Predição Casos de Dengue em Séries Temporais II

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

Faculdade de Engenharia Elétrica
e de Computação, Unicamp

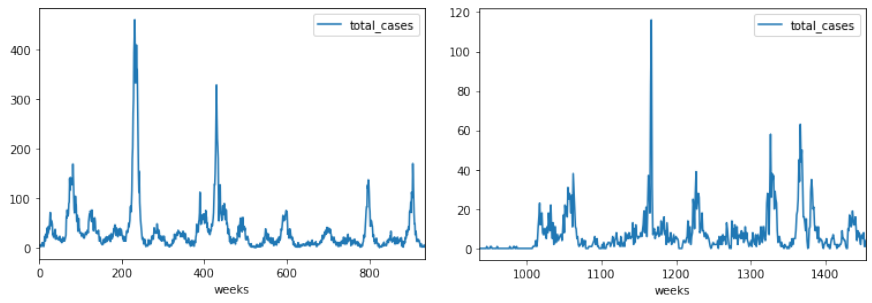
Introdução

Nesta etapa do projeto focamos na análise exploratória do problema, na familiarização com séries temporais e ferramentas de software relacionadas e no estabelecimento de *baselines*.

Series Temporais

Uma série temporal pode ser entendida como uma sequência de observações de um processo estocástico feitas ao longo do tempo. Em nosso caso, o resultado desse processo é número de casos de dengue ocorridos semanalmente nas cidades de Iquitos no Peru e San Juan em Porto Rico, conforme ilustrado na Figura 1.

Figura 1: Casos Reportados Semanalmente



(a) San Juan (b) Iquitos

As séries temporais são composta por combinações de quatro padrões: tendência, ciclos, sazonalidade e irregularidades. Os modelos clássicos de estimadores para séries temporais buscam, em geral, detectar a existência destes padrões.

A tendência refere-se ao comportamento a longo prazo da série, por exemplo um indicativo de que o número de casos de dengue tende a diminuir com o passar dos anos. Assim um série pode possuir tendência crescente, decrescente ou mesmo não possuir tendência. Nesse último caso definem-se as chamadas séries estacionárias, em que, de forma simplificada, seu comportamento não se altera com o passar do tempo. Assim, as propriedades estatísticas de sequências de observações em tempos distintos são semelhantes.

Ciclos e sazonalidade são comportamentos que se repetem com certa periodicidade. A diferença entre ciclo e sazonalidade está relacionada ao tempo de observação e sua previsibilidade. A sazonalidade é o comportamento periódico observados dentro de um ano, por exemplo, o aumento dos casos de dengue durante o verão. Os

ciclos, em contrapartida, não possuem sua periodicidade tão bem definida e, em geral, possuem tempo de observação superior a um ano. Um exemplo hipotético de ciclo seria a ocorrência de grandes epidemias de dengue em intervalos de tempo aproximados de 3 a 4 anos.

As variações irregulares são as alterações inexplicáveis ou não esperadas. Normalmente ocorrem devido a fatores externos, como catástrofes naturais e intervenções humanas.

