Forecasting of COVID-19 Cases Using Time-Series Analysis of Google Mobility Data

LUIS GUZMAN, guzma102@umn.edu
ISAAC KASAHARA, kasah011@umn.edu
EMILY MULHALL, mulha024@umn.edu

The spread of COVID-19 has been linked to social gatherings and places where large numbers of people visit. In this paper we study the connection between Google's COVID-19 Community Mobility Data and the number of COVID-19 cases per 100,000 people in any given state. We model this problem using ARIMA and an LSTM network to measure trends in and predict future mobility information. As a baseline, we used a number of different algorithms to model the growth of COVID-19 cases in order to predict the number of cases in the future. Our findings show that both the ARIMA and LSTM methods significantly outperform the baseline methods in the medium term (20-day) to long term (40-day), with the ARIMA slightly outperforming the LSTM as the length of the forecast increases.

1 PROBLEM DESCRIPTION

COVID-19 has drastically changed our way of life over the last nine months, resulting in lockdowns, hospital crowding, and the loss of over 300,000 American lives. In order to curb coronavirus cases, states often shut down public gathering places, such as restaurants and bars, workplaces, and more. Throughout the pandemic, Google has been tracking traffic at various public location types, including retail and recreation (restaurants, bars, malls, etc.), groceries and pharmacies, parks, transit stations, workplaces, and residential places. They represent the change in traffic at a particular location type with a percent increase or decrease from a baseline taken January 3rd-February 6th, 2020. Using this mobility data, we hope to predict the number of COVID-19 cases up to 40 days in the future from a particular date in a given state.

We propose two different time-series techniques for projecting new case numbers from mobility data. The first simply uses ARIMA to predict trends in the mobility data. A multi-layer perceptron (MLP) is then used to learn the optimal combination of predicted mobility data and current case numbers to make accurate predictions about the future number of cases. The second method we propose accomplishes the same task using a Long Short-Term Memory (LSTM) network. In the remainder of this paper, we compare both of these methods to the state-of-the-art in forecasting COVID-19 case numbers. We believe that including mobility data into these models can help improve the accuracy of predictions when forecasting into the medium term (20 days) and long term (40 days).

2 RELATED WORK

Though COVID-19 is a relatively new strain of virus, due to its severity, it has been a popular topic of research. Previous work by Ribeiro et al. compared various regression techniques to predict COVID-19 infections up to 6 days in advance [1]. Of the methods evaluated, Ribeiro et al. found support vector regression (SVR) and the stacking ensemble method to perform the best, with an error rate of between one and seven percent when predicting on the cumulative case numbers six days ahead. The ensemble method they tested consisted of cubist regression, random forest, ridge regression, and SVR as base-learners, and they used a Gaussian process as a meta-learner. Additionally, Car et al. used a neural network to predict infection and death rate from location and days since initial infection [2]. Multiple studies have demonstrated that the logistic equation and the ARIMA model can be an effective tool for predicting cases of COVID-19 [3][4][5]. In our baseline model, we implement and compare many of these techniques.

Previous work has also been done regarding predictions of diseases using an LSTM[6]. LSTMs are a form of recurrent neural network that allow for the retention of memory in order to help form future predictions. To get the best results, formatting the data to be stationary allows for faster and more accurate learning[7]. The LSTM is built to predict one step in advance. To predict multiple time-steps ahead, the two main methods used are a recursive strategy, and a direct strategy[8]. These methods each have advantages and disadvantages, so we tested both in our initial testing to compare them for our problem.

Authors' addresses: Luis Guzman, guzma102@umn.edu; Isaac Kasahara, kasah011@umn.edu; Emily Mulhall, mulha024@umn.edu.

3 METHODS

3.1 The Dataset

Our first step was building our dataset from the COVID-19 Community Mobility Reports provided by Google, and the United States COVID-19 Cases and Deaths by State over Time provided by the CDC[9][10]. The mobility dataset shows how traffic in places of interest, such as grocery stores, parks, work places, etc. changes in each state[9]. These increases or decreases in different types of locations are the features given to the models. The infection data from the CDC was used to assign our features a corresponding y-value. This value is the moving average over the last week of the number of new cases plus the number of probable cases. Because we are comparing data over states with large population differences we needed to normalize this number. Thus, it is the moving average per 100,000 people in that state. If there were any missing feature data, it was filled using linear interpolation.

