[Handwritten annotation at top: "→ also some of the Consequences of not stopping infectious dis."]

Integrating socioeconomic and geographic data to enhance infectious disease prediction in Brazilian cities

L. Lober,[1, *] K. Oliveira Roster,[2] and F. A. Rodrigues[1]

[1]*Departamento de Matemática Aplicada e Estatística,*
Instituto de Ciências Matemáticas e de Computação,
Universidade de São Paulo—Campus de São Carlos,
Caixa Postal 668, 13560-970 São Carlos, São Paulo, Brazil
[2]*Harvard T. H. Chan School of Public Health, Boston, MA.*
(Dated: May 3, 2024)

Supervised machine learning models and public surveillance data has been employed for infectious disease forecasting in many settings. These models leverage various data sources capturing drivers of disease spread, such as climate conditions or human behavior. However, few models have incorporated the organizational structure of different geographic locations for forecasting. Traveling waves of seasonal outbreaks have been reported for dengue, influenza, and other infectious diseases, and many of the drivers of infectious disease dynamics may be shared across different cities, either due to their geographic or socioeconomic proximity. In this study, we developed a machine learning model to predict case counts of four infectious diseases across Brazilian cities one week ahead by incorporating information from related cities. We compared selecting related cities using both geographic distance and GDP per capita. Incorporating information from geographically proximate cities improved predictive performance for two of the four diseases, specifically COVID-19 and Zika. We also discuss the impact on forecasts in the presence of anomalous contagion patterns and the limitations of the proposed methodology.

I. INTRODUCTION

Data driven models are increasingly used for disease forecasting given the availability of public health records and advances in machine learning algorithms and their applications to epidemiology [1, 2], with Brazil's *Data-SUS* [3] reporting over 50 unique diseases and health conditions being actively monitored, including multiple endemic diseases and also COVID-19, which had widespread repercussions in the country and beyond in recent years.

When employing theoretical procedures, one has the possibility of understanding the dynamics and properties of the infectious agents through compartmental models, such as done analysing the different COVID-19 variants [4] and Dengue's serotypes [5], with works also considering the co-circulation of diseases transmitted by the *Aedes* mosquito [6]. However, the main advantage of current supervised machine learning approaches over those approaches lies in their ability to directly infer the relationship between features of interest without the necessary knowledge of the epidemic's dynamics to arrive at reasonable predictions [7].

In terms of computational costs and explainability, decision trees ensembles, such as the Random Forests or the XGBoost algorithms, often proves to be a reliable method, specially when the neural networks' requirement of data availability is in the tens of thousands of measurements, which could also bias these models towards over-fitting.

A range of data sources has been used for forecasting, including active and passive surveillance records (cases, hospitalizations, serosurveys) [8], meteorological and environmental conditions [9], human behavior and internet searches [10]. The importance of different data sources varies among infectious diseases.

As a vector-borne disease, dengue cases depend on vector suitability conditions and human behavior, such as socioeconomic factors affecting vector suitability, mobility, and susceptibility to infection[11], with Brazil currently accounting for the most cases reported in Latin America, with an increase of 13% in these registers when compared to the past year [12].

COVID-19 on the other hand is a respiratory condition and thus cases are more closely correlated with measures of human interactions, such as crowding in public spaces and traveling, with early lock-downs, mask and latter vaccination adoption also had significant implications to its development, with the rate, delays and general adoption of those methods varying significantly between countries [13, 14].

Concurrently, other endemic infectious diseases also affects the country, such as Zika virus, which was shown to be linked to Dengue and Chikungunya [15]; and also Influenza as another influential airborne disease, both which also amplify the effects of Dengue and COVID-19 on the Brazilian population.

Studying those socioeconomic components as possible means of understanding the main factors of propagation could in turn increase the accurateness of predictions for those diseases, which would likely enable faster responses to new outbursts, as demonstrated in the case of malaria forecasting in Guyana [16]. Previous studies on the spreading of dengue in Brazil shown that geo-

graphic proximity and hierarchical levels of influence between cities are impactful in the transmission process, with highly influential cities with many transport links having increased odds of an outbreak. [17]. Also, the readiness for a given country to respond and its vulnerability to the effects of epidemic diseases have also been quantified in terms of several socioeconomic indicators [18, 19]. For this study, the gross domestic product (GDP) of each municipality will be considered as the chosen indicator of economic growth.

