Forecasting of Zika space-time spread with topological data analysis

Marwah Soliman[1], Vyacheslav Lyubchich[2], and Yulia R. Gel*[1]

[1]Department of Mathematical Sciences, University of Texas at Dallas, Richardson, TX, USA
[2]Chesapeake Biological Laboratory, University of Maryland Center for Environmental Science, Solomons, MD, USA

Abstract

The first confirmed cases of Zika virus infection in Brazil were registered in 2015, causing an epidemic with 440,000 to 1,300,000 people infected during 2015, as reported by the World Health Organization. In addition, the disease has spread rapidly to many other countries in South America, North America, and East Asia. Zika virus is transmitted among humans through the bite of an infected mosquito of Aedes species (*Aedes aegypti* and *Aedes albopictus*). Mosquito abundance and cases of Zika are prevalent in areas with high temperature, high precipitation, and high population density. In this project, we introduce topological data analysis in conjunction with three machine learning models – random forest, generalized boosting regression, and back propagation neural network – to forecast Zika space-time spread in Brazil in year 2018. The topological data analysis is implemented by utilizing persistent homology features through cumulative Betti numbers.

1 Introduction

Zika virus was first identified in Uganda in 1947 from the serum of a rhesus monkey (Dick et al, 1952; Dick, 1952). Zika belongs to the family of *Flaviviridae*; other flaviviruses are dengue virus, West Nile virus, yellow fever virus, Japanese encephalitis virus (Goeijenbier et al, 2016). During 2015, Zika disease reached an estimated 440,000–1,300,000 cases in Brazil (Heukelbach et al, 2016; Malone et al, 2016) and has spread from Brazil to other South American, North American and East Asian countries. Zika was declared an epidemic disease during the years of 2015 and 2016 by the World Health Organization (WHO, 2016).

Zika is spread through the bite of an infected Aedes species mosquito, *Aedes aegypti* and *Aedes albopictus* (CDC, 2018). The symptoms include fever, rash, headache, joint pain,

*Corresponding author. E-mail: ygl@utdallas.edu

red eyes, and muscle pain. Infection of pregnant women can lead to certain birth defects. Weather conditions favorable for the spread of mosquitoes, such as high temperature and high precipitation, may facilitate the spread of Zika (Tjaden et al, 2013; Rees et al, 2018; Muñoz et al, 2017).

Several studies were performed on Zika prediction and modelling. Jiang et al (2018) predicted Zika cases and mapped the probability of Zika outbreak on a global scale using three machine learning models, namely, back propagation neural network (BPNN), gradient boosting machine (GBM), and random forest (RF). Based on the area under the curve (AUC), the predictive ability of BPNN was the best among the three models. Box–Jenkins time series modeling was used by McGough et al (2017), where an autoregressive model with exogenous predictors was applied to predict Zika rate up to three weeks ahead in Colombia, El Salvador, Honduras, Venezuela, and Martinique. The exogenous predictors included Google trends, data from Twitter microblogs, HealthMap digital surveillance system, and historical official cases counts. Teng et al (2017) used autoregressive integrated moving average (ARIMA) model for predicting Zika cases worldwide with an exogenous predictor being the Google search activity for the keyword "Zika" from 12 February 2016 to 9 November 2016 (epidemiological weeks 6 to 45). Pearson product-moment correlation was used as a measure of temporal correlation between the accumulative volumes of Zika-related search queries and the cumulative numbers of reported cases.

In addition, mathematical models were used in the following studies. Suparit et al (2018) developed a vector-borne compartmental model to analyze the spread of Zika virus during the 2015–2016 outbreaks in Bahia, Brazil. Muñoz et al (2017) applied a two-vector one-host susceptible-infected-removed (SIR) model to forecast the areas in South America with high Zika rates. The analysis facilitated the discussion on how climate factors like temperature and rainfall might affect the transmission of Zika, chikungunya, and dengue disease. The results showed high spread rate in the northeastern Brazil, which is the epicenter of the epidemic. Mordecai et al (2017) used mechanistic transmission models to find how the probability and magnitude of transmission for Zika, chikungunya, and dengue change with average air temperature. The study concluded that the maximal transmission occurs in the temperature range of 26–29 °C, controlling other factors such as population size and socioeconomic factors.

In our project we introduce the use of topological data analysis (TDA). The fundamental advantage of TDA over the traditional models referenced above is that topological models use spatial information in a computationally efficient manner. For example, areas with greater amounts of precipitation are generally associated with higher mosquito abundance (Lo and Park, 2018; Muñoz et al, 2017). TDA explores the structure of the precipitation data by utilizing computational geometry and topology that provides us with information about the data shape useful for predicting Zika rates. In addition, applying TDA in bio-surveillance projects is still a new approach.

Lo and Park (2018) used persistent homology features such as H0, H1, and maximum H1 lifetime of Zika population locations in each state in Brazil as predictors. In addition, population density and average temperature were used as predictors in a linear regression model. The results showed an improvement of the model performance compared to the

model without persistent features through higher adjusted R^2 and lower leave-p-out cross-validation mean squared error. Costa and Škraba (2014) applied persistent homology to compare time series of influenza-like illness (ILI) patients in Portugal and Italy for the flu seasons of 2008–2013. Torres et al (2016) used Ayasdi platform to create topological models that cluster data. In a similar manner, Siddiqui et al (2018) used TDA for data clustering.

