Forecasting Covid-19 Confirmed Cases in Virginia with ARIMA and LSTM: A Comparative Study

Zhewei Wu
Department of Computer Science
Virginia Tech
Blacksburg, VA 24060
zhewei6@vt.edu

1 Introduction

The new coronavirus called COVID-19 appeared in Wuhan, China at the end of 2019, and then rapidly spread all over the world. Although China's brutal yet effective lockdown strategy worked quite well and got situation under control since March 2020, some other countries didn't really follow the same strategy due to different national situations. Till Oct 23, 2020, the Coronavirus Resource Center at Johns Hopkins University has recorded more than 42 million global cases, and nearly 1.15 million global deaths. That is so unfortunate and sorrowful. Since the COVID-19 pandemic led to a historical challenge to our society, we also would like to contribute myself to this field and make some changes. In this project, we will examine and compare how the two models, Auto Regressive Integrated Moving Average (ARIMA) and Long Short-Term Memory (LSTM), perform based on forecasting the COVID-19 confirmed cases in Virginia.

2 Related Work

Today, many engineers, researchers and scientists, all across the public, private, academic and industry sectors are confronting the COVID-19 pandemic together. In order to protect people and the economy, many researchers have done some decent work lately trying to slow down the increasing trends of the spread of the disease. Thus, to understand how the pandemic spreads and to provide further insights on how the next step the governments should take, diverse modeling and forecasting methods have been introduced. For instance, in [1], three mathematical models haven been applied to predict the expected total cases and total death toll, which is by using Logistic, Bertalanffy, and Gompertz models. It turned out that the logistic model had the best result, but all three models have a limitation of acquiring enough data to perform well. In [2], a discrete stochastic epidemic model with binomial distributions have been developed to predict when the total confirmed cases in China will reach peak and to simulate the impact of timing of people returning to work, which provides insights in how to implement the control measures with different strength. Also, very recently, in [3], a generalized Susceptible Exposed Infectious Recovered (SEIR) model has been proposed to predict the inflection point and the possible ending times for 5 different regions based on categorizing the flows of people into 4 states, including people who are susceptible, who are exposed, who are infected, and last but not least, who are recovered. Some other time-series models have also been studied to predict the number of confirmed cases in a certain region. For example, in [4], a traditional ARIMA model has been applied to show the potential infected cases in India within a certain amount of period (30 days) in the worst-case scenario and in the most optimistic scenario. In [5], a seasonal ARIMA package with R statistical model has been used to forecast the Italian registered cases and recovered cases if assuming the lockdown will remain for sixty days. Besides various studies done based on the traditional time-series models, accurate confirmed cases forecasting has also been achieved by some widely used machine learning and deep learning models in this pandemic. In [6], a novel model based on machine learning and cloud computing has been designed for real-time prediction of the

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

growth of the epidemic in a proactive way. In [7], a Long short-term memory (LSTM) model is used to forecast new infections over time with a relatively small dataset yet it achieves a good performance. Some other deep learning models including Variation Auto Encoder (VAE), Gated Recurrent Units (GRUs), Bidirectional LSTM (BiLSTM), etc. are also applied in many studies and have attractive performances.

