IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 7, 2020, accepted March 6, 2020, date of publication March 13, 2020, date of current version March 25, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2980634

A Multi-Stage Machine Learning Approach to Predict Dengue Incidence: A Case Study in Mexico

ANNALISA APPICE[1,2], YULIA R. GEL[3], ILIYAN ILIEV[4], VYACHESLAV LYUBCHICH[5], AND DONATO MALERBA[1,2], (Member, IEEE)

[1]Department of Computer Science, University of Bari Aldo Moro, I-70125 Bari, Italy
[2]Consorzio Interuniversitario Nazionale per l'Informatica (CINI), I-70125 Bari, Italy
[3]Department of Mathematical Sciences, University of Texas at Dallas, Richardson, TX 75080, USA
[4]School of Social Science and Global Studies, University of Southern Mississippi, Hattiesburg, MS 39406, USA
[5]Chesapeake Biological Laboratory, University of Maryland Center for Environmental Science, Solomons, MD 20688, USA

Corresponding author: Annalisa Appice (annalisa.appice@uniba.it)

This work was partially supported by the research project CLOSE – Close to the Earth (ID ARS01_00141) within Bando PON Ricerca e Innovazione 2014–2020 funded by the Italian Ministry for Universities and Research (MIUR). The work of Yulia R. Gel was partially supported by NSF grant numbers IIS 1633331 and NSF DMS 1925346.

ABSTRACT The mosquito-borne dengue fever is a major public health problem in tropical countries, where it is strongly conditioned by climate factors such as temperature. In this paper, we formulate a holistic machine learning strategy to analyze the temporal dynamics of temperature and dengue data and use this knowledge to produce accurate predictions of dengue, based on temperature on an annual scale. The temporal dynamics are extracted from historical data by utilizing a novel multi-stage combination of auto-encoding, window-based data representation and trend-based temporal clustering. The prediction is performed with a trend association-based nearest neighbour predictor. The effectiveness of the proposed strategy is evaluated in a case study that comprises the number of dengue and dengue hemorrhagic fever cases collected over the period 1985–2010 in 32 federal states of Mexico. The empirical study proves the viability of the proposed strategy and confirms that it outperforms various state-of-the-art competitor methods formulated both in regression and in time series forecasting analysis.

INDEX TERMS Clustering, machine learning, time-series analysis, predictive analysis.

I. INTRODUCTION

Dengue fever is a mosquito-borne disease caused by the dengue virus. It is a global problem that affects numerous tropical countries, with tens to hundreds of millions of people infected annually. Dengue can be deadly – annually there are thousands of deaths attributed to it. The mosquitoes carrying dengue are tropical and subtropical species that mostly live between latitudes 35 °N and 35 °S and usually in altitudes below 1000 meters. The disease-carrying mosquitoes generally grow in water-filled habitats close to human dwellings and thus dengue can be transmitted from community to community not just by the mosquitoes themselves, but also by humans [1, p. 14].

Dengue has received considerable attention in the data analysis literature [2]–[12]. The existing studies have focused on assessing the association between a number of factors (e.g., climatic, demographic and/or socio-economic variables) and the transmission properties of dengue. The data analysis techniques mainly considered in these investigations include regression analysis, correlation studies, and time series analysis. The studied data are often collected in Mexico, which has states with vastly different geography, climate, economies and demographics. For example, the authors of [2] conduct a time series analysis that uses an auto-regressive model to evaluate the role of climatic factors (e.g., temperature and precipitation) on the incidence of dengue over the period 1995–2005 on the Texas-Mexico border. The authors of [4] fit multiple linear regression models to look for associations between changes in the incidence rate of dengue fever and climate variability in the warm and humid region of 12 Mexican states, over the period 1985–2007. The authors of [8] apply wavelet analysis to identify time- and frequency-specific associations between temperature and dengue on a multi-year scale in Mexico, Puerto Rico and Thailand. The authors of [9] explore the effect of temperature on the fluctuation of population immunity and hyperendemicity in Singapore over the period 1980–2009. The authors

The associate editor coordinating the review of this manuscript and approving it for publication was Benyun Shi.

of [5] study the epidemiology of dengue fever in Mexico over the period 2000–2011 by looking for the age distribution pattern of the dengue disease, while the authors of [6] conduct a Bayesian phylogenetic analysis to determine the origin, persistence and geographical dispersion of various serotypes of the dengue virus in Mexico over the period 1980–2002.

In this paper, we investigate the dengue virus activity by exploring the dynamics of dengue incidence data in various states of Mexico.

From an application point of view, we consider the dengue incidence data of the Mexican ministries of health, described in [13]. We investigate the dynamics of temperature-related patterns in this database and evaluate whether these patterns exhibit predictive utility for dengue incidence in Mexico. From a methodological point of view, we formulate a machine learning strategy that performs cluster analysis of historical data to learn a model of the associations between temperature and dengue-related data. This model is used for predictive purposes. We investigate the effectiveness of this strategy compared to the state-of-the-art methods formulated in both regression analysis and time series forecasting.

Contrary to previous studies, which have mainly explored the idea of discovering associations between temperature and dengue [2], [8], [14] collected at a single site, we account for data collected at multiple sites (i.e., several Mexican states). In particular, we try to model the temporal dynamics of associations between temperature and dengue as they are spanned across space. For this purpose, the cluster analysis is performed over spatio-temporal data, which represent monthly measurements of both temperature and dengue, collected over consecutive years in various states of Mexico. The cluster analysis is performed to discover clusters of yearlong state co-located time series of temperature and dengue, so that their trends are continuously co-associated on the yearlong scale. An auto-encoding representation of temperature and dengue data is learned for mapping existing associations between temperature and dengue from the bivariate space to the univariate space. A univariate clustering technique is then applied to process the transformed data. Finally, the knowledge enclosed in the cluster model is used in a simple time series nearest neighbour predictor to accurately predict dengue from temperature on the yearlong scale.

The knowledge that we expect to derive with the cluster analysis is the extension of the temporal continuity of the trend pattern in the associations between the historical data of temperature and dengue. Therefore, we intend to discover clusters, which highlight the existence of distinctive yearlong trend patterns of co-located temperature and dengue, spanned over various states and, possibly, in different (not consecutive) years. The discontinuity points, which may occur over space and/or time in the extension of these patterns will be isolated into separate clusters. If enclosed in a global model, this may help in the finer modeling of the changes occurring over space and time in preventing the outlier association patterns from diminishing the accuracy of the predictions.

It is noteworthy that the cluster analysis represents a crucial phase of the proposed machine learning predictive strategy. Cluster analysis has received considerable attention in its application to political behavior [15], ecological trends [16], geophysical data streams [17]–[19], and many others [20]–[22]. In epidemiology, the authors of [23] use the Kulldorff space-time scan statistic (STSS) to identify statistically significant space-time clusters of chikungunya and dengue fever in Colombia from 2015 to 2016. The authors of [24] investigate a parallel computing approach for scaling the computation of space-time kernel density with epidemiological data of increasing size, diversity and availability. The proposed approach is used to perform clustering of a high volume of dengue fever cases from 2010 to 2011 in the city of Cali, Colombia. In this paper, we adopt the clustering algorithm introduced in [17]. One motivation for this decision is that, as shown in the empirical study described in [17], this algorithm scales well with large streams of geophysical data. In addition, the approach allows us to discover trends in addition to clusters as the method groups geographical sites (e.g., states of Mexico) around distinct trends, which depict how the geo-referenced data, measured at the clustered sites, evolve over time.

In summary, the contributions of this paper are:

- The definition of a holistic strategy, where the clustering is used to drive the historical data modeling for yielding accurate time series predictions.
- The use of auto-encoding as a univariate modeler of associations between temperature and dengue to ease the discovery of a model of the main trends in such bivariate associations through a univariate algorithm.
- The use of a nearest neighbour time series predictor with the temporal trend of associations between temperature and dengue as they are discovered in historical data spread across space.
- The evaluation of the effectiveness of the presented strategy in a case study that comprises monthly data on temperature and the number of dengue and dengue hemorrhagic fever cases collected over the period 1985–2010 in 32 federal states of Mexico.

This paper is organized as follows. The data are presented in the next Section. The proposed machine learning method and its competitors are described in Section III. The experimental setting is illustrated in Section IV, while the findings in the evaluation are discussed in Section V. Finally, Section VI draws conclusions and proposes future developments.