Figura 2: Decomposição San Juan

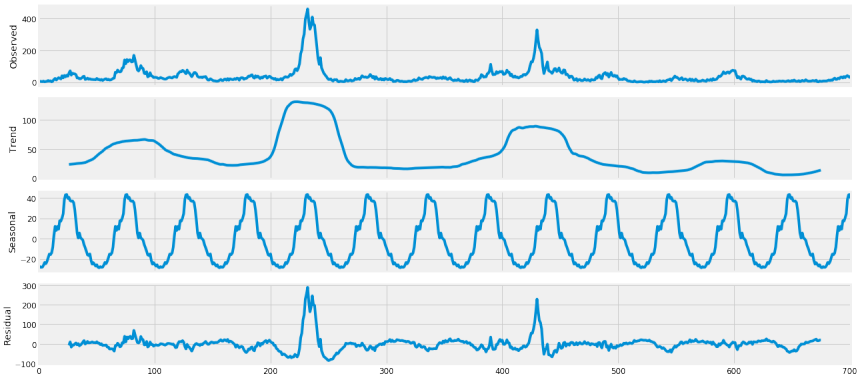
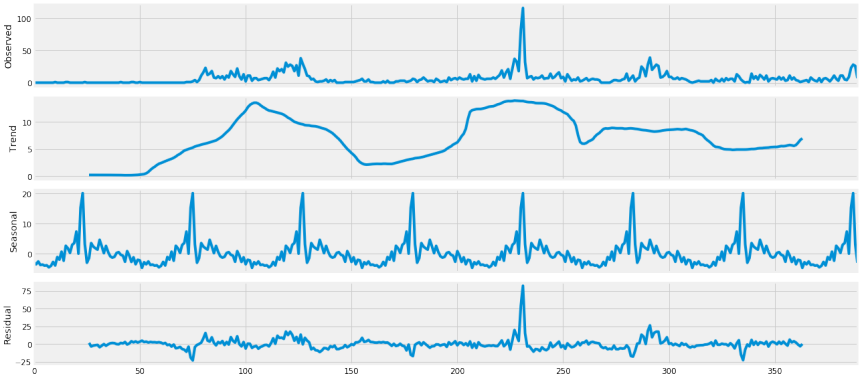


Figura 3: Decomposição Iquitos



Análise Exploratoria

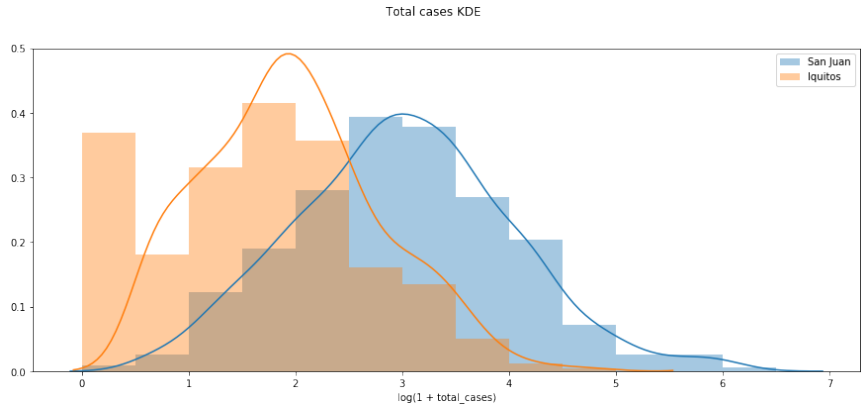
Entendendo os Dados

O conjunto de dados da competição é constituido por uma série temporal com resolução semanal dos casos reportados de dengue, além de 20 atributos contendo condições climáticas, como temperatura, nível de precipitação, umidade e índice de vegetação. A descrição dos dados e uma breve análise podem ser encontrados nas Tabelas 1, 2 e 3 e a descrição das fontes pode ser encontrada no Apêndice. São considerados dados da cidade de San Juan por aproximadamente 19 anos (1990 a 2008) e da cidade de Iquitos por cerca de 11 anos (2000 a 2010).

As duas cidades são bem distintas. A cidade de San Juan é a capital e o município mais populoso de Porto Rico, com cerca de 347 mil habitantes (dados de 2016). É uma cidade litorânea localizada no Caribe, sendo o centro industrial, financeiro, cultural e turístico da ilha de Porto Rico. Por sua vez, a cidade de Iquitos destaca-se por ser a cidade com maior número de habitantes no mundo que não pode ser alcançada por rotas terrestres. São cerca de 466 mil habitantes no que é conhecida como a Capital da Amazônia Peruana.

Na Figura 4 observa-se o histograma em escala logarítimica do número de casos de dengue para esses duas cidades. Pode-se notar que o comportamento das séries temporais para cada cidade são bem distintos, com uma incidência média bem maior em San Juan (38.0) do que em Iquitos(6.6).

Figura 4: Distribuição das duas Séries temporais



Análise de corelação

A análise dos dados têm por objetivo determinar se a série temporal é estacionária, perceber a ocorrência de sazonalidades, determinar a existência de atributos redundantes (isto é, muito correlacionados) e descobrir quais atributos são os mais preditivos em relação ao número de casos de dengue. Nesse estudo, os dados referentes a cada cidade foram separados e serão tratados como séries temporais distintas. Para realizar essa etapa utilizou-se principalmente do cálculo de correlação entre as variáveis.