In this work, the aim was to develop and apply a strategy utilizing the underlying information contained in socioeconomic and geographic data from Brazilian cities as a way to increase the effectiveness of predictions for diseases, while also verifying the impact of this protocol for intrinsically different conditions. Multiple decision tree frameworks were employed and analyzed under a criteria that includes the seasonal naïve baseline, selecting the best model for each city via cross-validation and then evaluating on a hold-out test set. Predictions for COVID-19 notably benefits from this methodology, while Dengue, Zika and Influenza benefits less from socioeconomic and geographic associations.

II. MATERIALS AND METHODS

A. Data

Weekly cases for the diseases were sourced by the official Brazilian government database. Dengue, Zika and Influenza registers can be found on the System of Information on Aggravations and Notifications (*Sistema de Informação de Agravos de Notificação, SINAN*) [3] and also on the official government panel for COVID-19 data [20]. As for geographic and economic data, the information utilized to generate the correlations between cities can be found on IBGE platform (*Instituto Brasileiro de Geografia e Estatística*) [21], with data ranging from 2014 to 2020. The latitude, longitude, and GDP per capita of all available cities in this time frame were considered, with the geographic distance between municipalities calculated through the euclidean distances between latitude and longitude, while the yearly GDP data for each city was treated as time series and similarities were defined as discussed in Section II C.

After combining the available cities on DataSUS with IBGE's database, and also accounting for all years in the time range considered for Dengue, Zika and Influenza, the analysis respectively accounted for 1804, 211 and 274 cities that fit this criteria, while the COVID-19 database encompass 5565 unique cities: notice that, although with much larger coverage on Brazilian cities, the endemic diseases have measurements on a greater array of years. Data was then split into a training set (Dengue: Jan. 2014 - May. 2020; Zika: Jan 2016 - May. 2022 ; COVID-19: Mar. 2020 - Fev. 2023; Influenza: Jan 2013 - Jul. 2018) and a hold-out test set (Dengue: Jun. 2020 - Dec.

2021; Zika: Jun. 2022 - Dec. 2023 ; COVID-19: Mar. 2023 - Dec. 2023; Influenza: Aug. 2018 - Dec. 2019). The data was normalized to have the maximum cases count of one for each trained model.

For the four diseases included in this project, the skewness, mean and standard deviation of data was also included on Table I.

TABLE I: Basic statistic description for the weekly number of cases to the investigated diseases.

Disease	Mean	Maximum	Skewness
Dengue	178 ± 566	13540	5.3 ± 1.9
Zika	85 ± 244	2472	5.3 ± 2.0
COVID-19	552 ± 2680	107057	4.5 ± 2.5
Influenza	18 ± 51	885	3.8 ± 1.2

B. Methods

We employed the predictive algorithms Random Forests and XGBoost for all predictive tasks, where the training set used 5 delays (or lags) for the disease's time series. These two models function as ensembles of decision trees with intrinsically different methods of achieving forecasts. In all of them, the trees are generated by selecting subsets of observations through sampling with replacement methods, in this work using a random subset of features in each splitting process and selecting the best fit using Mean Absolute Error (MAE). Random Forests [22] determine the final result using a combination of multiple decision trees, being an approach with few parameters to tune throughout the training processes and thus presenting a robustness to over-fitting. As for the Gradient Boosting Regression method XGBoost [23] chosen, the trees are built individually, where weights are added depending on the performance shown by that tree to a given example, that is, the ensemble will be able to account for higher variability in data by evaluating trees based on the difficulty of prediction for those examples.

Mean absolute scaled error (MASE) was used to evaluate the performance of predictions, due to its interpretability and scale-invariant properties [24].

$$MASE = \frac{\frac{1}{J}\sum_j |\hat{y}_j - y_j|}{\frac{1}{T-m}\sum_{t=m+1}^{T} |y_t - y_{t-m}|}, \qquad (1)$$

where the upper term is just the mean average error (MAE).