3.2 Methodologies

Having built our dataset, we moved on to establishing a proof of concept (POC) model that shows the mobility data adds useful information that we can use to predict future case numbers. Our POC model implements simple linear regression using the mobility data as a feature vector and cases as the output value. We started by analyzing our data and computing correlations between the number of cases and the mobility data, seeking the optimal offset in the number of cases. Because COVID-19 can take up to two weeks for symptoms to appear, the rise in cases following spreading events tends to trail behind. Thus, we needed to establish the optimal offset in order to get meaningful results from our models.

It was found that the optimal offset was 11 days for the linear correlation and 13 days for the logarithmic correlation, as seen in the below plots. Additionally, we removed any days that had zero cases. Both linear correlation and logarithmic correlation were analyzed, resulting in similar offsets. We calculated the offsets for all states, but only ran the POC model on Minnesota data for simplicity and time efficiency. As shown in Figure 1, our POC model could predict the amount of new COVID-19 cases with average error rates of 27% one day ahead, 32% seven days ahead, and 40% one month ahead. This indicates that mobility data can be used to infer the number of new cases. This model ignores all temporal dependencies and predicts from mobility data alone.

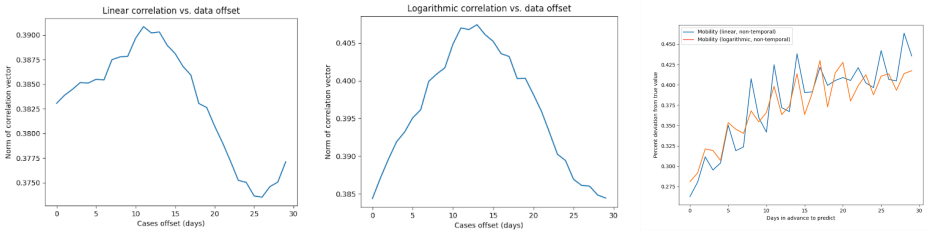

Fig. 1. The linear and logarithmic correlations (left, middle) and our POC model accuracy (right)

Next, we established a baseline in order to properly evaluate the performance of our time-series models. We trained our model on the past two months of CDC cases data, and tested it by predicting the number of new cases up to forty days in advance. We followed the procedure in Ribeiro et al. by using a variety of regression techniques as base-learners, and using a Gaussian process as a meta-learner [1]. Our base-learners include linear ridge regression using cross-validation, ridge regression using cross-validation on a log-log scale, fitting a logistic function, a stacking-ensemble method, and ARIMA. Grid search was performed to determine ARIMA order. The stacking regressor consisted of a linear regression model, linear SVR model, and random forest model. It used a gaussian process regressor as its final estimator. Of the regression techniques we evaluated, ARIMA and the gaussian process meta-learner performed the best, so we chose these as our baseline to compare to. Exact performance metrics are given and analyzed in the results section.

Having established our baseline, we moved on to adding the mobility information into our model. The first time-series model we tested uses ARIMA to predict future mobility and a multi-layer perceptron (MLP) to give a final case prediction based on ARIMA's projection. We again applied grid search to determine ARIMA order and trained a model on the past two months of mobility data. Each mobility feature was analyzed separately and the best-performing order was chosen for each one. We opted for this method over more complex forms of Vector Autoregression (VAR) due to our familiarity with the SARIMAX package and some fatal errors that we encountered when trying to perform VAR on our dataset. Figure 2 shows that ARIMA is able to accurately capture both the long-term and periodic trends in the mobility data when predicting up to 30 days in advance.