II. DATA

We investigate dengue-related data collected monthly at the scale of the federal states of Mexico. We note that this spatial and temporal scale of data is used in various studies of dengue incidence in Mexico [7], [13]. As the risk of dengue epidemics is higher when the weather is suitable for mosquitoes that transmit the virus, we supplement the dengue data with air temperature information. This decision is based

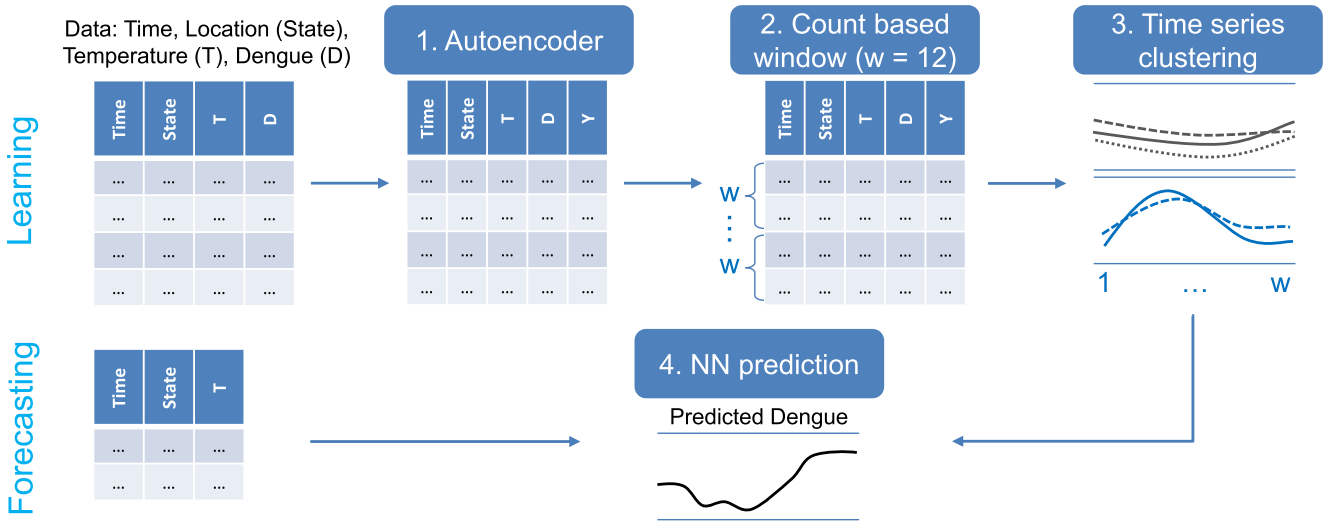


FIGURE 1. **Machine learning methods of** AutoTiC-NN**. (1) The auto-encoding method builds a 1-dimensional representation of** Dengue **and** Temperature **on the training data collected at various states. (2) The count-window operator decomposes the training data into a yearlong time series geo-located at the considered states. (3) The time series clustering method builds clusters of states whose encoding representation of the measured values of** Dengue **and** Temperature **evolve with a similar trend along the considered yearlong training windows. For each cluster, the prototype time series of both** Dengue **and** Temperature **is determined by aggregating training data collected at the cluster scale. (4) The NN method is used to predict any testing time series of** Dengue **based on both the co-located known time series of** Temperature **and the detected cluster-based prototypes built for both** Dengue **and** Temperature **from the training data.**

on several studies [2], [4], [8], [9], [14], [25], [26], which have shown that additional climate covariates can improve predictive accuracy of dengue-related phenomena.

In particular, we focus our attention on the task of predicting monthly dengue incidence one year ahead, based on monthly temperature. We highlight that the decision of considering the monthly data aggregation level is coherent with the analysis formulated in [14], who show that the highest correlation of the incidence of dengue can be found with the temperature at a lag of one month. On the other hand, the use of the year scale in the association search follows the considerations formulated by the authors of [8], who highlight that associations between temperature and dengue cannot be stably observed in Mexico on the multi-year scale. The study suggests that the continuity of the associations between temperature and dengue should be analyzed at the annual scale instead.

Therefore, based upon the premises formulated above, we study the dengue incidence in Mexico based on the monthly number of dengue and dengue hemorrhagic fever cases. These data were collected from January 1985 to December 2010 by the [27] for 32 Mexican federal states. Similar data are investigated in the recent literature [13], [25], [28].

Finally, the average air temperature (in °C) was obtained from conventional and automatic stations [29] at the monthly level for each of the states of interest in the matching period.

III. METHODS AND MATERIALS

We investigate the use of machine learning to yield accurate yearlong predictions of dependent Dengue time series on independent (or ancillary) Temperature time series. To this aim, we propose a novel multi-stage machine learning strategy (see Section III-A), denoted as AutoTiC-NN (AUTOencoding based TIme series Clustering with Nearest Neighbour) that is formulated to learn a bivariate predictive model of Dengue from Temperature. In addition, we illustrate well-known predictive methods (Section III-B), formulated in both time-series forecasting and regression theory, which will be evaluated as possible competitors of AutoTiC-NN in Section V.

A. AutoTiC-NN *METHOD*

AutoTiC-NN is composed of four stages, i.e., auto-encoding, count-based window, time series clustering, and nearest-neighbour prediction (see Figure 1). The auto-encoding, count-based window and time series clustering methods are cascaded in order to learn a clustering model of the training historical measurements of Dengue and Temperature. The count-based window decomposes the training data, geo-located in each state, into consecutive yearlong time series. The time series clustering method learns a clustering model of the time series representation of the training data. The auto-encoding stage allows us to derive a univariate representation of training data by also capturing possible associations hidden in the historical measurements of Dengue and Temperature. Therefore, the clustering step can be attempted with a univariate algorithm, without giving away the opportunity of accounting for patterns describing associations between Dengue and Temperature. Finally, the learned clustering model can then be used in the predictive stage. In particular, the nearest neighbour prediction method allows us to predict any testing time series of Dengue based on both the co-located time series of Temperature and the

cluster model of both Dengue and Temperature learned from the historical training data. A detailed description of auto-encoding, count-based window, time series clustering and nearest-neighbour prediction stages is given in Sections III-A.1–III-A.4.

1) AUTO-ENCODING

In the first stage, we train the auto-encoder of the training data comprising both Dengue and Temperature. Auto-encoding is an unsupervised deep learning algorithm that produces codifications for input data. The model is trained so that the decodification resembles the input data as closely as possible [30]. The basic structure of an auto-encoder is defined by the encoder-decoder architecture that is used to map an input x onto the encoding y via an encoder network so that $y = \sigma(\mathbf{W}x + \mathbf{b})$, where σ is an element-wise activation function, $\mathbf{W}$ is a weight matrix and $\mathbf{b}$ is a bias vector. Weights and biases are usually initialized randomly and then updated iteratively during training through back-propagation. The encoding is in turn mapped to the reconstruction r by means of a decoder network, so that $r = \sigma'(\mathbf{W}'y + \mathbf{b}')$. In particular, the auto-encoder is trained to minimize the loss $||x - r||^2$ through a feedforward neural network that reproduces the input data on the output layer. Both x and r have the same dimension, while the auto-encoder has as many layers as needed, placed symmetrically in the encoder and the decoder. Every unit located in any of the hidden layers receives several inputs from the preceding layer. The unit computes the weighted sum of these inputs and applies the activation function to produce the output.

In this study, the input x is the collection of bivariate historical data points (Dengue and Temperature) recorded at each state and at every monthly time point in the training period. These data are scaled between 0 and 1 using the Exponential Linear Unit (ELU) function. This is a recent, popular activation function in deep learning that is selected as it speeds up learning and improves learning characteristics, compared with other activation functions [31]. ADAM, an adaptive learning rate optimization algorithm, which was designed specifically for training deep neural networks, is adopted as the algorithm for optimizing weights and biases with the mean square error as the loss function [32]. The encoder is made up of three layers of sizes 4, 2 and 1, including the middle encoding one, while the decoder starts in the middle one and also spans three layers (see Figure 2). This defines an undercomplete neural network, where the inner encoding layer (size 1) has a lower dimensionality than the input one (size 2). The smaller number of units in the inner layer imposes a restriction, so during training the auto-encoder is forced to learn a more compact representation. This is achieved by fusing the features according to the weights assigned through the learning process.

Based upon the above theory, the output of the inner encoder layer reduces the dimensionality of the input data to a high capacity model that captures interactions and dependencies hidden in the input, while rejecting noise from the input [33]. It acts in a manner similar to a non-linear combination of the input. The rationale of using this high capacity model in this study is that the encoder layer allows us to identify the latent variable that expresses some very fundamental information about a possible interaction between Dengue and Temperature. In particular, the analysis of this information in the clustering stage allows us to group the data based on the specific interaction model they comply with. In addition, the use of the univariate encoder representation of Dengue and Temperature allows us to make the time series clustering process simpler to perform. In fact, we halve the number of learning parameters as a consequence of halving the number of variates in the data representation.

