Intelligent Epidemiological Surveillance in the Brazilian Semiarid

omitted to revision

***Abstract*—Right after the Chinese example in conducting COVID-19 epidemic originated in Wuhan, the readiness to detect and respond by health authorities to local (sometimes global) epidemics has become central lately. Within the idea of health 4.0, information about the individual is essential in supporting public community health policies. This paper presents a proposal for an epidemiological surveillance system applied to arboviruses. Data mining techniques and Machine Learning (ML) are used to design mathematical models for detecting epidemics enhanced by *Aedes Aegypti* (vector for dengue, chikungunaya, yellow fever and zica). A Prove of Concept (PoC) is presented for dengue epidemics detection, a common endemic disease in the semiarid region of Brazil.**

***Index Terms*—Brazilian health data, arbovirus, CRISP-DM, data mining, public health system.**

I. Introduction

The 2020 year was decisive for bring epidemiology terms to lives of bilions around the globe. A viral epidemic started in Wuhan, a province located in central China, changed the routine and influences the human behavior in large scale. The corona virus (COVID-19 [1]) was detected by Chinese health authorities and triggered a combination of actions that contained contamination within the country, so far totalling five thousand dead according estimatives [2]. Despite the warning from Chinese authorities, world was not prepared to contain what is today the biggest epidemic in human history.

The implications of an epidemic/pandemic has different dimensions. The bigger and more deadly it is, the more damaging it has proved to be for individuals and their relationships, with profound symptoms in the economy. Mechanisms that produce a local epidemic outbreak, which can evolve to epidemic, or even extrapolate country borders and cause a pandemic, has been extensively studied. There are tools in the prescription of good practices that allow infectious disease containment, ranging from social distancing to lock-downs in more serious interventions.

In countries that has a semi-arid climate, arboviruses like dengue, chikungunya and zica may become endemic, maintaining a minimum level of cases per year that could evolve into an epidemic. Because that Ministério da Saúde do Brasil (MSB) track the mosquito proliferation in urban areas using a system called Levantamento Rápido de Índices para Aedes Aegypti (LIRAa). Another Brazilian government tool is the Sistema de Informação de Agravos de Notificação (SINAN), a system which gather information about those infected.

Brazilian health authoroty Sistema Único de Saúde (SUS) proposed at the end of 2019 the Rede Nacional de Dados em Saúde (RNDS). This is a health data integration initiative and is part of the Connect SUS program developed by Departamento de Informática do SUS (DATASUS) which aims to create a platform to promote advances in public health care. The pilot project is being implemented in the state of Alagoas. They aim that each federation state has a similar system in the next few years [3].

The architecture defined by Connect SUS includes a module that promotes intelligence in healthcare. Inserted in this proposal the system **omitted to revision** (**otr**^{TM}) owned by **omitted to revision**^{TM} company already employs artificial intelligence analysing risk in maternal/infant scope, supporting individualized care actions [4]. A new tool that is central theme in this work is the prediction of arboviruses cases. This enables forecast epidemics in advance so that actions in public health can be taken in both front: against the vector mosquito and preparing the health network to serve the population efficiently.

This paper presents the concept of epidemiological surveillance employed in **otr**^{TM}. It uses ML techniques to quantitatively predict the number of infected people to be confirmed by SINAN. Additionally, an Application Programming Interface (API) developed with the Representational State Transfer (REST) concept used by the **otr**^{TM} portal is presented. The epidemiological surveillance model is part of a set of artificial intelligence services available to partner companies.

Section II presents the approach used in modeling and forecasting epidemics applied by related works, section III reviews some important concepts in problem analysis, section IV relates the phases of the project and ML models evaluation, section V discuss the surveilance system results and finally in section VI higlights applications, limitations and future works.

II. Related Work

Since it is a stochastic processes, we naturally report several factors external to arbovirus epidemics, especially meteorological factors that directly influence disease vector (*Aedes Aegypti*). Naturally it is important to add meteorological information to the model [5].