We aim to predict the Zika rate in each of the 26 states in Brazil in January–August 2018, by incorporating TDA through using persistent homology features of precipitation in year 2017 in conjunction with three machine learning models: random forest, generalized boosting regression, and back propagation neural network.

Our contribution is as follows: we introduce cumulative Betti numbers of the Vietoris–Rips filtration of the precipitation point cloud in year 2017 in each state as predictors in the models. Cumulative Betti 0 numbers is considered a robust method. We also calculate the number of H0 features and use it as a predictor. In addition, we use flexible machine learning models for prediction, not only linear regression as in Lo and Park (2018). Meteorological factors (average annual temperature and average annual precipitation) and socioeconomic factor (population density) were used as the exogenous predictors in our three machine learning models.

The remainder of the paper is organized in five major sections: data description, methodology for epidemiological forecasting, validation metrics, results, and discussion. In Section 2, we provide information on the collected Zika rate, population density, and atmospheric data. We introduce our statistical and machine learning forecasting methodology and topological data analysis in Section 3. Section 4 lists the validation metrics, Section 5 is devoted to validation of the proposed modeling approaches to prediction of Zika rate in Brazil. Finally, the paper is concluded with discussion and future work in Section 6.

2 Data description

Data on the Zika rate per 100,000 inhabitants from January 2017 up to August 2018 were obtained for each of the 26 states of Brazil from monthly reports published by the Ministry of Health of Brazil (MHB, 2018).

Since Zika virus transmission is associated with air temperature and precipitation (e.g., see Tjaden et al, 2013; Rees et al, 2018; Muñoz et al, 2017), we collected monthly data of air temperature, precipitation and found the average per year in each state. The data were collected from World-Weather-Online (2018) in each state for January 2017–August 2018.

Also, Zika is anticipated to spread easily in highly populated areas (Lo and Park, 2018; Gardner et al, 2018). Thus, we added the population density of each state as a predictor. We used World-Atlas (2018) to calculate the population density, through dividing the total population in each state, according to 2014 census, by its area (km^2).

3 Methodology for epidemiological forecasting

3.1 Topological data analysis

Topological data analysis (TDA) is a rapidly emerging methodology which appeared in the works of Edelsbrunner et al (2000), Zomorodian and Carlsson (2005), and Carlsson (2009) on persistent homology. TDA explores the structure of the data by utilizing computational geometry and topology that provide us with information about the data shape at multiple resolution levels in Euclidean space or other metric spaces (Chazal and Michel, 2017).

To start TDA, we build a simplicial complex on top of the point cloud data. Simplicial complexes is the basis in investigating the shape of the data and they are mathematically and computationally tractable (Singh et al, 2007; Carlsson, 2009). Next are the definitions of basic topological concepts that we use in our analysis. Specifically, we use Vietoris–Rips (VR) filtration with Euclidean metric.

Definition 1 (Vietoris–Rips complex) *For a point cloud P with dimension m as a subset of $\mathbb{R}^m$, the Vietoris–Rips complex $V_\epsilon(P)$ (the VR complex over the point cloud P with scale ϵ) is defined as $V_\epsilon(P) = \{\sigma \subseteq P | d(q,p) \leq \epsilon,\ \forall p \neq q \in \sigma\}$.*

The definition implies that a VR complex at ϵ consists of all subsets σ in P such that the pairwise distance of any non-identical points in σ is less than or equal to ϵ.

Definition 2 (k-simplex) *A k-simplex is the convex hull of k+1 points in Euclidean space, where a set of points is said to be convex if it has line segments connecting each pair of points. So the convex hull is the intersection of all convex sets containing k+1 points. A vertex is defined as 0-simplex, edge as 1-simplex, triangle as 2-simplex, and tetrahedron as 3-simplex.*

Using the simplex definition we define the simplicial complex which is a collection of simplices, to look at the shape of the data.

Definition 3 (Simplicial complexes) *Let σ, τ, and δ be simplices, then a simplicial complex is a finite collection of simplices K such that*
1) $\sigma \in K$ and $\tau \subset \sigma$ implies $\tau \in K$, and
2) $\sigma, \delta \in K$ implies $\sigma \cap \delta$ is either empty or a common face of both.

We build the simplexes on the top of the point clouds, then we extract the topological features of the data like connected components and holes. H0 represent the connected components and H1 is the holes. In this project our point cloud dimension is in $\mathbb{R}$.

