DIA-2019

Predição Casos de Dengue em Séries Temporais

Antonio José Pinheiro Prado, Juliano Siloto Assine e Luiz Eduardo Pita Mercês Almeida

Faculdade de Engenharia Elétrica
e de Computação, Unicamp

1 Introdução

A dengue é uma doença infecciosa transmitida pela picada de mosquitos do gênero *Aedes*, tendo se tornado um problema de saúde global nas últimas décadas do século 20 [1]. Estudos sobre o impacto econômico apontam prejuízos de bilhões de dólares anuais decorrentes da doença [2, 3]. Apenas na cidade de Campinas, Brasil, foram registrados mais de 20 mil casos de de dengue no primeiro semestre de 2019 [4].

O entendimento das variáveis que levam ao aumento no casos de dengue e possivelmente a capacidade de prever epidemias é de grande interesse para órgãos de saúde pública que podem investir de forma mais localizada e eficiente seus recursos. Baseado nisso, a plataforma DrivenData [5] oferece uma competição aberta de objetivo educacional para predição de casos de dengue. A descrição completa encontra-se disponível no site oficial[1]. O desafio consiste predizer o número de casos de dengue em duas localidades, utilizando-se para isso da série temporal de casos registrados e de atributos auxiliares, como dados climáticos e de vegetação.

Pelo tamanho e natureza mista dos dados são muitas as abordagens que podem ser utilizadas para resolver esse problema. Dentre as técnicas que foram consideradas como candidatas encontram-se desde técnicas de predição de séries temporais, como os modelos ARIMA/SARIMA, a modelos mais genéricos, como Multilayer Perceptron, Convolutional Neural Networks, Recurrent Neural Networks e Random Forests.

Após uma exploração inicial das soluções, com a procura de um resultado que pudesse ser submetido para o desafio, o foco do trabalho consistiu na uniformização do tratamento do problema, tal como descrito na seção 3 e implementado no repositório[2]. O objetivo com isso foi facilitar a comparação metódica de modelos para a previsão dos casos de dengue, no que consiste um possível direcionamento para trabalhos futuros.

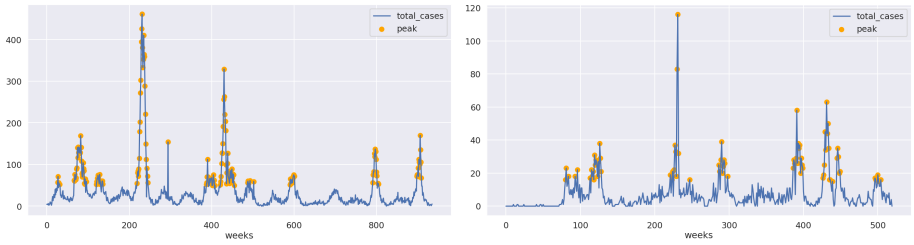
2 Os Dados

Os dados utilizados fornecidos pelo desafio abrangem um período de 5 anos para a localidade de San Juan, em Porto Rico, e de 3 anos para a localidade de Iquitos, no Peru. Além do número de casos de dengue reportados semanalmente, que podem ser vistos na Figura 1, existem outros 20 atributos coletados que podem estar relacionados ao número de casos, e que estão apresentados na Tabela 2.

Podemos observar que o número total de casos e a distribuição média dos casos (Figura 2) são bem diferentes nas duas localidade, o que também é verdade para os atributos climáticos (Tabelas 3 e 4). Isso não é surpreendente, levando em conta que

[1] https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/
[2] https://github.com/jsiloto/dengAI/

Figura 1: Casos Reportados Semanalmente



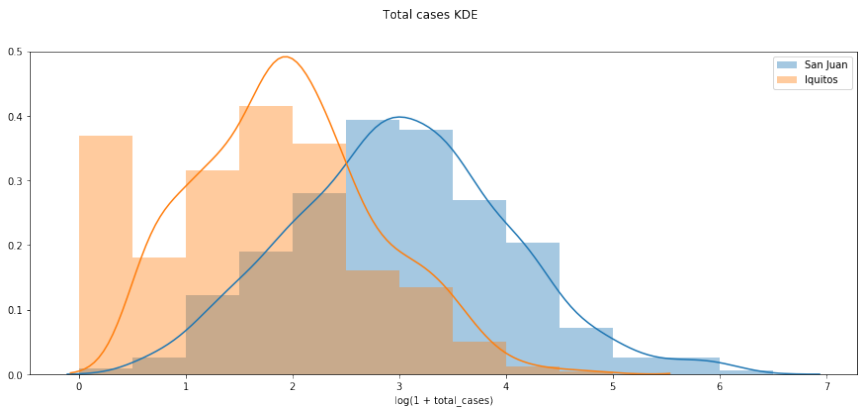
(a) San Juan (b) Iquitos

as características geográficas das duas cidades são bem distintas. Tal informação foi crucial na escolha da modelagem das soluções, de tal modo que cada cidade foi considerada um conjunto de dados independentes e o treinamento dos modelos foi feito separadamente.