To predict a given event (outbreak or epidemic), the Susceptible, Infected, Recovery (SIR) [6], [7] model is commonly employd. Some restrictions are inherent to the model, such as knowing the amount of susceptible people, infection rate and amount of recovereds. In another analysis [8], logistic function is used to approximate the accumulated infected curve behavior (Figure 1, cummulative cases) and to predict infection rates. Both methods requiring a certain amount of week measurements to present a consistent result. Then, deriving

to obtain the number of new cases per week (Figure 1, cases per week).

In [9] is developed a study correlating population data, number of infected and meteorology in different regions of Colombia: island and continent. A sister country and partner in Latin America, it presents a climate and conditions quite different from those found in the Brazilian Northeast, especially the backwoods climate. The study proposes a comparison between the Multilayer Perceptron (MLP) and Random Forest (RF) model highlighting the potential of using the latter one given its simplicity. The study proposes 12 models to predict the cumulative incidence in the following months. Based on the chosen inputs measurements, the results using national data proved to be more interesting with less Mean Absolute Error (MAE) compared to specific models generated from local data. Additionally, it was identified that socio-demographic variables such as the Gini coefficient has some effect on the forecasting and highlights opportunities in the application of Recurrent neural network (RNN).

III. Implementation Aspects

The idea that the accumulated cases infection curve follows the logistic function is quite experienced in literature [8]. The first derivative then represents the amount of infected per time. The second derivative shows new cases rate. This last measure is key in the process of identifying the epidemic fase in order to produce more appropriate forecasts.

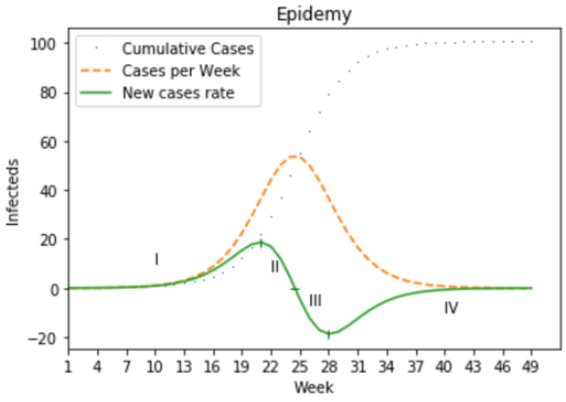

Fig. 1. **Cumulative cases**, **cases per week** and **new cases rate** behaviour.

The SIR or even the Susceptible, Exposed, Infected, Recovery (SEIR) model identifies exposure rates, both infected and recovered, and predicts the evolution of an epidemic based on the population considered susceptible to infection. In another approach, this work aims to identify the number of cases of infection in the near future and to adapt the predicted infected curve based on simple regression (equation 1).

Considering outbreaks or epidemics with a minimum duration of 7 and a maximum duration of 60 weeks. Given the average duration of the events being 30 weeks, we opted for 10 MLP models to predict the amount of infected in the weeks $1, 2, 3, 4, 5, 6, 7, 8, 9$ and 10 ahead . During the experiments it became clear a trade off between choosing many models for prediction and the resulting MAE. The more forward the forecast, the lower the observed accuracy and instability as well.

After forecast results given 10 subsequents weeks it's performed a simple regression using *curve_fit()* [10] method applied to the derivative of the logistic function in (1).

$$f(x) = L \times \frac{1}{1 + e^{-k(x-x_p)}}$$

$$f'(x) = Lk \times \frac{e^{-k(x-x_p)}}{(1 + e^{-k(x-x_p)})^2} \quad (1)$$

Where x is the week, x_p is the week where the registered cases peak occurs, L is the total number of infecteds accumulated for the analyzed event and k is proportional to the infection rate.