Definition 4 (Betti number) *The k-th Betti number is the rank of the k-th homology group of a topological space. The β_k counts the number of k-dimensional features of a simplicial complex. For example, β_0 gives the number of connected components, β_1 gives the number of holes, and β_2 gives the number of voids.*

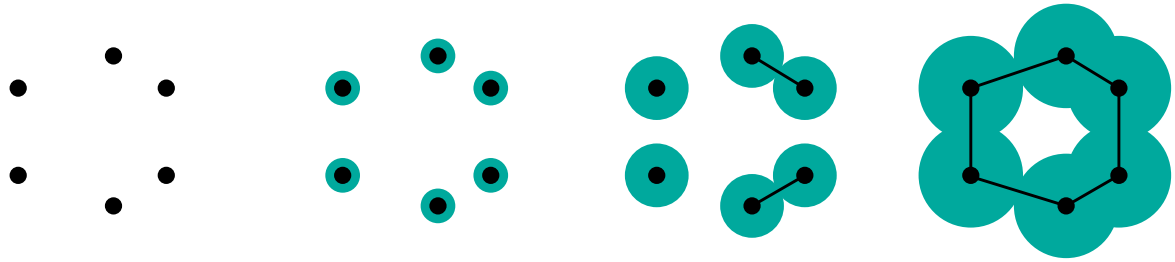
We have only connected components, thus we calculated the β_0 numbers for varying ϵ, which gives us the number of connected components at each of these ϵ values.

In our project we introduce cumulative Betti numbers, which is a more robust method for generating predictors in Zika rates forecasting models.

Definition 5 (Cumulative Betti numbers) *The sum of Betti numbers β_k up to a specified ϵ value. For n values of ϵ, the n cumulative k-th Betti numbers are $\sum_{i=1}^{j} \beta_k(\epsilon_i)$, where $j = 1, \ldots, n$.*

Now we explain how the Vietoris–Rips filtration is performed. The filtration starts with a ball of radius 0 that is called ϵ-ball around each point. As we increase ϵ, the balls grow bigger which allows the balls to intersect and form a simplicial complex for each ϵ value. A simplicial complex for a specified value of ϵ is a subset of the simplicial complexes of larger ϵ.

Figure 1 is an illustration of a toy example of Vietoris–Rips filtration with increase of ϵ. Figure 1(a) is the point cloud representing 0-simplices with $\epsilon = 0$. With further increase of ϵ to 0.2 (Figure 1(b)) and to 0.5 we get connected components as 1-simplices (Figure 1(c)). The increase of ϵ up to 1 in Figure 1(d) gives a loop.



(a) $\epsilon = 0$ (b) $\epsilon = 0.2$ (c) $\epsilon = 0.5$ (d) $\epsilon = 1$

Figure 1: An illustration of Vietoris–Rips filtration with an increase of ϵ from 0 and a disconnected point cloud in (a); to $\epsilon = 0.5$ and connected components in (c); to $\epsilon = 1$ and a loop in (d).

Cumulative Betti 0 numbers aggregate information over different values of ϵ, thus, this cumulative feature is robust to selection of the threshold. In Section 5, we show that cumulative Betti 0 numbers improve prediction performance of our models. The idea of cumulative Betti numbers came from that each specified ϵ simplicial complex is a subset of the larger ϵ simplicial complex. Figure 2 shows the cumulative Betti 0 of the state of Piaui.

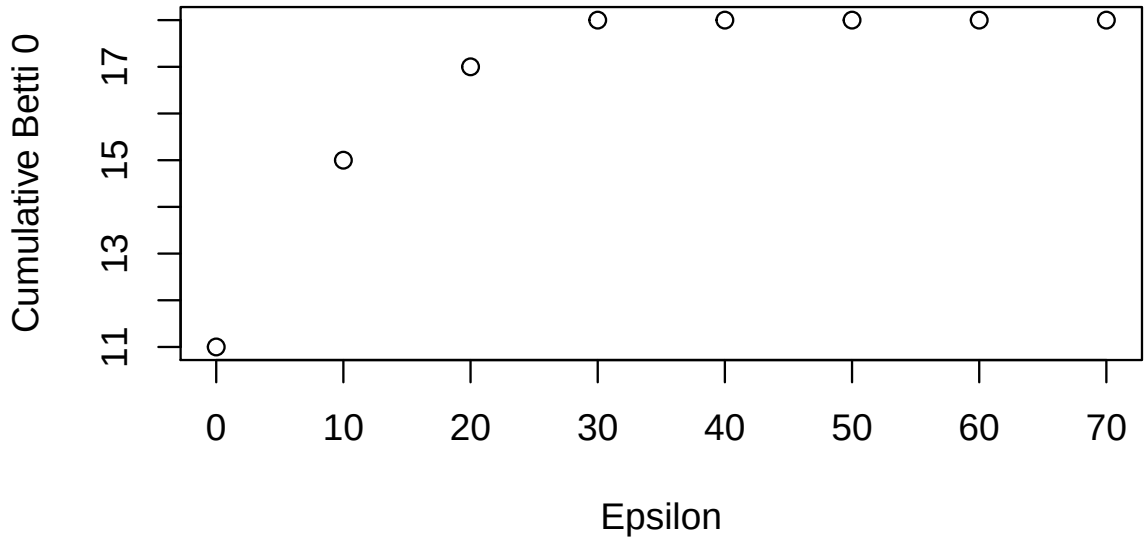


Figure 2: Cumulative Betti 0 of the state of Piaui.

3.2 Models

We applied three machine learning models to our prediction task. In addition, Bayesian model averaging (BMA) was applied to combine predictions from the different models.

