OneQuietNight Covid-19 Forecast

Authors	Areum Jo (areumjo1@gmail.com), Jae Cho (jaehun.cho@gmail.com)
Date	2020-11-18
Last Update	2021-01-10

Introduction

Our ability to contain the coronavirus pandemic depends on being able to forecast potential outbreaks.

In this work, we develop scientifically-driven machine learning models to accurately predict the spread of Covid-19 infections using real-time data.

We collect and organize various data sets that may bear on the spread of Covid-19 -- daily case reports, movement trends, weather reports, and economic changes. Our models use this data to make predictions about future increases in Covid-19 cases at the county, state, and national levels in the United States.

Problem Description

The official CDC Covid-19 forecast[1] uses an ensemble of models to predict the number of new cases that are likely to arise in different geographic locations. The CDC Covid-19 forecast predicts the number of new Covid-19 cases per week for the next 4 weeks at the national, state, and county levels. It currently combines the forecasts from dozens of modeling groups.

To aid in this effort, we develop and operationalize an accurate Covid-19 forecast based on data from Delphi COVIDcast[9], JHU CSSE[10], The COVID Tracking Project[11], Apple Mobility Trend Reports[12], Google COVID-19 Community Mobility Reports[13], and C3 AI Covid 19 Data Lake[14]. Our forecast is competitive and outperforms some well-established models in backtests. We visualize this data using an interactive web application.

Broad Approach

Coronavirus is thought to spread from person to person. A typical case starts with a person coming into contact with a patient, who may not have symptoms. The virus has a chance to spread to the person during each contact. When the virus is successful, the person becomes infected and infectious to other people. The virus spreads exponentially in this way.

Epidemiological models use the structural knowledge of the spread of the virus to make predictions using the number of infected individuals and the number of transmittive contact. But it is difficult to measure how many infected individuals there really are and with whom they had close contact. Instead, we only have some imperfect measurements of a set of inputs that may have bearing on these components.

In order to learn the useful relations between variables with limited data, we use machine learning models with scientifically-driven features. We find that temporal and spatial features of the daily case reports and movement trends data predicts future Covid-19 cases. Our models use these to make predictions for all counties, states, and the country for the next 4 weeks.

Technical Details of the Approach

We model the number of new cases per week in the next week for region i, $y_i(t+1)$, using scientifically-driven features of the data, $X(t)$.

We develop the following features:

- Sum of new cases per 100,000 people per week[9, 10, 11].
- Sum of new hospitalizations per licensed beds per week[9,10,11].
- Average movement trends[12, 13].
- Social distancing metrics[9].
- Google search trends[9].

We impute features with top-down and bottom-up hierarchical aggregations and Census Core-Based Statistical Area aggregations[13]. We stabilize features by applying winsorizations across hierarchies. We also use one-week lagged versions of each of these features to capture their dynamics. We did not include other features because they did not work consistently or intuitively or because they did not help improve the predictions when used with the features above.

We develop forecasts using different machine learning models. The models are optimized using the data set of $X(t)$ and $y_i(t+1)$ from a moving window of N_{train} training weeks, $[t-N_{train}-N_{test},\ t-N_{test})$, and evaluated on a moving window of N_{test} testing weeks,

$[t - N_{test},\ t]$ in a walk-forward backtest from 2020-06-15 to 2020-09-14. We reserve 8 instances from 2020-09-14 to 2020-11-02 for validation. We find that a linear regression of the features against new cases per 100,000 people for each horizon and for each level generates the best results in the back tests. These predictions are multiplied by the population in the region and combined together to produce the final forecast. The final forecast uses the optimized models trained on $t \in [t - N_{train},\ t]$ to predict $y_i(t + 1)$.

Results

We compare our forecast to all the models from the CDC Covid-19 forecasts and show that our forecast is competitive and outperforms some well-established forecasts in the backtest. We backfilled our forecast by training on data up to the forecast date and making predictions with the inputs available on the forecast date in the validation period from 2020-09-14 to 2020-11-02. We used the historical forecasts from the Covid-19 Forecast Hub[2] in the same period.