The model including 10 weeks forecast shows to be more stable empirically. Therefore, when two epidemics caused by two different strains are ongoing simultaneously it suppress curve of those infected by a minor event (outbreak or epidemic). In this scenario a solution comes analysing new cases rate (Figure 1, green line). Every regular event follows these phases:

I. Exponential growth in new cases rate ending up in maximum positive;
II. Accelerated reduction in new cases rate up to zero;
III. New cases rate inversion at a fast pace ending up in minimum (negative);
IV. Exponential reduction in number of new cases up to zero;

Each phase suggests to the epidemiological surveillance model a set of predictions sufficient for the current event, not necessarily a fixed set of predictions.

IV. Methodology

The prediction model for epidemiological surveillance applied to dengue follows two stages: (1) prediction of the future weekly incidence and (2) adaptation of this data to the function derived from the logistic function (equation 1). Since the project workflow focus on the data mining process from data collection to the deployment, it is applied Cross Industry Standard Process for Data Mining (CRISP-DM) [11] methodology.

A. Business Understanding

Arboviruses are spread by mosquitoes such as *Aedes Aegypti* and depends on existence of circulating virus and environmental conditions for mosquito's proliferation. The first condition is given by urban agglomerations, how larger it is implies in virus circulating as an endemic disease. The second refers to climatic factors as the accumulation of clean and still water (key condition for mosquito proliferation) multiplies throughout the city on rainy days. Temperatures between $25°C$ and $30°C$ further favor the insect's eggs hatching, a condition easily found in the Brazilian semiarid region.

The information related to arboviruses applied in this project is maintained and made available by Instituto Nacional de Meteorologia (INMET) [12], Instituto Brasileiro de Geografia e Estatistica (IBGE) [13] and SINAN [14]. They are part of the information structure of the Brazilian federal government. The INMET is responsible by maintain and collect meteorological information from bases spread throughout the national territory, IBGE relates population characteristics in each region and SINAN makes available the quantity infected by tracking compulsory notifications.

In this work is collected data from Fortaleza, Caucaia, Quixeramobim, Pacatuba, Sobral and Tauá cites, in Ceará state. In all, 65 events were observed among outbreaks and epidemics distributed over $27,895$ samples. The measurements were performed daily between January 2007 and December 2019. Only the municipality of Tauá didn't have complete data in the analysed period.

B. Data Preparation

After agregate data it is performed Exploratory Data Analysis (EDA) to verify the generated datasets. It allows verifying data variables quality by graphics visualization and statistical measurements. All these actions aim to prevent biased and overfitting models.

In the data extraction phase some aggregation measures were performed such as arithmetic averages and accumulations in the last 7, 14, 21, 28 and 35 days. These variables allow the model to infer context more accurately. In this perspective due usage of deterministic methods it's important to add information so that the prediction process occurs with minimum MAE. Based on correlation matrix it's manually chosen which measures best represent the context avoiding redundancy and information loss. After repeated experiments and evaluations considering last 7 days Simple Moving Average (SMA) arrives at the final set of *features*.

- F_1: Infecteds SMA;
- F_2: Weekly cumulative raining precipitation (mm);
- F_3: Weekly cumulative evaporation (mm);
- F_4: Daily maximum temperature SMA (Celsius);
- F_5: Daily average temperature SMA (Celsius);
- F_6: Daily minimum temperature SMA (Celsius);
- F_7: Daily insolation SMA (hours);
- F_8: Daily wind speed SMA (m/s);
- F_9: City population;
- F_{10}: City population density.

When experiencing the correlation between the inputs and the target variables (Figure 2), it is noticed that the more weeks ahead is predicted the lower general correlation between features and target is observed. This follows the empirical observation that the transmitting mosquito requires clean water for reproduction. Predicting cases of dengue in the Brazilian semi-period requires indirectly to predict rain (still water environments) and the existence of infected people in the region for mass contamination to occur.

Analysis points out that the presence of infected in the previous weeks correlates with the number of infected in the future, but this is diluted over the weeks. However the size of the population and the demographic density have a greater influence. In larger populations, represented here by samples from Fortaleza, the virus remains in circulation in a phenomenon known as the minimum plateau of cases. Precipitation has a positive correlation with the future cases, however it reduces importance compared to population and demographic density variables due to this endemic behavior.