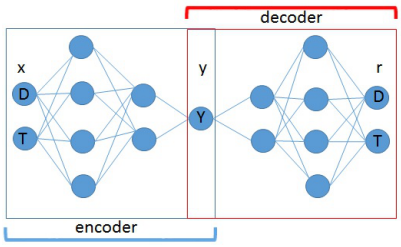


FIGURE 2. The autoencoder architecture trained to derive the univariate model of Dengue and Temperature. It consists of an encoder function mapping the input x to a hidden code r and a decoder producing the reconstructed input r. It is learned by minimizing a loss function $||r - x||^2$. The encoder network adopted in this study is composed of 2 × 4 × 2 × 1 layers, while the decoder is composed of 2 × 4 × 2 × 1 layers. With this architecture the hidden code r defines a univariate representation of Dengue and Temperature.

2) COUNT-BASED WINDOW

In the second stage, we use the count-based window operator [34] to prepare the training data for the time-series clustering stage. We decompose the historical training data, collected at each state, into yearlong consecutive windows of 12 monthly-spaced time points considered in series. This corresponds to building one state-window unit of analysis, denoted as SW, for every state S and for every yearlong window W in the training dataset. For example, the count-based window method, applied to the data of Dengue and Temperature described in Section II, transforms monthly measurements collected at 32 states over one yearlong window into 32 state-window units of analysis. The consecutive monthly time points are locally enumerated within SW. Every point $SW[i]$ $(i = 1, \ldots, 12)$ is one-to-one associated with the triple $\langle D_{SW}[i], T_{SW}[i], Y_{SW}[i]\rangle$, where $D_{SW}[i]$ and $T_{SW}[i]$ are the measurements of Dengue and Temperature, respectively, as they are geo-located at state S and collected at time point i of window W. The $Y_{SW}[i]$ is the encoder combination of $D_{SW}[i]$ and $T_{SW}[i]$, as it is computed at the encoder layer y of the auto-encoding stage.

3) TIME SERIES CLUSTERING

In the third stage, we perform the clustering of the state-window units of analysis, which are built from the training dataset by applying the count-based window operator. For

each state-window unit of analysis, we consider the univariate time series associated with the encoder information Y and process these univariate data with the trend cluster discovery method formulated by [18]. This method is selected as it implements an efficient threshold-based clustering algorithm that partitions the univariate time series into trend-based clusters.

By replicating the theory reported in [18], each discovered trend cluster allows us to identify a group of state-window units of analysis so that the cluster-spanned univariate encoder representation of both Dengue and Temperature depicts values which are homogeneous at the same time point and evolve with a similar trend along the window time horizon. Formally, every trend cluster $\mathcal{C}_k$ is a set of training state-window units SW, which satisfy a homogeneity condition $h(\mathcal{C}_k)$. This homogeneity condition is formulated on the encoder variate Y. Formally,

$$h(\mathcal{C}_k) = \begin{cases} true & \text{if } \left(\max_{SW \in \mathcal{C}_k} Y_{SW}[i] - \min_{SW \in \mathcal{C}_k} Y_{SW}[i]\right) \leq \delta \\ & \text{for all } i = 1, 2, \ldots, 12 \\ false & \text{otherwise,} \end{cases} \tag{1}$$

where δ is a user-defined threshold that controls the granularity of the clustering. It is noteworthy that, as Y is a univariate model of Dengue and Temperature, trend clusters that satisfy Equation 1 distinguish emerging trends in the interaction between co-located values of Dengue and Temperature.

At the completion of the construction of a trend cluster $\mathcal{C}_k$, we determine the cluster trend prototype time series of both Dengue ($\mathcal{D}_k$) and Temperature ($\mathcal{T}_k$), associated with $\mathcal{C}_k$. For each i ranging between 1 and 12, we compute:

$$\begin{aligned} \mathcal{D}_k(i) &= \frac{1}{|\mathcal{C}_k|} \sum_{SW \in \mathcal{C}_k} D_{SW}[i], \\ \mathcal{T}_k(i) &= \frac{1}{|\mathcal{C}_k|} \sum_{SW \in \mathcal{C}_k} T_{SW}[i], \end{aligned} \tag{2}$$

where $|\mathcal{C}_k|$ denotes the cardinality of cluster $\mathcal{C}_k$. The time series couple composed of both $\mathcal{D}_k$ and $\mathcal{T}_k$ describes the interaction of Dengue and Temperature spanned across the trend cluster $\mathcal{C}_k$. We note that a trend cluster may group units of analysis, which comprise data collected at the same state, but over various windows, as well as data collected at the same window, but over various states. Therefore, this cluster-level model of the training data can reveal spatial and temporal extensions of similar interactions of Dengue and Temperature. These interactions, modeled at the cluster level, define the training samples for the NN prediction method.

4) NEAREST NEIGHBOUR PREDICTION

In the final stage, the nearest neighbour (NN) method is applied to predict any yearlong testing time series of Dengue (dependent time series variable), based on the yearlong co-located time series of Temperature (ancillary time series variable). The NN method [35] is a non-parametric approach to the multivariate prediction of a dependent variable. NN is based on the similarity in the ancillary variable space between a (target sample) unit, for which a prediction is desired, and a set of reference units (training samples), for which an observation of the dependent variable is available. That is, NN is based on the computation of the Euclidean distance between a test sample and the specified training samples. In this study, the training samples comprise the trend time series prototypes $\mathcal{D}_k$ and $\mathcal{T}_k$, built during the clustering stage (see Equation 2).

Every target sample to be predicted is composed of the ancillary yearlong time series T', which is the series of the known monthly measurements of Temperature as they are observed along one testing year in one state. The NN prediction task aims to yield an accurate estimate of the dependent yearlong time series D', which is the series of the unknown monthly measurements of Dengue as they are co-located at the testing time in the same state as T'. To this purpose, we first identify the trend cluster whose trend time series prototype of Temperature $\hat{\mathcal{T}}$ is the closest to T', that is:

$$\mathcal{C}_k = \operatorname*{argmax}_{\mathcal{T}_k} distance(\mathcal{T}_k, T'), \tag{3}$$

where $distance(\mathcal{T}_k, T') = \sum_{i=1}^{12} \left(\mathcal{T}_k(i) - T'(i)\right)^2$ is the Euclidean distance. Based upon Equation 3, we set the predicted dependent time series D' equal to the trend time series prototype of Dengue – $\mathcal{D}_k$ – associated with the NN-selected trend cluster $\mathcal{C}_k$ that is identified with Equation 3.

B. COMPETITOR METHODS

1) TIME SERIES FORECASTING METHODS

The time series forecasting analysis is a machine learning baseline that can be considered when the time series of Dengue and Temperature, collected in every state, are dealt with as separate training sets. Specifically, time series forecasting can be performed for each state, to take a distinct model fit on the time series of the historical observations of Dengue and Temperature as they are collected at the consecutive training time points of the specific state. This time series model is then exploited to predict future observations of Dengue for the state under consideration. We note that we learn a distinct time series model for each state. For this study we evaluate two well-known time series forecasting method competitors, that is, Automatic AutoRegressive Integrated Moving Average (auto.arima) [36] and Vector AutoRegression (VAR) [37].

auto.arima is one of the most powerful univariate auto-regressive forecasting methods. For each state, the Dengue forecasts are computed based on historical data of Dengue observed in that state. Therefore, this method ignores the ancillary temperature information. It differentiates the original time series for an appropriate number of times, until a test for the presence of unit roots ceases to provide statistically significant signals and the transformed data

can finally be considered (at least approximately) stationary. The optimal parameters (i.e., the orders of auto-regressive and moving average components) of the fitting model are selected according to a stepwise procedure. In the analysis of dengue data, we train seasonal ARIMA models as reported in [13], with the seasonal terms of the model involving backshifts of the seasonal period (12 months).

VAR is a multivariate auto-regressive method to analyze a system of multiple variables. That is, for each state, the fitting model to forecast Dengue is computed considering the historical data of Dengue and Temperature, co-located in the state under consideration. Each variable has an equation explaining its evolution, based on its own previous lagged values, the previous lagged values of the other model variable and an error term. The optimal number of lagged values is selected with a sequential increase in the lag order, based on the same sample size. The implementation of auto.arima and VAR is available in the R packages forecast and vars, respectively.

