Project Presentation

Dynamic Graph Neural Networks For Monitoring Pandemics And Risk Factors

By - Sripath Mishra, Vishnu Devarakonda, Jason Kao, Boya Ouyang

Data Source

- CDC State wise daily infection data
  - COVID-19 Case Surveillance Public Use Data | Data | Centers for Disease Control and Prevention (cdc.gov)
  - Submission_date,state,tot_cases,conf_cases,prob_cases,new_case,pnew_case,tot_death,conf_death,prob_death,new_death,pnew_death,created_at,consent_cases,consent_deaths
  - 17,641 rows of data.
- Nytimes county level Covid infection data
  - nytimes/covid-19-data: An ongoing repository of data on coronavirus cases and deaths in the U.S. (github.com)
  - date,county,state,fips,cases,deaths
  - 803,684 rows of data.

Data Source (Continued)

- Safegraph State visitor monthly data
  - Places Schema (safegraph.com)
  - Year,month,state,num_visits,num_unique_visitors
  - 602 rows of data.
- Safegraph state population weekly data
  - Places Schema (safegraph.com)
  - Date_range_start,date_range_end,state,census_block_group,number_devices_residing
  - 5,062,411 rows of data.

Data merging and cleaning

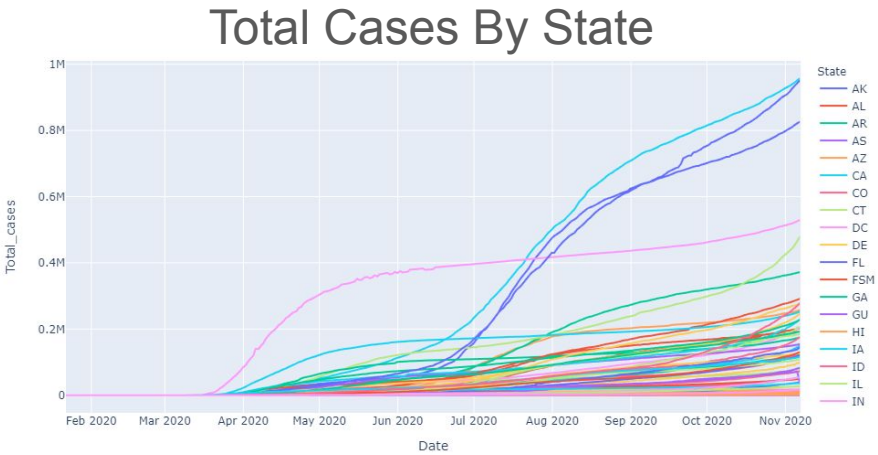
- Removed all data related to probability.
- Removed the columns of data which have a granularity lower than states.
- Made the formats same between dates and other common fields.
- Final data columns for states data(17,461 rows of data):
  Date,State,Total_cases,Confirmed_cases,New_cases,Total_deaths,Confirmed_death,New_deaths,State_resident_number,State_visitor_number
- Final data columns for county data(803,684 rows of data):
  Date,State,County_Name,County_Fips,Cases,Deaths

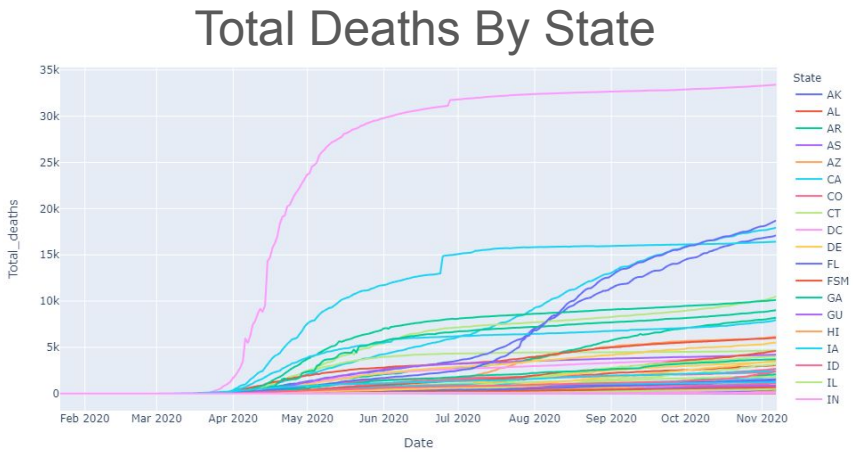
Approach

- The goal is to forecast the possible new cases in each state/USA.
- There are multiple ways to make forecasts.
- Measure our performance by comparing with other approaches for forecasting data.
- ARIMA - Autoregressive integrated moving averages.
- LSTM - Long Short Term Memory Network.
- GNN - Forecasting with Graph Neural Network.

