

FUNDAÇÃO GETULIO VARGAS
ESCOLA DE MATEMÁTICA APLICADA – FGV/EMAp
CURSO DE GRADUAÇÃO EM MATEMÁTICA APLICADA

Validação de modelos preditivos para Dengue

Beatriz Macedo Coimbra dos Santos

Rio de Janeiro
2019

FUNDAÇÃO GETÚLIO VARGAS
ESCOLA DE MATEMÁTICA APLICADA – FGV/EMAp
CURSO DE GRADUAÇÃO EM MATEMÁTICA APLICADA

Validação de modelos preditivos para Dengue

“Declaro ser o único autor do presente projeto de monografia que se refere ao plano de trabalho a ser executado para continuidade da monografia e ressalto que não recorri a qualquer forma de colaboração ou auxílio de terceiros para realizá-lo a não ser nos casos e para os fins autorizados pelo professor orientador.”

Beatriz Macedo Coimbra dos Santos

Orientador: Flávio Codeço Coelho

Rio de Janeiro
2019

**Rio de Janeiro
2019**

Beatriz Macedo Coimbra dos Santos

Validação de modelos preditivos para Dengue

“Trabalho de Conclusão apresentado à Escola de Matemática Aplicada como requisito para a obtenção parcial do grau de bacharel em Matemática Aplicada.”

Aprovado em ____de ____de ____.

Grau atribuído ao Trabalho de Conclusão: ____.

Professor Orientador: Flávio Codeço Coelho
Escola de Matemática Aplicada - FGV/EMAp
Fundação Getulio Vargas

Agradecimentos

Agradeço, primeiramente, aos meus pais Maria Clara e José Wellington, que durante toda a minha vida fizeram o impossível para que eu pudesse me concentrar nos meus estudos, respeitando minhas escolhas e me fortalecendo com muito amor e carinho. Dedico essa conquista a eles.

Agradeço imensamente ao meu orientador e amigo Flavio Codeço Coelho, que esteve comigo desde meu primeiro ano de graduação e em nenhum momento sequer duvidou da minha capacidade de alcançar tudo o que eu desejasse. Flavio foi imprescindível para que eu atravessasse esta etapa de uma maneira mais leve e, ao mesmo tempo, me inspirava todos os dias com seu comprometimento e prazer em fazer pesquisa da maior qualidade e relevância.

Agradeço com todo o meu coração aos grandes amigos que a música me deu, meus companheiros Ordinarius: Antonia, Augusto, Fabiano, Maíra, Mateus e Matias. Tenho plena consciência de que, sem a musicalidade, a sensibilidade e o apoio de cada um deles, não teria sido tão feliz durante o período em que fiz este trabalho.

Agradeço aos meus queridos professores: Renato Rocha, Yuri Saporito, Rodrigo Targino, Hugo de la Cruz, Eduardo Wagner, Asla Sá, Roger Behling e Eduardo Massad. Tê-los como exemplo durante todo este período de aprendizagem foi um privilégio sem tamanho. Além de profissionais altamente qualificados, são também pessoas admiráveis e com quem espero conviver e aprender por muito mais tempo.

Agradeço com muito amor à minha amiga de longa data Isabela Pyrrho. Estivemos juntas na creche e no colégio e seguimos nos apoiando, vibrando em conjunto a cada passo que damos. Sem perceber, ela fez com que eu crescesse muito a partir da sua presença em minha vida, com suas demonstrações de afeto e, ao mesmo tempo, racionalidade admirável. Ainda que os encontros, fisicamente, não sejam constantes, consolidamos um vínculo muito forte e que tenho certeza de que levarei para sempre comigo.

Agradeço, por fim, aos meus colegas de turma, especialmente Alifer, Fernanda, Muniz, Fifi e Matheus. Construímos relações muito bonitas de amizade, baseadas em coletividade e apoio mútuo. Nunca pensávamos apenas em nós mesmos, e essa rede de amizades tornava nosso aprendizado mais natural e proveitoso. Obrigada

pela prazerosa convivência de todos esses anos e por compartilharem comigo esta experiência de cursar Matemática Aplicada.

Conteúdo

1	Introdução	7
2	Dados e Metodologia	9
2.1	Modelos	9
2.1.1	Lasso	9
2.1.2	<i>Random Forest</i>	10
2.1.3	<i>Quantile Regression Forest</i>	11
2.1.4	LSTM	12
2.2	Transformação quantílica	13
3	Resultados	15
3.1	Análise dos Resíduos	15
3.1.1	Lasso	19
3.1.2	<i>Random Forest</i>	22
3.1.3	<i>Quantile Regression Forest</i>	22
3.1.4	LSTM	23
4	Discussão	24
4.1	População	25
4.2	Semanas epidemiológicas	26
4.3	<i>Quantile Regression Forest</i>	27
4.3.1	Rede de cidades	27
5	Conclusão	30

Resumo

Existe um interesse atual muito forte na implementação de novos modelos e métodos matemáticos na análise em tempo real de doenças transmissíveis com potencial epidêmico. Esses modelos incluem diferentes abordagens: sistemas dinâmicos, estatísticos, aprendizagem de máquina. O presente projeto visa aplicar metodologias matemáticas para predição de curto prazo de tendências em séries temporais de arboviroses. Analisamos os resíduos de quatro modelos diferentes de regressão e exploramos sua relação com as *features* dos modelos. Os resultados serão integrados no sistema Infodengue, em operação em 790 cidades brasileiras.

1 Introdução

A dengue se tornou um problema global desde o período da Segunda Guerra Mundial e é endêmica em inúmeros países, principalmente na Ásia e na América do Sul. Ela consiste numa arbovirose transmitida pelo mosquito *Aedes Aegypti* e, em termos sintomáticos, é caracterizada por febre bifásica, dor de cabeça, dor em várias partes do corpo, prostração, erupção cutânea, linfadenopatia e leucopenia [1]. Ela é tida como uma das mais importantes infecções virais humanas causadas por artrópodes, [2] e sabe-se que entre 50 e 528 milhões de pessoas são infectadas por ano, enquanto que aproximadamente 40.000 morrem [3]. Os locais nos quais encontra-se a transmissão da dengue têm se expandido nos últimos anos, e todos os quatro sorotipos do vírus (DENV-1-4) circulam na Ásia, África e nas Américas [2]. A estrutura molecular desses sorotipos tem sido estudada afim de entender suas relações evolutivas [4] e, nesse contexto, se evidencia a necessidade de estudar esta arbovirose mais profundamente a partir de diferentes abordagens, na tentativa de entender o que afeta predominantemente sua transmissão.