Apesar da análise a nível de predição ser independente para as cidades, algumas características comuns da doença podem ser notadas. Observando a autocorrelação dos casos (Figura 3) pode-se perceber uma sazonalidade anual. De fato, o ciclo reprodutivo do mosquito Aedes está relacionado a épocas chuvosas, em que as fêmeas podem depositar ovos em recipientes com água acumulada. Percebe-se também a existência de picos epidêmicos em intervalos de alguns anos. Apesar deles, em geral, não ocorrerem em anos consecutivos, provavelmente devido à imunização da população, não é claro se existe informação nos dados disponíveis sobre quando eles ocorrem, fazendo da previsão desses picos uma das maiores dificuldades do desafio.

Foi feita a análise de correlação, utilizando correlação de Pearson, dos atributos com os casos de dengue, o que pode ser visto na Tabela 6. Nenhum dos atributos aparenta ser especialmente eficiente na predição do número de casos. Pensando que o efeito de um atributo pode ser sentido apenas num instante posterior do tempo foi também analisada a correlação do número de casos de dengue com o valor de cada atributo com um atraso variável. Pode ser visto na Figura 4 que os atributos

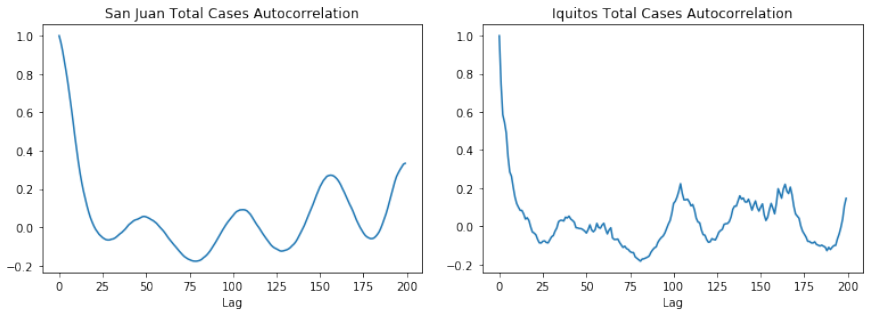
Figura 2: Distribuição das duas Séries temporais



mais preditivos para San Juan se encontram com um atraso entre 80 e 90 semanas, enquanto Iquitos provém uma correlação temporal muito mais caótica.

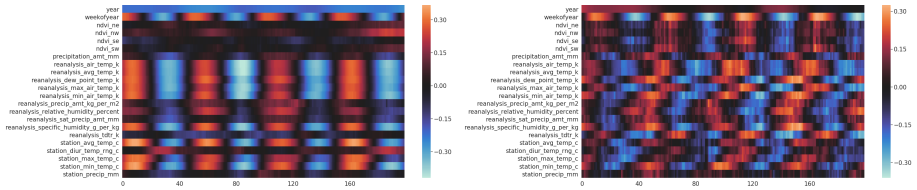
Uma análise mais extensa dos dados pode ser encontrada no repositório[3] [4].

Figura 3: Autocorrelação da série temporal



(a) San Juan (b) Iquitos

Figura 4: Correlação com atraso



(a) San Juan (b) Iquitos

3 O Problema

3.1 Forecasting

O problema descrito no desafio é um problema de predição em séries temporais conhecido como *forecasting*.

Dada uma sequência conhecida $x[i], i \in \{0, \ldots, N-1\}$ queremos predizer os próximos M valores ou seja, $x[N] \ldots x[N+M-1]$. Em geral essa predição é feita a partir de um modelo autorregressivo $f()$ que estima o valor $x[i+p]$ baseado nas p observações prévias

$$\hat{x}[i] = f(x[i-1], \ldots, x[i-p])$$

Para estimar valores fora da janela conhecida, ou seja, onde $i > N$, os valores estimados são realimentados no modelo, por exemplo:

$$\hat{x}[N+1] = f(\hat{x}[N], x[N-1], \ldots, x[i-p])$$

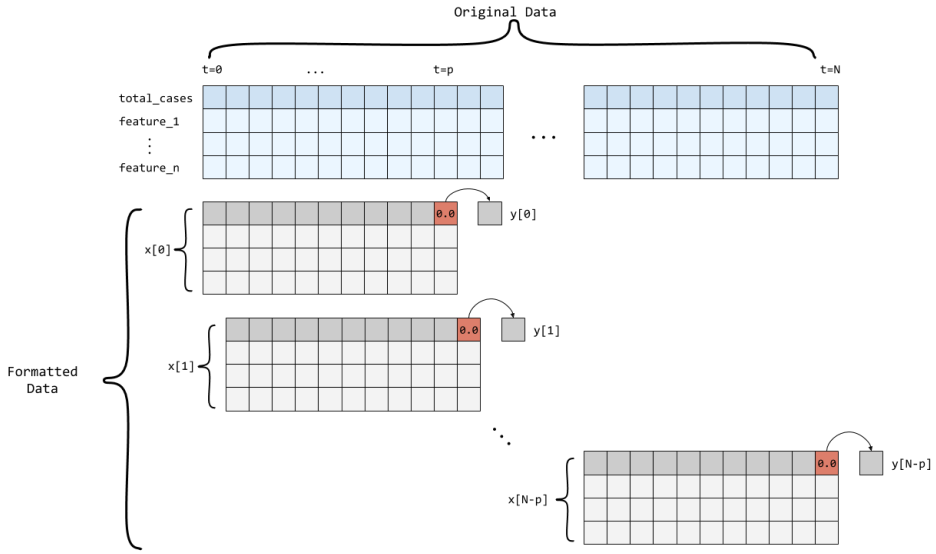
[3] https://github.com/jsiloto/dengAI/blob/master/analysis/data_analysis.ipynb
[4] https://github.com/jsiloto/dengAI/blob/master/analysis/correlation_analysis.ipynb

3.2 Long Term Forecasting

Em geral o problema de *forecasting* é realizado para poucas amostras no futuro. Como as predições são realimentadas no modelo, existe uma realimentação positiva nos erros de predição. No entanto o problema em questão apresenta um período longuíssimo de predição. Enquanto o trabalho em [6] não explora períodos maiores que 16 semanas para predição usando o mesmo dataset, o desafio escolhido requer predição de **260 semanas** para San Juan e **156 semanas** para Iquitos.

Para possibilitar esse cenário, são providas para o período de teste todas os *features* auxiliares. Em problemas de *forecasting*, variáveis que são conhecidas no futuro e utilizadas na predição são conhecidas como variáveis exógenas[7]. Variáveis exógenas são geralmente utilizadas em modelos econômicos em que existem variáveis de controle humano e, portanto, são conhecidas para períodos futuros. No nosso caso porém, elas são principalmente dados de sensores, que em um problema real nunca seriam conhecidos com antecedência, dando uma característica única aos problema em questão.

Figura 5: Formatação dos dados para autorregressão



3.3 Modelagem Utilizada

Para tratar o problema acima mencionado, dividimos os dados em duas formas de tratamento.

1 Consideramos todos os dados, incluindo a própria variável alvo em uma janela de p posições como entrada para o modelo, zerando a posição a ser predita para não haver contaminação. A Figura 5 apresenta descrição gráfica dessa abordagem que será chamada doravante de **autorregressão**.

2 Para evitar o problema de realimentação dos erros, uma outra alternativa é a de utilizar apenas os *features*. Essa opção se difere da abordagem anterior apenas em que a matriz de entrada não possui a primeira linha. Chamaremos essa abordagem de **regressão**.

4 Os Modelos

Na literatura há uma prevalência de modelos da família ARMA para predição de casos de dengue (ver Tabela 1). Fizemos uma exploração com esses modelos, mas sem grandes sucessos, em geral constatamos que os resultados eram altamente dependentes da janela de autorregressão e dos *splits* de treino/validação, além do baixo suporte de software na linguagem `python`.

Para a modelagem do problema, como descrito na seção 4 escolhemos duas alternativas baseadas em redes neurais devido à generalidade de suas aplicações, as redes MLP(*Multilayer Perceptron*) e RNN(*Recurrent Neural Networks*).

4.1 MLP

Multilayer Perceptrons (MLP) são a classe mais simples de redes neurais artificiais. Um MLP é construído a partir da composição de unidades menores geralmente chamadas de camadas da rede neural. Se considerarmos um MLP com N camadas e $i \in \{0, \ldots, N-1\}$ temos que na i-ésima camada a entrada x_i é transformada por uma função linear e depois por uma função não linear de ativação σ_i resultando na entrada da camada seguinte, ou seja $x_{i+1} = \sigma(W_i x_i + bi)$. Temos como casos especiais que a entrada da rede é x_0 e a saída da rede $y = x_N$ então:

$$y = \sigma_{N-1}(W_{N-1}(\sigma_{N-2}(W_{N-2}(\ldots \sigma_0(W_0 x_0 + b_0)\ldots) + b_{N-2}) + b_{N-1})$$

Nos nossos experimentos a função de ativação σ_i utilizada foi a função ReLU[8] para $i \in \{0, \ldots, N-2\}$ e a função identidade para a ultima camada da rede $\sigma_{N-1}(x) = x$.