2) REGRESSION METHODS

The regression analysis is the additional machine learning baseline that is considered in this study as a method for learning the relationships between Dengue (dependent variable) and Temperature (ancillary, independent variable) by disregarding the temporal information. This regression relationship can then be used to predict an unknown value of Dengue from the co-located observations of Temperature. In this study, we consider a suite of well-known regression methods, that is, M5′ [38], Support Vector Regression – SVR [39], and k-Nearest Neighbourhood – kNN [40].

M5′ induces a structured tree that is composed of non-terminal nodes, each one representing a test over Temperature, and linking edges that partition the data according to the test result. At the bottom of the tree the terminal nodes hold linear regression models, which are formulated according to the data that reach each given node. Hence, for predicting the Dengue value for a given Temperature, we walk along the tree from the root node to the bottom until a terminal node is reached and then we apply the corresponding linear model. Model trees result in a clear knowledge representation, providing the user with information on how the output was reached.

SVR is formulated as an optimization method by first defining a convex ϵ-insensitive loss function to be minimized, and finding the flattest tube that contains most of the training data. Hence, a multi-objective function is constructed from the loss function and the geometrical properties of the tube. Then, the convex optimization is solved, using appropriate numerical optimization algorithms. The hyperplane is represented in terms of support vectors, which are training samples that lie outside the boundary of the tube. The support vectors are the most influential training samples that affect the shape of the tube, while the training and test data are assumed to be independent and identically distributed, drawn from the same fixed but unknown probability distribution function. This method has excellent generalization capacity, with high prediction accuracy. In this study, SVR is used with the Gaussian kernel rule, while its parameters are optimally selected according to a grid-search method.

kNN is a generalization of the nearest neighbour method. The input consists of the k closest training samples in the ancillary space. The output is the average of the values of its k nearest neighbours. This method is well known for its simplicity. It is independent of any data distribution, and it only needs to be adjusted or assigned an integer parameter k. M5′, SVR and kNN are implemented in Java in the software toolkit Weka [41].

IV. EXPERIMENTAL SETTING

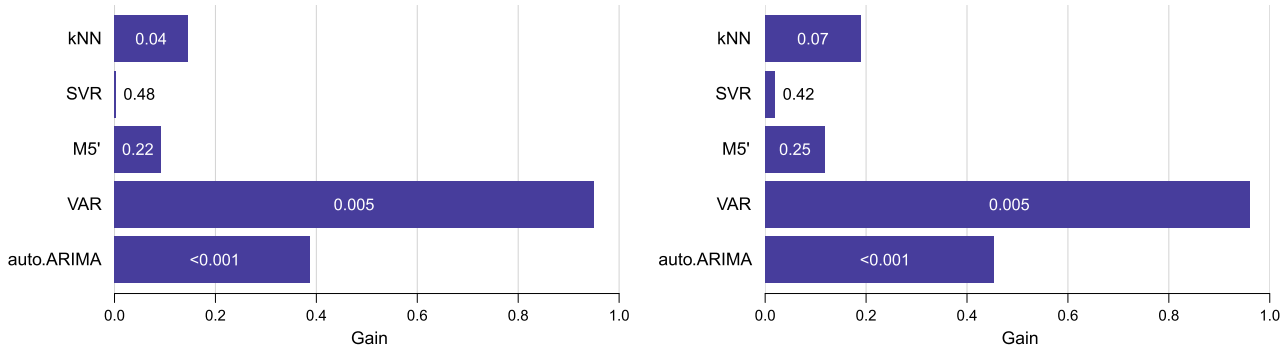
In the experiments, we split the state time series data collected for Dengue and Temperature into training and testing datasets. The training dataset comprises the data observed between January 1985 and December 2009, while the testing dataset comprises the data observed between January 2010 and December 2010.

The data for the case study were collected across 32 federal states of Mexico, and one experimental setting is certainly defined by considering all federal states together. However, [13] point out that if the primary interest of studying these data is to assess models for forecasting dengue incidence in endemic locations, the analysis may be restricted to the 17 federal states where median monthly incidence of dengue is greater than zero during the training period.[1]

Based upon these considerations, we evaluate the predictive accuracy of AutoTiC-NN, auto.arima, VAR, M5′, SVR and kNN in both the complete and restricted experimental settings of the case study. The evaluation consists of using the training dataset to learn the model that is used to predict unseen data of Dengue in the testing dataset, based on the Temperature. We use the Root Mean Square Error (RMSE) to measure the accuracy of the predictions. The RMSE is computed as the square root of the second sample moment (quadratic mean) of the differences between observed and predicted values of Dengue in the testing dataset. To compare the competitive models with the suggested method, AutoTiC-NN, we also calculate the accuracy gains as $Gain_j = 1 - RMSE^2_{\text{AutoTiC-NN}}/RMSE^2_j$, where j denotes a competitive model ($-\infty < Gain_j < 1$, where $Gain_j = 0$ means equally good performance of AutoTiC-NN and the jth model, and $Gain_j \to 1$ means superiority of AutoTiC-NN over the jth model).

In addition, we evaluate the performance of the methods based on the computational time spent completing both the training and testing phases. This estimate of the computational time, repeated on both the complete and restricted data settings, allows us to explore how the considered methods scale with the amount of data.

[1]The states selected according to these considerations of [13] are: Campache, Chiapas, Colima, Guerrero, Jalisco, Michoacan, Morelos, Nayarit, Nuevo Leon, Oxaca, Puebla, Quintano Roo, Sinaloa, Tabasco, Tamaulipas, Veracruz, and Yucatan.



(a) 32 federal states (b) 17 federal states

FIGURE 3. Dengue cases analysis: bar graph of gain values for comparing the competitive models with AutoTiC-NN. The *p*-values of Student's t-tests, computed for evaluating the statistical difference between squared residuals of AutoTiC-NN and competitive models, are reported in each bar.

V. RESULTS

We analyze the performance of the methods illustrated in Section III in the case study on dengue incidence in Mexico presented in Section II and with the experimental setting described in Section IV. We note that this experimental study aims at validating the effectiveness of AutoTiC-NN compared to that of competitor methods (time series forecasting methods and regression methods) in terms of accuracy and efficiency. Specifically, it aims to seek answers to the following questions:

1) Can the cluster knowledge extracted through a multi-stage combination of auto-encoding, window-based data representation and trend-based temporal clustering be effectively exploited to empower an accurate time series nearest neighbour predictor?
2) Can the trend association-based nearest neighbour predictor be more accurate than the predictive models discovered through the competitor methods?
3) How efficient are the training phase and the prediction phase performed by the compared methods?
4) Is the descriptive skill of trend-based temporal clustering actually able to highlight the existence of temporal (and possibly spatial) dynamics in training data, by explaining the effectiveness of a time series nearest neighbour predictor that uses this cluster knowledge?

The presentation of the results is organized as follows. Initially, the accuracy performance is evaluated to answer questions 1–2. To this purpose, we analyze the overall gain in accuracy achieved by AutoTiC-NN compared with competitor methods and explore the gain results more in depth by analyzing the accuracy of the compared methods state-by-state. Subsequently, to answer question 3, we evaluate the computation time spent completing the training and testing for each compared method. Finally, we analyze the cluster model built by AutoTiC-NN – the only method in this study equipped with a clustering descriptive skill in addition to the predictive one – to answer question 4.

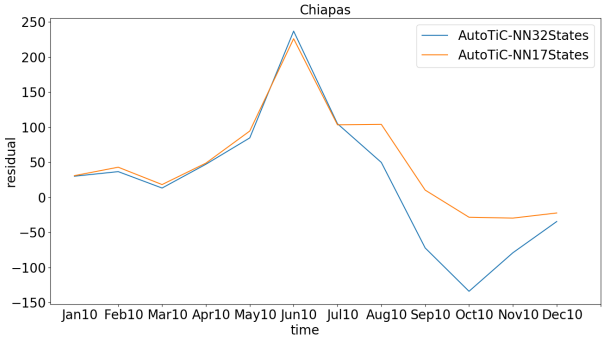


FIGURE 4. The time series of the residuals (computed as the difference between the ground truth value of the dengue cases and their prediction) collected at Chiapas with AutoTiC-NN run in the settings with 32 states and 17 states.

A. PREDICTIVE ACCURACY

We start analyzing the overall accuracy performance of the compared methods. Based on the gain in accuracy achieved by AutoTiC-NN in comparison with other considered methods, the AutoTiC-NN outperforms its competitors in both full and reduced data settings (see Figures 3(a) and 3(b)). The *p*-values of Student's t-tests for testing the difference of mean squared residuals of AutoTiC-NN and each other *j*th model (reported in the bars in Figure 3) also show that the better performance of AutoTiC-NN is commonly statistically significant with VAR, auto.arima, and kNN.