Random forest (RF) we used `randomForest` function in R package "randomForest" (Breiman et al, 2018).

Generalized boosting regression model (GBM) The model is built on two techniques: decision tree algorithms and boosting methods (Breiman, 1997; Friedman, 2001, 2002). Generalized boosting models repeatedly fit many decision trees to improve the accuracy of the model.

Gradient boosting at the m-th step would fit a decision tree $h_m(x)$ to pseudo-residuals. The tree partitions the input space X into J_m disjoint regions $R_{1_m}, \ldots, R_{J_m}$ and predicts a constant value in each region

$$\sum_{i=1}^{J_m} b_{j_m} I(x \in R_{j_m})$$

where b_{j_m} is the value predicted in the region R_{j_m}. Then the update will be

$$F_m(x) = F_{m-1}(x) + \rho_m \sum_{i=1}^{J} b_{jm} \mathbf{1}_{R_{jm}(x)}$$

where ρ_m is a scaling factor is the solution to the "line search". The b_{jm} are the corresponding least-squares coefficients, Here $\{R_{jm}\}_1^J$are the regions defined by the terminal nodes of the tree at the m-th iteration.

Let γ_{jm}= $b_{jm}\rho_m$, Friedman proposes to choose a separate optimal value γ_{jm} for each of the tree's regions, instead of a single γ_m for the whole tree. The model update rule becomes:

$$F_m(x) = F_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} \mathbf{1}_{R_{jm}(x),}$$

$$\gamma_{jm} = \arg\min_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, F_{m-1}(x_i) + \gamma).$$

Which is the optimal constant update in each terminal node region, based on the loss function L, given the current approximation $F_{m-1}(x)$.

See Ridgeway (2019); Friedman (2001) for more details of the boosting algorithm. We have used the `gbm` function in R in package "gbm" for the generalized boosting regression model (Greenwell et al, 2018; Ridgeway, 2019). Since the Zika rate is continuous, we used Gaussian distribution with number of trees in the model being 100 and interaction depth of 1 selected using cross-validation. The regressors included in the model are: the population density, temperature in 2017, precipitation in 2017, number of H0 features at the initial filtration, and the cumulative Betti 0 numbers at $\epsilon = (10, 20, 30, 40, 50, 60, 70)$.

Back propagation neural network (BPNN) BPNN algorithm performs as follows the data moves from input to hidden layers to output then calculate the error based on the cost function which reflect how far the output with respect to the actual data. Our goal is to minimize the error.

Thus the BPNN algorithm recalculate the weights for the layers starting from the first layer which will eventually change the weights for the other layers then calculate the error each time and keep repeating this process to finally reach optimization.

let w_{jk}^l be the weight for the connection from the kth neuron in the (l-1)th layer to the jth neuron in the lth, if we done a change in the weights Δw_{jk}^l, then that change in weight will cause a change in the output activation from the corresponding neuron Δa_j^l and this will change in all the activations in all next layers ending to a change in the cost function which will change the error.