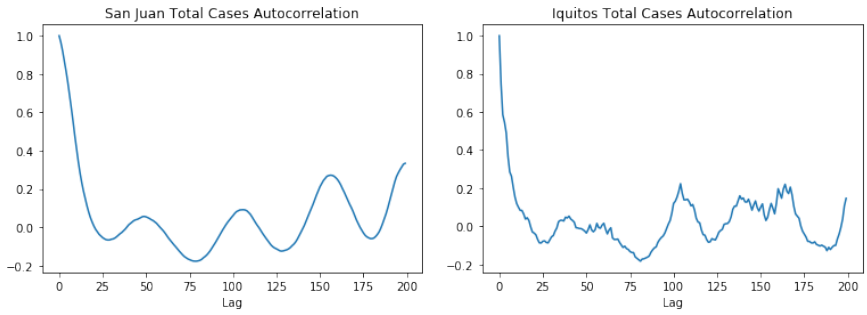
Séries temporais estacionárias possuem um menor grau de complexidade para a criação de um estimador, uma vez que seus dados estatísticos se mantém semelhantes com o passar do tempo. Existem alguns testes que podem ser realizados para determinar se uma série é ou não estacionária, entre eles o teste de Dickey-Fuller. Nesse teste ambas séries discutidas foram consideradas estacionárias.

Ao observar as séries temporais nas Figuras 2 e 3 percebe-se a existência de um possível ciclo periódico em torno de 3 a 4 anos para a cidade de San Juan e de 2 em 2 anos para a cidade de Iquitos. Uma explicação plausível para esse comportamento é a imunização da população após uma epidemia, levando a redução do número de casos nos anos seguintes.

O cálculo da autocorrelação da variável a ser predita permitiu a confirmação da existência de uma sazonalidade anual nas duas série temporais observadas. Na

Figura 5 podemos observar que em ambas as séries possuem picos de autocorrelação em intervalos de aproximadamente 50 meses ou seja um ano. Além disso, destacou a diferença de comportamento entre as duas séries uma vez que a cidade de San Juan apresentou uma autocorrelação mais comportada.

Figura 5: Autocorrelação da série temporal



(a) San Juan (b) Iquitos

A próxima etapa da análise dos dados consiste em observar a correlação entre os atributos. Essa medida serve para identificar dados estatisticamente semelhantes que podem ocasionar sobreajuste em nossos modelos. Assim os atributos mais correlacionados deverão ser agregados em um novo atributo de modo a reduzir a dimensionalidade dos dados.

Nas Figuras 6 e 7 estão apresentados os diagramas de correlação entre os atributos para cada uma das cidades. Neles podemos observar que os atributos relacionados a vegetação aparentemente possuem pouca relação com os demais atributos. Em contrapartida, os atributos relacionados a temperatura apresentam uma alta correlação.

Interessante destacar a diferença entre os diagramas das duas cidades, mostrando que o problema da predição do número de casos de dengue será melhor abordado sobre o preceito de que os dados de cada cidade não devem ser misturados. Observa-se também que os dados de vegetação e os dados climáticos estão mais correlacionados na cidade de Iquitos. É importante notar que nenhum dos atributos possui elevada correlação com o numero de casos de dengue, ou seja, nenhum deles é excepcionalmente bom em predizer os dados.

Da Tabela 6 tem-se que os atributos mais preditivos foram: a umidade específica e a temperatura do ponto de orvalho, seguidos de outros atributos de temperatura. No caso de Iquitos a umidade relativa também possui algum destaque. Em ambos os índices de vegetação possuem as mais baixas correlações com o número de casos de dengue.