Table 1. Forecasting accuracy of forecasts between 2020-09-14 and 2020-11-02. We computed the mean absolute error using the daily reports containing the cases data from the JHU CSSE group as the gold standard reference for the cases in the US. We normalized all the numbers by the COVIDhub-baseline number. The normalized mean absolute error numbers for each of the forecast horizons are shown below (lower is better).

	country				**county**				**state**			
target	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
OneQuietNight	0.72	0.93	0.97	0.91	0.95	0.95	0.93	0.93	0.86	0.83	0.84	0.88
COVIDhub-baseline	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
CEID-Walk	1.00	1.00	1.00	0.99	1.01	1.01	1.01	1.00	1.00	1.01	1.01	1.01
CMU-TimeSeries	nan	nan	nan	nan	6.24	6.51	7.19	7.73	nan	nan	nan	nan
CU-nochange	0.99	0.70	0.66	0.62	1.19	1.23	1.32	1.45	1.09	1.01	1.02	1.04
CU-scenario_high	0.98	0.62	0.56	0.61	1.19	1.24	1.33	1.41	1.09	0.98	0.96	0.99
CU-scenario_low	1.03	0.83	0.84	0.87	1.19	1.24	1.31	1.37	1.11	1.08	1.12	1.19
CU-scenario_mid	0.98	0.66	0.75	0.75	1.19	1.23	1.25	1.23	1.09	0.99	0.99	0.98
CU-select	0.98	0.66	0.75	0.75	1.19	1.23	1.25	1.23	1.09	0.99	0.99	0.98
Columbia_UNC-SurvCon	0.61	0.53	0.63	0.87	nan	nan	nan	nan	nan	nan	nan	nan
Covid19Sim-Simulator	1.21	1.09	1.03	0.99	nan	nan	nan	nan	1.24	1.16	1.13	1.12
CovidAnalytics-DELPHI	2.91	1.96	1.62	1.46	nan	nan	nan	nan	2.68	1.97	1.72	1.64
DDS-NBDS	0.65	0.56	0.64	0.80	nan	nan	nan	nan	1.08	0.87	0.92	1.14
Geneva-DetGrowth	0.82	nan	nan	nan	nan	nan	nan	nan	0.89	nan	nan	nan

IowaStateLW-STEM	1.10	1.17	1.16	1.14	1.26	1.14	1.12	1.11	1.36	1.33	1.32	1.31
JCB-PRM	0.91	0.88	0.89	0.90	nan	nan	nan	nan	1.05	0.97	0.96	0.98
JHUAPL-Bucky	1.01	0.96	0.93	0.93	1.22	1.18	1.17	1.17	1.13	1.02	1.01	1.02
JHU_IDD-CovidSP	2.43	1.61	1.31	1.18	1.71	1.36	1.21	1.13	2.19	1.54	1.30	1.19
Karlen-pypm	0.80	0.72	0.68	0.63	nan	nan	nan	nan	1.05	0.89	0.85	0.90
LANL-GrowthRate	1.31	1.34	1.29	1.29	1.17	1.22	1.23	1.26	1.19	1.24	1.30	1.35
LNQ-ens1	0.54	0.67	0.75	0.83	0.87	0.89	0.91	0.93	0.77	0.80	0.84	0.91
OliverWyman-Navigator	0.65	0.61	0.59	nan	1.05	1.03	1.01	nan	0.98	0.87	0.82	nan
PandemicCentral-USCounty	nan	nan	nan	nan	1.85	1.53	nan	nan	nan	nan	nan	nan
QJHong-Encounter	0.78	0.74	0.76	0.77	nan	nan	nan	nan	nan	nan	nan	nan
RobertWalraven-ESG	nan	nan	nan	nan	nan	nan	nan	nan	1.39	1.26	1.27	1.30
UCLA-SuEIR	1.16	1.28	1.27	1.26	2.43	2.90	2.97	2.95	1.30	1.50	1.49	1.46
UMass-MechBayes	nan	nan	nan	nan	3.81	4.06	4.24	4.63	nan	nan	nan	nan
UMich-RidgeTfReg	0.69	0.76	0.84	0.91	nan	nan	nan	nan	1.14	1.03	1.01	1.06
USC-SI_kJalpha	0.88	0.97	1.04	1.11	1.18	1.15	1.17	1.19	0.99	1.00	1.06	1.13
UVA-Ensemble	nan	nan	nan	nan	1.34	1.18	1.08	1.08	nan	nan	nan	nan
COVIDhub-ensemble	1.00	0.95	0.94	0.95	0.99	1.03	1.05	1.06	1.04	1.04	1.04	1.06

As shown in Table 1, our OneQuietNight forecast generates accurate results across all horizons in the backtest. Our approach is different from the empirical models and dynamical models that are commonly used in the Covid-19 forecasts in that it does not make any forward-looking assumptions about the factors affecting transmission. Instead, it uses the historical dynamics between the number of cases and people's movement levels to make the forecasts. This tends to produce waves of Covid-19 peak cases rather than a continued increase over a four week time frame based on the historical patterns.

