Data Augmentation Techniques And Clustering To Improve Deep Learning Forecasts Of Dengue Cases

Juan V. Bogado[1], Diego H. Stalder[2], Christian E. Schaerer[3]

[1] *Universidad Nacional de Caaguazú, jbogado@unca.edu.py*
[2] *Universidad Nacional de Asunción, Facultad Politécnica, cschaer@pol.una.py*
[3] *Universidad Nacional de Asunción, Facultad de ingeniería, dstalder@ing.una.py*

INTRODUCTION

Dengue fever represents a public health problem and accurate forecasts can help governments take the best preventive actions. As the volume of data provided continuously increases, machine learning and deep learning (DL) models have become an attractive approach. However, it is difficult to perform accurate predictions in areas with fewer cases or with a lack of available data. Several city models may present heterogeneous behaviors and poor accuracy. To mitigate this problem, we propose to enrich the data already available through clustering and data augmentation techniques and compare those techniques with an LSTM model considering weekly dengue incidence and climate, in 217 cities in Paraguay

PROBLEM

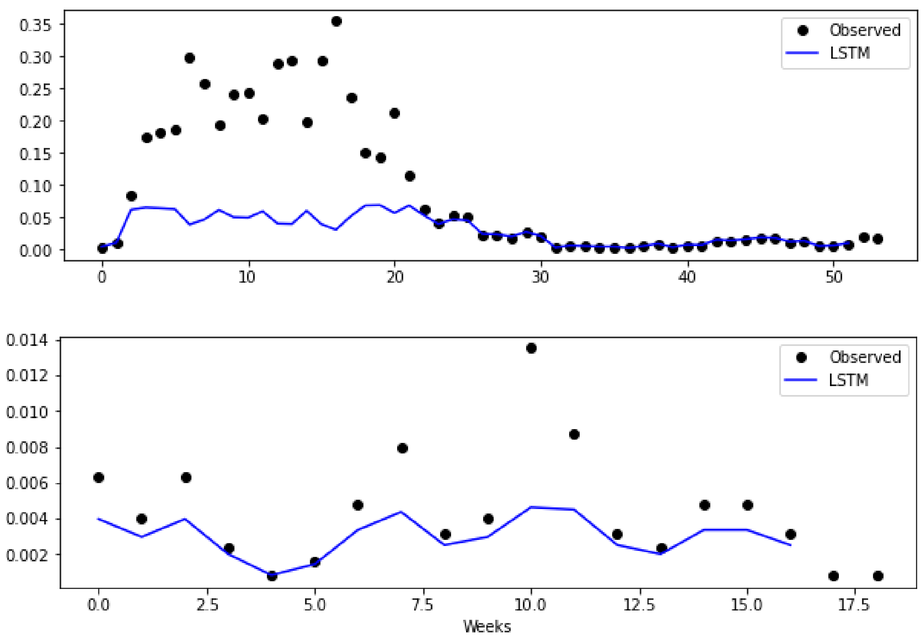

Figure 1. Failed predictions using LSTM models to predict Dengue cases in cities of Paraguay. From top to bottom San Lorenzo and Encarnación. The cause of this failure seems to be the lack of data available to train the model.

METHODOLOGY

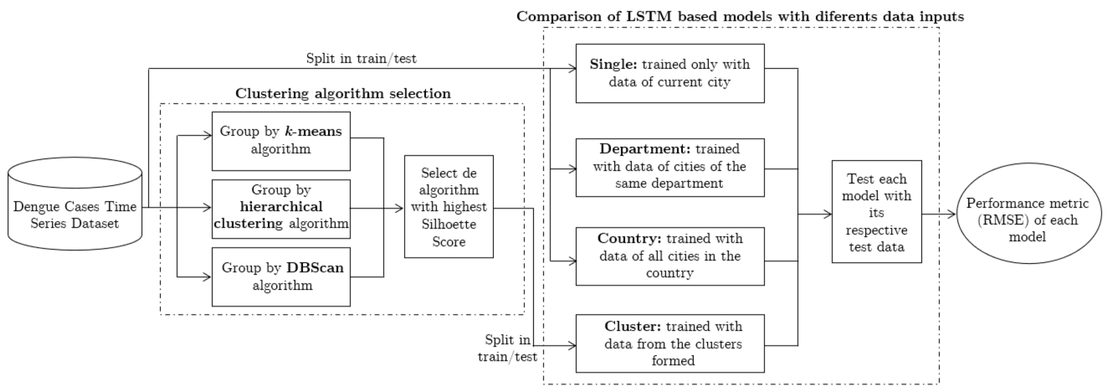

Figure 2. Summarized workflow of the clustering technique in order to obtain more data: reading the databases, carrying out the tests to determine the most suitable clustering technique, group the data by department, country and clusters obtained, train and test the clustered data and evaluate the model performance.