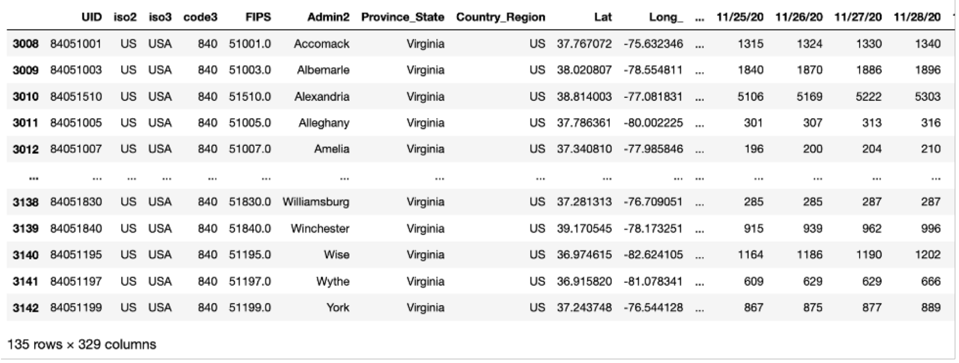
3 Data

3.1 Data Source

The dataset we collected is obtained from the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University [7]. The whole dataset contains many different CSV files which provide various perspectives such as global and U.S. cases data, or confirmed, deaths, and recovered data, etc. We chose the time series data that has all the U.S. daily-updated confirmed cases data specifically. We believe this is the most trustworthy and well-built data we may find. It is of high quality and up to date, and it also includes both the data labels that we need and don't need such as state, country, latitude and longitude, dates, etc.

3.2 Data Formatting and Cleaning

Since we're doing a forecasting for the state of Virginia specifically, we will need to filter the data by the data label "Province_State" first as shown in Fig. 1.

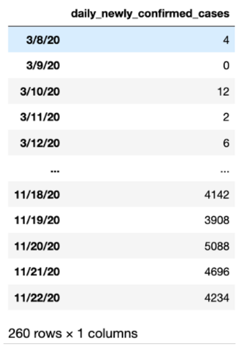


	UID	iso2	iso3	code3	FIPS	Admin2	Province_State	Country_Region	Lat	Long_	...	11/25/20	11/26/20	11/27/20	11/28/20
3008	84051001	US	USA	840	51001.0	Accomack	Virginia	US	37.767072	-75.632346	...	1315	1324	1330	1340
3009	84051003	US	USA	840	51003.0	Albemarle	Virginia	US	38.020807	-78.554811	...	1840	1870	1886	1896
3010	84051510	US	USA	840	51510.0	Alexandria	Virginia	US	38.814003	-77.081831	...	5106	5169	5222	5303
3011	84051005	US	USA	840	51005.0	Alleghany	Virginia	US	37.786361	-80.002225	...	301	307	313	316
3012	84051007	US	USA	840	51007.0	Amelia	Virginia	US	37.340810	-77.985846	...	196	200	204	210
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3138	84051830	US	USA	840	51830.0	Williamsburg	Virginia	US	37.281313	-76.709051	...	285	285	287	287
3139	84051840	US	USA	840	51840.0	Winchester	Virginia	US	39.170545	-78.173251	...	915	939	962	996
3140	84051195	US	USA	840	51195.0	Wise	Virginia	US	36.974615	-82.624105	...	1164	1186	1190	1202
3141	84051197	US	USA	840	51197.0	Wythe	Virginia	US	36.915820	-81.078341	...	609	629	629	666
3142	84051199	US	USA	840	51199.0	York	Virginia	US	37.243748	-76.544128	...	867	875	877	889

135 rows × 329 columns

Fig. 1. A peek at the whole data set of Virginia confirmed cases.

But the data is still not what we wanted as it only recorded the daily total confirmed cases for each state region of Virginia, so we dropped the unnecessary columns at the beginning such as UID, iso2, so on and so forth, and then summed up all the confirmed cases for each date from Jan 22, 2020 to Nov 29, 2020. The last thing we did is to get rid of the dates which the sum of the confirmed cases is zero. So eventually, the real dates we will be using to build my models ranges from Mar 8, 2020 to Nov 22, 2020, and the transformed data for daily increased confirmed cases in Virginia is shown in Fig. 2 and Fig. 3.



	daily_newly_confirmed_cases
3/8/20	4
3/9/20	0
3/10/20	12
3/11/20	2
3/12/20	6
...	...
11/18/20	4142
11/19/20	3908
11/20/20	5088
11/21/20	4696
11/22/20	4234

260 rows × 1 columns

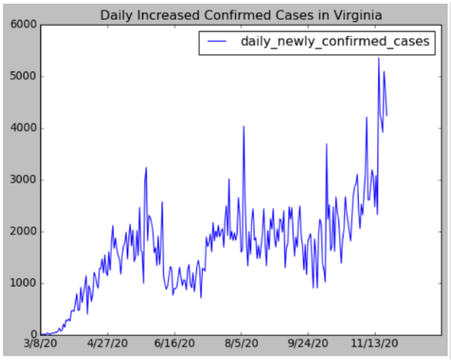


Fig. 2. Daily newly confirmed cases in Pandas DataFrame, and Fig. 3. in line graph.