4.2 RNN

Recurrent Neural Networks(RNN) são uma família de redes neurais capazes de lidar com sequências de dados. Elas são arquitetadas de modo a compartilhar informações entre o evento atual e os anteriores. Nessa arquitetura as informações persistem devido a realimentação (*loop*) dos dados passados, atuando como um sistema dinâmico. Assim são muito utilizadas no processamento de dados de fala, texto, vídeo e séries temporais.

A forma mais simples de redes recorrentes é uma extensão do modelo MLP para incluir uma camada com realimentação. Uma camada com realimentação além de entrada x e saída y possui um estado h_t (para simplificar a notação não utilizaremos o índice da camada). A cada instante t temos uma entrada x_t que é utilizada para computar um novo estado da camada h_t, que por sua vez é utilizado para computar a saída y_t essa relação de recorrência é formalizada por:

$$h_t = \sigma(W_h x_t + U h_{t-1} + b_h)$$

$$y_t = \sigma(W_y h_t + b_y)$$

Onde W_h, W_y, U e b são parâmetros aprendidos durante o treinamento da rede neural e σ é a função de ativação.

Observando a equação percebe-se que o valor do estado da unidade recorrente dependerá de todos os valores dos estados anteriores. Assim, durante a retroalimentação do gradiente descente, principalmente em sequências com muitos eventos, ocorrerá um grande número de derivações. A arquitetura clássica de RNN, por esse motivo, sofre com os problemas de saturação do gradiente: sumiço (*vanishing*) e explosão (*explosion*). Além disso, essa arquitetura é muito suscetível a dados ruidosos e valores atípicos (*outliers*) por guardar todos os dados sem nenhum tipo de seleção.

Buscando resolver esses problemas, foi proposta a arquitetura LSTM (*Long Short-Term Memory*) por Hochreiter e Schmidhuber [9]. Ela introduz na unidade recorrente três portões (*gates*): portão de esquecimento (*forget gate*), portão de entrada (*input gate*) e portão de saída (*output gate*). A LSTM introduz também o conceito de estado da célula que representa a informação a ser mantida.

Cada portão atua sobre o estado da célula de uma forma distinta. Enquanto portão do esquecimento remove informações que não são mais úteis do estado da célula, o portão da entrada adiciona novas informações úteis. Por sua vez, o portão de saída tem a função de extrair do estado da célula as informações que serão apresentadas como saída. Para uma apresentação didática sobre o tema recomendamos o tutorial em [10].

5 Experimentos

Os experimentos foram realizados com os modelos MLP e LSTM para ambos os casos de regressão e autorregressão. Foi feito uma separação de 65% das amostras para treino e 35% para validação. Essas quantidades foram escolhidas para que houvessem picos no conjunto de validação (Ver discussão na seção 7). Os experimentos foram comparados com duas baselines: um processo gaussiano i.i.d com mesma média e desvio padrão do conjunto de treino e uma abordagem de *naive forecasting*[11] onde os dados do ultimo período conhecido (no nosso caso 52 semanas) são repetidos para os próximos períodos. Para todos os modelos foram utilizadas janelas de regressão de 170 amostras, esses períodos foram escolhidos por englobar pelo menos 3 anos (período tipico entre epidemias) e em geral terem bons resultados nos testes realizados.

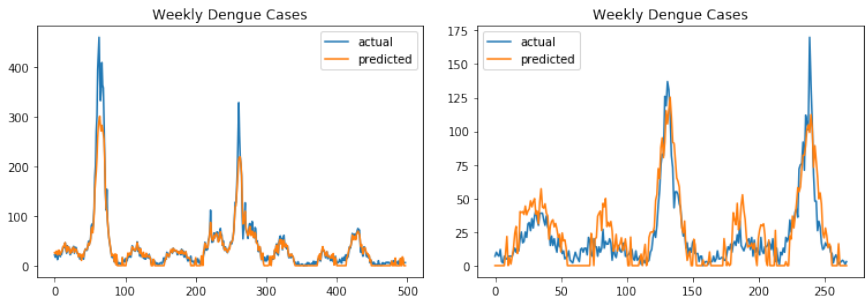
Segundo o requerimento do desafio o erro de predição utilizado foi o erro absoluto médio (MAE). Foram calculados 3 métricas para cada experimento: **MAE Treino** e **MAE Validação** são calculados baseado nas predições ponto a ponto dos conjuntos de treino e validação enquanto **MAE Forecast** é calculado baseado no *forecast* com realimentação (apresentado na seção 3) no conjunto de validação.