Out of Sample Results (Update on 2021-01-10)

We update the national model on 2021-01-10 based on out of sample results. The county and state level models were left as is.

We compare our out of sample forecasts to all models from the CDC Covid-19 forecasts in the time period from 2020-11-21 to 2020-01-09 in Table 2. These consists of four weeks of observations for the four week horizon, five weeks of observations for the three week horizon, and so on. While our model retained performance at the county and at the state level, our model had big misses at the country level.

Table 2. Forecasting accuracy of forecasts between 2020-11-21 and 2020-01-09. We computed the mean absolute error using the daily reports containing the cases data from the JHU CSSE

group as the gold standard reference for the cases in the US. We normalized all the numbers by the COVIDhub-baseline number. The normalized mean absolute error numbers for each of the forecast horizons are shown below (lower is better).

	country				**county**				**state**			
target	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
OneQuietNight-ML	0.82	1.85	3.93	5.28	0.93	0.87	0.92	0.98	1.03	0.91	1.11	1.30
COVIDhub-baseline	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
BPagano-RtDriven	1.26	1.09	1.36	0.76	nan	nan	nan	nan	1.44	1.42	1.34	1.14
CEID-Walk	1.06	0.97	nan	nan	1.01	1.01	nan	nan	1.00	1.01	nan	nan
CMU-TimeSeries	nan	nan	nan	nan	9.78	8.77	9.40	9.63	nan	nan	nan	nan
CU-nochange	1.01	1.16	1.87	1.60	1.17	1.26	1.34	1.29	1.05	1.11	1.13	0.97
CU-scenario_high	0.99	1.16	1.87	1.45	1.17	1.27	1.37	1.34	1.05	1.11	1.14	0.99
CU-scenario_low	1.01	1.12	1.66	1.18	1.17	1.26	1.21	1.03	1.05	1.09	1.07	0.87
CU-scenario_mid	0.98	1.13	1.83	1.20	1.17	1.28	1.39	1.28	1.05	1.10	1.16	0.97
CU-select	0.99	1.12	1.66	1.18	1.17	1.26	1.21	1.03	1.05	1.09	1.07	0.87
Columbia_UNC-SurvCon	1.87	2.17	3.97	4.14	nan	nan	nan	nan	nan	nan	nan	nan
Covid19Sim-Simulator	0.97	0.92	1.32	0.85	nan	nan	nan	nan	1.18	1.13	1.06	0.98
CovidAnalytics-DELPHI	1.98	1.54	2.63	2.59	nan	nan	nan	nan	2.05	1.87	1.99	1.92
DDS-NBDS	3.44	5.42	2.20	4.04	nan	nan	nan	nan	3.67	4.19	1.76	2.05
FAIR-NRAR	nan	nan	nan	nan	1.66	1.09	nan	nan	nan	nan	nan	nan
Geneva-DetGrowth	0.98	nan	nan	nan	nan	nan	nan	nan	1.11	nan	nan	nan
Google_Harvard-CPF	nan	nan	nan	nan	1.93	1.74	1.75	1.35	1.63	1.61	1.69	1.66
IBF-TimeSeries	1.29	0.93	1.09	0.77	nan	nan	nan	nan	nan	nan	nan	nan
IowaStateLW-STEM	1.17	0.86	0.90	0.88	1.31	1.14	1.11	1.10	1.28	1.14	1.20	1.19
JCB-PRM	nan	1.14	2.25	2.79	nan	nan	nan	nan	nan	1.38	1.50	1.51
JHUAPL-Bucky	1.06	1.44	1.94	2.71	1.48	1.75	1.86	1.93	1.22	1.53	1.56	1.52
JHU_CSSE-DECOM	nan	nan	nan	nan	nan	nan	nan	nan	1.11	1.07	1.10	1.26
JHU_IDD-CovidSP	3.14	1.79	1.64	1.30	1.61	1.15	0.98	0.94	1.98	1.21	0.94	0.85
Karlen-pypm	1.61	1.73	3.16	4.33	nan	nan	nan	nan	1.53	1.65	1.74	1.92
LANL-GrowthRate	0.88	0.90	1.60	0.85	1.16	1.18	1.10	0.93	1.19	1.12	1.07	0.85
LNQ-ens1	1.12	1.11	1.15	0.86	0.93	0.95	0.87	0.83	0.94	0.94	0.87	0.74
OliverWyman-Navigator	1.21	1.02	1.24	nan	1.15	1.08	1.04	nan	1.19	1.03	0.95	nan
QJHong-Encounter	0.78	0.72	0.69	0.84	nan	nan	nan	nan	nan	nan	nan	nan
RobertWalraven-ESG	1.84	2.33	3.06	2.83	nan	nan	nan	nan	1.33	1.24	1.08	0.96