Exploratory Data Analysis

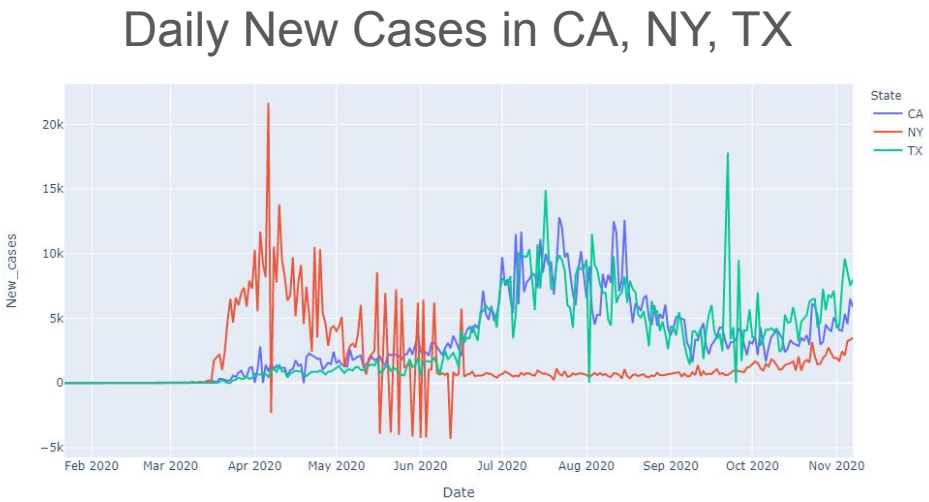
- First, let's get an idea as to how the data looks.

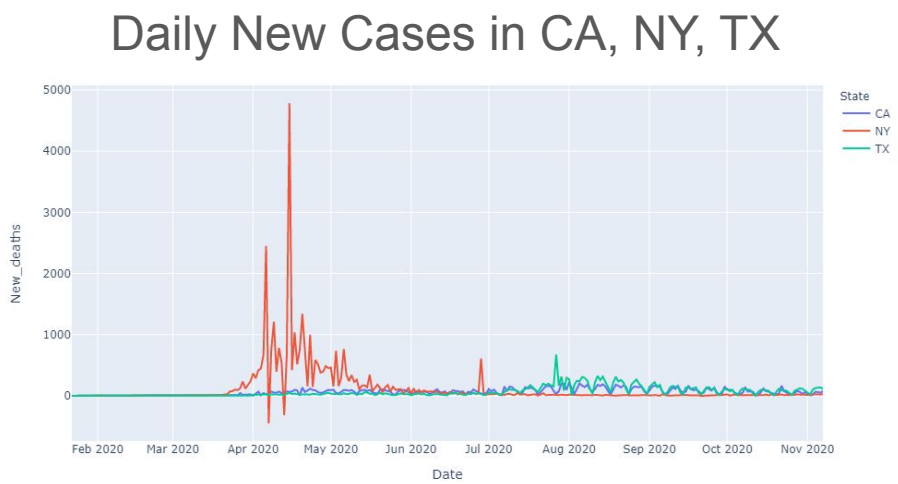




EDA - Continued

- Cases in the most populous states.





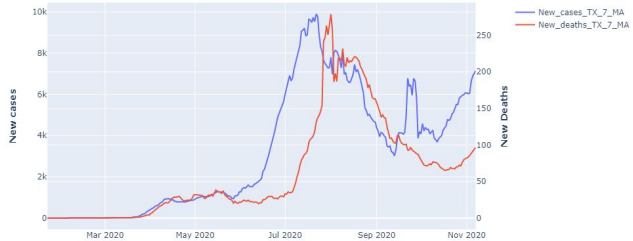
EDA - Continued

- Rolling Averages in each state.