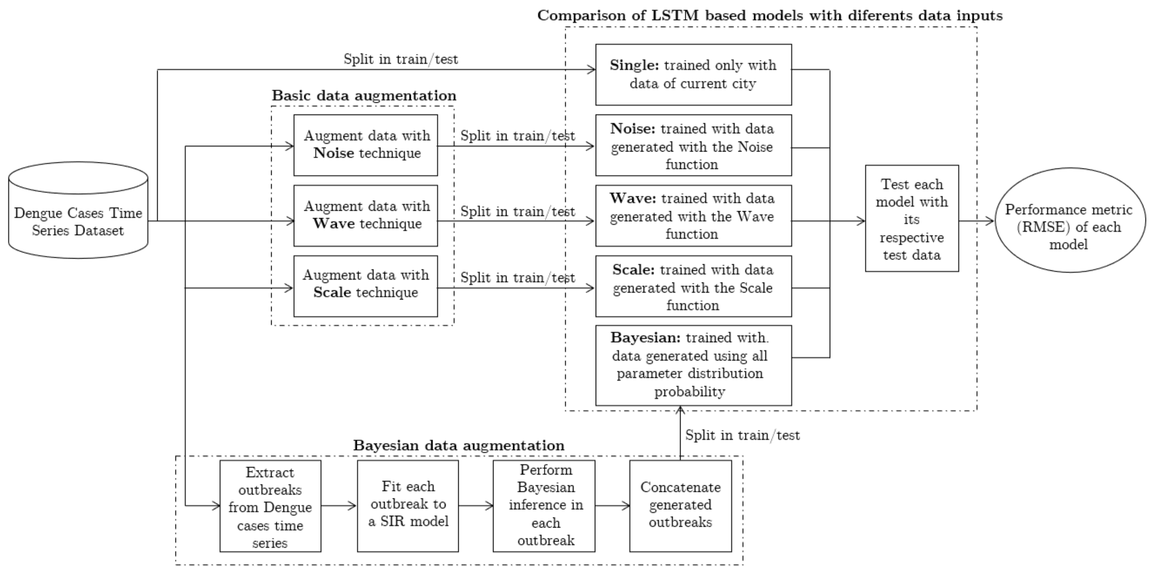

Figure 3. Summarized workflow of the data augmentation technique in order to obtain more data: reading the databases, carrying out the tests to determine the most suitable data augmentation technique, add the synthetic data obtained to dataset, train and test the augmented data and evaluate the model performance.

RESULTS

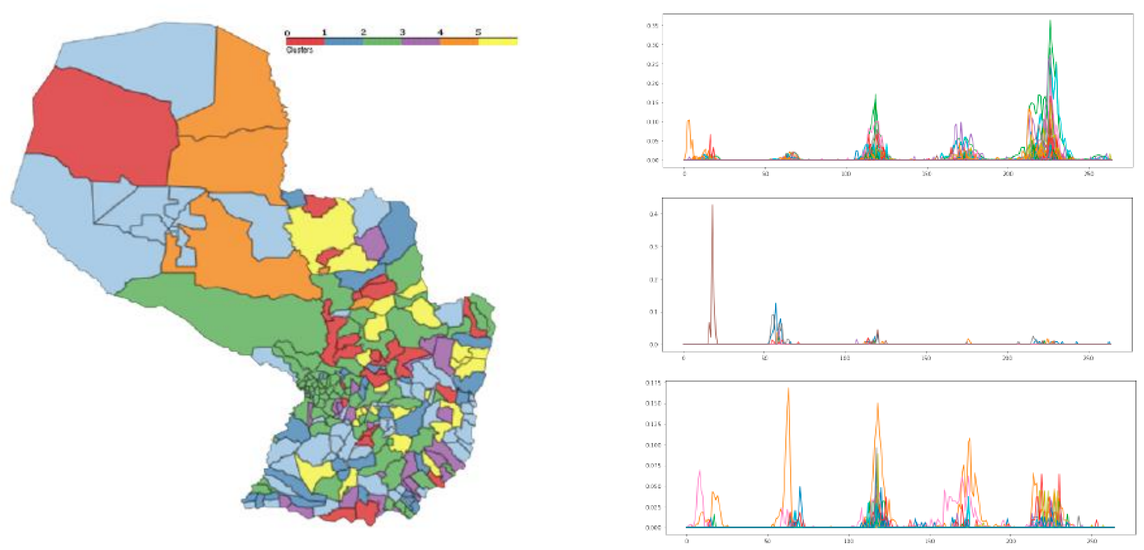

Figure 4. From left to right, the clusters formed are shown on the Paraguay map by color codes, cities with the same color correspond to the same cluster. Sample of three of the cluster formed. Note that clusters are not necessarily geographically adjacent.