Modelos Preditivos

Levando em consideração as referências levantadas durante a primeira parte do projeto fizemos explorações com modelos (S)ARIMA em comparação com uma alternativa "ingênua" de Regressão Linear. Não foi possível tirar nenhuma conclusão sobre a melhor alternativa de configuração dos hiperparâmetros, mas as referências

Figura 6: Correlação entre atributos: San Juan

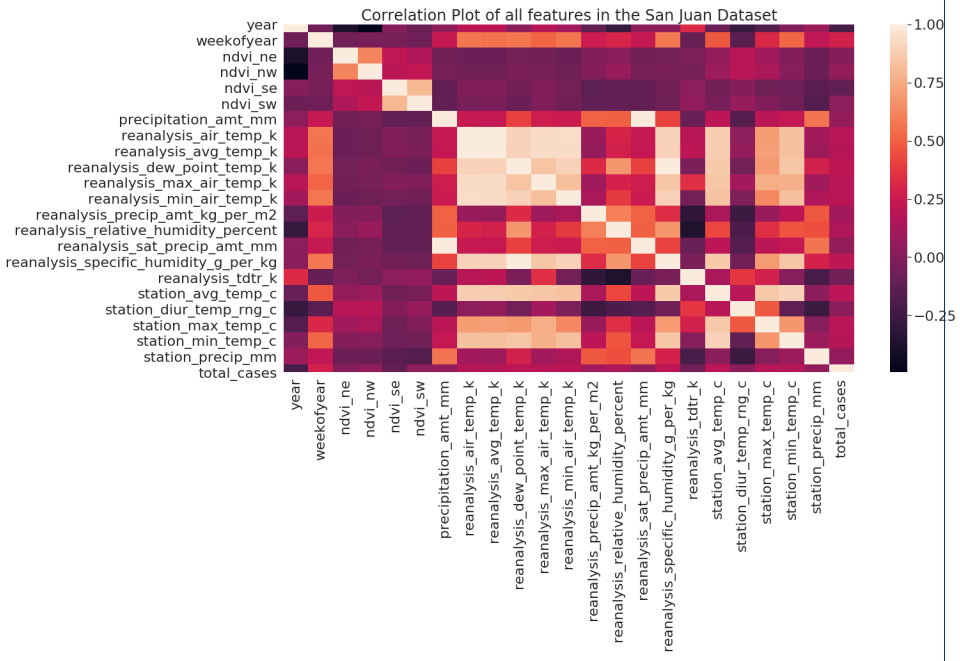
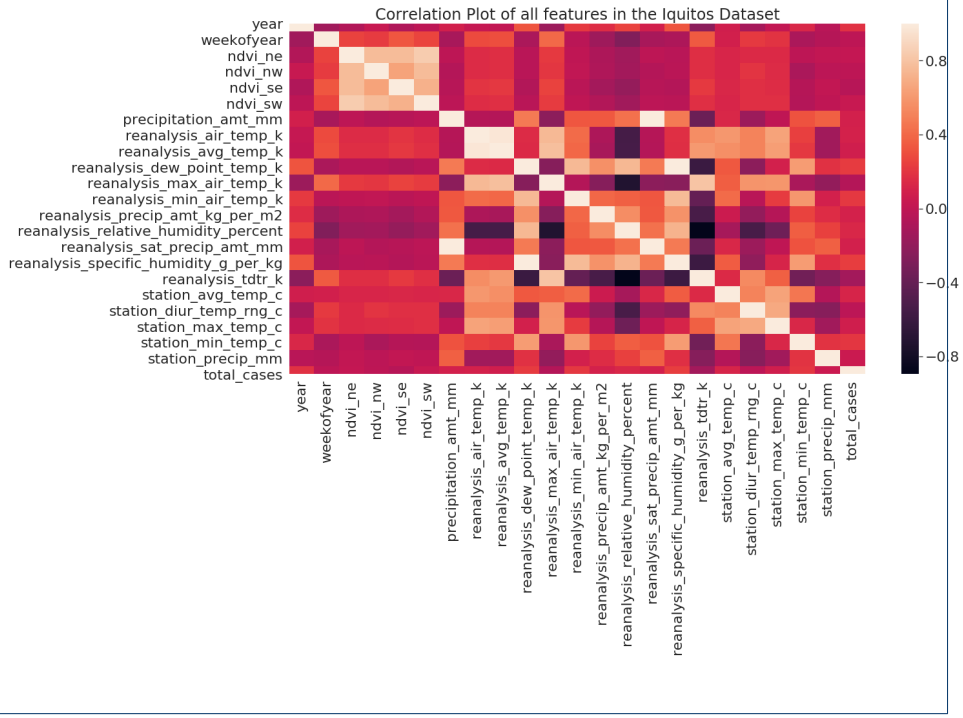


Figura 7: Correlação entre atributos: Iquitos



proveram uma boa base para entendimento de boas práticas da modelagem. Dito isso, nossa abordagem não se diferencia do que está presente na literatura e nosso

trabalho foi com o enfoque em estabelecimento de baselines para a próxima etapa do trabalho que será focada em modelagem.