NY

CA

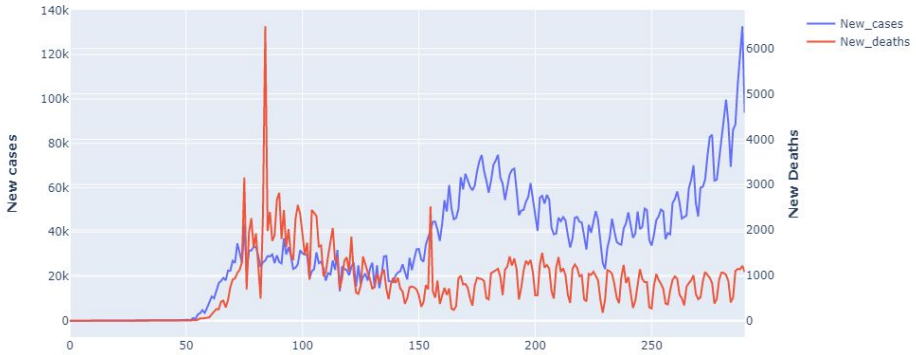
TX



EDA - Continued

- Pandemic in the USA

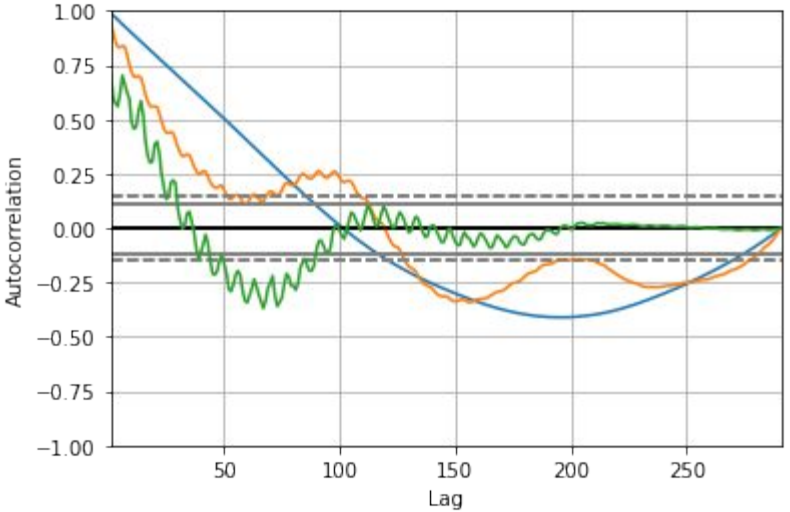
Daily Cases in USA



Daily Deaths in USA

Autoregressive Integrated Moving Averages

- Autoregressive integrated moving averages or ARIMA typically operates on time series data.
- It is designed to capture the patterns in time series data and consists of 3 parts.
- Autoregressive model.
- Moving-average model.
- Differencing.

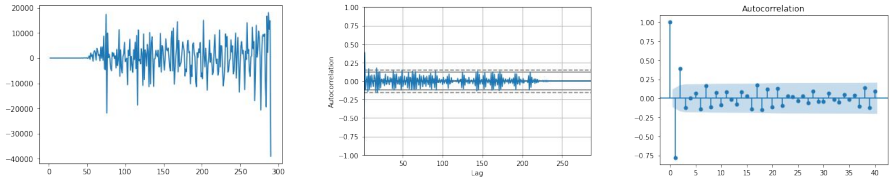


Autocorrelation plot for USA total cases, new cases, and new deaths.

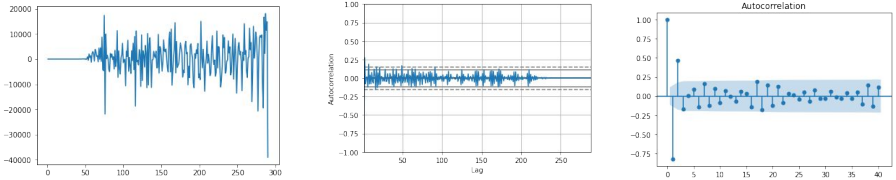
Forecasting With ARIMA

- Forecasting with ARIMA requires us to identify the appropriate parameters for the 3 models described earlier.
- The plots on the right help us determine these values.
- Left plot is difference.
- Middle plot is for autocorrelation.
- Right plot is for partial autocorrelation.