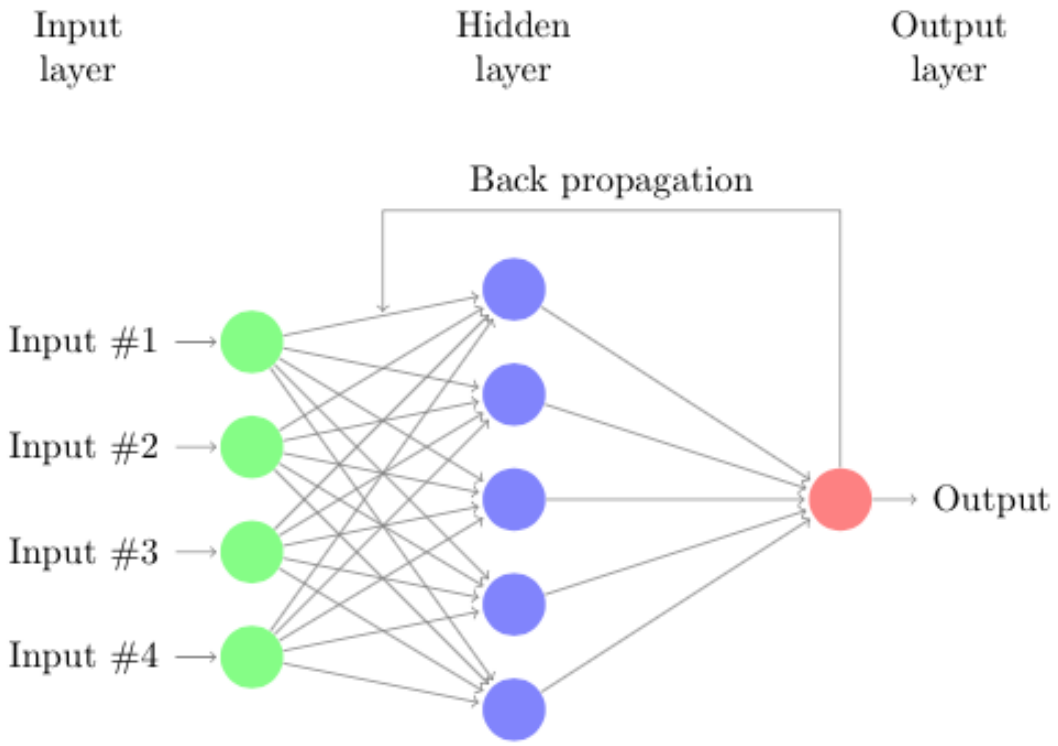


Figure 3: Example of BPNN architecture

The input are the population density, temperature of 2017, precipitation of 2017, number of H0 features at the initial filtration and the cumulative Betti 0 numbers at epsilon values (10, 20, 30, 40, 50, 60, 70). We trained BPNNs using R package using the function neuralnet in R in package "neuralnet" Fritsch and Guenther (2016). The optimal BPNN structure selected using cross-validation. The number of hidden nodes is 5, the number of hidden layers is 2, and the learning rate is 0.01.

Bayesian model averaging (BMA) To evaluate prediction uncertainty we used a weighted multi-model ensemble of future epidemiological cases. That is, we applied the Bayesian model averaging (BMA), which allows us to combine multiple models with weights corresponding to their most recent prediction performance.

Our BMA approach is to define model weights via root mean square error (RMSE) of fitted $\tilde{y}_t$ by each model on the training set of data y_t $(t = 1, \ldots, n)$, i.e.,

$$RMSE(k) = \sqrt{\frac{\sum_{i=1}^{n}(y_t - \tilde{y}_t)^2}{n}},$$

where n is the size of training data set; RMSE(k) corresponds to the k-th model, namely, for random forest $k = 1$; for Generalized boosting regression $k = 2$; for Back propagation

neural network $k = 3$. The resulting RMSE(k) leads to the following corresponding weight of the k-th model in BMA:

$$\frac{1/RMSE(k)}{\sum_{i=1}^{3} 1/RMSE(i)}.$$

Let the predicted values from each model be $Pred(k)$ for $k = 1, 2, 3$. Then the BMA forecast with RMSE as weights is given by

$$Pred_{RMSE} = \sum_{k=1}^{3} \frac{1/RMSE(k)}{\sum_{i=1}^{3} 1/RMSE(i)} Pred(k).$$

3.3 Persistent homology features

In addition, we use topological data analysis in terms of persistent homology (Edelsbrunner and Harer, 2010; Chazal and Michel, 2017). Vietoris-Rips filtration is used to build a (simplicial complexes) using euclidean distance, from which we can extract topological information from our point cloud data. First, we determined the locations of weather stations in each state, then the precipitation level in these location is calculated. Thus the point cloud is precipitation level at each weather location within each state.

We calculated the number of H0 features at the initial of the filtration for each state using the function ripsDiag in the R package "TDA" Fasy et al (2018). In addition, we calculated the cumulative Betti 0 numbers in each state using the same function in R by plotting the barcode as follows, we divided the barcode for each state into intervals of ϵ values from 0 to 70 by an increment of 10 so it is ($\leq$0,$\leq$10, $\leq$20,...,$\leq$70), in each interval we calculated the cumulative Betti 0 numbers. Where Betti 0 numbers is the number of horizontal lines in the barcode that intersect with the vertical line from each ϵ value, this Betti 0 number represents the number of connected components at this filtration. For example, state Piaui has the below barcode

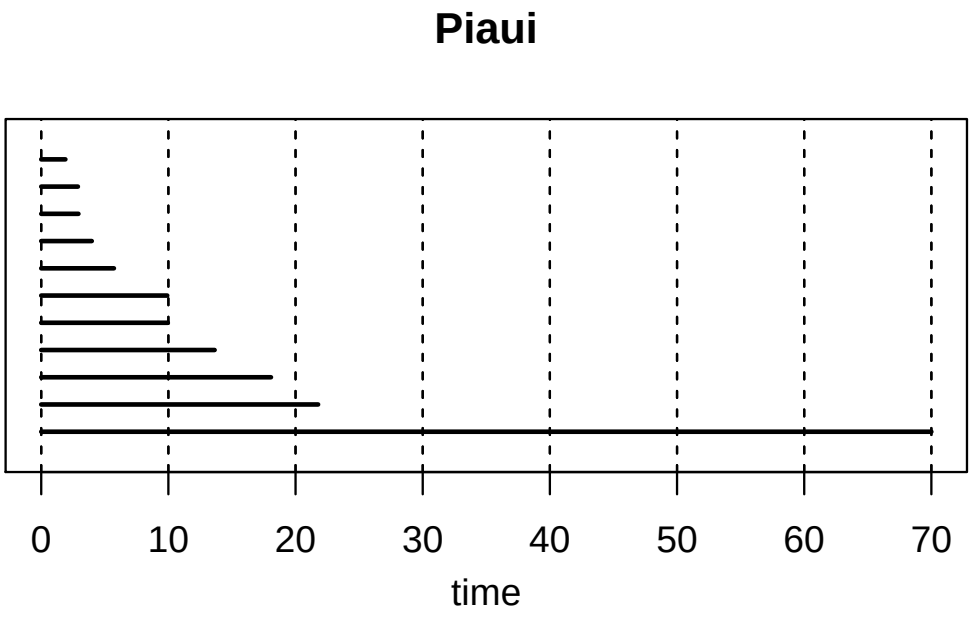


Figure 4: Barcode of H0 features for state of Piaui

In the interval (≤ 0) is 11 Betti 0 numbers, ($\leq$10) is 15 Betti 0 number, in the interval ($\leq$20) we have 17 Betti 0 number, in the interval ($\leq$30) we have 18 and the the rest of intervals ($\leq$40), ($\leq$50), ($\leq$60), and ($\leq$70) is also 18.