MASE's main advantage is its interpretability, which is directly linked to the performance of the seasonal naïve model for a given city, which is defined in the denominator of Eq. 1. Models with $MASE < 1$ will outperform the naïve forecast. Moreover, MASE is scale-invariant, a known limitation of MAE when comparing multiple

time series with varying amplitudes, as was observed for all data and illustrated on Figure 2. It is also symmetric and robust when the predicted value $\hat{y}$ approaches 0, which is often the case in epidemiological records. The algorithm employed will take m to be the same length of the prediction window for each disease.

In order to evaluate the performance of the implemented models, we began by defining the hyperparameters of each model and using an exhaustive search method over those to define the optimal combination for the trained models, alongside employing cross-validation with four splits of the train data in each case. These parameters are listed in the following table.

TABLE II: Parameters for each regression model. For both diseases, four splits of the train set were used for cross-validation.

Algorithm	Hyperparameter	Values
RF	number of trees:	[25,50,100,150,200]
	maximum tree depth:	[2,4,None]
XGBoost	number of trees:	[25,50,100,150,200]
	maximum tree depth:	[2,4,None]
	learning rate:	[0.001, 0.005, 0.01]

It is important to note that the cross-validation model in this approach selects the best hyperparameters for each city independently from the rest of the data set, both for the direct prediction baseline and for the methodology described in Sec. II C.

The data was filtered to exclude cities with anomalies, defined as cases outside the expected distribution in the training with z-score $z > 4$ observed on the hold-out test set of each disease (see Sec. II A). Despite all diseases displaying non-normal distribution of cases, according to the mean skewness on Table I, the second moment of these distributions can still present an informed criteria according to Chebyshev's inequality, where at least 94% of data would have $\sigma = 4$ distance from the mean. As a deviation measurement, the implications to forecasting on cities with patterns that would not be seen by the models will be discussed in Sec. IV. Table III and Fig. 1 describes the precision of the evaluated regression models for time series with and without anomalies present.

C.   Feature Engineering

The main objective of this study was to evaluate the predictive performance of different approaches for selecting features from related cities. We compared the baseline model described above to models that incorporate features from cities with correlated time series, defined below.

1. Geographic proximity between cities, with the euclidean distances being calculated from IBGE's data described in Sec. II A. This allows one to generate a network of municipalities using only distances as the connection criteria, that is, the closest city to a given target will be part of its neighborhood;

2. Optimal match distance calculated through the use of Dynamic Time Warping [25] for both yearly GDP data for each city from the IBGE database (see Sec. II A), and the diseases' time series. Applying this algorithm to the latter data is done as a way to compare the patterns of contagions between cities, generating a non-informed baseline for the other two (geographic distances and GDP) selection criteria.

To quantify the optimal match distances for diseases cases and GDP data, dynamic time warping is employed as a method to evaluate which of the time series will have the minimum traversal cost to a given target. Defining such traversals as $T = ((i_1, j_1), ..., (i_t, j_t))$, with $i_1 = j_1, i_t = n, j_t = m$ for each time step $k \in 1, ..., t-1$, the algorithm will then minimize the function:

$$\sum_{k=1}^{t} d\left(p_{i_k}, q_{j_k}\right), \qquad (2)$$

where p_{i_k} and q_{j_k} are points in the curves $P = (p_1, ..., p_n)$ and $Q = (q_1, ..., q_m)$ that are being analyzed. The optimal match will be denoted as $\mathrm{DTW}(P, Q)$. Notice that Eq. 2 requires a choice of distance metric $d(.,.)$, which in this work was taken to be the Euclidean metric, where $d(x, y) = ||x - y||_p$.

For all methods, new features made from the time series of cities that fulfilled the criteria above were added to the training sets, using up to the top three cities' time series with optimal, or minimal, distances with respect to a given target.

III.   RESULTS

To first generate a baseline, the selection of optimal hyperparameters for each individual city was performed with two regression models: Random Forests and XGBoost. The resulting best model on this validation stage was then evaluated on the test set slice of the city being studied. Table III show the average MASE performance of this approach for all cities.