Os resultados podem ser observados na Tabela 5. Eles são melhores do que aqueles obtidos com técnicas tradicionais em etapas anteriores do projeto, mas as técnicas baseadas em redes neurais se provaram bem difíceis de utilizar, tendo resultados altamente dependentes da inicialização da rede e sujeitos facilmente a *overfitting*. Para mitigar esses problemas foram utilizadas técnicas como *Batch Normalization* [12] e *Dropout* [13], mas a melhor forma de utilizar essas técnicas não é nem um pouco clara e sua utilização foi feita na base da tentativa e erro. Um exemplo visual dos resultados obtidos para a cidade de San Juan pode ser visto nas Figuras 6 e 7

Os resultados apresentados foram escolhidos dentre várias rodadas de experimento segundo critérios subjetivos, não foi adotada uma abordagem metódica de escolha

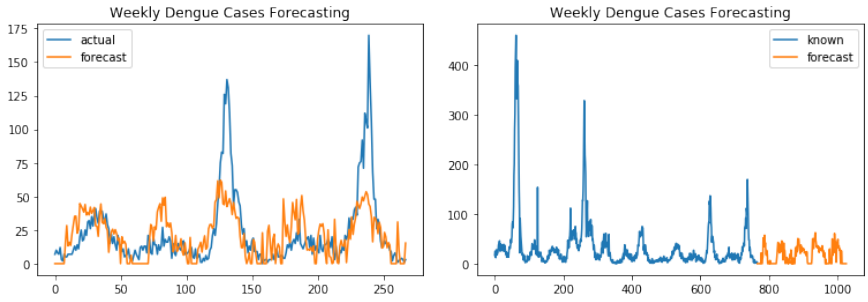
de hiperparâmetros e resultados devido à alta volatilidade dos experimentos, um ponto claro a ser abordado em trabalhos futuros.

Figura 6: Predições por MLP Autorregressivo em San Juan



(a) Treino (b) Validação

Figura 7: Forecasting por MLP Autorregressivo em San Juan



(a) Validação (b) Teste

6 Submissão ao DataDriven

Foram realizadas 3 submissões ao desafio baseadas nos resultados reportados. Uma submissão utilizando *Naive Forecasting* ($MAE = 37.6010$), outra com os melhores modelos MLP com menor erro de *forecast* ($MAE = 29.6659$), e similarmente uma para os modelos LSTM ($MAE = 59.5841$), todos muito distantes do líder da competição ($MAE = 10.30$).

7 Prevendo Picos da Doença

Um dos maiores desafios que enfrentamos foi o de prever os picos da doença, o que é a causa mais provável de grandes erros no conjunto de teste. Em geral, a maior parte dos modelo testados consegue capturar a sazonalidade dos episódios da doença, porém raramente os picos são modelados. Em geral os dois comportamentos observados são:

- os picos são atenuados e os modelos agem como filtros passa-baixa;
- os modelos "alucinam" picos realistas, mas em lugares errados.

Para analisar esse fenômeno, observamos que as correlações dos *features* com o número de casos muda bastante quando calculada apenas nos picos, mas não houve nenhuma conclusão a partir disso. Os picos podem ser observados na Figura 2 e resultados da análise realizada na Tabela 6)

8 Conclusão e Trabalhos Futuros

Concluímos este trabalho com fracasso em relação à competição, mas sem arrependimentos. O tema de previsão de séries temporais é um tema vasto e abordado por diferentes comunidades (Aprendizado de Maquina, Processamento de Sinais, Economia, etc) com nomenclaturas e suposições diferentes sobre os dados. Fomos capazes de testar várias técnicas diferentes com nuances diferentes e atacar o problema em largura. Esperamos que trabalhos futuros identifiquem modelos mais estáveis e reprodutíveis, permitindo uma análise mais sistemática dos hiperparâmetros utilizados e que tenham uma fundamentação teórica mais embasada, em especial, abordagens no domínio da frequência podem ser úteis para lidar com o problema dos picos.

Finalmente concluímos notando que o software e os conhecimentos desenvolvidos são amplamente aplicáveis para problemas semelhantes, em especial no Brasil que possui muitas cidades afetadas pela doença. O uso de dados de forma preditiva é um grande potencial inexplorado na administração publica de saúde.

