Modeling and Predicting COVID-19 Infection Trends

Desmond Cheong, Natalie Rshaidat, Thomas Phillipoff, Rahul Mani

Goal:

Forecast COVID-19 infection trends for up to a week in locations affected by COVID-19 given relevant data including current infections, susceptible population, recovered cases, deaths, mobility score, and other factors.

Data:

Global COVID-19 data

- Source: Center for Systems Science and Engineering (CSSE) at Johns Hopkins
- Time frame: January 23, 2020 to May 4, 2020
- 59 different global locations, 104 days worth of data each. 6136 data points compiled into database

Population Densities

- Source: 2019 Revision of World Population Prospects, United Nations
- Locations: all 59 of the above

Presence of lockdown

- Source: ACAPs
- Locations: all 59 of the above
- Date range: January 1 2020 - May 5, 2020

Social Mobility based on smartphone data

- Source: Descartes Lab, a geospatial analysis company
- Locations: US states and counties only
- Date range: March 1, 2020 - April 9, 2020

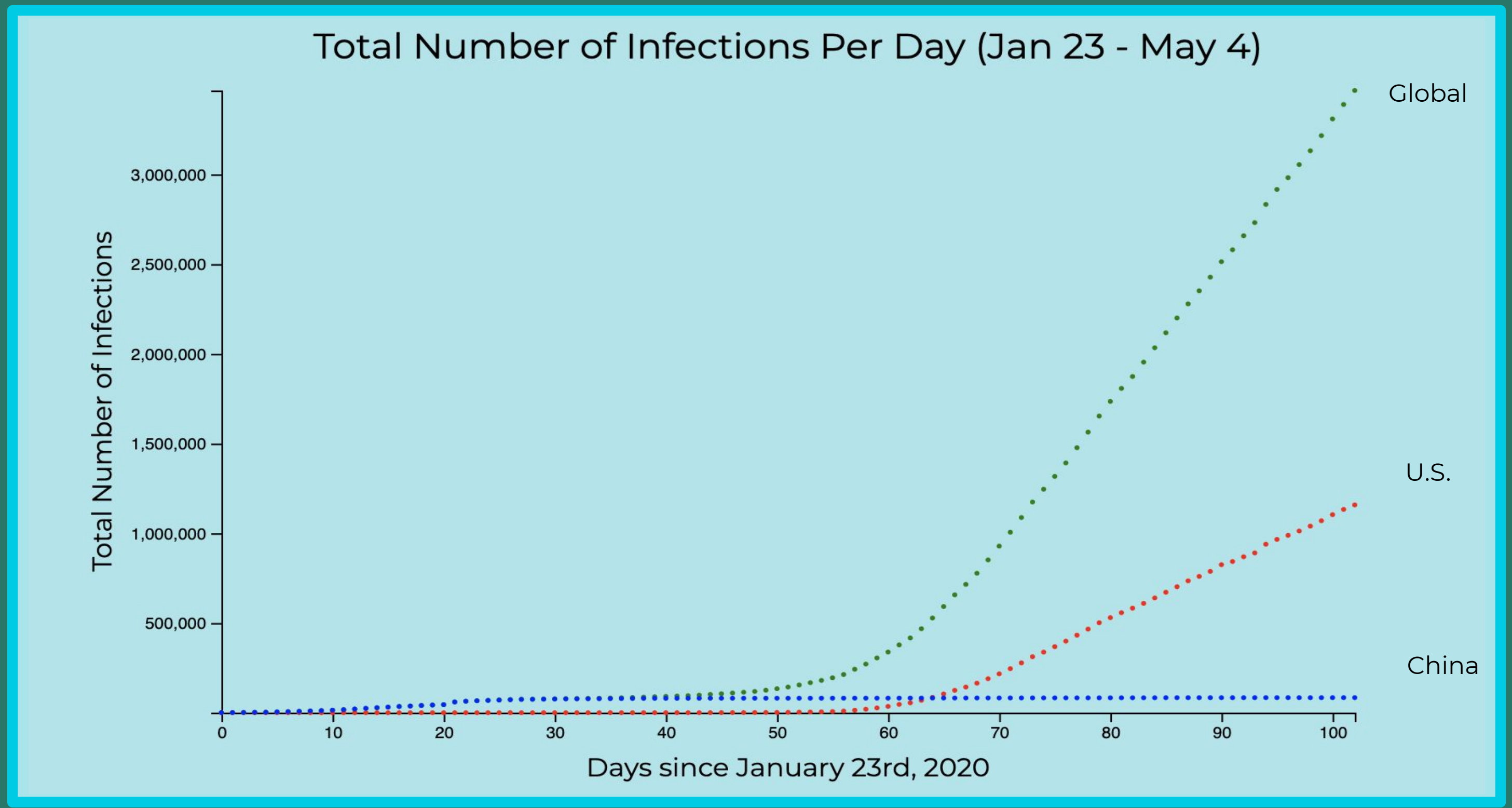

Figure 1. *Graph of total infections over time with data retrieved by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University: [https://github.com/CSSEGISandData/COVID-19 accessed on 4/04/2020].*