As shown in Table III, for Dengue and Zika the Random Forest algorithm performed best both in cities displaying anomalies or those without then in the test set; while for COVID-19 and Influenza data, XGBoost was the best overall model for cities. Nonetheless, excluding Zika, applying the models for the dataset that included anomalous cities significantly reduced the observed accuracy of the prediction.

The results shown in Figs. 1 used the best regression models from the baseline, respectively Random Forests

TABLE III: MASE performance on the train and hold-out test set of each regression framework for both diseases.

Disease	Algorithm	$z < 4$		$z \geqslant 4$	
		Train	Test	Train	Test
Dengue	**Random Forests**	0.697	**0.552**	0.694	1.148
	XGBoost	0.871	0.586	0.857	1.228
Zika	**Random Forests**	1.100	**0.604**	1.040	1.044
	XGBoost	1.287	0.703	1.199	1.134
COVID-19	**Random Forests**	0.605	**0.362**	0.605	0.423
	XGBoost	0.643	0.508	0.643	0.566
Influenza	Random Forests	0.798	1.070	0.786	1.274
	XGBoost	0.866	**1.061**	0.847	1.277

for Dengue, Zika and COVID-19, and XGBoost for Influenza, along with the expanded train sets generated through the methodology described in Sec. II C for the three different association criteria. The parameter optimization and cross-validation were executed independently from the baseline selection.

When considering the augmentation of the training set with the proposed methodology, simulations performed for Dengue and Influenza cases do not display considerable benefits of increasing the features past the initial five lags from the target time series, given that all results are within the uncertainty range and are comparable to the baseline performance, shown in Figures 1a, 1b, 1d.

COVID-19 and Zika predictions including such features shown in Fig. 1c, on the other hand, notably increases the regression effectiveness, and more evidently for the hold-out test set for the pandemic, with geographical associations resulting in the best performance for both cases. Overall, aside from Influenza, all results without including anomalies in the test set displayed higher accuracy than the seasonal naïve model.

Fig. 2 illustrates the forecasts, and also the mentioned advantages of using MASE as a evaluation metric instead of MAE, as discussed in Section II B, for selected municipalities using the best models for each disease according to the results of Fig. 1.

IV. DISCUSSION

The main proposal of this method is to verify the potential of including variables that are known to be linked to a given disease's spread dynamic, thus not only creating predictions that are robust to reporting fluctuations, but also identifying "signaling cities" and help understand the wave-like patterns of disease transmission across Brazil. The positive outcome of including disease's data from geographically close cities for Zika and COVID-19 can be interpreted as a reflection of such patterns, indicating that regional-level health policies also
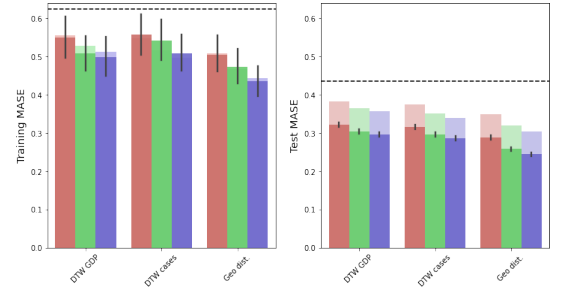
FIG. 1: MASE performance using varying number of features on the training set, with the amount represented by the numbers associated with the bar colors, for all investigated diseases. Brighter colors indicate the use of the dataset without anomalous cities ($z < 4$).
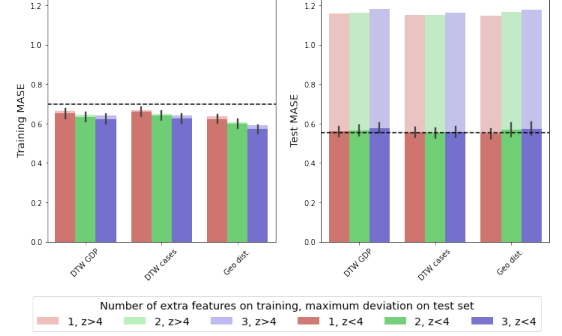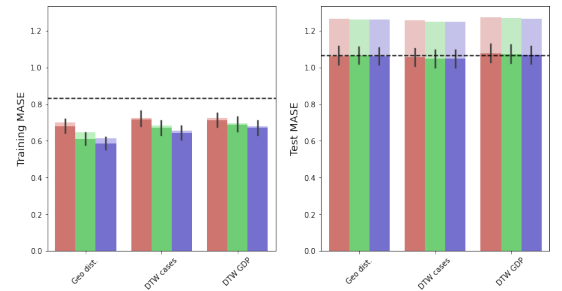