O projeto Infodengue, coordenado pelo PROCC/Fiocruz e a Escola de Matemática Aplicada (FGV) oferece modelagem online do risco da dengue desde 2014 [5], atualmente em operação em 790 municípios brasileiros em cinco estados diferentes. O projeto serve de plataforma de pesquisa de métodos quantitativos em epidemiologia, e suas metodologias estão em contínua evolução. Ultimamente, modelos preditivos têm sido o foco de pesquisadores ligados ao Infodengue [6, 7, 8]. Estes modelos — que têm sido desenvolvidos tendo como base séries temporais de arboviroses — requerem validação que justifique a utilização de suas previsões na gestão da saúde pública.

Evidentemente, a previsão do número de casos de dengue em uma localidade não é um problema trivial, visto que muitas variáveis de naturezas diferentes afetam sua disseminação. Estas englobam desde aspectos urbanos, como abastecimento regular de água, coleta pública de lixo e velocidade do transporte público [9], a aspectos climáticos, como temperatura e umidade. Além disso, visto que o vírus da dengue é transmitido aos humanos por meio de mosquitos da espécie *Aedes Aegypti*, seus índices de transmissão são altamente variáveis de acordo com a presença destes

insetos no ambiente. Em outras palavras, a necessidade do mosquito no processo de infecção nos traz uma forte sazonalidade aos seus padrões de incidência e, ao mesmo tempo, faz com que estes sejam altamente sensíveis à localidade em questão.

Idealmente, visando a predição do número de casos de dengue, procuramos desenvolver um modelo robusto, isto é, que apresente boa performance em contextos geográficos diferentes. Isso quer dizer que o modelo deveria ser capaz de prever o número de casos tanto em cidades grandes como em cidades menores. Um dos desafios para a validação de um modelo desse tipo é a grande heterogeneidade na escala dos dados devido aos tamanhos das populações em cada cidade.

O objetivo do presente trabalho é propor um método que torne os erros preditivos comparáveis através de ordens de grandeza de tamanhos de população. Partimos dos modelos desenvolvidos em [7] — que propôs que os erros dos modelos preditivos fossem comparados em escalas quantílicas — para explorar o problema mencionado. Gostaríamos de construir uma métrica de avaliação dos erros que levasse em conta o tipo de série temporal em questão: extremamente sazonal, não-estacionária e cujas distribuições possuem caudas pesadas.

Em suma, o objetivo é a comparação de modelos ao longo de várias escalas de valores esperados, que nos fornecem erros de magnitudes diferentes. Com isso, buscamos desenvolver métricas representativas de séries temporais de arboviroses.

2 Dados e Metodologia

Utilizamos aproximadamente 10 anos de dados de incidência semanal de Dengue e variáveis obtidas do projeto Infodengue. Focamos as análises em cidades do Rio de Janeiro e do Ceará, partindo dos modelos desenvolvidos em [7] para dar início à análise dos resíduos. Exploramos, portanto, três modelos de regressão: Lasso, *Random Forest* e LSTM. Além destes três modelos descritos em [7], desenvolvemos um outro modelo de predição que consiste numa generalização das Random Forests, que são as *Quantile Regression Forests*. A partir delas, começamos a explorar os quantis e, conseqüentemente, as distribuições acumuladas do número de casos de dengue para várias cidades.

2.1 Modelos

2.1.1 Lasso

O Lasso foi originalmente introduzido na literatura geofísica em 1986 e posteriormente popularizado em 1996 por Robert Tibshirani, que forneceu mais informações sobre seu desempenho [10].

Este modelo é um método de análise de regressão que é utilizado tanto para seleção de variáveis quanto para regularização, afim de melhorar a acurácia preditiva e interpretabilidade de um modelo estatístico. A regularização é uma forte ferramenta contra o sobreajuste (ou *overfitting*) de modelos de *machine learning*. Isso porque sua essência está na construção de uma medida para a complexidade do modelo e, em vez de minimizarmos o valor esperado do erro dentro da amostra, minimiza-se uma combinação dele e desta medida de complexidade. Isso evita o *overfitting* à medida em que restringe o algoritmo de aprendizado para ajustar os dados usando uma modelo relativamente simples. [11]

Considere um modelo de aprendizado em que H é o conjunto de polinômios em uma variável $x \in [-1, 1]$. Em vez de expressar os polinômios em termos de potências de x , nós os expressamos como uma combinação de polinômios de Legendre com boas propriedades analíticas que nos fornecem derivações mais simples [12]. O polinômio de Legendre de ordem zero é a constante $L_0(x) = 1$ e o de ordem 1 é dado por $L_1(x) = x$. Quando a ordem do polinômio de Legendre aumenta, a curva

se torna mais complexa. Os polinômios de Legendre são ortogonais entre si em $x \in [-1, 1]$, e qualquer polinômio regular pode ser escrito como uma combinação linear de polinômios de Legendre.

Modelos polinomiais são um caso especial de modelos lineares em um espaço \mathcal{Z} , sob uma transformação não linear $\Phi : \mathcal{X} \mapsto \mathcal{Z}$. Aqui, para o modelo polinomial de ordem Q , Φ transforma o vetor x em um vetor z dos polinômios de Legendre, como vemos abaixo

$$z = \begin{bmatrix} L_0(x) \\ L_1(x) \\ \vdots \\ L_Q(x) \end{bmatrix} \quad (1)$$

e, assim, segue que o conjunto de hipóteses do modelo é uma combinação linear desses polinômios de ordem 0 a Q . Temos:

$$H_Q = \{h \mid h(x) = w^T x = \sum_{q=0}^Q w_q L_q(x)\}_{w \in \mathbb{R}^{Q+1}} \quad (2)$$

onde $L_q(x)$ é o polinômio de Legendre de ordem q e w é o vetor de pesos relativo a x que buscamos encontrar. No caso do Lasso, temos $Q = 1$. [11]

2.1.2 *Random Forest*

Árvores de Decisão

As árvores de decisão são um tipo de método de aprendizado supervisionado que constrói modelos de classificação e regressão por meio de uma estrutura em árvore. O modelo aprende regras de decisão que são usadas na predição da variável resposta. A árvore de decisão é criada a partir da divisão do conjunto de dados em subconjuntos homogêneos menores e associando-os com base em alguma regra inferida.