4 Validation metrics

We use an adaptive form of out of sampling forecast where the model is trained using data of the year 2017. The data is year 2018 up to August is used for an out-of-sample forecast validation.We use two standard statistical measures of accuracy: root mean square error (RSME), and mean absolute error (MAE) (Liang et al, 2018; Lu et al, 2018). Let $\hat{y}_i$ be the forecast value and y_i be the corresponding real values, then

$$\begin{aligned} RMSE &= \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}, \\ MAE &= \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right|. \end{aligned}$$

In addition, we measured the correlation between the observed and forecast values using Pearson correlation coefficient

$$r(y, \hat{y}) = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

Lower RMSE, MAE and higher Pearson correlation coefficient imply better forecasting performance.

5 Results

In our models, we used logarithmic transformation of Zika rate as a response variable since it produce normally distributed normal probability plot to the residuals of the generalized boosting regression as shown in Figure 5. Our predictors are population density, average temperature, average precipitation, number of H0 features of precipitation in year the 2017 and the cumulative Betti 0 numbers.

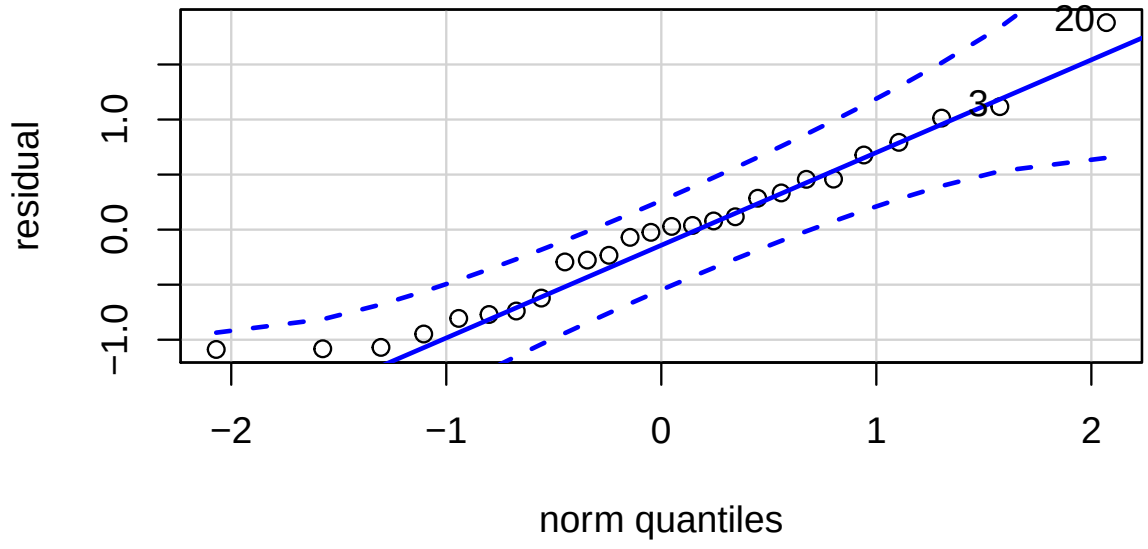


Figure 5: Normal probability plot of residual in Boosting regression model

Table 1: Performance summary comparison for the four models

	With persistent features		Without persistent features	
Method	RMSE	MAE	RMSE	MAE
Randomforest	1.216	0.995	1.262	1.032
Boosting regression	1.287	1.065	1.327	1.089
BPNN	1.617	1.323	1.948	1.560
BMA-RMSE	1.175	0.902	1.310	1.069

Table 2: Pearson correlation coefficient between predicted and actual of 2018 Zika rate with and without persistent features models

	Cor(with persistent features)	Cor(without persistent features)
Randomforest	0.759	0.738
Boosting regression	0.665	0.618
BPNN	0.549	0.318
BMA-RMSE	0.710	0.641

We compared the predicted values from each model to the log-transformed Zika rate in 2018 up to August. From Tables 1 and 2 The models with the persistent features performed better compared to the models without the persistent features. BMA-RMSE with persistent features performed the best in predicting Zika rate of year 2018 with the lowest RMSE and MAE. The next is Random forest model.

The model of BMA-RMSE improved by 9.4%, next the model of random forest improved by 3.6%, Generalized boosting regression with 3% improvement and then Back propagation neural network with 8.9% improvement compared to the model without persistent features.

Comparing with Jiang et al (2018) who used the same three models, their conclusion was that BPNN has the best result in prediction. That is because their study has larger simulated sample size 1224 as Jiang et al (2018) prediction was on a global level.