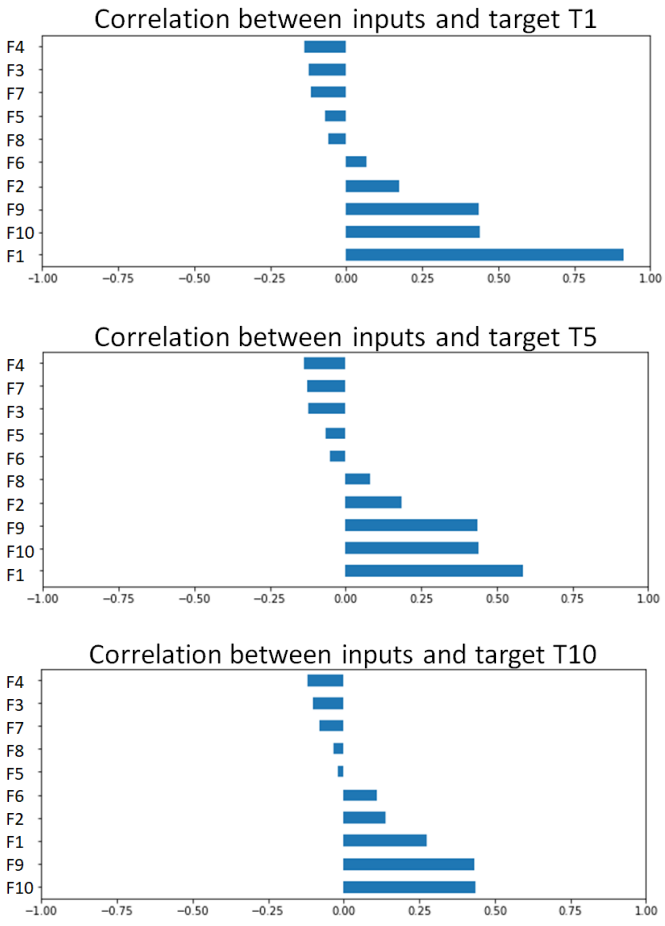

Fig. 2. Correlations considering targets for 1, 5 and 10 weeks ahead.

To train supervised machine learning algorithms samples are divided into three groups: training (64%), validation (14%) and testing (20%). For each group to be statistically similar it is separate the samples into two groups: infecteds and normals. As a cutoff point it's adopted that above 5 dengue infection cases accumulated in the last 7 days there is an event in progress otherwise it is considered to be normal samples. Table I lists the number of samples for each city participating in the research. Infected samples are shuffled and separated keeping the ratio of 64%, 16% and 20%, the same is done for the normal sample set resulting in three sets of data according to Table II.

After this phase, each dataset is standardized with *StandardScaler* class, available at *Scikit-Learn* library [15]. This oper-

TABLE I
DATASETS COMPOSITION

Cities	Infected Samples	Normal Samples
Fortaleza	4,707	41
Caucaia	1,638	3,110
Sobral	844	3,904
Quixeramobim	617	4,131
Pacatuba	620	4,128
Tauá	1,019	3,136
Total	**9.445**	**18.450**

TABLE II
TRAIN, VALIDATION AND TEST SPLIT

Set	Infected	Normal
Train (64%)	6,045	11.808
Validation (16%)	1,511	2,952
Test (20%)	1.889	3,690

ation results in zero mean and scaling data to unit variance, considering each feature separately in all samples selected. Each value is scaled by the expression: $x_{scaled} = (x - \mu)/\sigma$, where μ and σ represent, respectively, the mean and standard deviation for a given feature in dataset.

C. Modeling

Some models of ML were analyzed for regression, such as Logistic Regression (LR), Support Vector Machine for Regression (SVR) and a MLP. Model based on this last algorithm class showed the best sensitivity and the lowest MAE when evaluated by the validation and test sets.