(a) Dengue forecasts using Random Forests.
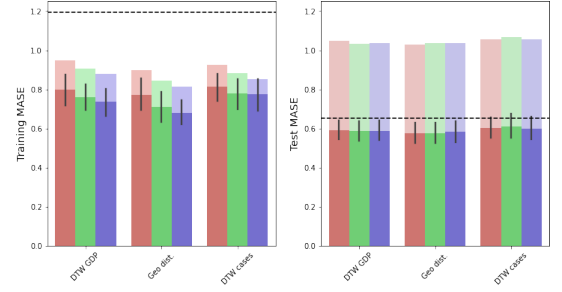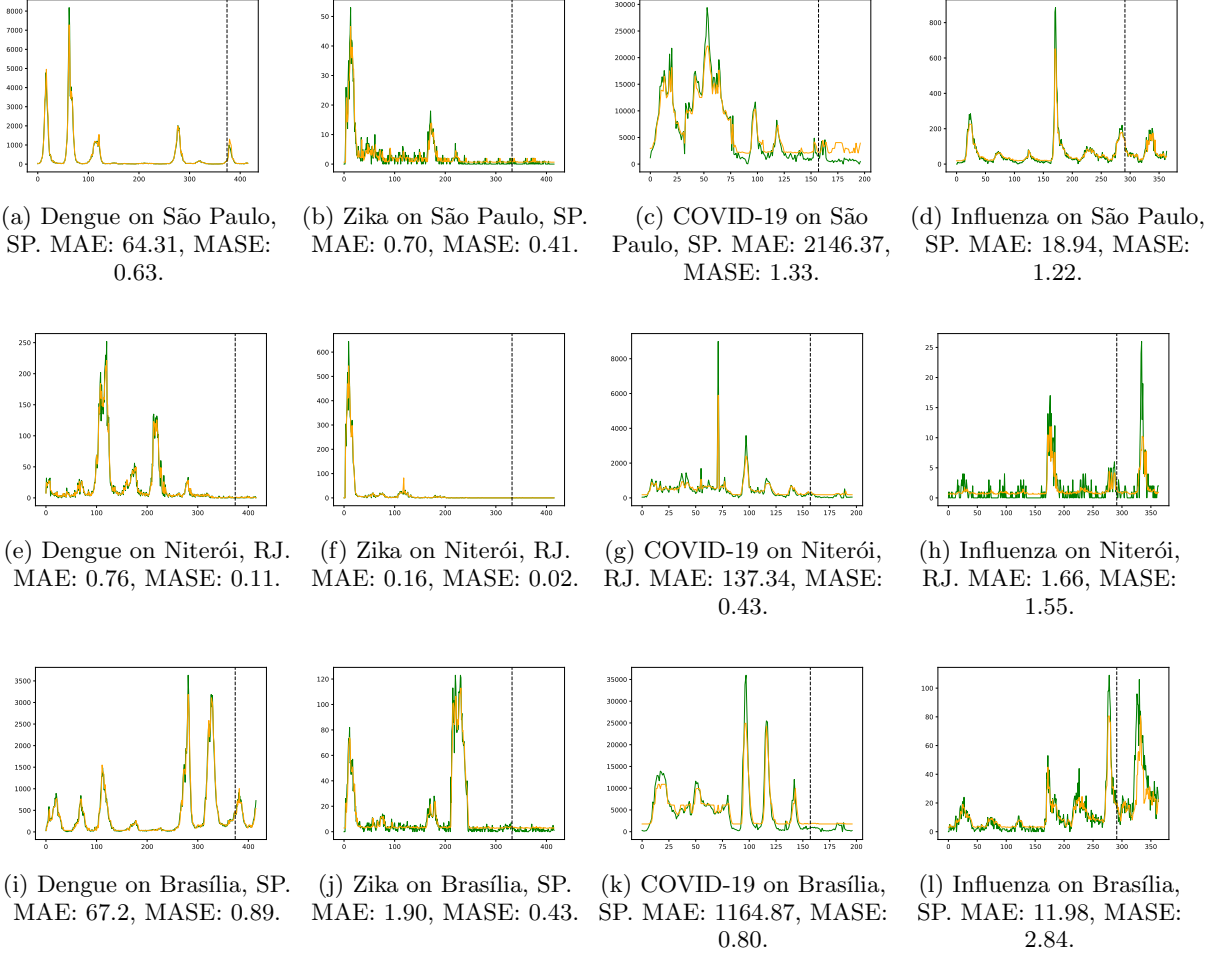