Regressão Linear

A regressão linear é talvez o mais comum dos modelos de regressão. Ela é expressa da pela equação

$$y = \boldsymbol{x}^T \boldsymbol{w} + b$$

Onde $\boldsymbol{x} = [x_1, x_2, \ldots, x_N]$ é o conjunto das N variáveis preditivas (atributos), $\boldsymbol{w} = [w_1, w_2, \ldots, w_N]$ e b são os parâmetros do modelo e y é a variável alvo da regressão. Em nossos experimentos utilizamos a regressão linear de duas formas:

1. Variáveis exógenas como variáveis preditivas: $y_{t=0} = \boldsymbol{x}_{t=0}^T \boldsymbol{w} + b$ onde $\boldsymbol{x} = [x_1, x_2, \ldots, x_{21}]$ são os atributos descritos na Tabela 1
2. Modelo autoregressivo(AR) linear $y_{t=0} = \boldsymbol{x}^T \boldsymbol{w} + b$ onde $\boldsymbol{x} = [y_{t=-1}, y_{t=-2}, \ldots, y_{t=-p}]$ e p é um hiperparâmetro.

SARIMA

Na análise de séries temporais temos uma classe muito importante de modelos chamada ARMA (Autoregressive Moving Average). Um modelo ARMA é uma extensão do modelo autoregressivo já apresentado com um componente de média móvel ($\boldsymbol{\theta}^T \boldsymbol{b}$) :

$$y_{t=0} = \boldsymbol{x}^T \boldsymbol{w} + \boldsymbol{\theta}^T \boldsymbol{b}$$

Onde $\boldsymbol{\theta}$ são outros parâmetros e $\boldsymbol{b} = [b_1, \ldots, b_q]$ em que cada b_i é um termo aleatório geralmente assumido como gaussiano. Além disso, o modelo ARMA pode ser estendido para casos nos quais existe uma tendência clara. Podemos "diferenciar" a série ao substituir o vetor $\boldsymbol{x}$ por $\boldsymbol{x} - \boldsymbol{x_d}$ onde $\boldsymbol{x_d} = [y_{t=-1-d} \ldots, y_{t=-p-d}]$, removendo tendências de primeira ordem. Na nossa forma compacta, temos que o modelo ARIMA definido pelos hiperparâmetros p, q e d é dado por:

$$\text{ARIMA(p, q, d):} \quad y_{t=0} = (\boldsymbol{x} - \boldsymbol{x_d})^T \boldsymbol{w} + \boldsymbol{\theta}^T \boldsymbol{b}$$

Nesse trabalho utilizaremos duas extensões do modelo ARIMA, o SARIMA (Seasonal ARIMA), que combina dois modelos originais para levar em conta séries com dependências de longa duração e o SARIMAX (Seasonal ARIMA extended), sua extensão para incluir atributos adicionais na predição. A descrição desses modelos é consideravelmente mais complexa do que foi apresentado até agora e está fora do escopo desse relatório. Para mais informações recomendamos a referência[1]

Considerações práticas

Por questões pragmáticas, a implementação das análises e modelos foi realizada em **python** e o código se encontra disponível em: https://github.com/jsiloto/dengAI. O modelo de regressão linear foi implementado usando a biblioteca `sklearn` com uso do algorítimo de mínimos quadrados para minimização da norma l_2. Os

modelos SARIMA foram implementados utilizando a biblioteca `statsmodels`, que utiliza necessariamente o critério de máxima verossimilhança baseado em filtros de Kalman, e por consequência otimiza a norma l_2[2].

