIR model:

$$
\begin{aligned}
&model(cases,\ population, NNMAR[county], days - ahead):\\
&\quad Y = expsmooth(diff(cases)/population, 0.45)\\
&\quad X = [0, ..len(Y)]\\
&\quad Days = len(X)\\
&\quad AttackRate = cases[-1]/population\\
&\quad seir - model(y, x, \alpha, \beta, \gamma):\\
&\qquad S = -\alpha * (y[0] * y[2]) + 0.00025\\
&\qquad E = \alpha * (y[0] * y[2]) - \beta * y[1] + (0.0005 * NNMAR[county])\\
&\qquad I = \beta * y[1] - \gamma * y[2] + (0.0005 * NNMAR[county])\\
&\qquad R = \gamma * y[2]\\
&\qquad return\ S, E, I, R\\
&\quad fitodeint(x, alpha, beta, gamma):\\
&\qquad curves = integrate(seir - model, (S0, E0, I0, R0), x, args = (alpha, beta, gamma))\\
&\qquad return\ curves[infected]\\
&\quad N = 1\\
&\quad I0 = (Y[0] + Y[1] + Y[2])/3\\
&\quad E0 = 5 * AttackRate * I0\\
&\quad R0 = 3 * I0\\
&\quad S0 = N - I0 - E0 - R0\\
&\quad param = optimizecurve(fitodeint, X, Y)\\
&\quad fitted = fitodeint([0..len(X) + daysahead)
\end{aligned}
$$

Data: The primary dataset can be attributed to the NYTimes, COVID-19 Cases by US County are

Made available and updated daily. [2] The county specific population data is provided by the US Census Bureau [3]. The final dataframe was constructed by joining the base COVID-19 infection counts with US census bureau population data and Tableau generated longitude and latitude locations for each county.

Results: To evaluate the model we tested it on a 3 week prediction task. We selected the top 1250 most populous counties. We took a NNMAR curve for each county after excluding the last 3 weeks of case counts. Then, for each county we calculated the RMSE between its 3 separate weekly predictions and the true confirmed case counts. Finally we took the average of these over all 1250 counties. This prediction task was done on both a baseline SEIR model and the NNMAR-SEIR model in order to compare results. The RMSE is shown for the two prediction tasks.

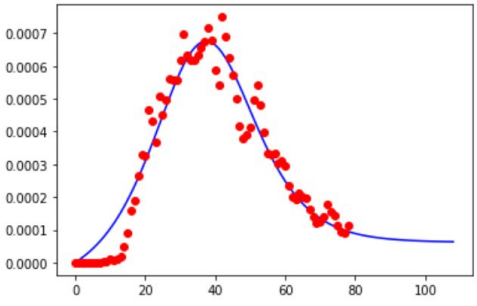
Results Table

Task	Weeks Ahead	Algorithm	Average RMSE
Prediction	1	NNMAR-SEIR	28.48
Prediction	1	Baseline SEIR	63.42
Prediction	2	NNMAR-SEIR	78.90
Prediction	2	Baseline SEIR	128.23
Prediction	3	NNMAR-SEIR	39.71
Prediction	3	Baseline-SEIR	165.26

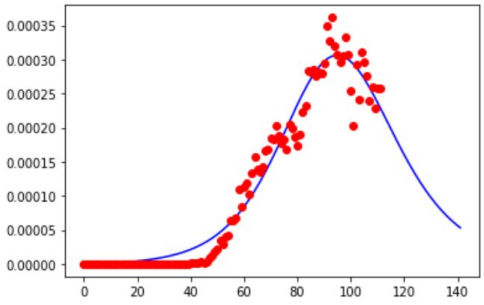
Figures

The figures below display the closeness of the NNMAR model fit. It fits tightly for those counties who have seen a serious decrease in the number of cases (like New York City) and gives a more uncertain model for the highly variable outbreaks in cities like Los Angeles.

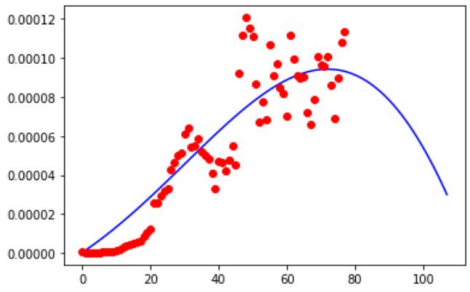
New York City, NY



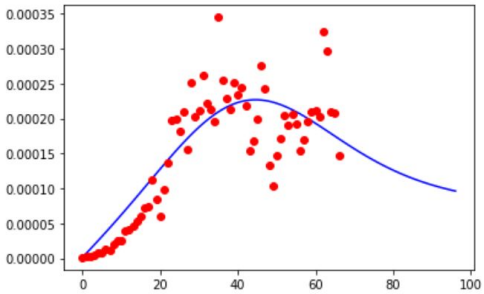
Cook County, Ill



Los Angeles, CA



Hartford, CT



Discussion: The NNMAR-SEIR model provides a strong case for incorporating localized outbreak information within traditional epidemiological models. The NNMAR-SEIR outperformed the baseline SEIR model in every weekly prediction task when using a RMSE metric. In order to improve these results the various parameters within the differential equations (ex. 0.0005*nearest_attack_rate in the infection equation) should be fine tuned through rigorous cross validation. Furthermore to truly test the efficacy of these models we would need more data and a more difficult benchmark to beat. Small heuristics can have a large effect when incorporated into robust models such as the SEIR family. It would be interesting to test a variety or combination or localized heuristics in this regard such as "Average Drive to Hospital" or "Median Household Income".

References:

[1] Vynnycky, E.; White, R. G., eds. (2010). *An Introduction to Infectious Disease Modelling*. Oxford: Oxford University Press. ISBN 978-0-19-856576-5.

[2] https://github.com/nytimes/covid-19-data

[3] https://www.census.gov/data/datasets/time-series/demo/popest/2010s-counties-total.html