As árvores de decisão são extremamente rápidas de ajustar e fáceis de interpretar. A grande desvantagem das árvores de decisão é sua propensão a *overfitting*. Uma tentativa de minimizar este efeito, é combinar os *outputs* de mais de uma árvore de decisão.

O modelo *Random Forest* faz parte de uma classe de métodos denominada *ensemble learning*. A característica principal de métodos deste tipo consiste na agregação dos resultados de vários classificadores [13]. Elas são um método de aprendizado para classificação e regressão construído a partir do conjunto de múltiplas árvores de decisão.

Breiman, em 2001, propôs as *Random Forests* [14]. Elas alteram a forma como as árvores de classificação ou regressão são construídas. Nas árvores padrão, cada nó é dividido a partir da melhor divisão entre todas as variáveis do problema, enquanto que em uma *Random Forest*, cada nó é dividido a partir de um subconjunto de preditores escolhidos aleatoriamente nesse nó.

2.1.3 *Quantile Regression Forest*

As clássicas *Random Forests* buscam aproximar a média condicional $E(Y|X = x)$, que é dada por uma média ponderada sobre todas as observações da variável resposta Y . Mas, além disso, esse processo também nos fornece uma boa aproximação de toda a sua distribuição condicional [15]. Temos

$$F(y|X = x) = P(Y \leq y|X = x) = E(1_{\{Y \leq y\}}|X = x). \quad (3)$$

A última expressão facilita a analogia com o *output* das *Random Forests*. Definimos uma aproximação para $E(1_{\{Y \leq y\}}|X = x)$ através da média ponderada sobre as observações de $1_{\{Y \leq y\}}$:

$$\hat{F}(y|X = x) = \sum_{i=1}^n w_i(x) 1_{\{Y \leq y\}}, \quad (4)$$

utilizando os mesmos pesos $w_i(x)$ do modelo de *Random Forests*. Essa aproximação é a chave para o desenvolvimento do algoritmo da *Quantile Regression Forest*. A sua principal diferença para as *Random Forests* é que, para cada nó em cada árvore, as *Random Forests* mantêm apenas a média das observações que caem nesse nó e negligenciam qualquer outra informação. Por outro lado, as *Quantile Regression Forests* mantêm o valor de todas as observações do nó (não apenas a média) e avaliam a distribuição condicional com base nessas informações.

2.1.4 LSTM

O modelo LSTM (*Long-short term memory*) faz parte de uma classe de modelos preditivos denominada Redes Neurais (*Neural Networks*).

As redes neurais são um modelo comumente utilizado na exploração de problemas que variam da previsão de séries temporais à classificação de imagens [11]. Redes neurais são uma generalização do perceptron, que consiste em um simples classificador linear binário. Elas são um modelo muito poderoso e flexível e, consequentemente, são um bom modelo candidato para aprender com os dados. Isso porque têm alto poder de aproximação de funções complexas e, durante o seu desenvolvimento, também foram desenvolvidos simultaneamente algoritmos muito robustos para lidar com os problemas de otimização resultantes de sua formulação. Na figura 1, encontramos um exemplo de como conseguimos construir preditores mais complexos a partir do encadeamento de modelos simples.

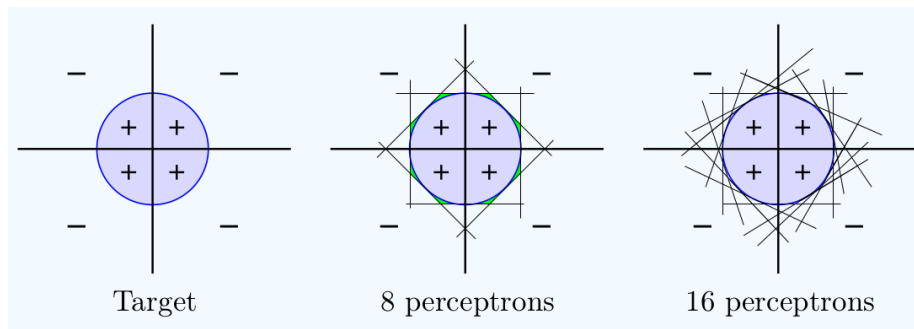


Figura 1: Gráficos extraídos de [11], que nos mostram a aproximação de uma função a partir de dois modelos diferentes.

Resumidamente, um modelo de redes neurais pode ser entendido como um classificador ou regressor que conta com algumas camadas de funções aproximadas pelos dados da amostra. Essas camadas são percorridas pelos dados quando queremos prever uma nova observação. O processo encontra-se exemplificado no esquema da figura 2.

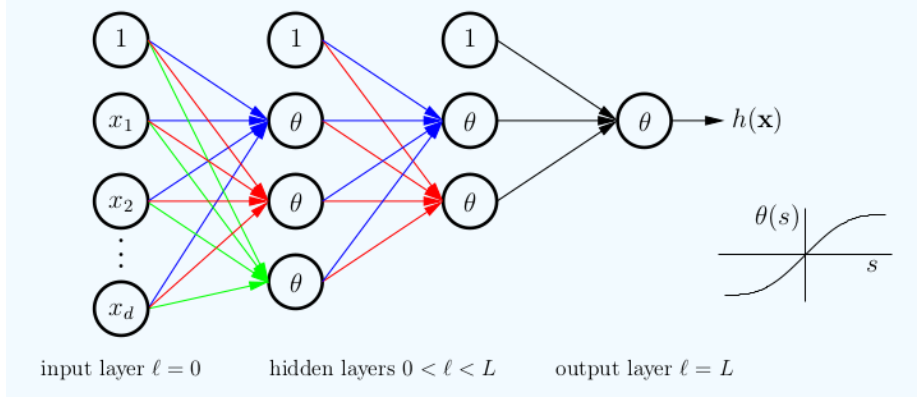


Figura 2: Esquema de uma rede neural extraído de [11].