Models & Methodology:

Forecasting with Linear Regressor/MLP Regressor

- Selected the top 5 most correlated variables based on Figure 2. Infected, recovered, previously infected, deaths, and days since the onset of the outbreak.
- Predict next-day infections, recovered, and deaths using:
  1. **Linear regressor** with Lasso regularization and normalized variables
  2. **Multi-layer Perceptron (MLP) regressor** with early stopping (10% of training data used for validation)
- Feed these predictions back into the model to extend forecast up to 7 days with student forcing
- Train/test split ratio of roughly 8:2. Trains on 47 global locations, and tests on 12 random locations with completely held-out data

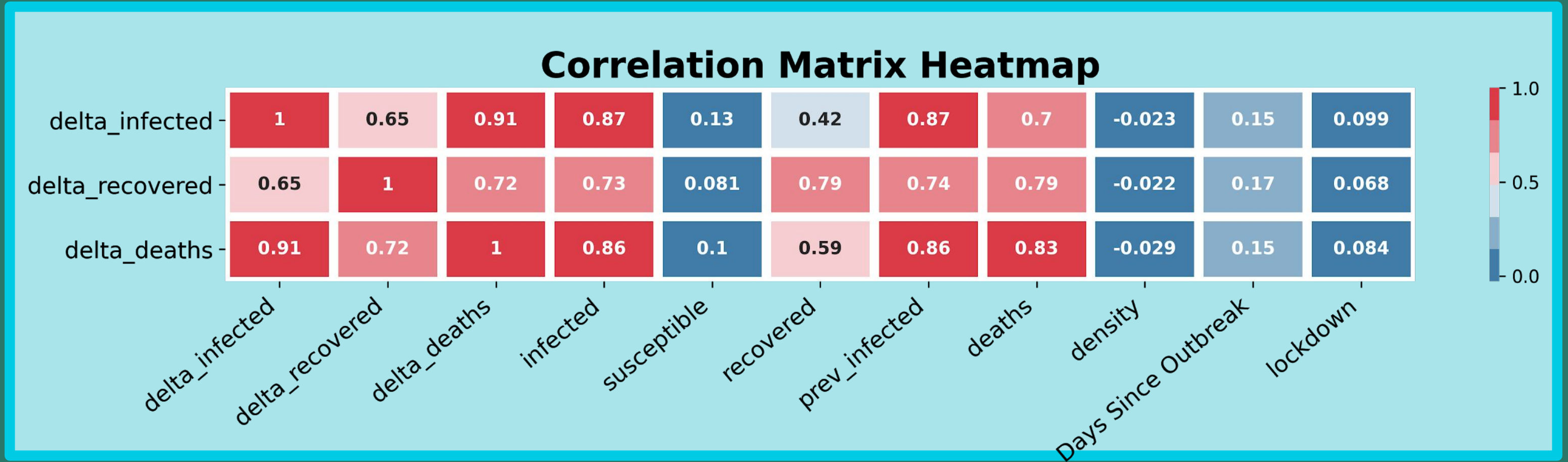

Figure 2. *A correlation matrix heatmap of variables for selection in our linear/MLP regression models. *delta_infected, delta_recovered, delta_deaths are the average number of new infections, recovered, deaths over a 3 day sliding window. Prev_infected is the number of infections from the day before.*

Forecasting with Time-Series Model

- Smaller scale analysis focused on 50 states
  - Uses additional data that is external to the virus i.e. movement of people
- Two variables: **infections** and **social mobility**
- Trains on 40 states by concatenating them, tests on remaining 10 states
- We predicts 7 days out for test states
- Sole parameter is number of lags. We chose 8 days. Why?
  - Delay between infection and infection detection of about this length
  - Takes time for mobility change to affect infections
  - 8 lags also minimizes penalty metrics like AIC, BIC, HQ on training data
- Model ran on 3 different ways of measuring infection rates. We call these different forecasts **Pure VAR, Percent VAR, and Log VAR.** They give different results.