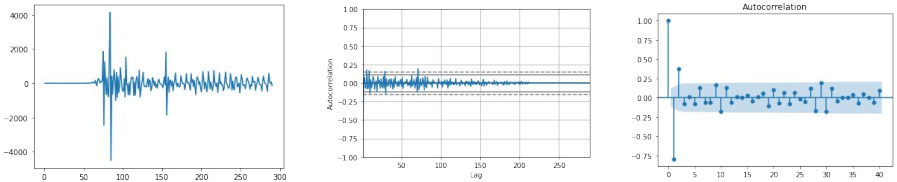
Total Cases Plots



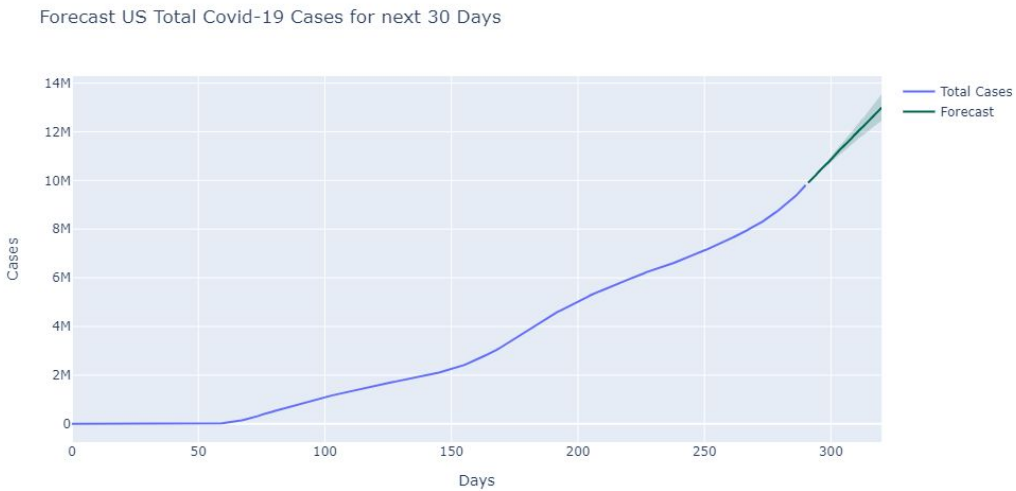
New Cases Plots

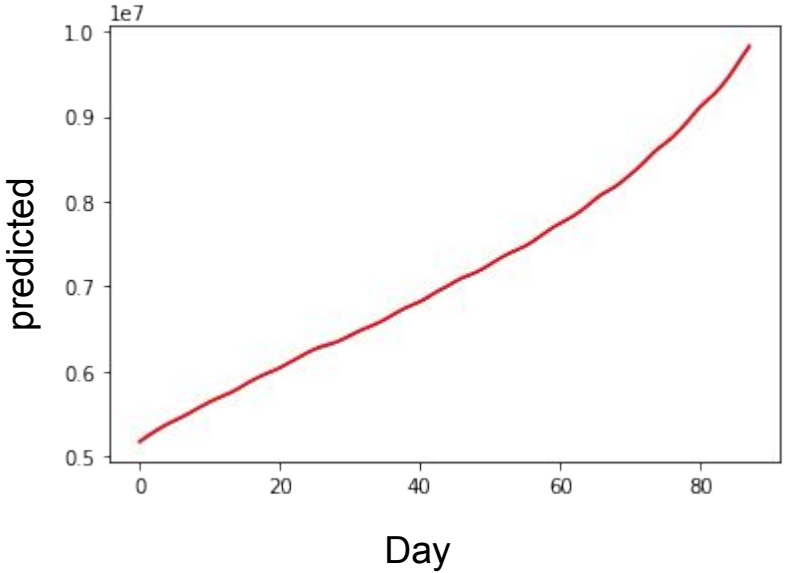


New Deaths Plots

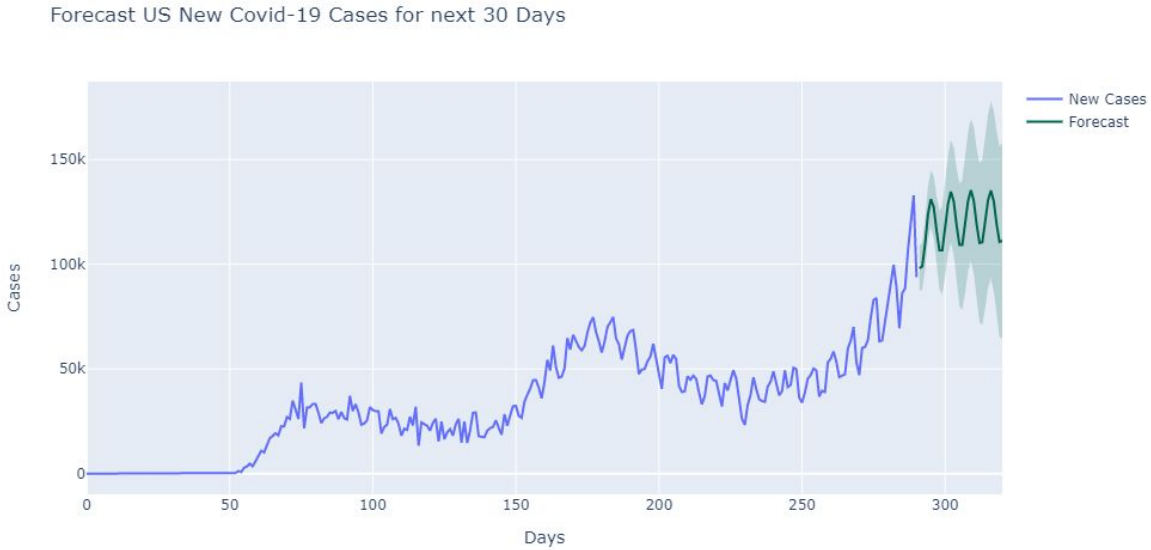


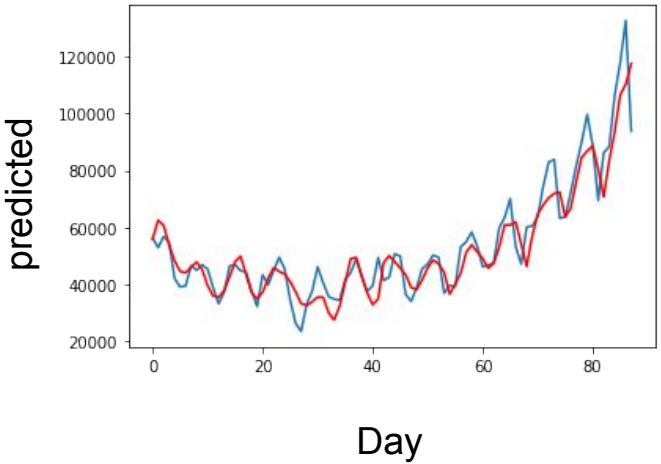
Forecasting Total Cases In USA



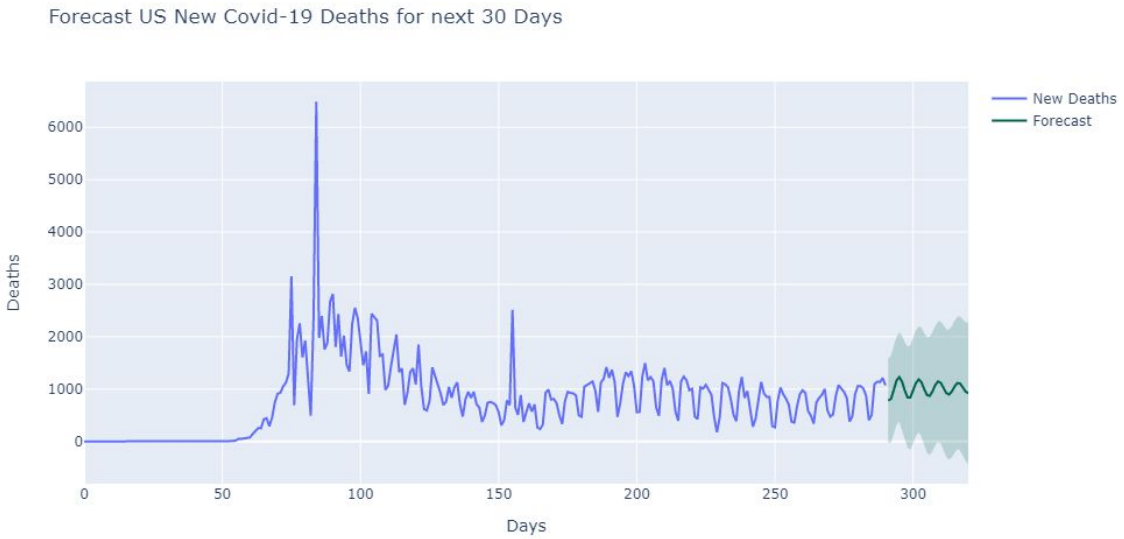