RNN

Redes neurais recorrentes (RNN) são um caso particular de redes neurais tradicionais. A diferença é que possuem *loops*, de maneira a permitir a persistência de informações. Elas podem ser vistas como várias cópias da mesma rede, cada uma passando uma mensagem para um sucessor. [7]

A RNN aprende a usar as informações passadas, ou seja, elas guardam informações fornecidas no tempo t_0 ao prever o tempo t_α . Comumente têm um bom desempenho quando a diferença entre t e t_α é pequena, mas têm dificuldades em lidar com dependências de longo prazo.

As redes *Long Short-Term Memory* (LSTM) são uma versão modificada de redes neurais recorrentes, o que facilita a lembrança de dados passados. Um maior aprofundamento sobre o funcionamento de redes neurais LSTM encontra-se em [7].

2.2 Transformação quantílica

Afim de tornar os dados comparáveis em termos de magnitude, realizamos uma transformação quantílica. Ela se baseia na construção de funções de distribuição acumulada a partir dos dados de casos de dengue e suas previsões.

Uma função de distribuição acumulada empírica (CDF) é um estimador não paramétrico da CDF de uma variável aleatória. Ele é construído a partir da ordenação dos dados do menor para o maior em valor e, em seguida, calcula a soma das pro-

habilidades atribuídas até cada valor. O resultado é uma *step function*, cujo valor em qualquer ponto especificado da variável em questão é a fração de observações menores ou iguais ao valor especificado.

Seja uma semana t e uma cidade c . Definimos, assim, o conjunto $X_t^c := \{x_{t,c}^{(1)}, \dots, x_{t,c}^{(n)}\}$ como os números de casos de dengue na semana t do ano em n anos diferentes. Seja $P_t^c := \{p_{t,c}^{(1)}, \dots, p_{t,c}^{(n)}\}$ o conjunto de predições para a mesma cidade c e semana t do ano em n anos.

Definimos $W_t^c := X_t^c \cup P_t^c$, como o suporte de uma variável aleatória W . Conseguimos, portanto, computar a função de distribuição acumulada empírica abaixo:

$$F_W(w) = P(W \leq w), \quad (5)$$

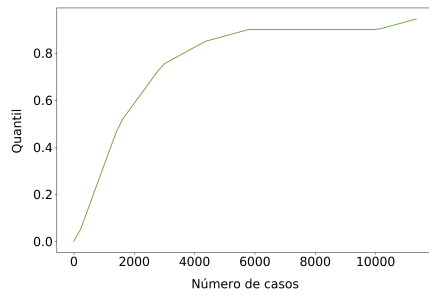
onde $w \in \mathbb{R}^{Q+1}$. Vale ressaltar que não precisamos ter $w \in W_t^c$, pois a função de distribuição acumulada assume valor igual a 0 quando avaliada em valores menores do que o mínimo encontrado no suporte e, analogamente, assume valor igual a 1 para valores maiores do que o máximo encontrado em seu suporte W_t^c . Segue que a função 5 está bem definida para todos os números reais.

Logo, para uma observação $x_{t,c}^{(k)}$, realizamos a transformação quantílica a partir da CDF construída para a cidade c na semana t . Assumindo que em nossa amostra trabalhamos com dados de n anos, temos

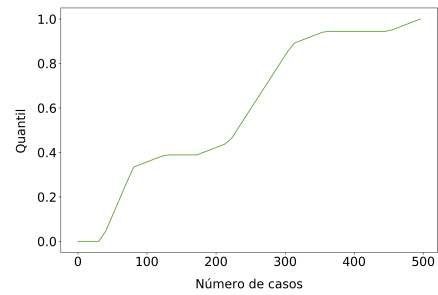
$$q(x_{t,c}^{(k)}) = P(W \leq x_{t,c}^{(k)}) := \frac{\sum_{i=1}^n 1_{\{x_{t,c}^{(i)} \leq x_{t,c}^{(k)}\}}}{n}. \quad (6)$$

A observação $x_{t,c}^{(k)}$, portanto, é representada por um quantil, construído a partir da distribuição acumulada de casos em sua respectiva cidade e semana epidemiológica do ano. O processo é análogo para obtermos o quantil da predição $p_{t,c}^{(k)}$.

Após a transformação quantílica, o erro é então definido como a distância entre o quantil predito e o quantil real. Temos, portanto, uma medida de erro que, necessariamente, sempre está no intervalo $[-1, 1]$.



Semana 20



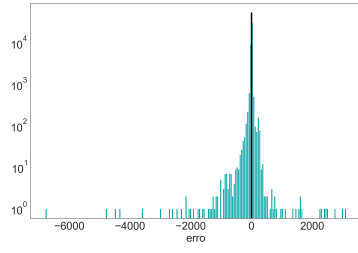
Semana 40

Figura 3: Exemplo de duas funções de distribuição acumulada de casos para a cidade do Rio de Janeiro. Ambas foram calculadas com base na união dos conjuntos de dados reais e predições. Aqui, utilizamos como exemplo os *outputs* do modelo Lasso. Note que, mesmo em uma mesma cidade, temos magnitudes de casos muito distantes para semanas diferentes do ano (eixo horizontal).

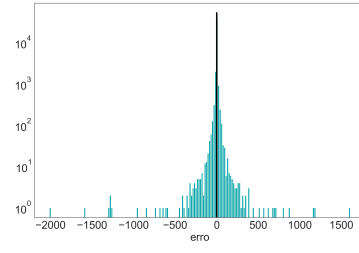
3 Resultados

3.1 Análise dos Resíduos

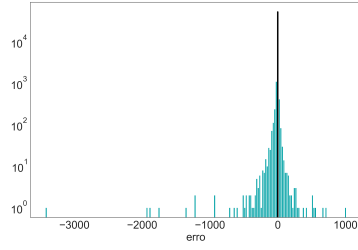
Analisamos os resíduos para os quatro modelos e, ao realizarmos transformações quantílicas, foi possível compará-los de maneira mais representativa em termos de performance. Nas próximas seções, encontramos algumas das análises que fizemos em torno dos erros, separadamente por modelo.