(b) Zika forecasts using Random Forests.

(c) COVID-19 forecasts using Random Forests.

(d) Influenza (pre-COVID-19) forecasts using XGBoost.

FIG. 2: Example predictions for three sample cities for both diseases, using the geographic distances as the aggregation method on the train set and including three new features from associated cities. From left to right: Dengue, Zika, COVID-19 and Influenza cases. None of these time series contain anomalous patterns in the hold-out test set. MAE and MASE results are presented for the test set.

(a) Dengue on São Paulo, SP. MAE: 64.31, MASE: 0.63.

(b) Zika on São Paulo, SP. MAE: 0.70, MASE: 0.41.

(c) COVID-19 on São Paulo, SP. MAE: 2146.37, MASE: 1.33.

(d) Influenza on São Paulo, SP. MAE: 18.94, MASE: 1.22.

(e) Dengue on Niterói, RJ. MAE: 0.76, MASE: 0.11.

(f) Zika on Niterói, RJ. MAE: 0.16, MASE: 0.02.

(g) COVID-19 on Niterói, RJ. MAE: 137.34, MASE: 0.43.

(h) Influenza on Niterói, RJ. MAE: 1.66, MASE: 1.55.

(i) Dengue on Brasília, SP. MAE: 67.2, MASE: 0.89.

(j) Zika on Brasília, SP. MAE: 1.90, MASE: 0.43.

(k) COVID-19 on Brasília, SP. MAE: 1164.87, MASE: 0.80.

(l) Influenza on Brasília, SP. MAE: 11.98, MASE: 2.84.

can be effective in the containment of those outbreaks.

It is important to also note the limitations of this study. The official reports on dengue, for example, could be instead a mislabel of Zika or Chikungunya, that shares some symptoms with Dengue, as shown by [15]. Furthermore, the predictions for Dengue and Influenza were not substantially enhanced through this method, which could imply that the variables selected are less impactful for these endemic diseases, or the trends included in the correlated time series does not contain information that can positively influence the model.

As for the use of GDP as the selection criteria for the training set, none of the considered diseases benefited significantly from including this metric. This implies that other, more specific measurements of social and economic aspects should be used instead. Chan et al. [18] demonstrated the effectiveness of using health and infrastructure indicators to the prediction of outbreaks in multiple countries, including Brazil.

Moreover, most common regression models, including decision trees, learn from seasonal patterns and trends, and as such would not be able to predict significant deviations from the training data should such pattern emerge in future measurements, as demonstrated by the notable difference in evaluating the performance on target series within or without the z-score threshold.

V. CONCLUSIONS

In this work, the predictive performance of machine learning models that incorporate information from related cities was evaluated by comparing three methods for selecting related cities, through similarity in geography, GDP and seasonal patterns in the data. We implemented these methods for four different diseases in Brazil. COVID-19 and Zika predictions improved when enriching the training data with features from geographically

proximate cities, while dengue and influenza forecasts did not benefit significantly from the same procedure. Moreover, forecasting is improved when applying the models for data that does not include unseen variations on the test set, with better performance than baseline in those cases. These results suggest that predictive models incorporating information from related cities can help infectious disease forecasts and create more robust early warning systems for public health departments.