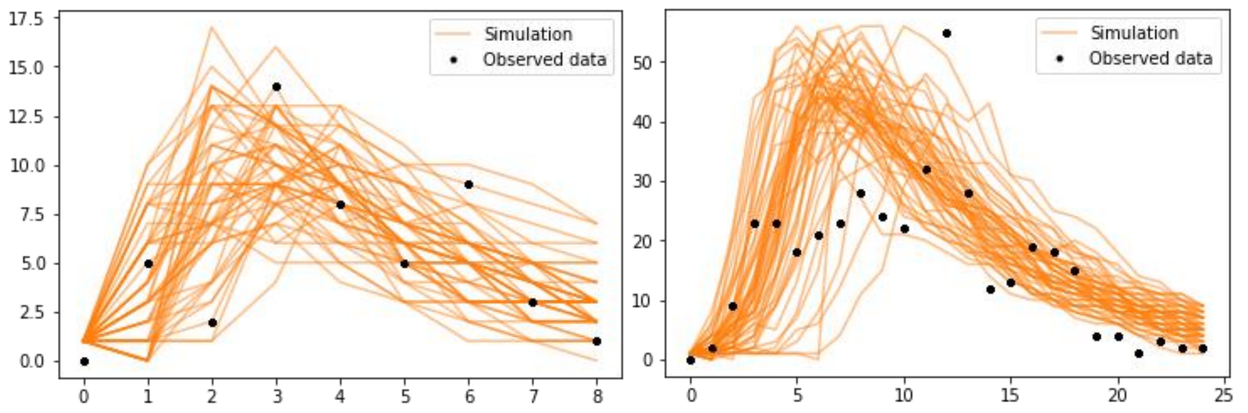

Figure 5. Detail of the synthetic series obtained with the Bayesian inference method for data augmentation.

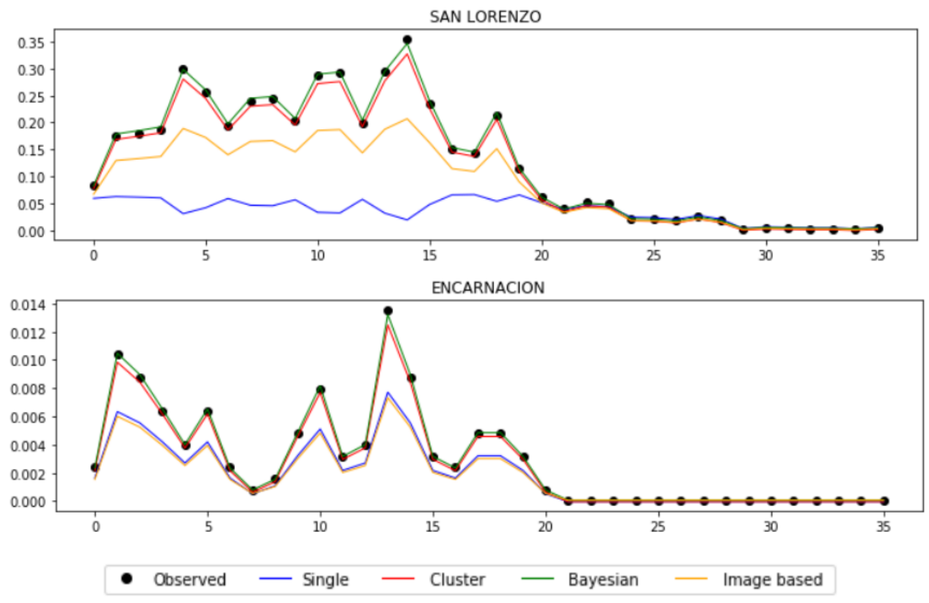

Figure 6. Comparison between LSTM models trained with data obtained by clustering (Cluster), with synthetic data obtained with Bayesian inference (Bayesian), with the best of the techniques inspired by image processing (ImageBased) and the reference model (Single) same as shown in Figure 1.

CONCLUSION

Through these techniques we have managed to improve the predictive models by expanding the dataset with synthetic data or grouping similar available series.

REFERENCES

[1] S. Aghabozorgi, A. S. Shirkhorshidi, and T. Y. Wah. “Time-series clustering – A decade review”. In: Information Systems 53 (2015), pp. 16–38. issn: 03064379. doi: 10.1016/j.is.2015.04.007.

[2] J. V. Bogado et al. “Time Series Clustering to Improve Dengue Cases Forecasting with Deep Learning”. In: 2021 XLVII Latin American Computing Conference (CLEI) (2021), pp. 1–10. doi: 10.1109/CLEI53233.2021.9640130.

[3] CDC. Centers for Disease Control and Prevention Official Site. Online. Accessed 10/02/2022, https://www.cdc.gov/.

[4] R. J. Hyndman, E. Wang, and N. Laptev. “Large-Scale Unusual Time Series Detection”. In: 2015 IEEE International Conference on Data Mining Workshop (ICDMW) (2015), pp. 1616–1619. doi: 10.1109/ICDMW.2015.104.