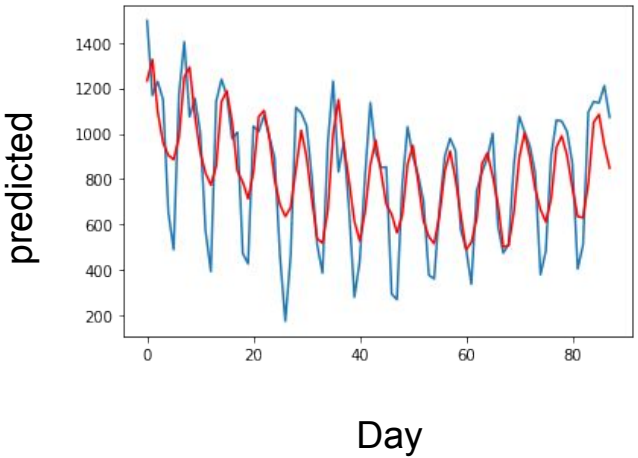
Forecasting New Cases In USA





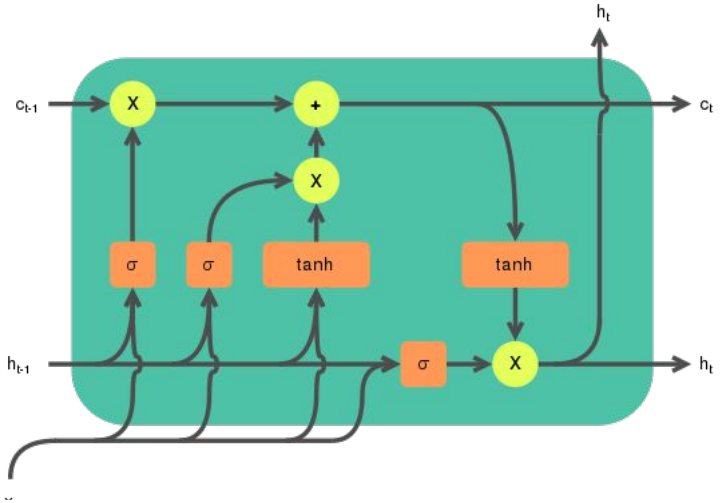
Forecasting New Deaths In The USA





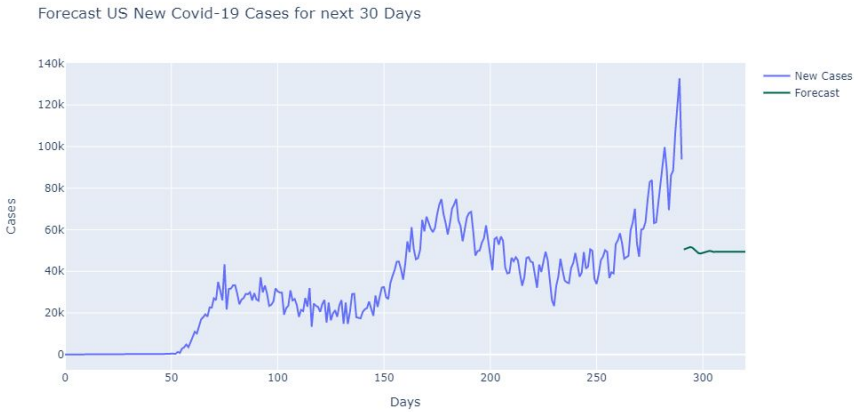
LSTM

- LSTM’s are deep learning models that are designed to work with sequential data sets.
- They store internal state information that influences predicts.