The last week's data that starts from Nov 23, 2020 will be used to evaluate the final accuracy of each model. When building the ARIMA model, we also made the data to be stationary.

4 Method

We considered to use two models for forecasting the confirmed cases in Virginia for the week starts from Nov 23, 2020 to Nov 29, 2020. The first model is a traditional statistical model called ARIMA, and the second model is called LSTM, an artificial recurrent neural network (RNN) architecture often used in deep learning. So, it will be very interesting to see which model will outperform the other one when it comes to univariate time series forecasting.

4.1 ARIMA

The autoregressive integrated moving average (ARIMA) has been explored in time series forecasting during the past three decades and has become very popular ever since. It is a statistical analysis model based on regression analysis, and its main goal is to better understand the data set or to predict trends such as future securities or financial market movements. In this case, we will use it to forecast the future trend of confirmed cases that is surging in Virginia.

The ARIMA model has three components. Autoregression (AR) refers to "to a model that shows a changing variable that regresses on its own lagged, or prior, values" [8]; Integrated (I) represents the idea of differencing the raw data to become stationary; Moving average (MA) shows a dependency between an observation and a residual error from a moving average model applied to lagged observations [9].

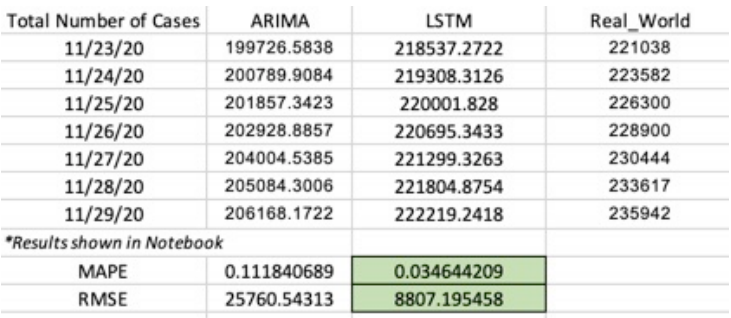
The unique strength for ARIMA is its temporal effects when it comes to deal with time series analysis, but it can be limited in forecasting extreme values as it is only supposed to be good at modeling trends and seasonality [9].

4.2 LSTM

The Long Short-Term Memory (LSTM) networks, as clearly described by Dr. Jason Brownlee in his article, are "a type of recurrent neural network capable of learning order dependence in sequence prediction problems [10]." In complex domains such as machine translation and speech recognition often require this kind of behavior. This model is a great tool for predicting anything that has a sequence. One of the advantages of LSTM is that it may "handle noise, distributed representations, and continuous values … and works well over a broad range of parameters such as learning rates, and input and output gate bias [11]." But it also has drawbacks. It does require more memory and take longer to train and is sensitive to inconsistent random weight initializations.

5 Results

After training and testing using ARIMA and LSTM models, in order to better understand how the two models perform and what is their accuracy, we calculated two error measures for both models in Microsoft Excel with predicted results obtained from Jupyter Notebook.



Total Number of Cases	ARIMA	LSTM	Real_World
11/23/20	199726.5838	218537.2722	221038
11/24/20	200789.9084	219308.3126	223582
11/25/20	201857.3423	220001.828	226300
11/26/20	202928.8857	220695.3433	228900
11/27/20	204004.5385	221299.3263	230444
11/28/20	205084.3006	221804.8754	233617
11/29/20	206168.1722	222219.2418	235942
**Results shown in Notebook*			
MAPE	0.111840689	0.034644209	
RMSE	25760.54313	8807.195458	

Fig. 4. Two error measures comparison for both models.