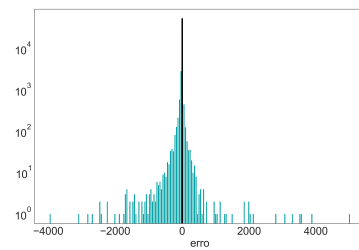
Lasso



Random Forest

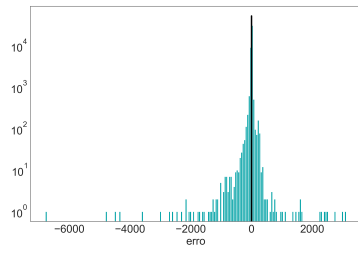


Quantile Forest

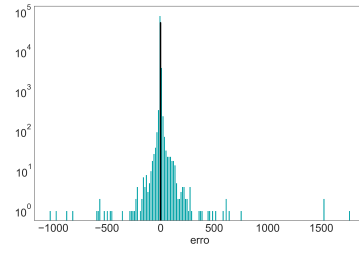


LSTM

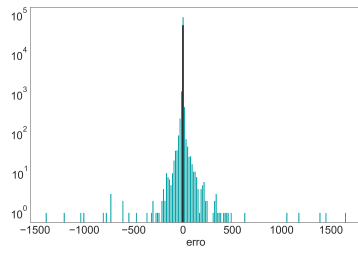
Figura 4: Histograma dos desvios dos modelos analisados, antes de aplicarmos a transformação quantílica. O desvio de uma predição p_t em t é dado por $p_t - x_t$, onde x_t é o dado real observado. Aqui, estão apenas os erros computados para as cidades do estado do **Rio de Janeiro**. As retas em preto marcam o quantil 0 e o eixo y está em escala logarítmica.



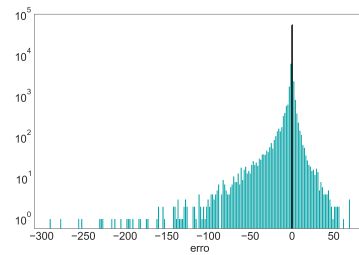
Lasso



Random Forest

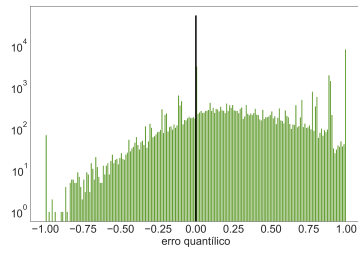


Quantile Forest

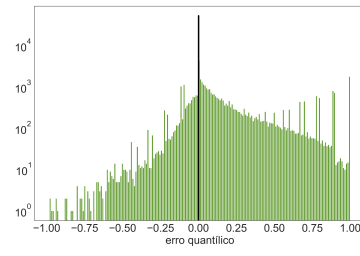


LSTM

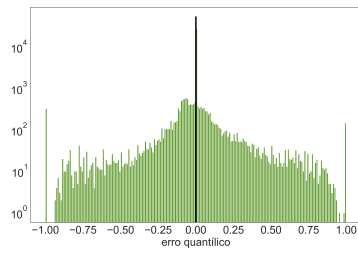
Figura 5: Histograma dos desvios dos modelos analisados, antes de aplicarmos a transformação quantílica. O desvio de uma predição p em t é dado por $p_t - x_t$, onde x_t é o dado real observado. Aqui, estão apenas os erros computados para as cidades do estado do **Ceará**. As retas em preto marcam o quantil 0 e o eixo y está em escala logarítmica.



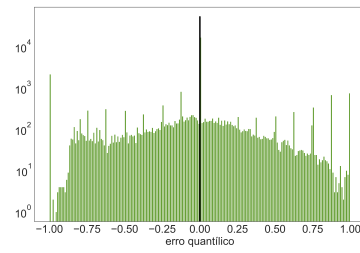
Lasso



Random Forest

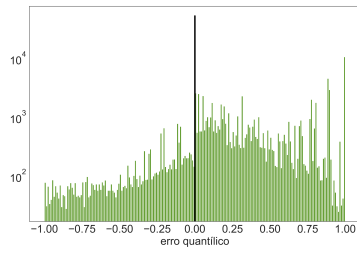


Quantile Forest

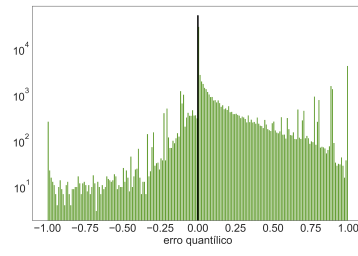


LSTM

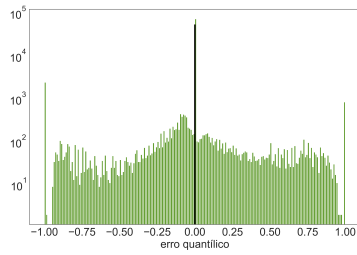
Figura 6: Histograma dos erros dos modelos analisados, após aplicarmos a transformação quantílica descrita na seção 2.2. Aqui, estão apenas os erros computados para as cidades do estado do **Rio de Janeiro**. As retas em preto marcam o quantil 0 e o eixo y está em escala logarítmica.



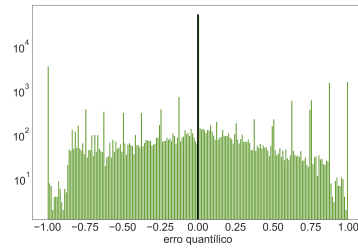
Lasso



Random Forest



Quantile Forest



LSTM

Figura 7: Histograma dos erros dos modelos analisados, após aplicarmos a transformação quantílica descrita na seção 2.2. Aqui, estão apenas os erros computados para as cidades do estado do **Ceará**. As retas em preto marcam o quantil 0 e o eixo y está em escala logarítmica.

3.1.1 Lasso

População

Uma análise importante que realizamos e que nos trouxe bastantes *insights* foi comparar os erros dos modelos a cada semana com as populações das respectivas cidades. Isso porque, comumente, cidades de populações menores também apresentam magnitudes de casos de dengue menores e, conseqüentemente, o conceito de "epidemia" varia de acordo com a cidade em questão. Com isso, buscamos avaliar os modelos em uma escala contínua de tamanhos de cidades.