- This should allow us to capture temporal features in these data sets.

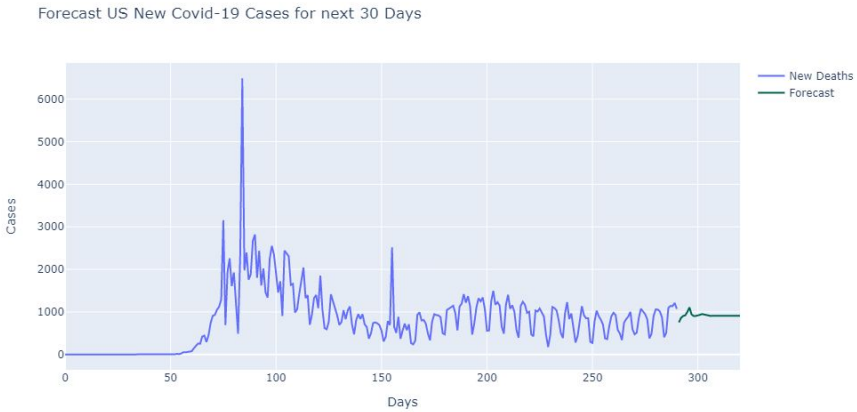


LSTM Forecasts

New Cases



New Deaths



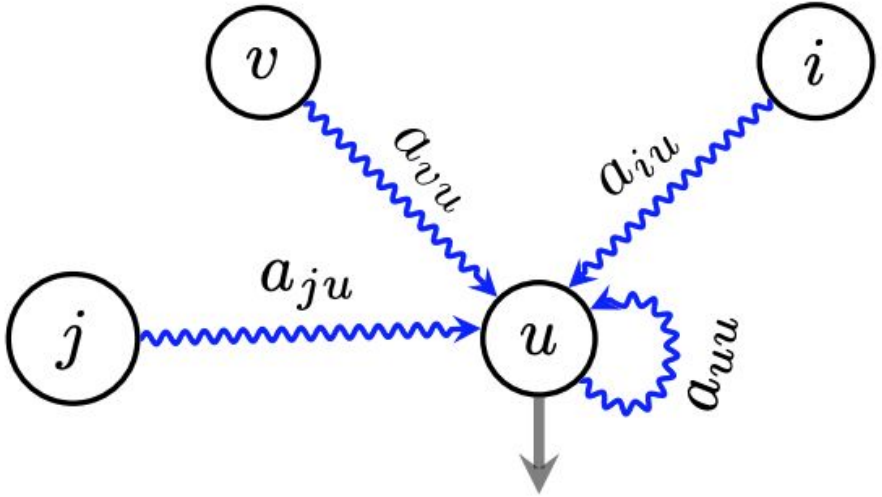
- ARIMA models are very good.
- LSTMs on the other hand fail to capture the
- But they only capture temporal features.
- Can we do better by including spatial features as well?