Expansions of this work could be done in multiple ways. First, data describing the travel flux between cities could help clarify the association of distances with the spreading of diseases, and consequently the impact observed of the use of this property in the prediction of COVID-19. Moreover, other indicators could be applied to this methodology, such as the Gini coefficient, the existence and investment levels in public health measures and sanitary services along with their coverage in a given city, hospitalization and mortality rates of the disease under study, and also other structural indicators, such as communications networks. It also could be useful to consider climate variables, specially for diseases transmitted through non-human vectors, as done in [26].

An alternative method that could be used to reduce the influence of outliers in the modeling process would be to use a *cellwise robust* filtering task, where it would flag cells in the data matrix as outlying and down-weight the influence of outliers, such as proposed on [27]. In the case of large datasets complying to sparsity requirements, a recent work could also provide further insights to the modeling process [28].

Furthermore, the explainability of a acute outbreak would require causal inference of multiple factors that goes beyond the scope of this project; while a halt, or a markedly decrease, of cases may be related to the implementation of lock-downs and other containment measures and should be taken into account when asserting the accuracy of predictions. In this context, the proposed methodology will then be effective in scenarios where the epidemic does not include such variations, unless those changes in the disease's pattern can be explained by known data. These models will then provide useful insights into the diseases dynamics by employing variables known to be linked to it that also improve forecastings, which in turn could contribute as an additional information source for public health decision-making.

CODE AVAILABILITY

All data and code used in this work are publicly available at `https://github.com/luizalober/epidemics-using-features`.

[1] N. Zhao, K. Charland, M. Carabali, E. O. Nsoesie, M. Maheu-Giroux, E. Rees, M. Yuan, C. G. Balaguera, G. J. Ramirez, and K. Zinszer, Machine learning and dengue forecasting: Comparing random forests and artificial neural networks for predicting dengue burden at national and sub-national scales in Colombia, PLoS Negl.Trop. Dis. **14**, e0008056 (2020).

[2] I. Rahimi, F. Chen, and A. H. Gandomi, A review on COVID-19 forecasting models, Neural Comput. &. Applic. **35**, 23671 (2023).

[3] M. da Saúde do Brasil, Datasus (2023), `https://datasus.saude.gov.br`.

[4] A. Dutta, COVID-19 waves: variant dynamics and control, Sci. Rep. **12**, 1 (2022).

[5] M. Andraud, N. Hens, C. Marais, and P. Beutels, Dynamic Epidemiological Models for Dengue Transmission: A Systematic Review of Structural Approaches, PLoS One **7**, e49085 (2012).

[6] F. M. R. Hirata, D. C. P. Jorge, F. A. C. Pereira, L. M. Skalinski, G. Cruz-Pacheco, M. L. M. Esteva, and S. T. R. Pinho, Co-circulation of Dengue and Zika viruses: A modelling approach applied to epidemics data, Chaos, Solitons Fractals **173**, 113599 (2023).

[7] K. Roster, C. Connaughton, and F. A. Rodrigues, Machine-Learning–Based Forecasting of Dengue Fever in Brazilian Cities Using Epidemiologic and Meteorological Variables, Am. J. Epidemiol. **191**, 1803 (2022).

[8] K. L. Ebi and J. Nealon, Dengue in a changing climate, Environ. Res. **151**, 115 (2016).

[9] Z. Xu, H. Bambrick, F. D. Frentiu, G. Devine, L. Yakob, G. Williams, and W. Hu, Projecting the future of dengue under climate change scenarios: Progress, uncertainties and research needs, PLoS Negl.Trop. Dis. **14**, e0008118 (2020).

[10] K. R. Moran, G. Fairchild, N. Generous, K. Hickmann, D. Osthus, R. Priedhorsky, J. Hyman, and S. Y. Del Valle, Epidemic Forecasting is Messier Than Weather Forecasting: The Role of Human Behavior and Internet Data Streams in Epidemic Forecast, J. Infect. Dis. **214**, S404 (2016).