Error Metrics

- We judge the forecasts based on three metrics: mean positive error, mean negative error, and mean percentage error
- Mean positive and negative errors are given in terms of infection numbers to help stakeholders plan for the future (e.g. number of hospital beds to prepare)
- Negative error is more important to minimize as underestimating infections causes more problems than overestimating

Results & Analysis:

Forecasting with Linear Regressor/MLP Regressor

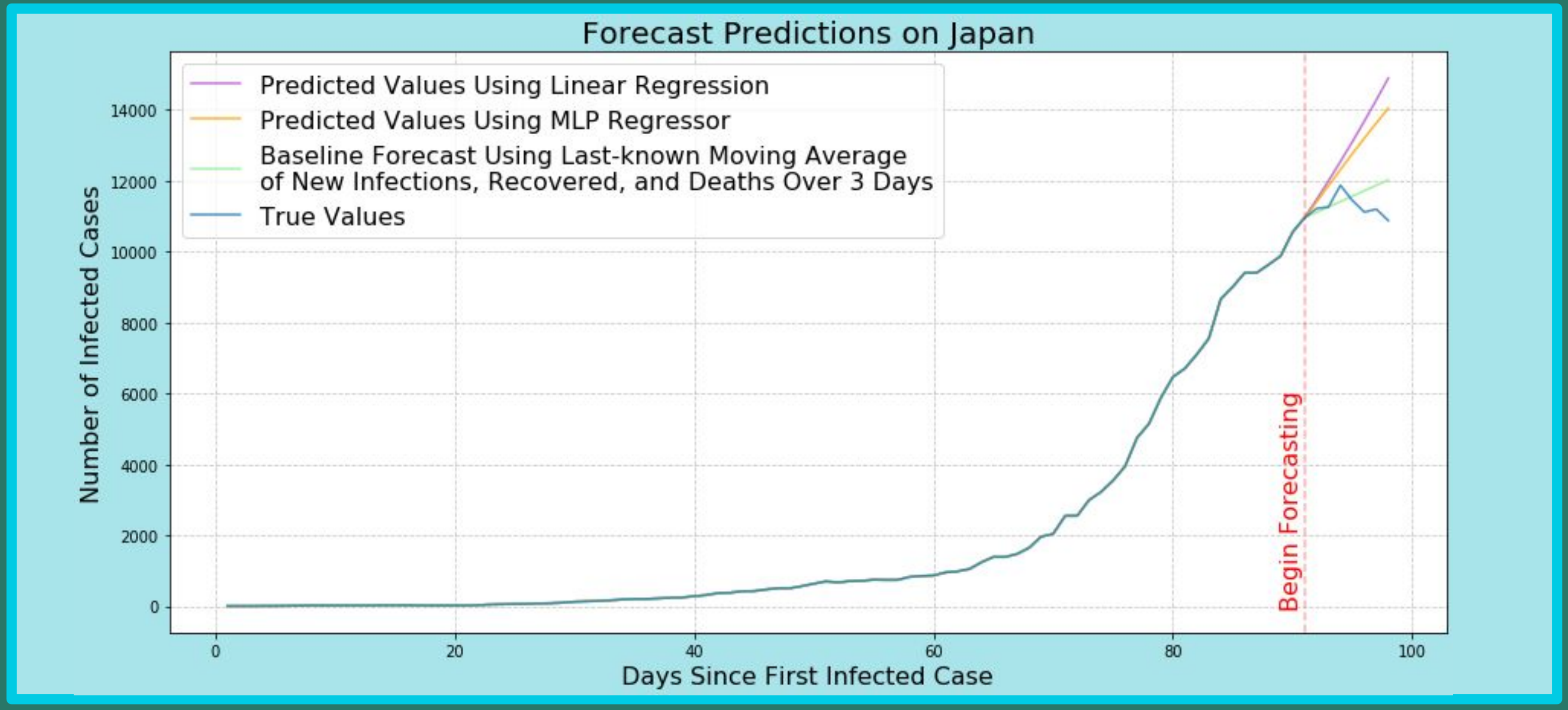

Figure 3. *Sample graph of forecast predictions on Japan (on of the test locations) over the last 7 days of available data.*

- Figure 3 shows a sample of the forecasting performance
- In Table 1, we see that across test countries for all error metrics, the baseline predictor performs best on average
- This is potentially a result of the short time-frame of the forecasts, as it takes time for infection patterns to change
- Additionally, inspecting forecasts on other test locations over different forecast periods, we noticed predictions tended to have higher error when there were downward trends in infection numbers
- This could be due to a lack of data that captures external factors, e.g. quarantine measures