	Total Cases MAE	New Cases MAE	New Deaths MAE
ARIMA	5337.5	5382.0	152.017
LSTM		2988.0	8

Message passing neural networks (MPNNs)

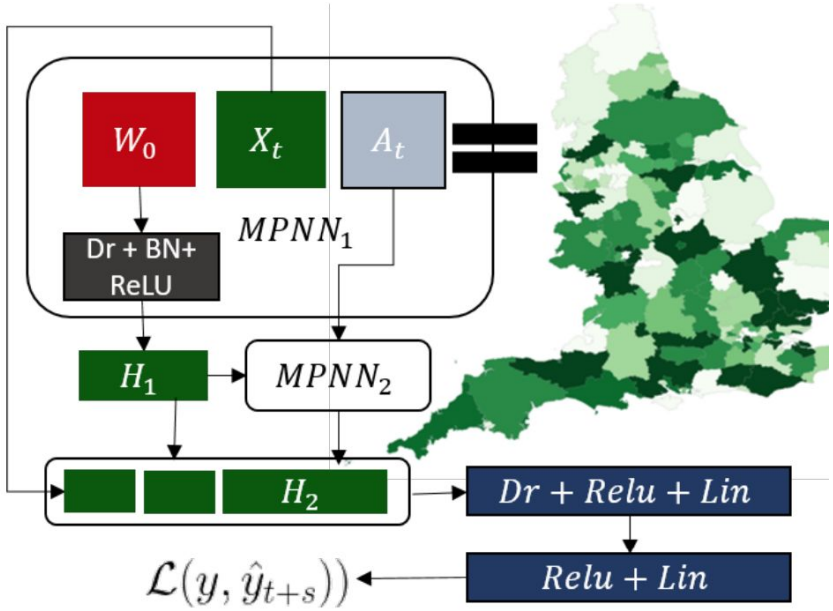
$$\mathbf{A}^{(t)}\,\mathbf{X}^{(t)} = \begin{bmatrix} w_{1,1}^{(t)} & w_{2,1}^{(t)} & \dots & w_{n,1}^{(t)} \\ w_{1,2}^{(t)} & w_{2,2}^{(t)} & \dots & w_{n,2}^{(t)} \\ \vdots & \vdots & \vdots & \vdots \\ w_{1,n}^{(t)} & w_{2,n}^{(t)} & \dots & w_{n,n}^{(t)} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1^{(t)} \\ \mathbf{x}_2^{(t)} \\ \vdots \\ \mathbf{x}_3^{(t)} \end{bmatrix} = \begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \\ \vdots \\ \mathbf{z}_3 \end{bmatrix}$$

Message passing neural networks (MPNNs)