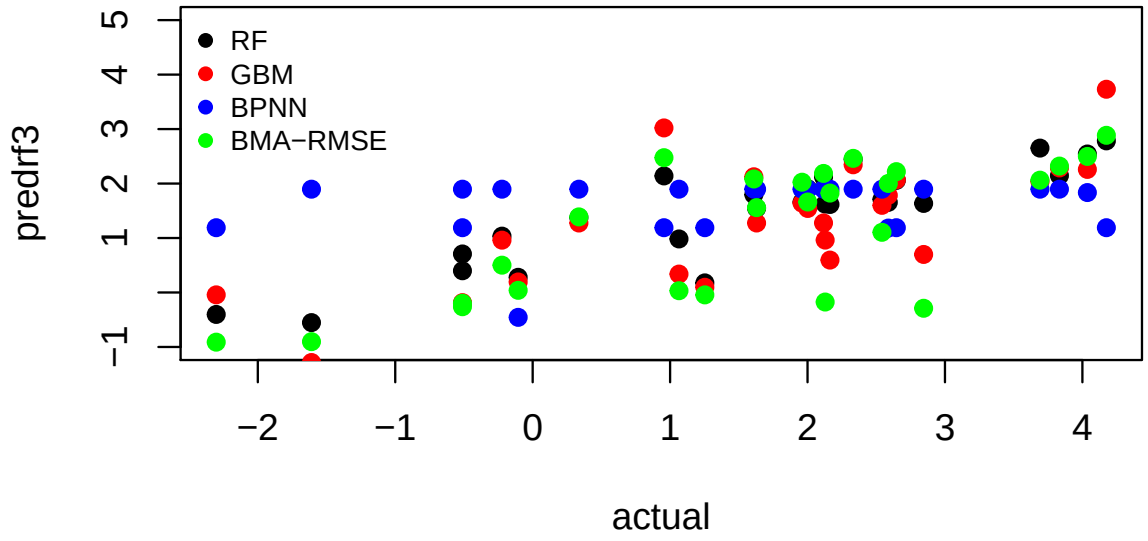


Figure 6: Plot of predicted values from RF, GBM, BPNN vs. actual Zika rate 2018

We can see from Figure 6 that the BMA-RMSE performs the best in forecasting the log Zika rate in 2018, next is Random forest.

6 Discussion

In this paper, we wanted to test the ability of persistent homology features in improving the prediction of Zika virus in the states of Brazil by using the H0 persistent features of precipitation and cumulative Betti 0 numbers.

High levels of precipitation can be a factor in increasing the mosquito population, we investigate the prediction performance using persistent features with three machine learning models they are random forest, generalized boosting regression and back propagation neural network. In addition to using Bayesian model average (BMA) with RMSE as weights to account for uncertainty.

The results shows that BMA-RMSE with persistent features cumulative Betti 0 numbers and number of H0 in the predictors performed the best in predicting the Zika rate in year 2018 up to August. Compared to the model with only population density, average temperature and average precipitation. Next best is random forest model.

Future work can be done on studying other climate-sensitive epidemic disease like chikungunya, and dengue in Brazil and other countries affected by such diseases and use the topo-

logical data analysis in studying the spread of the diseases by looking at H0 and H1 features.

References

Breiman L (1997) Arcing the edge. Technical Report 486 Statistics Department, University of California, Berkeley

Breiman L, Cutler A, Liaw A, Wiener M (2018) randomForest: Breiman and Cutler's Random Forests for Classification and Regression. URL https://CRAN.R-project.org/package=randomForest, R package version 4.6-14

Carlsson G (2009) Topology and data. Bulletin of the American Mathematical Society 46(2):255–308

CDC (2018) Centers for Disease Control and Prevention (overview of influenza surveillance in the United States). URL https://www.cdc.gov/zika/about/index.html

Chazal F, Michel B (2017) An introduction to topological data analysis: Fundamental and practical aspects for data scientists. arXiv Preprint arXiv:171004019

Costa JP, Škraba P (2014) A topological data analysis approach to epidemiology. In: European Conference of Complexity Science

Dick G (1952) Zika virus (ii). pathogenicity and physical properties. Transactions of the Royal Society of Tropical Medicine and Hygiene 46(5):521–534

Dick G, Kitchen S, Haddow A (1952) Zika virus (i). isolations and serological specificity. Transactions of the Royal Society of Tropical Medicine and Hygiene 46(5):509–520

Edelsbrunner H, Harer J (2010) Computational Topology: an Introduction. American Mathematical Soc.