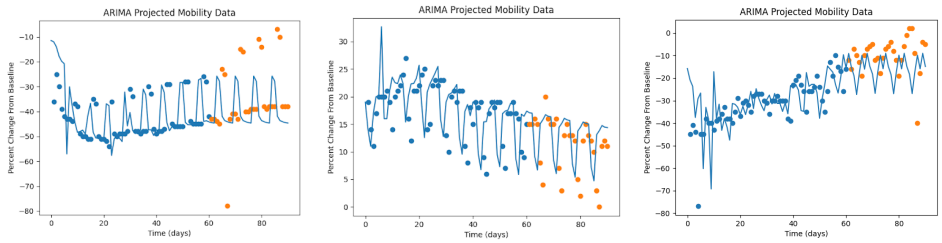

Fig. 2. ARIMA projections of mobility data showing past values (blue), future values (orange), and predicted (blue line)

The MLP was trained using the past 60 days of mobility data. The input feature vector consisted of six mobility features and one feature corresponding to the current number of cases. During training, the MLP learns the optimal combination of these features to predict future cases. From our testing, we found the best performing network to consist of 10 input nodes, 30 hidden nodes and a single output node, all of which used ReLU activation to give a continuous output. The number of days in advance that the network predicts was set to be 11 based on the result of our earlier correlation analysis. For completeness, we also tested other offset values and found 11 days to give the best results. In order to test the performance of the network, we used the ARIMA projections as our testing input and evaluated how well the network predicts new cases.

The next algorithm implemented was a deep learning network known as a Long Short Term Memory network, or LSTM. Since LSTMs are generally trained to do single time-step predictions, we would have to adapt the LSTM to predict multiple steps ahead in order to predict multiple days into the future. The two main approaches are a direct or a recursive LSTM. With a direct approach, the LSTM is trained to learn the time jump directly, using values for the desired prediction as the training label. In a recursive LSTM the network is trained to one time-step ahead, and then uses that prediction as an observation for predicting another time-step ahead. This process is repeated until the desired future step is reached.

After creating preliminary models for both algorithms, it became apparent that the recursive LSTM would heavily outperform the direct LSTM. The direct LSTM struggled to learn connections between the mobility data and the case numbers as the time jump it was attempting to learn was too high. In figure 3 below it can be seen that when case numbers increased faster than what the network had seen before, the direct LSTM would lag behind dramatically, as it had failed to learn connections between the mobility data and the case numbers.

However, there was also an initial concern with a recursive strategy: due to the nature of how it predicts, it would need to predict not only case numbers, but future mobility data as well. This would mean that the network has to learn far more than the direct LSTM, and it would be more susceptible to error if one of the other features predicted inaccurately. By adding another layer to the LSTM to increase its complexity and a dropout layer to prevent overfitting, the recursive LSTM accurately learned how to predict the trends for the mobility data features in the dataset. This is displayed in the figure below as the LSTM predicts the future residential mobility data.

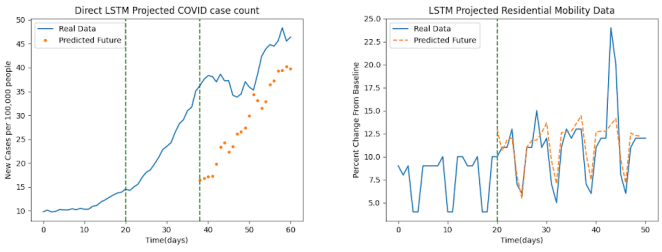

Fig. 3. Sample month of direct LSTM predictions (left) and recursive LSTM month of prediction values for one of the mobility features (right)

Another benefit the recursive model had over the direct model was that due to the consecutive nature of the input and predictions, we could make the cases number stationary. Making the case numbers stationary means instead of training the network on the number of cases, it is trained on the change in number of cases each day. This allows the LSTM to learn trends more easily. The prediction can then be converted back from change in cases to cases per 100,000. For testing the performance of the network, samples from the testing data were predicted and compared to their actual values.

4 RESULTS

	Baseline	ARIMA+MLP	LSTM
Short term error (5-day)	15%	18%	14%
Medium term error (20-day)	42%	21%	27%
Long term error (40-day)	71%	30%	33%

Table 1. Comparison of models

Our results show that the baseline models can predict COVID-19 case numbers with 15% mean absolute error (MAE) one-day-ahead, and this error slowly increases as the prediction time increases. This is to be expected because the daily fluctuation of our new cases is around 14%. Out of the baseline models, we found ARIMA to perform the best. The Gaussian Process meta-learner (adapted from [1]) matched ARIMA's performance in the short term, but was beaten out by other methods in the long term. This agrees with the results of [1] because their maximum prediction timeframe was only six days. Additionally, our error rates are higher than the work of Ribeiro et al. because we are predicting the number of new cases, rather than the cumulative cases.