Different MLP architecture was formulated and the one that offered the best results has an input layer with 45 neurons, a hidden layer with 45 neurons and an output layer with 1 neuron. All layers are densely connected and weights are randomized initialized with uniform distribution. The activation function broadly used in network is Rectified Linear Units (ReLU). The optimizer chosen is Adaptive Moment (Adam) and MAE as loss function. For each model is performd 500 training epochs evaluating 10 samples before adjusting weights and bias ending each training step. At the end of each epoch the model accuracy is assessed using data from booth training and validation set by evaluating test dataset once.

D. Evaluation

The same metric applied in training was used in models evaluate. All 10 models were trained and evaluated, however Figure 3 demostrate results for training (blue line) and validation (orange line) only for the 1, 5 and 10 weeks forecast models. After training fase, each final model is submitted to the test set, the result (single value) is represented by the red line.

It is noticed that the prediction models lose accuracy as it is intended to predict the behavior of the number of infected more weeks ahead. For 1 week, graphs show MAE of 3.18 infected by prediction, compared to 4.07 (5 weeks) and 5.62 (10 weeks).

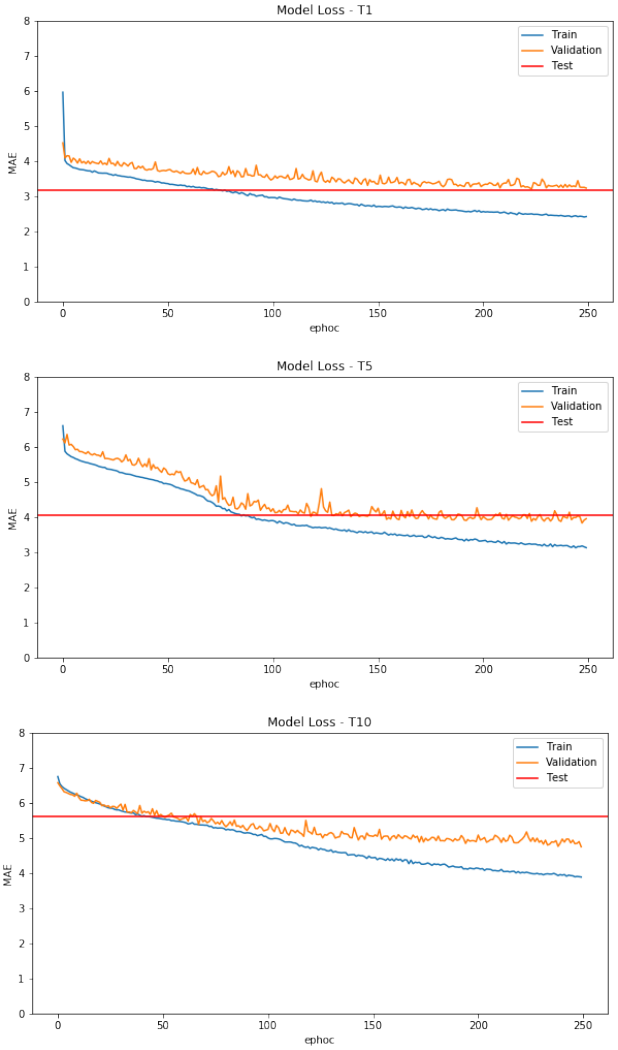

Fig. 3. Train and evaluation for 1, 5 and 10 weeks ahead for small cities (population up to 150 thousands).

Another aspect observed is that the model loses the ability to predict small events (below 60 cases daily) insofar as data from cities with different population quantities are used. Samples of infection events in Fortaleza (2.7 Million) are easily much larger than those in other cities such as Caucaia (361 thousand), Sobral (210 thousand), Pacatuba (83 thousand) or Tauá (59 thousand) [13]. MAE results for 1 week is 20.35 infected by prediction made, 31.02 (5 weeks) and 45.39 (10 weeks). Due this, two models are proposed: small cities (less than 150 thousand inhabitants) and large cities.