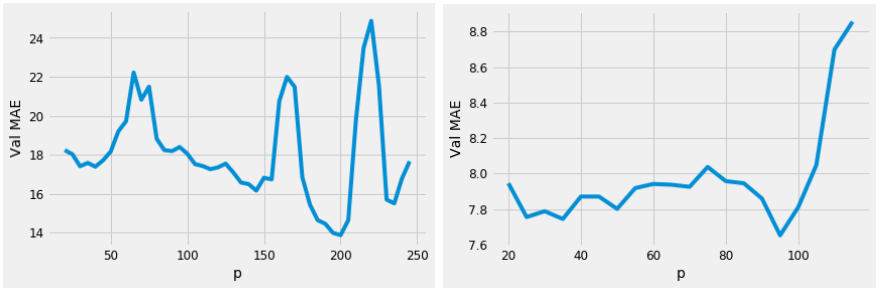
Apesar da facilidade de uso da biblioteca `statsmodels`, que é padrão em referências como blogs e tutoriais para tratamento de séries temporais em `python`, sua implementação é bem ineficiente, e ocasionalmente apresenta comportamento indefinido. Depois de vários testes com diferentes opções de otimização sem sucesso, foi utilizado o otimizador padrão "*LBFGS*", baseado em descida de gradiente de segunda ordem. Para ele apenas valores bem pequenos de p, q e d foram utilizados, devido à problemas de tempo de processamento e convergência.

Experimentos

Os experimentos foram realizados de forma independente para os dados das duas cidades, utilizando as últimas 25% amostras como conjunto de validação. Como hipótese nula nós utilizamos os resultados de variáveis gaussianas i.i.d. com mesma média e desvio padrão de cada série. Os resultados para as duas cidades estão relacionados na Tabela 4 onde pode-se ver que os modelos AR e SARIMA tem desempenho parecido entre si, mas bem melhor que os demais.

Para a escolha dos hiperparâmetros para o modelo autoregressivo foi analisado o erro de validação para cada valor de p como apresentado na Figura 8. Durante o desenvolvimento, observamos que o valor ideal de p é bem dependente da partição do conjunto de validação. Dessa forma, mesmo com resultados muito bons é difícil garantir que esse modelo tenha uma boa performance sem *overfitting*.

Figura 8: Erro de Validação vs Número de Amostras(p) no modelo AR



(a) San Juan (b) Iquitos

No caso dos modelos SARIMA o ajuste dos parâmetros foi bem mais problemático. O trabalho em [3] serviu como base para início das buscas devido à similaridade da formulação e variáveis. Foi realizado uma busca em todas as 64 combinações de $p, q, d, P, Q, D \in \{0, 1\}$ do modelo SARIMA$(p, q, d)(P, Q, D)_{52}$ baseado no critério de informação de Akaike, como é comum na metodologia Box-Jenkins, mas não foi notado muita diferença entre os resultados.

Também foram explorados diferentes pré-processamentos dos dados, como normalização e transformação em escala logarítmica. Apesar da escala logarítmica trazer grandes diferenças no resultado, a normalização em (0, 1) parece não fazer efeito, levando à crer que o próprio pacote `statsmodels` realiza algo semelhante de forma

implícita[1]. Além disso também foram testados a utilização de atributos extras (modelo SARIMAX), mas sem diferenças perceptiveis.

Submissão ao challenge

Foi feita uma submissão ao challenge utilizando os dois modelos de melhor resultado e o resultado total. Como requisito da competição a métrica de erro é o erro absoluto médio (MAE). O erro combinado dos dois modelos na validação é de $MAE = 11.71$ próximo ao líder da competição ($MAE = 10.30$) porém o erro obtido no conjunto de teste foi $MAE = 29.66$ indicando um claro *overfitting* ou erro de implementação que será investigado na próxima etapa do projeto.

Próximos Passos

Consideramos concluída a fase exploratória do projeto e para os próximos passos esperamos ter um foco no aprimoramento dos modelos existentes e desenvolvimento e teste de modelos alternativos. Em específico temos interesse nos pontos a seguir:

- Melhor engenharia de atributos
- Procurar uma implementação alternativa para os modelos SARIMA
- Testes com redes neurais MLP e redes convolucionais causais.