Edelsbrunner H, Letscher D, Zomorodian A (2000) Topological persistence and simplification. In: Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on, IEEE, pp 454–463

Fasy BT, Kim J, Lecci F, Maria C, Millman DL, included GUDHI is authored by Clement Maria VRT, by Dmitriy Morozov D, by Ulrich Bauer P, Kerber M, Reininghaus J (2018) TDA: Statistical Tools for Topological Data Analysis. URL https://CRAN.R-project.org/package=TDA, r package version 1.6.4

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Annals of Statistics pp 1189–1232

Friedman JH (2002) Stochastic gradient boosting. Computational Statistics & Data Analysis 38(4):367–378

Fritsch S, Guenther F (2016) neuralnet: Training of Neural Networks. URL https://CRAN.R-project.org/package=neuralnet, r package version 1.33

Gardner LM, Bóta A, Gangavarapu K, Kraemer MU, Grubaugh ND (2018) Inferring the risk factors behind the geographical spread and transmission of zika in the americas. PLoS neglected tropical diseases 12(1):e0006194

Goeijenbier M, Slobbe L, Van der Eijk A, de Mendonça Melo M, Koopmans M, Reusken C (2016) Zika virus and the current outbreak: an overview. Neth J Med 74(3):104–9

Greenwell B, Boehmke B, Cunningham J, Developers G (2018) gbm: Generalized Boosted Regression Models. URL https://CRAN.R-project.org/package=gbm, r package version 2.1.4

Heukelbach J, Alencar CH, Kelvin AA, de Oliveira WK, de Góes Cavalcanti LP (2016) Zika virus outbreak in brazil. The Journal of Infection in Developing Countries 10(02):116–120

Jiang D, Hao M, Ding F, Fu J, Li M (2018) Mapping the transmission risk of zika virus using machine learning models. Acta Tropica

Liang F, Guan P, Wu W, Huang D (2018) Forecasting influenza epidemics by integrating internet search queries and traditional surveillance data with the support vector machine regression model in Liaoning, from 2011 to 2015. PeerJ 6:e5134

Lo D, Park B (2018) Modeling the spread of the zika virus using topological data analysis. PloS One 13(2):e0192120

Lu FS, Hou S, Baltrusaitis K, Shah M, Leskovec J, Sosic R, Hawkins J, Brownstein J, Conidi G, Gunn J (2018) Accurate influenza monitoring and forecasting using novel internet data streams: A case study in the Boston Metropolis. JMIR Public Health and Surveillance 4(1)

Malone RW, Homan J, Callahan MV, Glasspool-Malone J, Damodaran L, Schneider ADB, Zimler R, Talton J, Cobb RR, Ruzic I, et al (2016) Zika virus: medical countermeasure development challenges. PLoS neglected tropical diseases 10(3):e0004530

McGough SF, Brownstein JS, Hawkins JB, Santillana M (2017) Forecasting zika incidence in the 2016 latin america outbreak combining traditional disease surveillance with search, social media, and news report data. PLoS Neglected Tropical Diseases 11(1):e0005295

MHB (2018) Ministry of Health of Brazil. http://portalms.saude.gov.br/boletins-epidemiologicos

Mordecai EA, Cohen JM, Evans MV, Gudapati P, Johnson LR, Lippi CA, Miazgowicz K, Murdock CC, Rohr JR, Ryan SJ, et al (2017) Detecting the impact of temperature on transmission of zika, dengue, and chikungunya using mechanistic models. PLoS Neglected Tropical Diseases 11(4):e0005568

Muñoz ÁG, Thomson MC, Stewart-Ibarra AM, Vecchi GA, Chourio X, Nájera P, Moran Z, Yang X (2017) Could the recent zika epidemic have been predicted? Frontiers in Microbiology 8:1291

Rees EE, Petukhova T, Mascarenhas M, Pelcat Y, Ogden NH (2018) Environmental and social determinants of population vulnerability to zika virus emergence at the local scale. Parasites & Vectors 11(1):290

Ridgeway G (2019) Generalized boosted models: A guide to the gbm package. Update 1(1):2019

Siddiqui S, Shikotra A, Richardson M, Doran E, Choy D, Bell A, Austin CD, Eastham-Anderson J, Hargadon B, Arron JR, et al (2018) Airway pathological heterogeneity in asthma: Visualization of disease microclusters using topological data analysis. Journal of Allergy and Clinical Immunology

Singh G, Mémoli F, Carlsson GE (2007) Topological methods for the analysis of high dimensional data sets and 3d object recognition. In: SPBG, pp 91–100

Suparit P, Wiratsudakul A, Modchang C (2018) A mathematical model for zika virus transmission dynamics with a time-dependent mosquito biting rate. Theoretical Biology and Medical Modelling 15(1):11

Teng Y, Bi D, Xie G, Jin Y, Huang Y, Lin B, An X, Feng D, Tong Y (2017) Dynamic forecasting of zika epidemics using google trends. PloS One 12(1):e0165085

Tjaden NB, Thomas SM, Fischer D, Beierkuhnlein C (2013) Extrinsic incubation period of dengue: Knowledge, backlog, and applications of temperature dependence. PLoS Neglected Tropical Diseases 7(6):e2207

Torres BY, Oliveira JHM, Tate AT, Rath P, Cumnock K, Schneider DS (2016) Tracking resilience to infections by mapping disease space. PLoS biology 14(4):e1002436

WHO (2016) World Health Organization. URL http://www.who.int/emergencies/diseases/zika/en/

World-Atlas (2018) URL https://www.worldatlas.com/articles/brazilian-states-by-population.html

World-Weather-Online (2018) URL https://www.worldweatheronline.com/

Zomorodian A, Carlsson G (2005) Computing persistent homology. Discrete & Computational Geometry 33(2):249–274