[11] D. S. Shepard, L. Coudeville, Y. A. Halasa, B. Zambrano, and G. H. Dayan, Economic Impact of Dengue Illness in the Americas, Am. J. Trop. Med. Hyg. **84**, 200 (2011).

[12] The World Health Organization, Dengue – the region of the americas (2023), [Online; accessed 31. Oct. 2023].

[13] P. Basak, T. Abir, A. Al Mamun, N. R. Zainol, M. Khanam, Md. R. Haque, A. H. Milton, and K. E. Agho, A Global Study on the Correlates of Gross Domestic Product (GDP) and COVID-19 Vaccine Distribution, Vaccines **10**, 266 (2022).

[14] B. Y. An, S. Porcher, S.-Y. Tang, and E. E. Kim, Policy Design for COVID-19: Worldwide Evidence on the Efficacies of Early Mask Mandates and Other Policy Interventions, Public Adm. Rev. **81**, 1157 (2021).

[15] R. Pessôa, J. V. Patriota, M. de Lourdes de Souza, A. C. Félix, N. Mamede, and S. S. Sanabani, Investigation into an outbreak of dengue-like illness in pernambuco, brazil, revealed a cocirculation of zika, chikungunya, and dengue virus type 1, Medicine **95** (2016).

[16] T. F. Menkir, H. Cox, C. Poirier, M. Saul, S. Jones-Weekes, C. Clementson, P. M. de Salazar, M. Santillana, and C. O. Buckee, A nowcasting framework for correcting for reporting delays in malaria surveillance, PLoS Comput. Biol. **17**, e1009570 (2021).

[17] S. A. Lee, T. Economou, R. de Castro Catão, C. Barcellos, and R. Lowe, The impact of climate suitability, urbanisation, and connectivity on the expansion of dengue in 21st century Brazil, PLoS Negl.Trop. Dis. **15**, e0009773 (2021).

[18] E. H. Chan, D. A. Scales, T. F. Brewer, L. C. Madoff, M. P. Pollack, A. G. Hoen, T. Choden, and J. S. Brownstein, Forecasting High-Priority Infectious Disease Surveillance Regions: A Socioeconomic Model, Clin. Infect. Dis. **56**, 517 (2013).

[19] R. Jain, S. Sontisirikit, S. Iamsirithaworn, and H. Prendinger, Prediction of dengue outbreaks based on disease surveillance, meteorological and socio-economic data, BMC Infect. Dis. **19**, 1 (2019).

[20] Covid-19 Casos e Óbitos (2023), [Online; accessed 27. Oct. 2023].

[21] Instituto Brasileiro de Geografia e Estatística, Ibge: Portal do ibge (2023), `https://www.ibge.gov.br/pt/inicio.html`.

[22] L. Breiman, Random Forests, Machine Learning **45**, 5 (2001).

[23] T. Chen and C. Guestrin, XGBoost: A Scalable Tree Boosting System, in *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, New York, NY, USA, 2016) pp. 785–794.

[24] R. J. Hyndman, Another look at measures of forecast accuracy, FORESIGHT , 46 (2006).

[25] N. L. Olsen, B. Markussen, and L. L. Raket, Simultaneous Inference for Misaligned Multivariate Functional Data, Journal of the Royal Statistical Society Series C: Applied Statistics **67**, 1147 (2018), https://academic.oup.com/jrsssc/article-pdf/67/5/1147/49337456/jrsssc_67_5_1147.pdf.

[26] L. M. Stolerman, P. D. Maia, and J. N. Kutz, Forecasting dengue fever in Brazil: An assessment of climate conditions, PLoS One **14**, e0220106 (2019).

[27] F. Alqallaf, S. Van Aelst, V. J. Yohai, and R. H. Zamar, Propagation of outliers in multivariate data, Ann. Stat. **37**, 311 (2009).

[28] L. Bottmer, C. Croux, and I. Wilms, Sparse regression for large data sets with outliers, Eur. J. Oper. Res. **297**, 782 (2022).