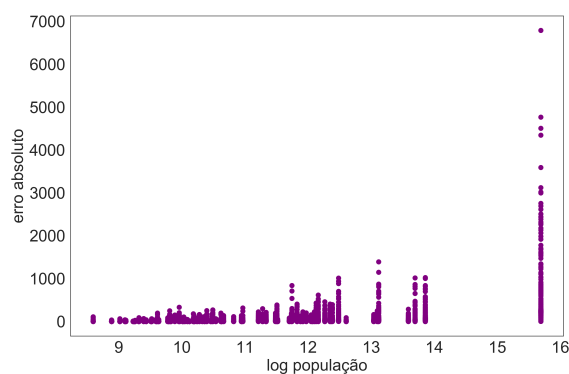
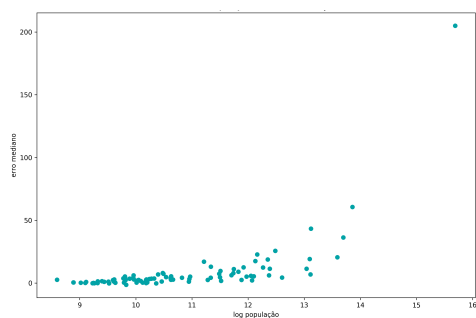
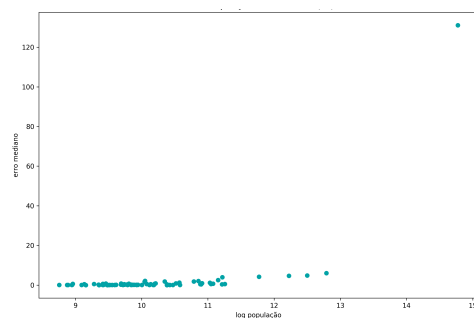


Figura 8: Erros absolutos semanais do modelo Lasso comparado às populações das respectivas cidades. Aqui, utilizamos o estado do Rio de Janeiro como exemplo. Cada ponto representa uma semana em uma cidade.

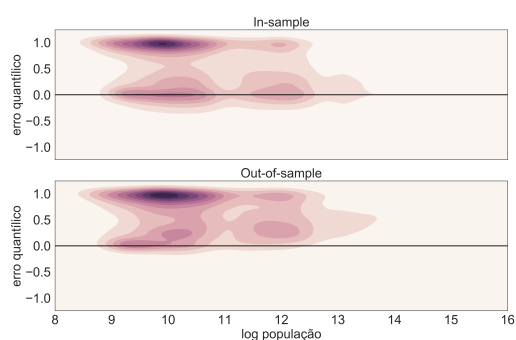


RJ

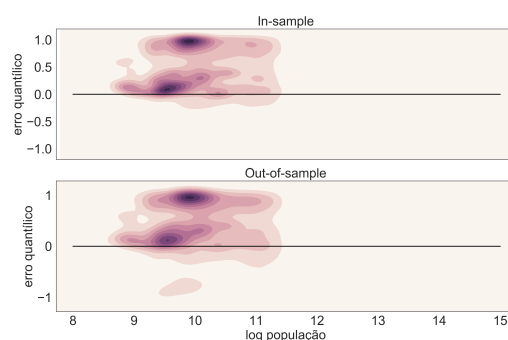


CE

Figura 9: Relação do erro mediano semanal do modelo com as populações das cidades do estado do Rio de Janeiro e do Ceará, agrupado por cidade.



RJ



CE

Figura 10: Relação do erro quantílico semanal do modelo com as populações das cidades do estado do Rio de Janeiro e do Ceará. Aqui, temos *scatterplots*, nos quais cada ponto representa uma semana em uma cidade. A reta em preto sinaliza o quantil 0.

Semanas

Afim de captar as tendências dos erros ao longo do ano, procuramos analisar sua dispersão nas semanas do ano. É, também, uma maneira de explorar as distribuições dos erros nos períodos epidêmicos e não-epidêmicos, visto que estamos trabalhando com séries extremamente sazonais.

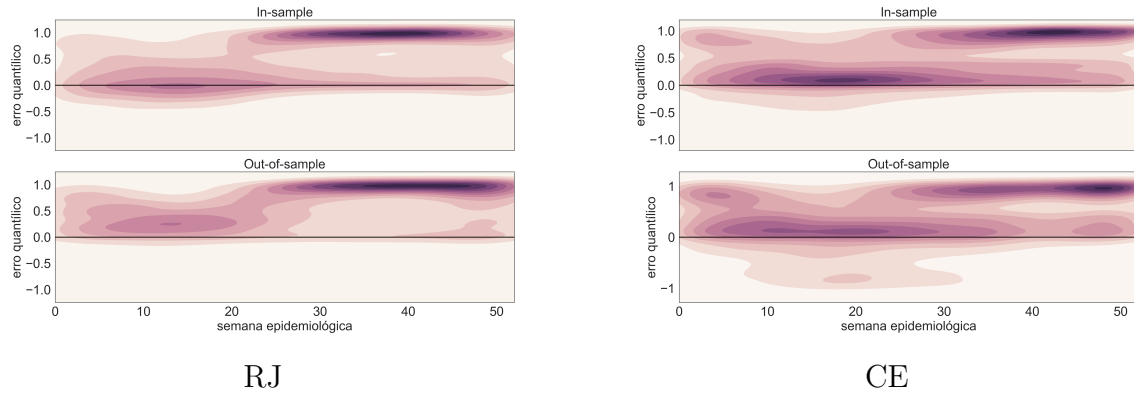


Figura 11: Relação do erro quantílico semanal do modelo em cidades do estado do Rio de Janeiro e do Ceará com as semanas do ano. Aqui, temos *scatterplots*, nos quais cada ponto representa uma semana em uma cidade.

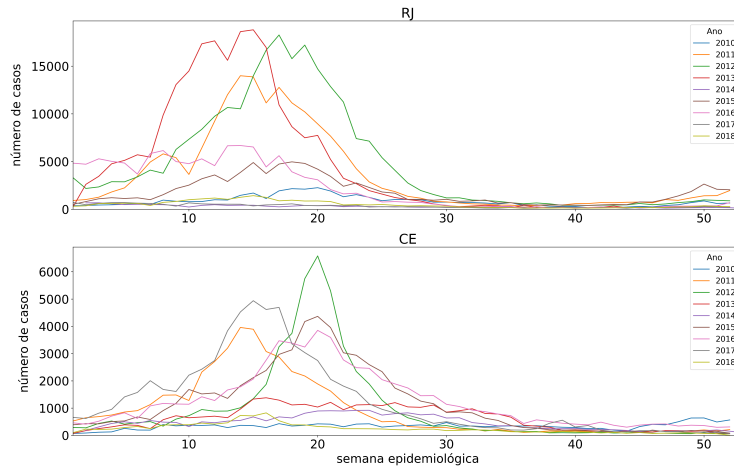


Figura 12: Séries temporais de casos de dengue no estado do Rio de Janeiro e do Ceará, separadamente.