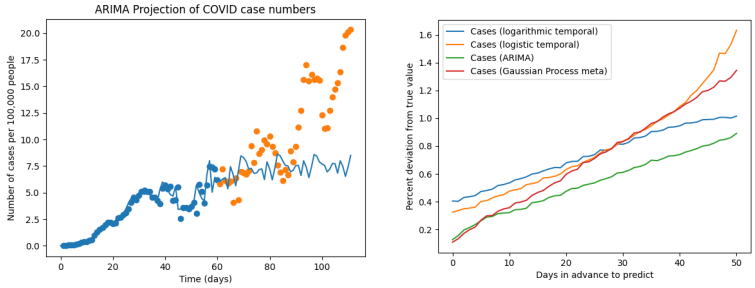

Fig. 4. A sample month of ARIMA prediction (left) and the baseline models' accuracy (right) Only the best performing baseline models are shown for clarity.

The ARIMA and MLP model did not perform as strongly as the baseline in the short term; however, this was more than made up for in the long-term predictions. The ARIMA+MLP model achieved 21% error when predicting twenty-days-ahead, whereas the baseline error was 42%. The recursive LSTM had a similar but slightly worse performance than the ARIMA and MLP model. It still performed much better than the baseline for medium to long terms, achieving an error of 27% for the medium term prediction. One of the better sample predictions can be seen below, showing the LSTM predict the relative trend of the case growth.

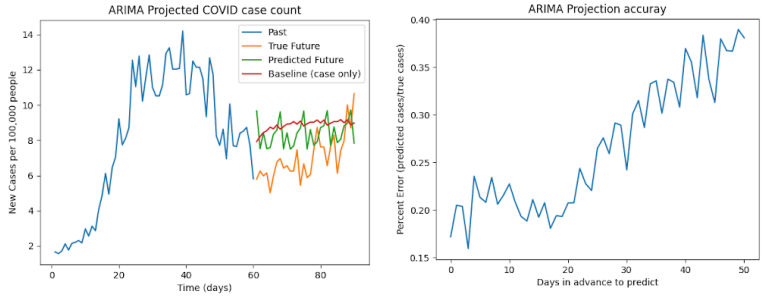

Fig. 5. A sample month of ARIMA predictions (left) and ARIMA testing accuracy (right)

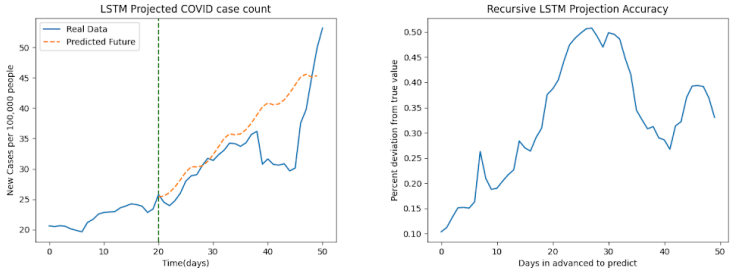

Fig. 6. A sample month of recursive LSTM predictions (left) and LSTM testing accuracy (right)

5 DISCUSSION

Overall, our results show that mobility data is an essential component to consider when attempting to predict the number COVID-19 cases in the United States. The baseline models, which do not use mobility data, fail completely when predicting more than a week or two in advance. Looking purely at the timeline, case spikes can occur at seemingly random intervals. Some of these spikes in case numbers are able to be predicted and quantified by trends in mobility data that the baseline is missing. When this additional data is taken into account, our medium and long term predictions greatly improve. Further proving the importance of this data addition, we believe that the medium to long term baseline error rates could have been largely thrown-off by a few months with very poor baseline predictions. For example, Figure 7 shows a particularly bad month for the ARIMA model where the long term error rate is over 120%. Since we chose to average the error rates over each month we trained on, a few large values like this could throw off the recorded error.

