United States COVID-19 Daily Cases Prediction in 2020 Using Bidirectional LSTM with Clustered Data

Lin Feng Zou[1]; Thesis Advisor: Stavroula Chrysanthopoulou[1]

[1]Brown University School of Public Health Department of Biostatistics, Providence, RI

Overview

We obtained natural clusters of US states based on similarities in trends of daily COVID-19 cases in 2020, modeled the trends within these clusters using a bidirectional long short term memory recurrent neural network, and evaluated the predictive accuracy, interpretability, and implications of the model.

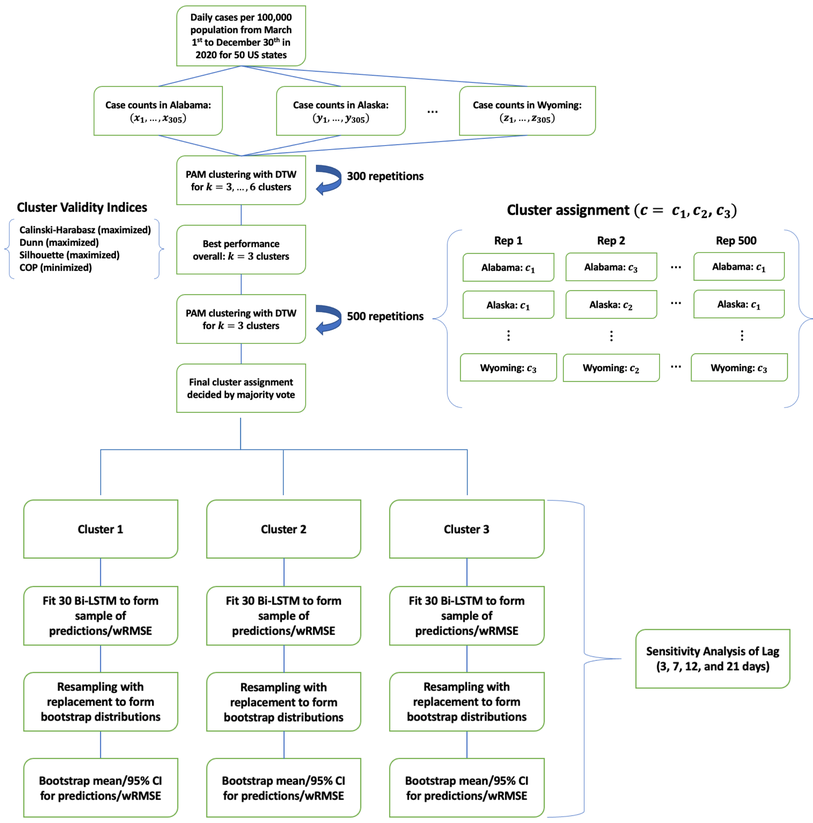
Background

- To help healthcare professionals be better prepared for spikes in cases, it would be beneficial for public health officials to develop intervention plans based on accurate predictions of daily COVID-19 cases
- The uncertainty surrounding the propagation of the pandemic in the US was further exacerbated by its political structure, as community mitigation strategies (or lack thereof) varied at the state-level and more granularly the county-level
- Previous studies have used the effective reproduction number R_t to cluster US states, but doing so could hide local variation, as R_t is a population average
  - To account for state-level variation, we clustered US states based on similarities in their incidence trends in 2020
- Known for its ability to learn long term dependencies, the bidirectional Long Short Term Memory (Bi-LSTM) model is a powerful method when data is abundant, and the prediction problem is nontrivial

Methodology

- Primary outcome: COVID-19 daily case count per 100,000 population
  - Data source: Johns Hopkins University
- Partitioning around medoids (PAM) clustering with dynamic time warping (DTW) was used to obtain natural clusters of US states based on similarities in trends of COVID-19 incidence in 2020
  - Sensitivity analysis to stabilize clustering assignment
- Bi-LSTM can access both past and future information, providing additional context when learning patterns during model training
  - Bootstrap procedure to estimate the effect of the random initialization of parameters on the predictions and cluster-weighted root mean squared error (wRMSE)
  - Sensitivity analysis of lag (prediction window) to assess model performance