TTU-squider	1.23	0.78	0.88	1.39	nan	nan	nan	nan	2.27	1.85	1.76	1.73
UCF-AEM	0.85	0.82	1.06	0.48	nan	nan	nan	nan	nan	nan	nan	nan
UCLA-SuEIR	1.08	0.98	1.51	1.54	2.52	3.05	3.26	3.38	1.10	1.19	1.25	1.26
UCSB-ACTS	nan	nan	nan	nan	nan	nan	nan	nan	2.56	2.07	2.07	2.18
UMass-MechBayes	nan	nan	nan	nan	6.27	5.55	6.05	6.50	nan	nan	nan	nan
UMich-RidgeTfReg	1.07	1.68	2.13	2.37	nan	nan	nan	nan	1.64	1.82	2.03	2.05
USC-SI_kJalpha	1.10	0.98	1.30	0.95	nan	1.11	1.20	1.22	1.05	0.98	1.08	0.92
UVA-Ensemble	3.22	1.33	1.23	0.81	2.87	1.75	1.55	1.44	3.02	1.57	1.42	1.29
COVIDhub-ensemble	0.96	0.94	1.29	0.70	0.95	0.92	0.87	0.85	0.97	0.90	0.90	0.80

The misses at the national level are driven by two problems. First, the model did not include the total number of cases and failed to account for the decrease in the number of susceptible people. Second, it modeled mobility linearly and when the mobility numbers shot up during holiday season, it naively forecasted a corresponding increase when it should have been clipped. We addressed these issues in the 2021-01-10 release of the national model by including a feature for the total number of cases, removing all mobility features except one in the national model, and increasing the training window to cover a larger range of outcomes.

Impact

We develop scientifically-driven machine learning models to accurately predict the spread of Covid-19 infections using real-time data. This generates accurate forecasts that are competitive with and different from the current set of models in the CDC Covid-19 ensemble. We operationalize the forecast to retrain the model and make predictions on new data. We publish this data through a web application to help slow the pandemic and prevent future ones.

References

1. Centers for Disease Control and Prevention. Covid-19 Mathematical Modeling. https://www.cdc.gov/coronavirus/2019-ncov/covid-data/mathematical-modeling.html. Accessed: 2020-11-18.
2. Covid-19 Forecast Hub. https://covid19forecasthub.org/. Accessed: 2020-11-18.
3. Pei, S. and J. Shaman. Simulation of SARS-CoV2 Spread and Intervention Effects in the Continental US with Variable Contact Rates, March 24, 2020.
4. Carnegie Mellon Delphi Group. https://delphi.cmu.edu/. Accessed: 2020-11-18.
5. Arık, S. Ö., et al. Interpretable sequence learning for Covid-19 forecasting. arXiv:2008.00646 (2020).

6. Castro, L., et al. COFFEE: COVID-19 Forecasts using Fast Evaluations and Estimation. https://covid-19.bsvgateway.org/. Accessed: 2020-11-18.
7. Zou, D., et al. Epidemic Model Guided Machine Learning for COVID-19 Forecasts in the United States medRxiv, doi:10.1101/2020.05.24.20111989.
8. Census.gov. Metropolitan and Micropolitan Statistical Area Reference Files. https://www.census.gov/geographies/reference-files/time-series/demo/metro-micro/delineation-files.html. Accessed: 2020-11-18
9. Data from Delphi COVIDcast. Obtained via the Delphi Epidata API. https://cmu-delphi.github.io/delphi-epidata/api/covidcast.html.
10. Ensheng Dong, H. D. and Gardner, L. An interactive web-based dashboard to track covid-19 in real time. The Lancet Infect. Dis. 20, 533–534 (2020).
11. The COVID Tracking Project. https://covidtracking.com/. Accessed: 2020-11-18.
12. Apple. Mobility Trend Reports. https://covid19.apple.com/mobility. Accessed: 2020-11-18.
13. Google. COVID-19 Community Mobility Reports. https://www.google.com/covid19/mobility/. Accessed: 2020-11-18.
14. C3 AI COVID-19 Data Lake. https://c3.ai/customers/covid-19-data-lake/. Accessed: 2020-11-18.