Referências

1. Gubler, D.J.: Dengue and dengue hemorrhagic fever. Clinical microbiology reviews **11**(3), 480–496 (1998)
2. Shepard, D.S., Coudeville, L., Halasa, Y.A., Zambrano, B., Dayan, G.H.: Economic impact of dengue illness in the americas. The American journal of tropical medicine and hygiene **84**(2), 200–207 (2011)
3. Martelli, C.M.T., Junior, J.B.S., Parente, M.P.P.D., Zara, A.L.d.S.A., Oliveira, C.S., Braga, C., Junior, F.G.P., Cortes, F., Lopez, J.G., Bahia, L.R., *et al.*: Economic impact of dengue: multicenter study across four brazilian regions. PLoS neglected tropical diseases **9**(9), 0004042 (2015)
4. DEVISA, D.d.V.e.S.: Informe Epidemiologico Arboviroses. Secretaria Municipal de Saude de Campinas (2019). http://www.saude.campinas.sp.gov.br/saude/vigilancia/informes/2019/Informe_Epid_Arboviroses_especial_dengue_01_07_2019.pdf
5. Bull, P., Slavitt, I., Lipstein, G.: Harnessing the power of the crowd to increase capacity for data science in the social sector. arXiv preprint arXiv:1606.07781 (2016)
6. Islas Abud, K., Emmanuel Vallejo Clemente, E., Sánchez Castellanos, H.: A novel deep recurrent neural network architecture for time series forecasting of mosquito-borne disease case counts. Master's thesis (May 2019)
7. Hyndman, R.J., Athanasopoulos, G.: Forecasting: Principles and Practice. OTexts, ebook (2018)
8. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10), pp. 807–814 (2010)
9. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
10. Olah, C.: Understanding lstm networks (2015)
11. What is naive forecasting? definition and meaning. http://www.businessdictionary.com/definition/na-ve-forecasting.html
12. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
13. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research **15**(1), 1929–1958 (2014)
14. Cortes, F., Martelli, C.M.T., de Alencar Ximenes, R.A., Montarroyos, U.R., Junior, J.B.S., Cruz, O.G., Alexander, N., de Souza, W.V.: Time series analysis of dengue surveillance data in two brazilian cities. Acta tropica **182**, 190–197 (2018)
15. Wu, Y., Yang, Y., Nishiura, H., Saitoh, M.: Deep learning for epidemiological predictions. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 1085–1088 (2018). ACM
16. Guo, P., Liu, T., Zhang, Q., Wang, L., Xiao, J., Zhang, Q., Luo, G., Li, Z., He, J., Zhang, Y., *et al.*: Developing a dengue forecast model using machine learning: A case study in china. PLoS neglected tropical diseases **11**(10), 0005973 (2017)
17. Anwar, M.Y., Lewnard, J.A., Parikh, S., Pitzer, V.E.: Time series analysis of malaria in afghanistan: using arima models to predict future trends in incidence. Malaria journal **15**(1), 566 (2016)
18. Midekisa, A., Senay, G., Henebry, G.M., Semuniguse, P., Wimberly, M.C.: Remote sensing-based time series models for malaria early warning in the highlands of ethiopia. Malaria journal **11**(1), 165 (2012)

19. Gharbi, M., Quenel, P., Gustave, J., Cassadou, S., La Ruche, G., Girdary, L., Marrama, L.: Time series analysis of dengue incidence in guadeloupe, french west indies: forecasting models using climate variables as predictors. BMC infectious diseases **11**(1), 166 (2011)
20. Wangdi, K., Singhasivanon, P., Silawan, T., Lawpoolsri, S., White, N.J., Kaewkungwal, J.: Development of temporal modelling for forecasting and prediction of malaria infections using time-series and arimax analyses: a case study in endemic districts of bhutan. Malaria Journal **9**(1), 251 (2010)
21. Lu, L., Lin, H., Tian, L., Yang, W., Sun, J., Liu, Q.: Time series analysis of dengue fever and weather in guangzhou, china. BMC Public Health **9**(1), 395 (2009)
22. Choudhury, Z.M., Banu, S., Islam, A.M.: Forecasting dengue incidence in dhaka, bangladesh: A time series analysis. (2008)
23. Luz, P.M., Mendes, B.V., Codeço, C.T., Struchiner, C.J., Galvani, A.P.: Time series analysis of dengue incidence in rio de janeiro, brazil. The American journal of tropical medicine and hygiene **79**(6), 933–939 (2008)

Apêndice

Fontes de dados

- **GHCN (Global Historical Climatology Network)** é uma base de dados da NOAA (National Oceanic and Atmospheric Administration) de estações meteorológicas terrestres com medições diárias;
- **NOAA NCEP (National Centers for Environmental Prediction)** é um sistema de previsão climática nos Estados Unidos que prover dados da reanálise de dados globais de tempo, água, clima, previsões entre outros;
- **NOAA CDR (Climate Data Record)** constitui de um conjunto de dados temporais obtidos por satélites e entre eles medições do Índice de Vegetação da Diferença Normalizada (NDVI) por região da cidade (noroeste, nordeste, sudeste e sudoeste);
- **PERSIANN-CDR (Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks-CDR)** constitui de um conjunto dados formados por de estimativas da quantidade de chuva diária;