Workflow



Conclusion

- Bi-LSTM can provide excellent interpolated predictions and moderately accurate extrapolated predictions after careful consideration is taken to simplify the prediction problem, such as fitting models within homogenous groups of US states with similar incidence trends in both magnitude and overall shape
- Shorter windows of prediction may not contain the relevant patterns necessary for predicting COVID-19 incidence, resulting in poor model performance, whereas longer windows of prediction are more likely to contain such patterns and are able to provide more accurate predictions
- Future work in this area may further simplify the prediction problem to attain improved model performance by including proxy measures, such as stringency index, mobility, and testing, that quantify the effect of community mitigation strategies in each US state

Results

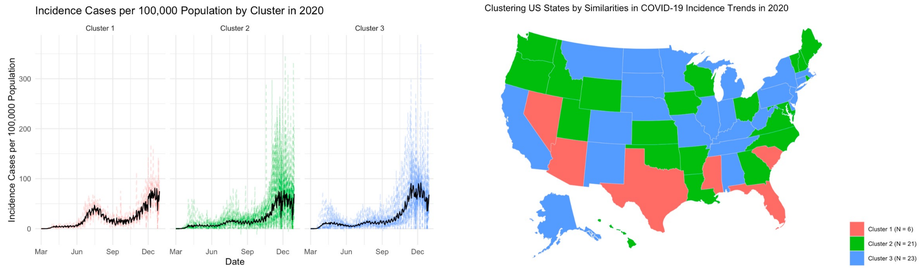


Figure 1. Clusters of US states based on similarities in incidence trends in 2020. Cluster 1 can be characterized as states that had a large spike in the summer and poor recovery in the winter. Clusters 2 and 3 more closely resemble the three-wave pattern observed in the US overall.

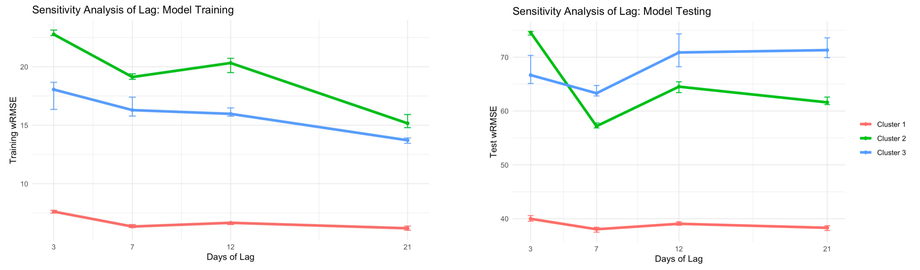


Figure 2. Sensitivity analysis of lag (prediction window) for model training (interpolation) and testing (extrapolation). Bi-LSTM fits the data very well and provides reasonable extrapolations. The "elbow" trend in model performance indicates the presence of an optimal lag window for COVID-19 incidence case prediction.

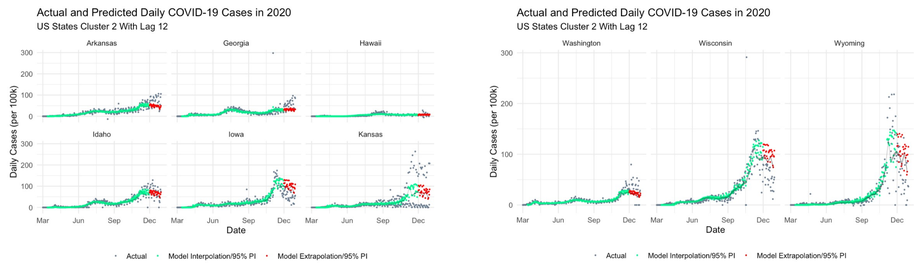


Figure 3. Actual versus predicted daily COVID-19 cases in 2020 for some of the states in Cluster 2 with a lag window of 12 days. The accuracy of the extrapolation suffered when predicting patterns never seen before during model training.

References

- Hochreiter, S. and J. Schmidhuber, *Long short-term memory.* Neural Comput, 1997. **9**(8): p. 1735-80.
- Muller, M., *Dynamic Time Warping*, in *Information Retrieval For Music and Motion*. 2007, Spring Nature: Switzerland. p. 69-84.