E. Deployment

The epidemiological surveillance model was designed to be included as a prediction and alert services within the **otr**^{TM} [16] platform. This web system is a tool developed by **omitted to revision**^{TM} for intelligent governance in public health systems. It consists of a collection of components that

allow integration and relevant information visualization to the decision-making process of health authorities in the different levels of government (local and state). Figure 4 presents the conceptual screen of the epidemiological surveillance service for dengue applied to Fortaleza given real current measurements.

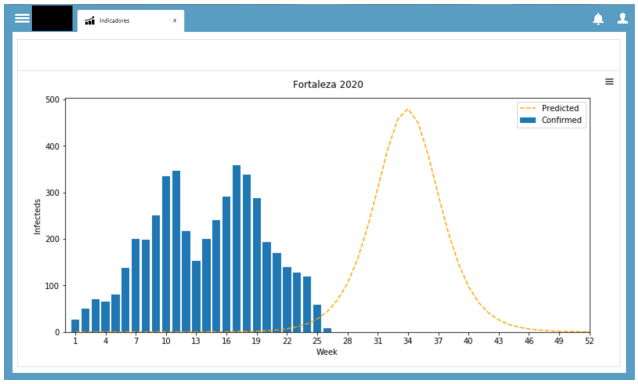

Fig. 4. PoC - 52 Week Graph for Dengue in Fortaleza (model for big cites).

V. Results

The epidemiological surveillance model applied to dengue suitable for warning health authorities. However, from the point of view of temporality, i.e., how many weeks before the epidemic is modeled accurately, depends on duration. To calculate the estimate model generalization, it's used the R^2 score [10]. Taking the example of an epidemic in the city of Sobral between the epidemiological weeks 10 and 35 in 2007 ($4,434$ infected), Figure 5 shows the score R^2 calculated for each week since the event beginning.

Each line in the graph corresponds to the adequacy of the forecast made using a subset of predictions. $MLP-10$ includes all of the 10 weeks predicted ahead, $MLP-9$ only 9 weeks ahead and so on to the dashed line $curve\ _fit$, which uses only the regression based function infected people registered until date. It is observed that the model using predictions of infected 10 weeks ahead is able to adjust to the epidemic after a week even with a small number of confirmed cases.

The second epidemic occurred in Fortaleza between 2 and 33 of 2008, totaling $67,857$ infecteds. In this event, the model takes 7 weeks to adjust to the actual infected curve (above 80% of the explained variance), as shown in Figure 5. This stems from the fact that this epidemic is more lasting and the peak will be reached more weeks ahead than the model is capable of evaluating. Another important note is that after a certain phase of the epidemic (near the peak, phases II and III) the number of weeks ahead can be reduced, maintaining the quality of the final forecast. This approach is essential to avoid shading minor events, when two epidemics occur very close together, sometimes overlapping.

Still using the epidemic that occurred in Fortaleza at 2008, Figure 6 shows the evolution of the model from the week

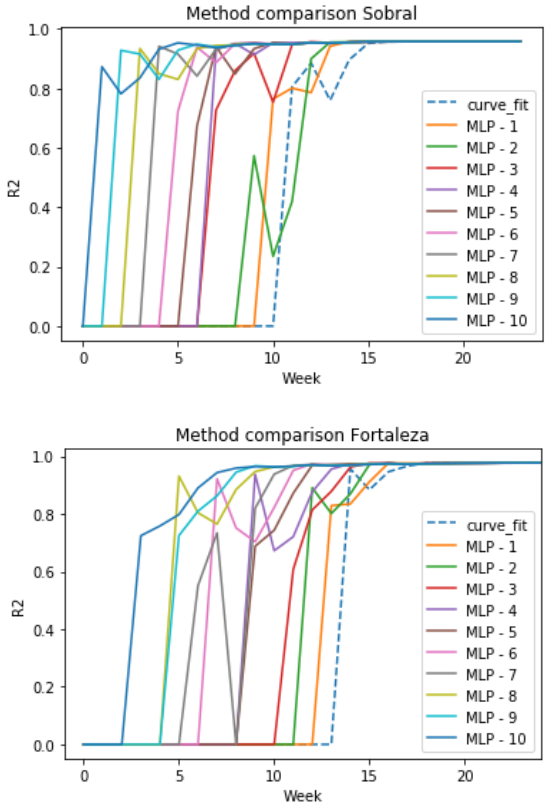

Fig. 5. Gráfico R^2 em semanas.