However, since the analysis illustrated by [13] is performed at the level of each federal state, we interpret the gain results more in depth by analyzing the accuracy of the compared methods state-by-state. In particular, Tables 1 and 2 show that AutoTiC-NN should be the preferred method for many of the states, both with low and high dengue rates. The last row of Tables 1 and 2 summarizes the results by counting how many times each method achieves the lowest error in the comparative analysis in the experimental setting with 32 federal states (Table 1) and 17 federal states (Table 2). In both settings, the number of federal states, where AutoTiC-NN achieves the lowest RMSE, is higher than the number of federal states

TABLE 1. RMSE of predictions of dengue incidence between January 2010 and December 2010 in all federal states of Mexico. The lowest error per state is in bold. The last row reports the number of federal states of Mexico where each method achieves the lowest RMSE in the comparative study.

State	AutoTiC-NN	auto.arima	VAR	M5'	SVR	NN
Aguascalientes	0.88	**0.57**	0.58	41.29	25.99	4.17
Baja California	**1.12**	1.12	1.49	59.03	26.50	5.32
Baja California Sur	196.74	219.03	207.97	**191.40**	212.96	229.88
Campeche	113.35	138.16	123.00	**101.75**	126.66	136.94
Coahuila	179.90	27.24	29.90	63.58	21.75	**11.44**
Colima	142.73	686.78	610.72	**106.84**	134.04	150.97
Chiapas	**86.38**	186.57	223.69	100.89	96.73	641.52
Chihuahua	7.81	**0.60**	1.21	55.16	25.79	2.27
Distrito Federal	22.24	**0.00**	0.05	34.04	26.17	0.29
Durango	**2.82**	4.81	7.32	39.73	24.15	121.87
Guanajuato	**0.10**	12.85	9.51	48.82	27.24	99.66
Guerrero	684.04	**474.27**	646.94	634.04	701.17	711.42
Hidalgo	**4.44**	26.24	38.59	34.21	22.26	27.36
Jalisco	156.06	1536.27	25396.71	671.08	**150.42**	199.13
México	**5.18**	12.23	24.11	115.34	19.62	8.20
Michoacán	151.44	263.39	426.92	**134.67**	137.57	156.21
Morelos	148.50	**98.74**	119.54	171.95	186.24	193.75
Nayarit	**159.97**	219.24	725.44	398.65	174.41	275.82
Nuevo León	**191.58**	206.44	133.30	227.45	259.39	266.84
Oaxaca	408.37	376.27	**307.55**	393.76	434.06	382.90
Puebla	152.93	138.41	132.11	**127.01**	136.00	150.32
Querétaro	**0.31**	0.70	1.52	41.69	26.89	2.20
Quintana Roo	235.83	**186.06**	163.03	196.10	235.99	256.41
San Luis Potosí	285.38	60.94	59.76	43.85	**16.43**	61.18
Sinaloa	**79.99**	86.13	96.05	90.90	111.21	143.51
Sonora	732.76	715.32	**674.25**	711.27	725.76	737.61
Tabasco	**27.37**	198.85	213.20	53.59	33.24	76.77
Tamaulipas	244.96	108.50	177.51	**92.09**	105.84	208.26
Tlaxcala	1.09	**0.03**	0.04	12.93	23.18	16.20
Veracruz	**80.55**	608.21	724.39	91.45	127.68	171.72
Yucatán	**148.11**	671.86	205.22	245.16	290.56	303.54
Zacatecas	0.66	**0.00**	1.11	36.53	25.32	120.89
Best	13	8	2	6	2	1

TABLE 2. RMSE of predictions of dengue incidence between January 2010 and December 2010 in the 17 federal states of Mexico, reporting dengue cases in at least half of the months in the training. The lowest error per state is in bold. The last row reports the number of federal states of Mexico where each method achieves the lowest RMSE in the comparative study.

State	AutoTiC-NN	auto.arima	VAR	M5'	SVR	NN
Campeche	**92.83**	138.16	123.00	112.94	126.67	136.90
Colima	141.57	686.78	610.72	**100.40**	134.11	150.97
Chiapas	**96.66**	186.57	223.69	130.12	96.77	641.21
Guerrero	657.32	**474.27**	646.94	629.30	701.28	701.33
Jalisco	**148.09**	1536.27	25396.71	643.06	150.74	200.28
Michoacán	129.50	263.39	426.92	**115.42**	138.03	145.81
Morelos	123.58	**98.74**	119.54	155.40	186.40	216.16
Nayarit	**169.18**	219.24	725.44	376.72	174.42	183.37
Nuevo León	280.59	206.44	**133.30**	217.80	259.49	264.47
Oaxaca	**232.67**	376.27	307.55	379.43	434.15	420.21
Puebla	121.40	138.41	132.11	**114.43**	136.25	148.29
Quintana Roo	225.86	186.06	**163.03**	178.66	236.02	256.61
Sinaloa	**64.55**	86.13	96.05	81.55	111.25	150.51
Tabasco	39.31	198.85	213.20	80.63	**33.23**	74.38
Tamaulipas	99.81	108.50	177.51	**87.31**	105.86	208.20
Veracruz	510.65	608.21	724.39	132.23	**127.77**	161.10
Yucatán	**174.45**	671.85	205.22	238.04	290.62	303.89
Best	7	2	2	4	2	0

where any other competitor is the best. In addition, Figure 4 plots the residuals computed, month-by-month, on the predictions of dengue cases in Chiapas as they are yielded by AutoTiC-NN in the two settings. This analysis shows two error peaks in the setting with 32 states (in June 2010 and in October 2010, respectively), while only one error peak (in June 2010) in the setting with 17 states.

Finally, we investigate the poor overall behaviour of the time series forecasting analysis. The state-level errors reported in Tables 1 and 2 show that there are federal states (e.g., Morelos) where both auto.arima and VAR are able to learn a good predictive model. On the other hand, there are also federal states (e.g., Jalisco) where these methods yield outlying predictions with high errors that decrease the

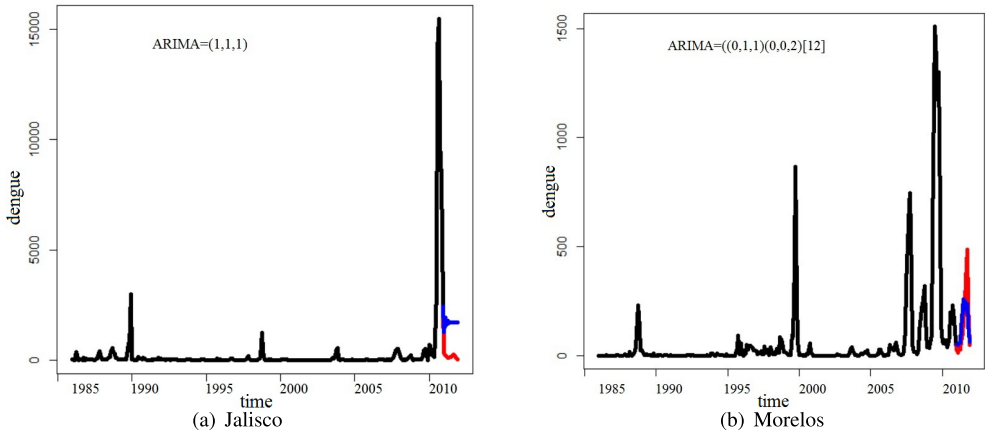


FIGURE 5. Time series of dengue cases collected at (a) Jalisco and (b) Morelos. The black line depicts the training data. The red line depicts the testing data. The blue line depicts the testing data predicted by the model learned with auto.arima.

overall estimate of the accuracy. To explain this phenomenon, we analyze the trend in the dengue time series for Jalisco and Morelos (see Figures 5(a) and 5(b), respectively). We note that the testing dengue data for Morelos resemble very well their closest past training data, while the testing dengue data for Jalisco drift with respect to the closer past training data (see, in particular, the training data in 2009 and the testing data in 2010 for both federal states). This explains the poor performance of the statistical model learned by both auto.arima and VAR for the state of Jalisco, as both models are affected by the presence of a sudden concept drift occurring at the testing time.