3.1.2 *Random Forest*

População

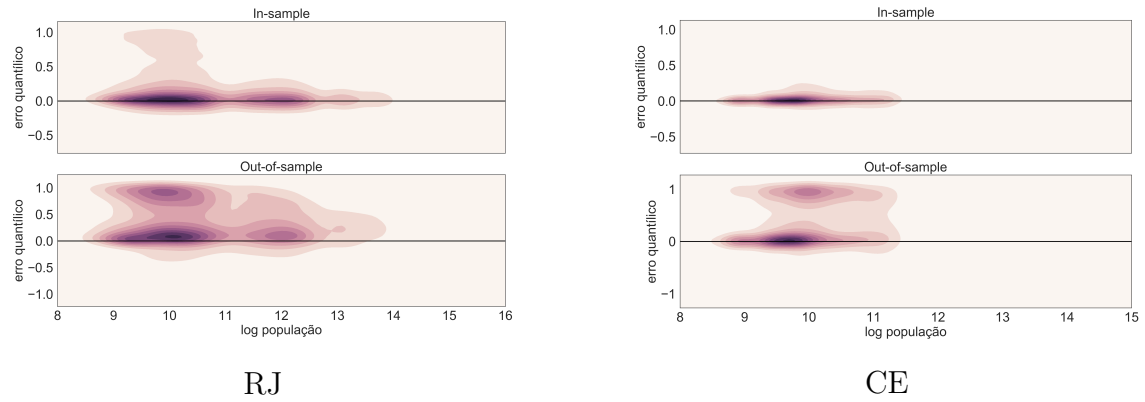


Figura 13: Relação do erro quantílico semanal do modelo com as populações das cidades do estado do Rio de Janeiro e do Ceará. Aqui, temos *scatterplots*, nos quais cada ponto representa uma semana em uma cidade. A reta em preto representa o quantil 0.

Semanas

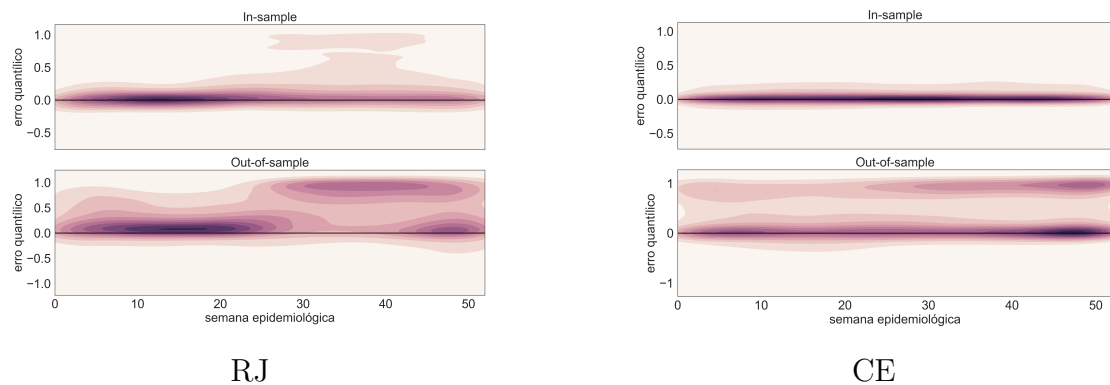


Figura 14: Relação do erro quantílico semanal do modelo em cidades do estado do Rio de Janeiro e do Ceará com as semanas do ano. Aqui, temos *scatterplots*, nos quais cada ponto representa uma semana em uma cidade. A reta em preto representa o quantil 0.

3.1.3 *Quantile Regression Forest*

População

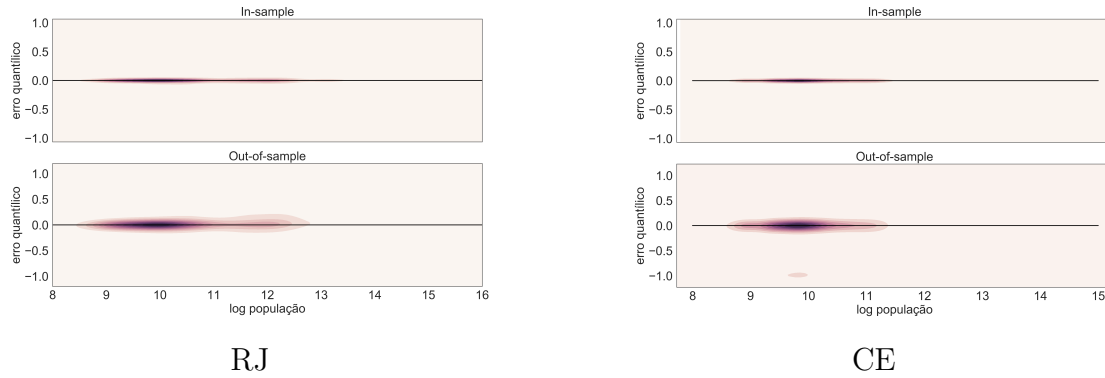


Figura 15: Relação do erro quantílico semanal do modelo com as populações das cidades do estado do Rio de Janeiro e do Ceará. Aqui, temos *scatterplots*, nos quais cada ponto representa uma semana em uma cidade. A reta em preto representa o quantil 0.