Tabelas

Tabela 1: Relação de trabalhos similares e características dos modelos utilizados

Doença	Outros Atributos	Modelo	Algoritimo	Obs	Trabalho
Dengue, Zica	Não	Vários	Vários	Tese de Mestrado com nosso dataset	[6]
Dengue	Não	ARIMA	Box Jenkins		[14]
Influenza	Não	ARMA-like, Processos Gaussianos, Deep Learning	N/A		[15]
Dengue	Temperatura, Precipitação, Humidade Relativa, Pesquisas no Baidu	Vários	Vários	Support Vector Regression teve os melhores resultados	[16]
Malaria	Temperatura, Precipitação, Humidade Relativa, Indice de Vegetação	ARIMA	Box Jenkins	Trabalha com escala logaritimica dos dados.	[17]
Malaria	Temperatura do chão, Precipitação, Indice de Vegetação, EvapoTranspiração	SARIMA	N/A	Trabalha com escala logaritimica dos dados.	[18]
Dengue	Temperatura Max/Min/Avg, Precipitação, Humidade Relativa	SARIMA	Box Jenkins	Variavel com lag de 3 meses e Temperatura são os atributos mais preditivo.	[19]
Malaria	Temperatura Max/Min/Avg, Precipitação, Humidade Relativa	ARIMAX	N/A	Variáveis não eram transferiveis para diferentes localizações	[20]
Dengue	Temperatura Max/Min, Precipitação, Humidade relativa, Vento	Regressão de Poisson	GEE/QICu	Temperatura minima/ Humidade são positivamente correlacionadas, vento é negativamente correlacionado	[21]
Dengue	Não	SARIMA	Normalized Bayesian Information Criteria		[22]
Dengue	Temperatura Max/Min, Precipitação diaria e anual	ARIMA	Box Jenkins	Trabalha com escala logaritimica dos dados	[23]

Tabela 2: Descrição dos atributos

Id	Descrição	Unidade	Fonte
ndvi_ne	NDVI da região nordeste	(-1, 1)	NOAA CDR NDVI
ndvi_nw	NDVI da região noroeste	(-1, 1)	NOAA CDR NDVI
ndvi_se	NDVI da região sudeste	(-1, 1)	NOAA CDR NDVI
ndvi_sw	NDVI da região sudoeste	(-1, 1)	NOAA CDR NDVI
prec_amt	Índice de precipitação	mm	PERSIANN
reanal_air_temp	Temperatura média do ar	Kelvin	NOAA NCEP
reanal_avg_temp	Temperatura média do ar	Kelvin	NOAA NCEP
reanal_max_temp	Temperatura máxima do ar	Kelvin	NOAA NCEP
reanal_min_temp	Temperatura mínima do ar	Kelvin	NOAA NCEP
reanal_tdtr	Amplitude térmica diurna	Kelvin	NOAA NCEP
reanal_dew_temp	Temperatura do ponto de orvalho	Kelvin	NOAA NCEP
reanal_sat_prec	Índice de precipitação determinado por satélite	mm	NOAA NCEP
reanal_prec	Índice de precipitação	Kg/m^2	NOAA NCEP
reanal_rel_hum	Umidade relativa média	%	NOAA NCEP
reanal_spec_hum	Umidade específica média	g/Kg	NOAA NCEP
st_avg_temp	Temperatura média da estação	Celsius	NOAA GHCN
st_max_temp	Temperatura máxima	Celsius	NOAA GHCN
st_min_temp	Temperatura mínima	Celsius	NOAA GHCN
st_tdtr	Amplitude térmica diurna	Celsius	NOAA GHCN
st_prec	Índice de precipitação	mm	NOAA GHCN

Tabela 3: Descrição estatística dos atributos para a cidade de San Juan

Id	Amostras Invalidas	Média	Desvio Padrão	Mediana	Minímo	Máximo
ndvi_ne	191	0,06	0,11	0,06	-0,41	0,49
ndvi_nw	49	0,07	0,09	0,07	-0,46	0,44
ndvi_se	19	0,18	0,06	0,18	-0,01	0,39
ndvi_sw	19	0,17	0,06	0,17	-0,06	0,38
prec_amt	9	35,5	44,6	20,8	0	391
reanal_air_temp	6	299,2	1,24	299,2	296	302
reanal_avg_temp	6	299,3	1,22	299,4	296	302
reanal_max_temp	6	301,4	1,26	301,5	297	304
reanal_min_temp	6	297,3	1,29	297,5	293	300
reanal_tdtr	6	2,52	0,5	2,46	1,36	4,43
reanal_dew_temp	6	295,1	1,57	295,46	290	298
reanal_sat_prec	6	35,5	44,6	20,8	0	391
reanal_prec	6	30,5	35,6	21,3	0	570
reanal_rel_hum	6	78,6	3,4	78,7	66,7	87,6
reanal_spec_hum	6	16,5	1,6	16,8	11,7	19,4
st_avg_temp	6	27	1,4	27,2	22,8	30,1
st_max_temp	6	31,6	1,7	31,7	26,7	35,6
st_min_temp	6	22,6	1,5	22,8	17,8	25,6
st_tdtr	6	6,76	0,83	6,76	4,53	9,91
st_prec	6	26,8	29,3	17,75	0	306