4. It is clear that the model detects the peak and the end of the epidemic, but did not predict the correct intensity. Advancing until the tenth week the model is already able to safely predict the total number of cases and the evolution of the epidemic until the end. In other words, the model predicts an upward trend in the number of new weekly cases, but it needs confirmed data to improve complete epidemic prediction, a fact observed between 7 and 10 weeks before the peak.

VI. Conclusion

It's proposed a matematical model supported by ML as a service for predicting events of population infection. It was evident that the ML models created showed different behaviors given the size of the local population. It is observed that cities with more than 150 thousand inhabitants in the semiarid region of Brazil experience greater epidemics and the forecast errors are proportionally greater. This evidence may indicates that models with greater granularity dividing large cities into neighborhoods may incur more accurate models.

Extending the number of predicted infected individuals to more than 10 weeks incurs instability and event shadowing when two epidemics of different sizes occur with peaks of infection nearby. This behavior is controlled by reducing the number of models for predicting the incidence of infection according epidemic stages: 10 weeks forecast for I, II and IV phases; 5 weeks forecast for III phase. New steps points

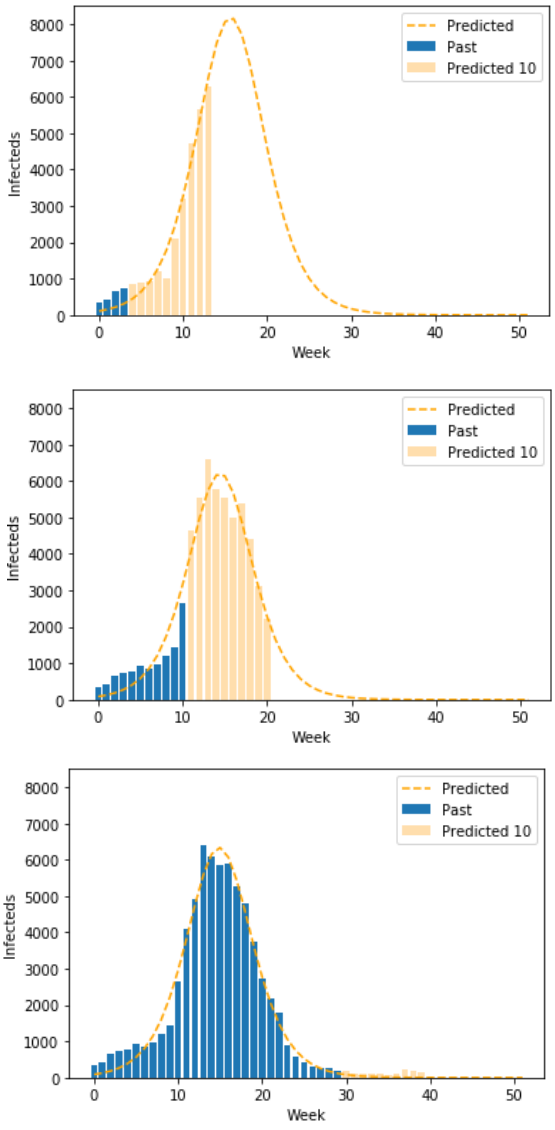

Fig. 6. Epidemic forecast evolution to Fortaleza in 2008.

out to apply the model for other arboviruses and carry out more analysis to create an adaptive model.

It should also be noted that this technique can be quickly adapted to other events of arboviruses such as chikungunaya, malaria, west Nile virus, yellow fever and zica, as it has the same mechanism of transmission through the vector mosquito. Additionally, mathematical modeling can be adapted to any epidemic (tuberculosis, cholera, COVID- 19, etc.), just by choosing variables that correlate with presence of the circulating virus and contamination mechanism.