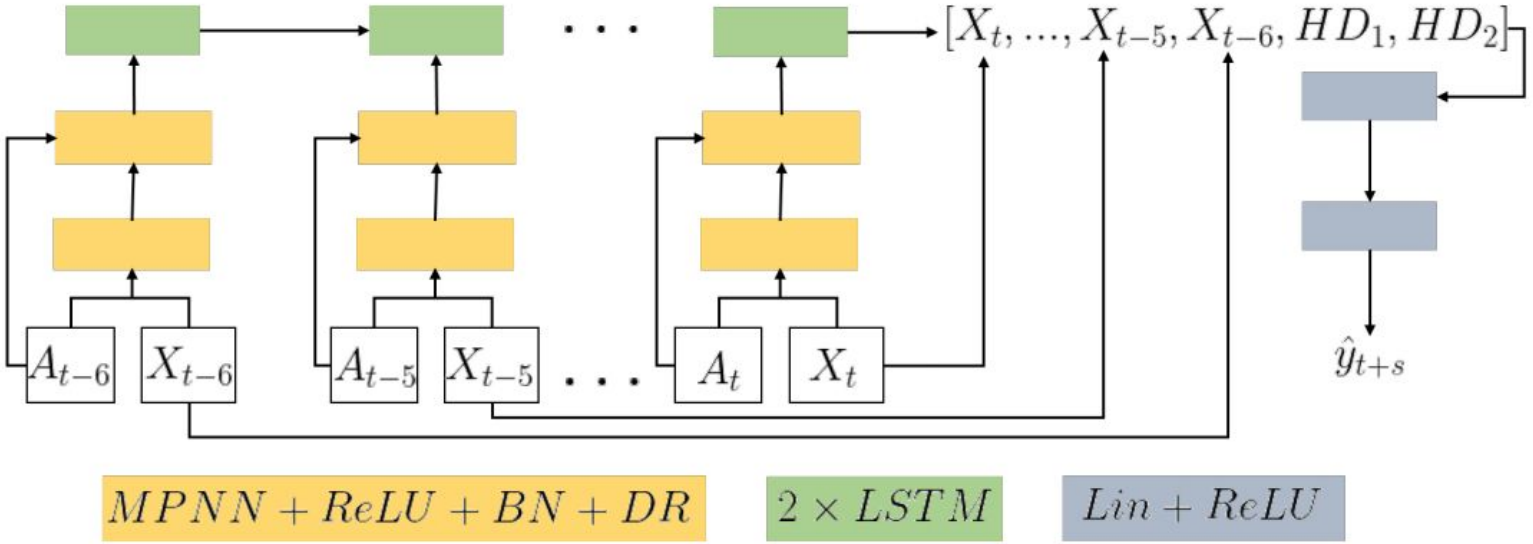


$$Z_u = (x_j a_{j,u} + x_i a_{i,u} + x_v a_{v,u}) + x_u a_{u,u}$$

Overview of the MPNN architecture



Overview of the MPNN LSTM architecture



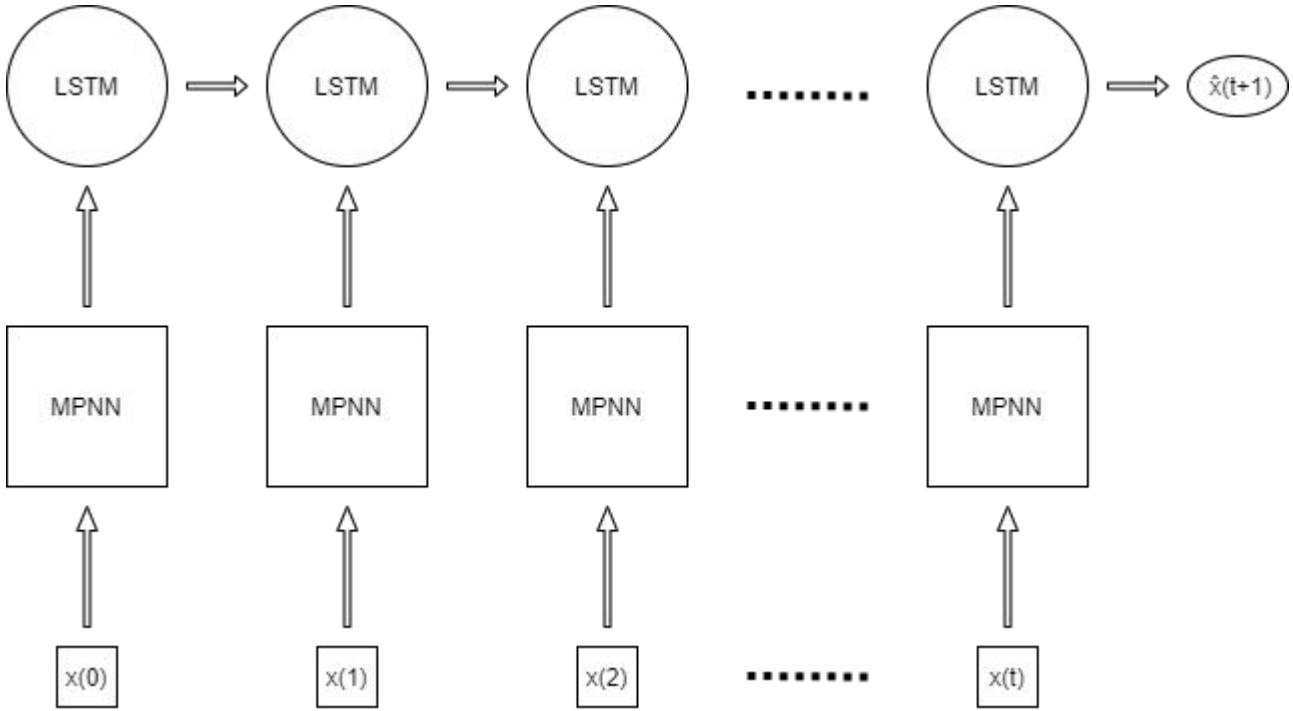
Our Implementation - Libraries

- Tensorflow/Keras - LSTM modules
- Spektral - MPNN modules
- SciPy - Data representation

Available Data

- Adj - 52x52 Adjacency matrix
  - 50 states, Washington DC, and US
  - Weighted
- X - 52x5 Node feature matrix
  - Number of cases in each node in the past 5 days
- County level data follows a similar structure with finer granularity.

Our Implementation - Network Structure



Project Status

- The model construction in code is still to be completed.
- Evaluation using ARIMA will be underway as soon as the model is ready.
- County data will also have to be evaluated.

Challenges

- Lack of mobility data between states.
- Modifying static graph neural network models to accommodate evolving data.

Thank you

Questions