Tabela 4: Descrição estatística dos atributos para a cidade de Iquitos

Id	Amostras Invalidas	Média	Desvio Padrão	Mediana	Minímo	Máximo
ndvi_ne	3	0,26	0,08	0,26	0,06	0,51
ndvi_nw	3	0,24	0,08	0,23	0,04	0,45
ndvi_se	3	0,25	0,08	0,25	0,03	0,54
ndvi_sw	3	0,27	0,09	0,26	0,06	0,55
prec_amt	4	64,2	35,2	60,5	0	211
reanal_air_temp	4	297,9	1,17	297,8	294	302
reanal_avg_temp	4	299,1	1,33	299,1	295	303
reanal_max_temp	4	307,1	2,38	307	300	314
reanal_min_temp	4	292,9	1,66	293	287	296
reanal_tdtr	4	9,21	2,45	8,96	3,71	16
reanal_dew_temp	4	295,5	1,42	295,8	290	298
reanal_sat_prec	4	64,2	35,2	60,5	0	211
reanal_prec	4	57,6	50,3	46,4	0	362
reanal_rel_hum	4	88,6	7,58	90,9	57,8	98,6
reanal_spec_hum	4	17,1	1,45	17,4	12,1	20,5
st_avg_temp	37	27,5	0,92	27,6	21,4	30,8
st_max_temp	14	34	1,32	34	30,1	42,2
st_min_temp	8	21,2	1,26	21,3	14,7	24,2
st_tdtr	37	10,57	1,54	10,62	5,2	15,8
st_prec	16	62,5	63,25	45,3	0	543

Tabela 5: Resultado dos modelos no conjunto de validação

Modelo	San Juan			Iquitos		
	Train	Val	Forecast	Train	Val	Forecast
Baseline Aleatorio	N/A	54.60	54.60	N/A	12.69	12.69
Naive Forecasting	N/A	18.57	18.57	N/A	6.93	6.93
MLP Autoregressão	**7.00**	11.13	14.25	4.59	6.59	6.92
MLP Regressão	9.93	16.23	15.94	5.85	6.61	6.66
LSTM Autoregressão	7.37	**9.62**	13.75	**4.13**	**6.10**	7.96
LSTM Regressão	11.24	17.18	**13.01**	4.22	8.04	**6.52**

Tabela 6: Correlação dos atributos com o número de casos para toda a série temporal e apenas para os picos

Feature	San Juan		Iquitos	
	Regular	Picos	Regular	Picos
reanalysis_tdtr_k	-0.165	-0.035	-0.199	-0.089
ndvi_se	-0.042	-0.364	-0.098	0.102
ndvi_ne	-0.030	-0.081	-0.044	0.147
ndvi_sw	0.003	0.092	-0.054	0.214
station_diur_temp_rng_c	0.022	0.068	-0.092	-0.365
station_precip_mm	0.100	0.057	0.098	-0.082
ndvi_nw	0.103	-0.005	-0.047	0.129
reanalysis_sat_precip_amt_mm	0.118	-0.059	0.130	-0.125
precipitation_amt_mm	0.118	-0.059	0.130	-0.125
station_max_temp_c	0.146	0.227	0.061	-0.418
station_avg_temp_c	0.180	0.182	0.152	-0.413
station_min_temp_c	0.186	0.121	0.250	-0.082
reanalysis_relative_humidity_percent	0.193	0.077	0.203	0.082
reanalysis_max_air_temp_k	0.194	0.083	-0.109	-0.079
reanalysis_avg_temp_k	0.194	0.036	0.094	-0.109
reanalysis_air_temp_k	0.202	0.043	0.117	-0.086
reanalysis_precip_amt_kg_per_m2	0.217	0.060	0.233	0.114
reanalysis_min_air_temp_k	0.230	0.056	0.318	0.012
reanalysis_dew_point_temp_k	0.249	0.075	0.369	0.039
reanalysis_specific_humidity_g_per_kg	0.255	0.070	0.373	0.031