This work is planned within a proposal of architecture of health intelligence using micro services to provide governance tools for systems of public health promotion and individual care. Future works include expanding the restful API system adding new predict models in Surveillance for Epidemics and risk analisys.

ACKNOWLEDGEMENT

Our sincere acknowledgements to the Financiadora de Estudos e Projetos (FINEP) for funding this research. We also thank the **omitted to revision**^{TM} team for provide access to data and support from IT technicians and health care specialists.

REFERENCES

[1] D. Benvenuto, M. Giovannetti, A. Ciccozzi, S. Spoto, S. Angeletti, and M. Ciccozzi, "The 2019-new coronavirus epidemic: evidence for virus evolution," *Journal of Medical Virology*, 1 2020.

[2] J. H. University. (2020) Johns hopkins coronavirus resource center. [Online]. Available: https://coronavirus.jhu.edu/map.html

[3] M. da Saúde do Basil. (2020) Conecte sus avança em todo país com a implantação da rede nacional de dados em saúde. [Online]. Available: https://www.saude.gov.br/noticias/agencia-saude/46988-conecte-sus-avanca-em-todo-pais-com-a-implantacao-da-rede-nacional-de-dados-em-saude

[4] R. Valter, S. Santiago, R. Ramos *et al.*, "Data mining and risk analysis supporting decision in brazilian public health systems," in *IEEE International Conference on E-health Networking, Application Services (HealthCom)*. Bogotá, Colombia: IEEE, Oct. 14–16 2019, pp. 1–6.

[5] W. L. F. A. Correia Filho, "Influence of meteorological variables on dengue incidence in the municipality of Arapiraca, Alagoas, Brazil," *Revista da Sociedade Brasileira de Medicina Tropical*, vol. 50, pp. 309 – 314, 06 2017.

[6] M. A. G. Kermack William Ogilvy and W. G. Thomas, "A contribution to the mathematical theory of epidemics," *Proceedings of the Royal Society*, p. 700 – 721, 08 1927. [Online]. Available: https://royalsocietypublishing.org/doi/10.1098/rspa.1927.0118

[7] D. Smith and L. Moore, "The SIR Model for Spread of Disease - The Differential Equation Model," *The Journal of Online Mathematics and its Applications*, 12 2004. [Online]. Available: https://www.maa.org/press/periodicals/loci/joma/the-sir-model-for-spread-of-disease-the-differential-equation-model

[8] m. batista, "Estimation of the final size of the covid-19 epidemic," *medRxiv*, 2020. [Online]. Available: https://www.medrxiv.org/content/early/2020/02/28/2020.02.16.20023606

[9] N. Zhao, K. Charland, M. Carabali, E. Nsoesie, M. Maher-Giroux, E. Rees, M. Yuan, C. Garcia Balaguera, G. Jaramillo Ramirez, and K. Zinszer, "Machine learning and dengue forecasting: Comparing random forests and artificial neural networks for predicting dengue burdens at the national sub-national scale in colombia," *bioRxiv*, 2020. [Online]. Available: https://www.biorxiv.org/content/early/2020/01/14/2020.01.14.906297

[10] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. Jarrod Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. Carey, İ. Polat, Y. Feng, E. W. Moore, J. Vand erPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and S. . . Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[11] R. Wirth and J. Hipp, "Crisp-dm: Towards a standard process model for data mining," in *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*. Citeseer, 2000, pp. 29–39.

[12] INMET. (2020) Instituto nacional de meteorologia. [Online]. Available: https://www.inmet.gov.br

[13] IBGE. (2020) Instituto brasileiro de geografia e estatística. [Online]. Available: https://www.ibge.gov.br

[14] SINAN. (2020) Sistema de informação de agravos de notificação. [Online]. Available: https://sinan.saude.gov.br

[15] F. Pedregosa, G. Varoquaux, A. Gramfort *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[16] Avicena. (2020) Governanca inteligente em serviços de saúde. [Online]. Available: https://gissa.avicena.in