So basically, the graph tells us that LSTM model absolutely outperforms the ARIMA model (the opposite result apart from my result shown in the Spotlight Video Presentation, as I improved and modified the training and testing process). The mean absolute percentage error, or MAPE, is a measure of prediction accuracy in percentage. The lower the MAPE, the better the performance of the forecasting algorithm, and the LSTM model scores about 3.46% compared to ARIMA model's 11.18%. The root mean squared error, or RMSE, is the standard deviation of the residuals which means it tells you how the data points are concentrated around the best fit line. Again, ARIMA's RMSE value scores more than twice the RMSE value of LSTM, which is very disappointing, and thus, leaves the LSTM model a huge win.

6 Conclusion

In this pandemic, many brave, precious and fragile lives have been lost, but we have to stay strong and carry on. I examined the past related work lately done by many researchers that are still working at the forefront of the battlefield with COVID-19, and described how I collect, format, and clean the original data set obtained from Johns Hopkins University in detail. I then explained the pros and cons of the two models I later used to forecast the total COVID-19 confirmed cases in Virginia from Nov 23 to Nov 29. Two error measures are carefully calculated to help us figure out which model has a better performance when it comes to COVID-19 time series data forecasting, and the result shows that the Long Short-Term Memory model surpassed the other one by a large lead. At the end, we wish a better future for all of us.

References

[1] Jia, Lin, Kewen Li, Yu Jiang, and Xin Guo. "Prediction and analysis of Coronavirus Disease 2019." arXiv preprint arXiv:2003.05447 (2020).

[2] He, Sha, Sanyi Tang, and Libin Rong. "A discrete stochastic model of the COVID-19 outbreak: Forecast and control." Math. Biosci. Eng 17 (2020): 2792-2804.

[3] Peng, Liangrong, Wuyue Yang, Dongyan Zhang, Changjing Zhuge, and Liu Hong. "Epidemic analysis of COVID-19 in China by dynamical modeling." arXiv preprint arXiv:2002.06563 (2020).

[4] Gupta, Rajan, and Saibal Kumar Pal. "Trend Analysis and Forecasting of COVID-19 outbreak in India." medRxiv (2020).

[5] Chintalapudi, Nalini, Gopi Battineni, and Francesco Amenta. "COVID-19 disease outbreak forecasting of registered and recovered cases after sixty day lockdown in Italy: A data driven model approach." Journal of Microbiology, Immunology and Infection (2020).

[6] Tuli, Shreshth, Shikhar Tuli, Rakesh Tuli, and Sukhpal Singh Gill. "Predicting the Growth and Trend of COVID-19 Pandemic using Machine Learning and Cloud Computing." Internet of Things (2020): 100222.

[7] Dong E, D. H., Gardner L. COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University, (2020).

[8] Chen, J. Autoregressive Integrated Moving Average (ARIMA). Investopedia (2019).

[9] Grogan, M. Limitations of ARIMA: Dealing with Outliers. towards data science (2020).

[10] Brownlee, J. A Gentle Introduction to Long Short-Term Memory Networks by the Experts. Machine Learning Mastery (2017).

[11] Sepp Hochreiter, Jürgen Schmidhuber: Long Short-Term Memory. Neural Computation 9(8): 1735-1780 (1997)

Contributions

The project proposal were collaboratively written by Stephen Yang, Richard Tran and Zhewei Wu. But after that, due to experiencing Fall 2020's stressful heavy workload, Stephen and Richard decided to drop both the class and the undergraduate research project. So, this will mainly be a solo research paper conducted and written by Zhewei Wu, and it is warmly supported and advised by Dr. Ming Jin.

Note: The whole report used "we" instead of "I".