B. COMPUTATIONAL EFFICIENCY

We analyze the computational time for both the training and testing of each compared method. For this analysis, we run AutoTiC-NN implemented in Java 8, except for the auto-encoder that is realized in Python 3.5; auto.arima and VAR implemented in R×64 3.4; M5′, SVR and kNN implemented in Weka 3.6 and run with Java 8. For each method, the computation time is measured in milliseconds on an Intel(R) Core(TM) i7-4720U CPU@2.60 GHz and 16 GB RAM running Microsoft Windows 8.1 (64 bits).

The computational times reported in Table 3 show that AutoTiC-NN scales well with the size of the training data, as it spends about 1.9 seconds computing the model of a training set of 5100 bivariate data points (12 monthly observations of dengue and temperature across 25 years at 17 federal states), and 3.5 seconds computing the model of a training set of 9600 bivariate data points (12 monthly observations of dengue and temperature across 25 years at 32 federal states). The prediction phase of AutoTiC-NN is very quick (no more than 1 millisecond) in both data settings. In addition, although there are a few competitors (i.e., M5′ and kNN) that complete the training phase spending less computational time than AutoTiC-NN, the competitors are slower in their use of the learning model to yield the testing predictions. The only exception is M5′ that, similarly to AutoTiC-NN, yields the testing predictions within 1 millisecond.

TABLE 3. Computational time (milliseconds). Training period corresponds to 1985–2009; testing – to 2010.

Method	32 federal states		17 federal states	
	training	testing	training	testing
AutoTiC-NN	3576	≤ 1	1967	≤ 1
auto.arima	40662	396	27369	248
VAR	3657	125	2562	31
M5'	1445	≤ 1	968	≤ 1
SVR	25427	58	7580	33
kNN	18	203	16	67

C. CLUSTER ANALYSIS

To complete this study, we inspect the cluster model built by AutoTiC-NN in the considered case study. The ability of building a clustering model of the historical measurements of Dengue and Temperature and using it as a base for the nearest neighbour prediction of the dengue incidence equips AutoTiC-NN with a descriptive skill, in addition to the predictive one. While regression and time series competitors can be used only for the predictive scope, AutoTiC-NN is also able to discover the spatial and temporal arrangement of the pattern according to Dengue and Temperature association in the historical data.

Table 4 reports the number of clusters detected in both settings, with the results for 17 states visualized in Figure 6. This model highlights the existence of a pattern of spatial and temporal continuity in the association between temperature and dengue incidence. For example, data in Campeche are grouped in the same cluster at the consecutive years except for the measurements collected on 2003 (see the blue cluster in Figure 6). In addition, data in the neighbouring states of Campeche, Quintana Roo and Yucatan are repeatedly grouped in the same cluster over time (see the blue cluster in Figure 6). On the other hand, the cluster model is also

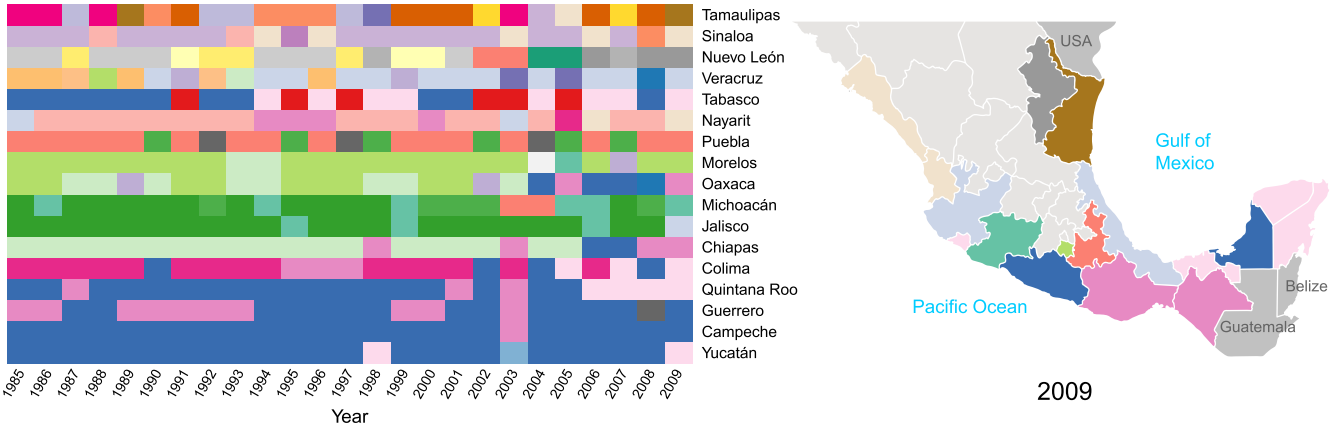


FIGURE 6. Cluster model discovered by AutoTiC-NN from the training data of dengue incidence and temperature monthly collected between January 1985 and December 2009 in 17 federal states of Mexico. The colors represent 37 clusters of the state-year units of analysis.

TABLE 4. Number of clusters discovered by AutoTiC-NN from the training data of dengue incidence and temperature collected monthly between January 1985 and December 2009.

Experimental setting	Training units of analysis	Number of clusters
32 federal states	$32 \times 25 = 800$	90
17 federal states	$17 \times 25 = 425$	37

able to identify possible discontinuity points in this pattern by isolating data changed with respect to the temporal and spatial surrounding in separate clusters (e.g., data on 1998, 2003 and 2009 in Yucatan, Figure 6). This helps in a finer modeling of the changes occurring over space and time in cluster-spanned associations preventing the outlier association patterns, if enclosed in a global model, from diminishing the accuracy of predictions. This skill can be considered one of the reasons of the higher overall accuracy of AutoTiC-NN with respect to the competitors in this study (see Figure 3).

VI. CONCLUSION AND FUTURE WORK

This paper proposes a new multi-stage machine learning strategy called AutoTiC-NN that combines auto-encoding, window-based data slicing and cluster analysis to discover the temporal dynamics in historical measurements of temperature and dengue variables. The cluster analysis allows us to model the spatial and temporal continuity of the trend pattern in the associations between historical data of temperature and the dengue variable. This cluster knowledge is then exploited to empower a time series nearest neighbour predictor.

We analyze the performance of the formulated method in a real case study that involves the number of dengue and dengue fever hemorrhagic cases collected monthly in several federal states of Mexico. The predictions account for both the co-located measurements of temperature and for the resembling trends in the associations between the temperature and the dengue variable discovered in the historical data. In our case study, we validate the predictive performance of the proposed new AutoTiC-NN procedure against a broad range of benchmark competitors. The results show that AutoTiC-NN tends to outperform all considered competing forecasting approaches. In particular, the proposed strategy has clear advantages (in terms of predictive accuracy), enabling us to account for association patterns in data collected over both space and time. In the considered case study, the ability to isolate specific association patterns in separate trend clusters is found to be closely connected to an observable gain in the predictive accuracy.

We note that the proposed AutoTiC-NN approach can then be employed for the continuous forecasting of future dengue outbreaks, while repeating the trend cluster discovery on the extended training data set to integrate newly available records of dengue incidences and environmental factors upon their arrival. A future development will be to properly frame the method in a data stream environment [17], [18] so that trend clusters may be adaptively incremented upon the arrival of new data records. However, to properly handle a data stream, one of our primary future research tasks will be to extend the proposed AutoTiC-NN approach by adding mechanisms for dealing with concept drift and incremental learning. This will discover and account for possible changes in data (e.g., climate changes), as well as adapt the cluster knowledge from past data to new data without repeating the entire learning phase from scratch.

When data are available, the transmission properties of dengue may be studied in association with various climatic, demographic and socio-economic factors, as well as microbiological data on dengue serotypes, which may be included as additional variables. Hence, another future research task will be to extend the proposed AutoTiC-NN approach from a bivariate to multivariate scenario, exploring the use of auto-encoding in modeling multivariate associations and studying the impact of these cross-domain associations in modeling dengue incidence.