Forecasting with Time-Series Model

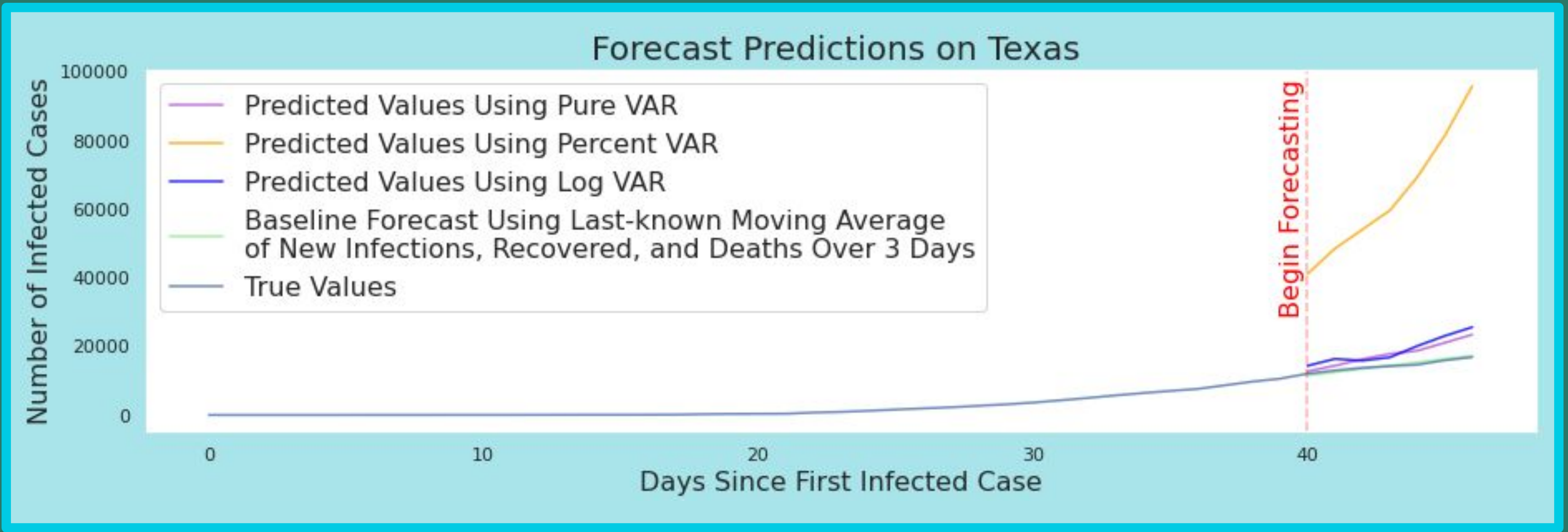

Figure 4. *Sample graph of forecast predictions on Japan (on of the test locations) over the last 7 days of available data.*

Causality:

- We performed a Granger causality test on the null hypothesis that coefficients on the lags of our variables are zero
- We reject the null hypothesis that mobility does not affect infections, because all 3 models had p-values of 0.000
- We did not reject the null hypothesis for the reverse relation, as only two models had p-values of 0.000, while Pure Var had a p-value of 0.710

Curve Fitting:

- Percent VAR fit the curves on testing states the worst, getting high mean positive error (23639) and low R-squared (0.20)
- Log VAR fit the curves on testing states the best, getting low mean positive error (2948) and high R-squared (0.95)

Predictor →→→ ---------------------- ↓ Error Metric*	Baseline predictor (last-known moving average of new infections, recovered, deaths)	Linear Regressor	MLP regressor	Baseline (last-known rate of new infections)	Time Series (Pure VAR)	Time Series (Percent VAR)	Time Series (Log VAR)
Mean positive error (infections)	107.2	607.2	309.7	242	3137	23639	2948
Mean negative error (infections)	-129.0	-322.8	-241.9	-174	0	0	0
Mean percentage error	29.1%	13112.0%	694.5%	0.3%	164%	530%	100%

*Error metrics are calculated as the mean error across all 7-day forecasts for all testing locations in the testing datasets.

Table 1. *Comparison table of all of our models' prediction errors.*

Limitations:

Forecasting with Linear Regressor / MLP Regressor

- Higher error when predicting downward trends in new number of infections
- Most training data showed upward trends, as many countries are not past their peak, possibly skewing predictions

Forecasting with Time-Series Model

- Successfully predicts slower rates of increased infections for decreased social mobility. However, unclear if it's learning or just copying other curves
- Potential for mobility data to help capture factors external to the virus to improve predictions of downward trends

Both models have higher error when there are downward infection trends, possibly due to inherent limitations in the dataset, or insufficient information about factors external to the virus e.g. social behaviour, government interventions etc. This reduces how far we can accurately forecast infections. Future directions would be to run our models on future data, where infection rates are well past their peaks in their respective location.

Since our models were unable to capture the full effect of social distancing in predicting future infection rates, one possible future direction would be to incorporate the recent social mobility dataset Apple released in hopes of improving accuracy. Another future direction could be to incorporate the differences in lockdown severity into the dataset.