Nota de Transparência

Deixamos claro que estamos cientes da existência de diversas outras implementações públicas para o mesmo desafio[2] e que fizemos uso desse recurso para aprendizado, além da utilização de pequenos trechos de código para análise dos dados. Entretanto nenhum desses recursos foi utilizado para escolha ou desenvolvimento dos modelos preditivos.

Referências

1. Hyndman, R.J., Athanasopoulos, G.: Forecasting: Principles and Practice. OTexts, ebook (2018)
2. Lacey, T.: Tutorial: The kalman filter
3. 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)

[1] Cuja natureza ainda não encontramos na documentação.
[2] Em 03/09/2019 a pesquisa por "dengai" no Github retorna 194 repositórios.

Apendice

Fontes de dados

GHCN (Global Historical Climatology Network) é uma base de dados da NOAA (National Oceanic and Atmospheric Administration) de estações metereológicas terrestres com medições diárias;

NOAA NCEP (National Centers for Environmental Prediction) é um sistema de previsão climática do nos Estados Unidos que provém 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: 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 2: Descrição dos 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 3: Descrição dos 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 4: Resultado dos modelos no conjunto de validação

San Juan	
Modelo	MAE
Baseline Aleatorio	54.60
Regressão Linear	29.82
AR(200)	**13.87**
SARIMA(1,1,1)(1,1,1)52	20.26

Iquitos	
Modelo	MAE
Baseline Aleatorio	12.69
Regressão Linear	8.06
AR(95)	7.96
SARIMA(1,1,1)(1,1,1)52	**7.83**

Tabela 5: Exploração de diferentes configurações do modelo SARIMA$(1,1,1)(1,1,1)_{52}$

San Juan				
Features	Log	Norm	AIC	MAE
N	N	N	4,967.46	25.38
N	N	Y	4,967.46	25.38
N	Y	N	541.35	**20.26**
N	Y	Y	541.35	**20.26**
Y	N	N	4,979.03	26.00
Y	N	Y	4,981.75	26.79
Y	Y	N	547.45	20.27
Y	Y	Y	547.51	20.27

Iquitos				
Features	Log	Norm	AIC	MAE
N	N	N	2,024.99	**7.83**
N	N	Y	2,024.99	**7.83**
N	Y	N	577.51	8.44
N	Y	Y	577.51	8.44
Y	N	N	2,030.89	8.14
Y	N	Y	2,031.48	8.03
Y	Y	N	577.85	8.47
Y	Y	Y	577.85	8.48

Tabela 6: Correlação dos atributos com numero de casos

San Juan	
ndvi_ne	0.004
station_diur_temp_rng_c	0.035
ndvi_sw	0.041
station_precip_mm	0.051
precipitation_amt_mm	0.057
reanalysis_sat_precip_amt_mm	0.057
ndvi_nw	0.059
reanalysis_tdtr_k	0.068
reanalysis_precip_amt_kg_per_m2	0.107
ndvi_se	0.120
reanalysis_relative_humidity_percent	0.142
reanalysis_avg_temp_k	0.173
station_min_temp_c	0.174
reanalysis_air_temp_k	0.179
reanalysis_min_air_temp_k	0.186
station_max_temp_c	0.188
reanalysis_max_air_temp_k	0.193
station_avg_temp_c	0.194
reanalysis_dew_point_temp_k	0.201
reanalysis_specific_humidity_g_per_kg	0.205
year	0.213
total_cases	1.000

Iquitos	
ndvi_nw	0.011
ndvi_ne	0.020
station_diur_temp_rng_c	0.021
ndvi_sw	0.031
ndvi_se	0.041
station_precip_mm	0.045
reanalysis_max_air_temp_k	0.053
station_max_temp_c	0.080
reanalysis_avg_temp_k	0.080
precipitation_amt_mm	0.089
reanalysis_sat_precip_amt_mm	0.089
reanalysis_air_temp_k	0.097
reanalysis_precip_amt_kg_per_m2	0.101
station_avg_temp_c	0.114
reanalysis_relative_humidity_percent	0.129
reanalysis_tdtr_k	0.131
year	0.179
station_min_temp_c	0.203
reanalysis_min_air_temp_k	0.211
reanalysis_dew_point_temp_k	0.229
reanalysis_specific_humidity_g_per_kg	0.235
total_cases	1.000