Another important question to address is the transferability of the proposed AutoTiC-NN approach to other geographical regions. While the question of a model transferable to

other regions, such as South Asia or sub-Saharan Africa, remains open until justified by the actual data analysis, we hypothesize that the proposed AutoTiC-NN approach is expected to exhibit a certain level of geographic transferability. Indeed, the AutoTiC-NN methodology is not explicitly based on the particular specifics of the Dengue spread in Mexico, but rather on data availability/data quality and on how well the data can be partitioned into clusters. Concerning the question of data availability/data quality, we performed an additional analysis of the proposed AutoTiC-NN methodology, based on the Google search queries data (i.e., Google Dengue Trend) instead of the official data of the dengue incidences. Our analysis produced similar conclusions on the predictive performance of AutoTiC-NN compared to the benchmark methods (the results are available from the authors upon request). Since Google search queries data are available wherever Google is in use, we believe that these findings offer a reasonable prospective to employ AutoTiC-NN for predicting dengue in other geographical regions. If Google is not in use, alternative web queries can be utilized instead, similarly to the Baidu flu queries which are used in China (for more details see, e.g., [42], [43]). A related topic is how well the data can be partitioned into clusters changing the geographic setting. This is more challenging and largely depends on local socio-demographic specifics – thereby requiring standalone validation for every geographical region. We leave this extension as a future research task. In any case, in terms of directions for future research, we think that new developments can also be fulfilled in this research topic by capitalizing on the recent achievements of transfer learning [44] in tasks of spatial and spatio-temporal prediction [45], [46]. A transfer learning method, specifically designed for the trend-based temporal clusters, discovered through the multi-stage machine learning methodology of AutoTiC-NN, may also allow us to transfer a cluster model, learned in a geographical area with adequate data, to a new area with few data.

Furthermore, motivated by the increasing interest in applications of deep learning in biosurveillance, we plan to explore the utility of deep learning in modelling and predicting the spread of dengue and other emerging climate-sensitive mosquito borne diseases. Indeed, recent research in remote sensing has highlighted the potential of convolutional learning in extracting spatio-temporal features (see, e.g., [47] and references therein). Following the same research direction, [48] have achieved promising results in spatio-temporal feature extraction by combining convolution neural networks and long short-term memory, and improving the accuracy of both classification and regression tasks. An interesting research direction is the tracking of trends in features extracted in massive data scenarios with sophisticated deep learning architectures and exploiting these trends to empower the accuracy of the considered prediction task.

We also consider the opportunity of developing an approach to forecast dengue spread via only the temperature data. In principle this is possible, but not via the verbatim application of AutoTiC-NN. It requires developing a connection between clustering dynamics and shape patterns of temperature and the dynamics of dengue spread. We are currently exploring this direction using topological data analysis tools.

Finally, we intend to explore the effectiveness of other algorithms for time series clustering such as the algorithms experimented in [23] and [24].

ACKNOWLEDGMENT

The authors would like to thank Lynn Rudd for her help in proof reading the manuscript and Mauricio Santillana for providing the official dengue incidence data.

REFERENCES

[1] World Health Organization. (2009). *Dengue: Guidelines for Diagnosis, Treatment, Prevention and Control*. [Online]. Available: https://apps.who.int/iris/bitstream/handle/10665/44188/9789241547871_eng.pdf;jsessionid=CC50DEFFCE9B8F70CB41F0B22A6661FC?sequence=1,2019-01-29

[2] J. M. Brunkard, E. Cifuentes, and S. J. Rothenberg, "Assessing the roles of temperature, precipitation, and ENSO in dengue re-emergence on the Texas-Mexico border region," *Salud Pública de México*, vol. 50, no. 3, pp. 227–234, 2008.

[3] E. H. Chan, V. Sahai, C. Conrad, and J. S. Brownstein, "Using Web search query data to monitor dengue epidemics: A new model for neglected tropical disease surveillance," *PLoS Neglected Tropical Diseases*, vol. 5, no. 5, p. e1206, 2011.

[4] F. J. Colón-González, G. Bentham, and I. R. Lake, "Climate variability and dengue fever in warm and humid mexico," *Amer. J. Tropical Med. Hygiene*, vol. 84, no. 5, pp. 757–763, May 2011.

[5] H. G. Dantés, J. A. Farfán-Ale, and E. Sarti, "Epidemiological trends of dengue disease in Mexico (2000–2011): A systematic literature search and analysis," *PLoS Neglected Tropical Diseases*, vol. 8, no. 11, p. e3158, 2014.

[6] F. J. Díaz, W. C. Black, J. A. Farfán-Ale, M. A. Loroño-Pino, K. E. Olson, and B. J. Beaty, "Dengue virus circulation and evolution in Mexico: A phylogenetic perspective," *Arch. Med. Res.*, vol. 37, no. 6, pp. 760–773, Aug. 2006.

[7] R. T. Gluskin, M. A. Johansson, M. Santillana, and J. S. Brownstein, "Evaluation of Internet-based dengue query data: Google dengue trends," *PLoS Neglected Tropical Diseases*, vol. 8, no. 2, p. e2713, 2014.

[8] M. A. Johansson, D. A. Cummings, and G. E. Glass, "Multiyear climate variability and dengue—El Nino southern oscillation, weather, and dengue incidence in Puerto Rico, Mexico, and Thailand: A longitudinal data analysis," *PLoS Med.*, vol. 6, no. 11, 2009, Art. no. e1000168.

[9] M. Oki and T. Yamamoto, "Climate change, population immunity, and hyperendemicity in the transmission threshold of dengue," *PLoS ONE*, vol. 7, no. 10, 2012, Art. no. e48258.

[10] R. A. Strauss, J. S. Castro, R. Reintjes, and J. R. Torres, "Google dengue trends: An indicator of epidemic behavior. The venezuelan case," *Int. J. Med. Informat.*, vol. 104, pp. 26–30, Aug. 2017.

[11] A. Husnayain, A. Fuad, and L. Lazuardi, "Correlation between Google trends on dengue fever and national surveillance report in indonesia," *Global Health Action*, vol. 12, no. 1, Jan. 2019, Art. no. 1552652.

[12] H. T. Ho, T. M. Carvajal, J. R. Bautista, J. D. R. Capistrano, K. M. Viacrusis, L. F. T. Hernandez, and K. Watanabe, "Using Google trends to examine the spatio-temporal incidence and behavioral patterns of dengue disease: A case study in Metropolitan Manila, Philippines," *Tropical Med. Infectious Disease*, vol. 3, no. 4, p. 118, 2018.

[13] M. A. Johansson, N. G. Reich, A. Hota, J. S. Brownstein, and M. Santillana, "Evaluating the performance of infectious disease forecasts: A comparison of climate-driven and seasonal dengue forecasts for Mexico," *Sci. Rep.*, vol. 6, no. 1, p. 33707, Dec. 2016.

[14] G. Chowell and F. Sanchez, "Climate-based descriptive models of dengue fever: The 2002 epidemic in Colima, Mexico," *J. Environ. Health*, vol. 68, no. 10, p. 40, 2006.