Additionally, simple regression methods and ARIMA are not adequately equipped to make long-term predictions on partially stochastic processes such as case numbers in a pandemic. This forces us to question whether there exists a baseline better suited to this particular problem. We have limited experience with modeling how pandemics spread, and consulting professionals of this subject matter could result in a more accurate baseline for us to compare our models to.

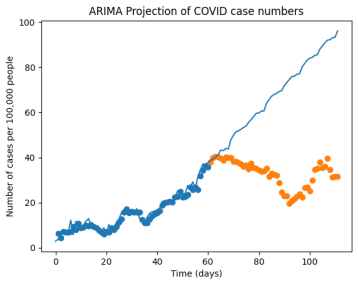

Fig. 7. A particularly bad month for our ARIMA baseline model

While our LSTM and ARIMA+MLP models far outperformed the baseline models, the lowest error rate was still above 10% deviation from the true value. As stated in the beginning of this paper, closing public gathering spaces has been a proven strategy in limiting the spread of COVID-19. However, there are many other factors that likely correlate with the number of cases. Mask wearing has proven quite effective, and it is likely that cases decrease following mandatory mask mandates. The availability of test kits and the administration of tests has increased over time, so it is likely that following an increase in widespread testing is an increase in cases. As the weather cools, people will have less outdoor gatherings, and more indoor ones, where spreading the virus is more likely. Political and societal events as well as holidays have increased the movement of people and has resulted in larger gatherings, also likely leading to a rise in cases. Clearly, there are many factors to be considered when predicting the number of COVID-19 cases in the United States. In future work, we hope to add some of these additional features to our model to achieve even better case prediction accuracy.

6 CONCLUSION

Our findings confirm the idea that mobility data can be utilized to predict the future trends of COVID-19 case numbers. Using multiple different methods, we found that an ARIMA+MLP model performed the most accurately for medium to long term predictions. This information can help further the research into the spread of COVID-19 and be generalizable for other diseases as well. We recognize the limitations of our research, as many outside factors besides mobility data will affect the growth of COVID-19 such as government restrictions, mask wearing, weather, etc. Future research could utilize global mobility data to obtain a broader model for the rise in case count, as well as take into consideration the previously mentioned factors such as weather to obtain higher accuracies of growth.

REFERENCES

[1] M. H. D. M. Ribeiro, R. G. da Silva, V. C. Mariani, and L. dos Santos Coelho, "Short-term forecasting covid-19 cumulative confirmed cases: Perspectives for brazil," *Chaos, Solitons & Fractals*, p. 109853, 2020.

[2] Z. Car, S. Baressi Šegota, N. Anđelić, I. Lorencin, and V. Mrzljak, "Modeling the spread of covid-19 infection using a multilayer perceptron," *Computational and Mathematical Methods in Medicine*, vol. 2020, 2020.

[3] E. Pelinovsky, A. Kurkin, O. Kurkina, M. Kokoulina, and A. Epifanova, "Logistic equation and covid-19," *Chaos, Solitons & Fractals*, vol. 140, p. 110241, 2020.

[4] T. M. Awan and F. Aslam, "Prediction of daily covid-19 cases in european countries using automatic arima model," *Journal of Public Health Research*, vol. 9, no. 3, 2020.

[5] P. Wang, X. Zheng, J. Li, and B. Zhu, "Prediction of epidemic trends in covid-19 with logistic model and machine learning technics," *Chaos, Solitons & Fractals*, vol. 139, p. 110058, 2020.

[6] N. Zheng, S. Du, J. Wang, H. Zhang, W. Cui, Z. Kang, T. Yang, B. Lou, Y. Chi, H. Long, *et al.*, "Predicting covid-19 in china using hybrid ai model," *IEEE Transactions on Cybernetics*, 2020.

[7] F. Qian and X. Chen, "Stock prediction based on lstm under different stability," in *2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, pp. 483–486, IEEE, 2019.

[8] J. Brownlee, "4 strategies for multi-step time series forecasting," *Machine Learning Mastery*, 2017.

[9] "Covid-19 community mobility reports." https://www.google.com/covid19/mobility/. (Accessed on 12/17/2020).

[10] "United states covid-19 cases and deaths by state over time | healthdata.gov." https://healthdata.gov/dataset/united-states-covid-19-cases-and-deaths-state-over-time. (Accessed on 12/17/2020).