[15] I. R. Iliev, X. Huang, and Y. R. Gel, "Political rhetoric through the lens of non-parametric statistics: Are our legislators that different?" *J. Roy. Stat. Soc., A, Statist. Soc.*, vol. 182, no. 2, pp. 583–604, Feb. 2019.
[16] X. Huang, I. R. Iliev, V. Lyubchich, and Y. R. Gel, "Riding down the bay: Space-time clustering of ecological trends," *Environmetrics*, vol. 29, nos. 5–6, p. e2455, Aug. 2018.
[17] A. Appice, P. Guccione, D. Malerba, and A. Ciampi, "Dealing with temporal and spatial correlations to classify outliers in geophysical data streams," *Inf. Sci.*, vol. 285, pp. 162–180, Nov. 2014.
[18] A. Appice, A. Ciampi, and D. Malerba, "Summarizing numeric spatial data streams by trend cluster discovery," *Data Mining Knowl. Discovery*, vol. 29, no. 1, pp. 84–136, Jan. 2015.
[19] S. Pravilovic, M. Bilancia, A. Appice, and D. Malerba, "Using multiple time series analysis for geosensor data forecasting," *Inf. Sci.*, vol. 380, pp. 31–52, Feb. 2017.
[20] D. F. Silva, V. M. A. De Souza, and G. E. A. P. A. Batista, "Time series classification using compression distance of recurrence plots," in *Proc. IEEE 13th Int. Conf. Data Mining*, Dec. 2013, pp. 687–696.
[21] A. Mueen, E. Keogh, and N. Young, "Logical-shapelets: An expressive primitive for time series classification," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2011, pp. 1154–1162.
[22] C. Euan, H. Ombao, and J. Ortega, "Spectral synchronicity in brain signals," 2015, *arXiv:1507.05018*. [Online]. Available: http://arxiv.org/abs/1507.05018
[23] M. R. Desjardins, A. Whiteman, I. Casas, and E. Delmelle, "Space-time clusters and co-occurrence of chikungunya and dengue fever in Colombia from 2015 to 2016," *Acta Tropica*, vol. 185, pp. 77–85, Sep. 2018.
[24] A. Hohl, E. Delmelle, W. Tang, and I. Casas, "Accelerating the discovery of space-time patterns of infectious diseases using parallel computing," *Spatial Spatio-Temporal Epidemiol.*, vol. 19, pp. 10–20, Nov. 2016.
[25] M. D. Eastin, E. Delmelle, I. Casas, J. Wexler, and C. Self, "Intra-and interseasonal autoregressive prediction of dengue outbreaks using local weather and regional climate for a tropical environment in Colombia," *Amer. J. Tropical Med. Hygiene*, vol. 91, no. 3, pp. 598–610, Sep. 2014.
[26] G. Sánchez-González, R. Condé, R. N. Moreno, and P. C. L. Vázquez, "Prediction of dengue outbreaks in Mexico based on entomological, meteorological and demographic data," *PLoS ONE*, vol. 13, no. 8, Aug. 2018, Art. no. e0196047.
[27] Mexico Secretariat of Health. (2013). *Anuarios de Morbilidad*. [Online]. Available: http://www.epidemiologia.salud.gob.mx/anuario/html/anuarios.html
[28] S. Yang, S. C. Kou, F. Lu, J. S. Brownstein, N. Brooke, and M. Santillana, "Advances in using Internet searches to track dengue," *PLOS Comput. Biol.*, vol. 13, no. 7, pp. 1–14, Jul. 2017.
[29] GOB.MX. (2017). *Temperatura Promedio de Conagua*. Accessed: Jan. 29, 2019. [Online]. Available: https://datos.gob.mx/busca/dataset/temperatura-promedio-excel
[30] D. Charte, F. Charte, S. García, M. J. del Jesus, and F. Herrera, "A practical tutorial on autoencoders for nonlinear feature fusion: Taxonomy, models, software and guidelines," *Inf. Fusion*, vol. 44, pp. 78–96, Nov. 2018.
[31] D. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," 2015, *arXiv:1511.07289*. [Online]. Available: https://arxiv.org/abs/1511.07289
[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980
[33] W. Han, G. Wang, and K. Tu, "Latent variable autoencoder," *IEEE Access*, vol. 7, pp. 48514–48523, 2019.
[34] M. M. Gaber, A. Zaslavsky, and S. Krishnaswamy, "Mining data streams: A review," *ACM SIGMOD Rec.*, vol. 34, no. 2, pp. 18–26, 2005.
[35] H. Alt, "The nearest neighbor," in *Computational Discrete Mathematics: Advanced Lectures*, H. Alt, Ed. New York, NY, USA: Springer, 2001, ch. 2, pp. 13–24.
[36] R. J. Hyndman and Y. Khandakar, "Automatic time series forecasting: The forecast package for R," *J. Stat. Softw.*, vol. 26, no. 3, pp. 1–22, 2008.
[37] R. S. Tsay, *Multivariate Time Series Analysis: With R and Financial Applications*. Hoboken, NJ, USA: Wiley, 2014.
[38] Y. Wang and I. Witten, "Induction of model trees for predicting continuous classes," in *Proc. 9th Eur. Conf. Mach. Learn.*, Prague, Czech Republic: Univ. Economics, 1997, pp. 128–137.
[39] M. Awad and R. Khanna, *Support Vector Regression*. Berkeley, CA, USA: Apress, 2015, pp. 67–80.
[40] M. Zhao and J. Chen, "Improvement and comparison of weighted k nearest neighbors classifiers for model selection," *J. Softw. Eng.*, vol. 10, no. 1, pp. 109–118, Jan. 2016.
[41] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Technique* (The Morgan Kaufmann Series in Data Management Systems), 3rd ed. Boston, MA, USA: Morgan Kaufmann, 2011.
[42] Q. Yuan, E. O. Nsoesie, B. Lv, G. Peng, R. Chunara, and J. S. Brownstein, "Monitoring influenza epidemics in China with search query from Baidu," *PLoS ONE*, vol. 8, no. 5, 2013, Art. no. e64323.
[43] Q. Xu, Y. R. Gel, L. L. R. Ramirez, K. Nezafati, Q. Zhang, and K.-L. Tsui, "Forecasting influenza in Hong Kong with Google search queries and statistical model fusion," *PLoS ONE*, vol. 12, no. 5, 2017, Art. no. e0176690.
[44] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.
[45] M. Bussas, C. Sawade, N. Kühn, T. Scheffer, and N. Landwehr, "Varying-coefficient models for geospatial transfer learning," *Mach. Learn.*, vol. 106, nos. 9–10, pp. 1419–1440, Oct. 2017.
[46] L. Wang, X. Geng, X. Ma, F. Liu, and Q. Yang, "Cross-city transfer learning for deep spatio-temporal prediction," in *Proc. 28th Int. Joint Conf. Artif. Intell. (IJCAI)*, S. Kraus, Ed., 2019, pp. 1893–1899.
[47] O. Costilla-Reyes, P. Scully, and K. B. Ozanyan, "Deep neural networks for learning spatio-temporal features from tomography sensors," *IEEE Trans. Ind. Electron.*, vol. 65, no. 1, pp. 645–653, Jan. 2018.
[48] H. Qiao, T. Wang, P. Wang, S. Qiao, and L. Zhang, "A time-distributed spatiotemporal feature learning method for machine health monitoring with multi-sensor time series," *Sensors*, vol. 18, no. 9, p. 2932, 2018.

ANNALISA APPICE received the Ph.D. degree in computer science from the University of Bari Aldo Moro. She was a Visiting Researcher with the University of Bristol, U.K., and also with the Jozef Stefan Institute, Slovenia. She is currently an Associate Professor with the Department of Computer Science, University of Bari Aldo Moro, Italy. She is a member of the "Consorzio Interuniversitario Nazionale per l'Informatica" (CINI). Her current research interests include data mining with spatio-temporal data, data streams, and event logs with applications to remote sensing, process mining, and cybersecurity. On these topics, she has published more than 135 articles in international journals and conferences. She is a member of the editorial board of *Data Mining and Knowledge Discovery* (DAMI) and the *Journal of Intelligent Information Systems* (JIIS). She has been Program Co-Chair of ECML-PKDD 2015 and ISMIS 2017. She has participated in the organization (as co-chair) of several workshops on the topics machine learning and data mining. She is the responsible of research units in two national projects (topic on remote sensing). She is serving/has served in the program committee for several international/national conferences.

YULIA R. GEL received the Ph.D. degree in mathematics from the University of Washington. She held a postdoctoral position in statistics at the University of Washington. She was a tenured Faculty Member with the University of Waterloo, Canada. She held visiting positions at Johns Hopkins University, the University of California at Berkeley, Berkeley, and the Isaac Newton Institute for Mathematical Sciences, Cambridge University, U.K. She is currently a Professor with the Department of Mathematical Science, The University of Texas at Dallas. Her research interests include statistical foundation of data science, inference for random graphs and complex networks, time series analysis, and predictive analytics. She is a Fellow of the American Statistical Association. She served as a Vice President of the International Society on Business and Industrial Statistics.

ILIYAN ILIEV received the Ph.D. degree in political science with a focus on data analytics from The University of Texas at Dallas. He held a visiting position at The University of Texas at Dallas. He is currently an Assistant Professor of political science with The University of Southern Mississippi. His research focuses on the various expressions of political behavior and the development of novel research methods to study such behavior. He specializes in natural language processing, time series analysis, and Bayesian analysis.

VYACHESLAV LYUBCHICH received the Ph.D. degree in statistics from Orenburg State University, Russia. He held postdoctoral positions at the Department of Statistics and Actuarial Science, University of Waterloo, Canada as a Government of Canada Postdoctoral Fellow and also with the Department of Mathematical Sciences, The University of Texas at Dallas, USA as a Visiting Scholar. He is currently an Assistant Research Professor with the University of Maryland Center for Environmental Science, USA.

DONATO MALERBA received the M.Sc. degree in computer science from the University of Bari, Italy, in 1987. He is currently a Full Professor with the Department of Computer Science, University of Bari Aldo Moro. He is also the Director of the Computer Science Department, University of Bari, and of the CINI Lab on Big Data. He has published more than 300 articles in international journals and conference proceedings. His research interests include machine learning, data mining, big data analytics, and their applications. He is on the Editorial Board of several international journals. He has been responsible for the local research unit of several European and national projects, and received an IBM Faculty Award in 2004. He has been in the Board of Directors of the Big Data Value Association and in the Partnership Board of the PPP Big Data Value. He was a Program Co-Chair of the International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems (IEA-AIE), in 2005, the International Symposium on Methodologies for Intelligent Systems (ISMIS), in 2006, SEBD 2007, the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPKDD), in 2011, and was the General Chair of ALT/DS 2016.