Semanas

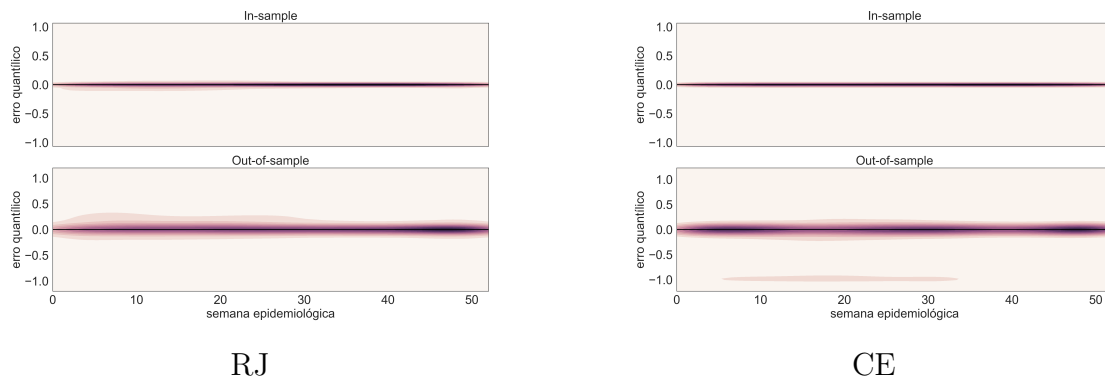


Figura 16: Relação do erro quantílico semanal do modelo em cidades do estado do Rio de Janeiro e do Ceará com as semanas do ano. Aqui, temos *scatterplots*, nos quais cada ponto representa uma semana em uma cidade. A reta em preto representa o quantil 0.

3.1.4 LSTM

População

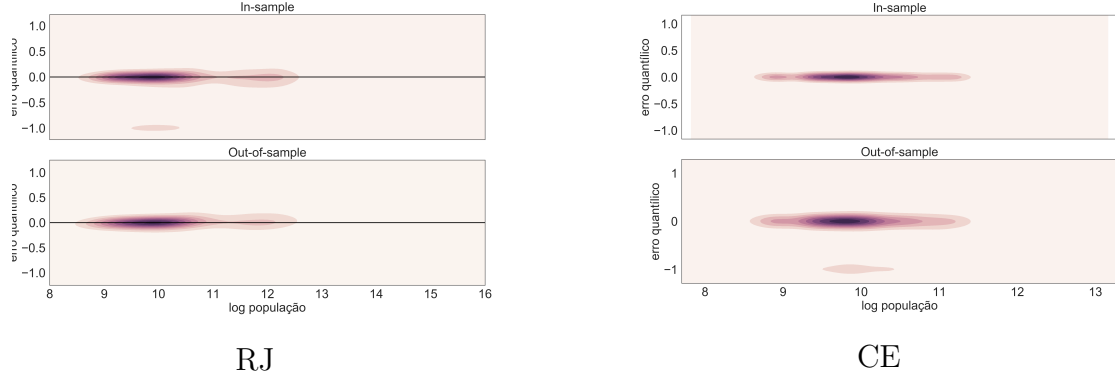


Figura 17: Relação do erro quantílico semanal do modelo com as populações das cidades do estado do Rio de Janeiro e do Ceará. Aqui, temos *scatterplots*, nos quais cada ponto representa uma semana em uma cidade. A reta em preto representa o quantil 0.

Semanas

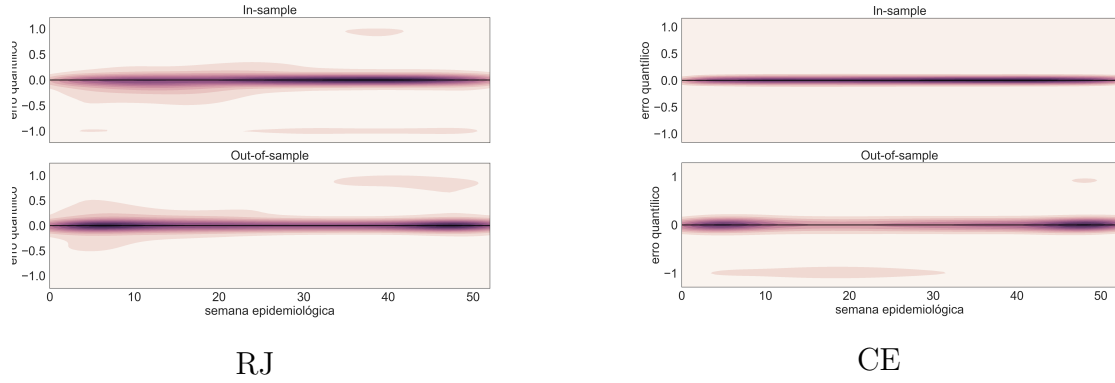


Figura 18: Relação do erro quantílico semanal do modelo em cidades do estado do Rio de Janeiro e do Ceará com as semanas do ano. Aqui, temos *scatterplots*, nos quais cada ponto representa uma semana em uma cidade. A reta em preto representa o quantil 0.

4 Discussão

Como mencionado na seção 1, o principal objetivo do presente trabalho é a comparação coerente e justa entre modelos de predição de casos de dengue, pois acreditamos que testes comparativos baseados em média são pouco informativos e estão longe de traduzir a natureza das séries temporais em questão. Queremos, com isso, ser capazes medir suas performances de uma maneira que faça sentido no dado contexto.

Se quiséssemos comparar modelos preditivos a partir de análises como as presentes na figuras 4 e 5, não seria algo trivial. Isso porque as distribuições de erro

convencionais de modelos consistentes comumente são centradas em zero e vêm de uma mesma família. Além disso, as diferenças entre elas não são estatisticamente significantes, visto que são altamente afetadas por valores extremos.

Por outro lado, se nos baseamos nas figuras 6 e 7, conseguimos diferenciar melhor os três modelos. Principalmente pelo fato de que os erros definidos a partir da diferença entre quantis estão limitados no intervalo $[-1, 1]$. Com isso, é possível perceber tendências de subestimação e superestimação e, ao mesmo tempo, temos uma melhor ideia de simetria e variância de suas respectivas distribuições.

4.1 População

Focamos boa parte de nossas análises na comparação dos erros com as populações das cidades. Isso se deve ao fato de que a magnitude de população de uma cidade nos dá uma ideia de seu tamanho e, conseqüentemente, afeta seus padrões de incidência.

Na figura 8, podemos ver um *heatmap* dos erros absolutos do modelo Lasso comparados aos tamanhos de população das respectivas cidades. Observamos um aumento da variância dos erros à medida em que a população cresce, o que é de se esperar. Um de nossos objetivos é conseguir comparar esses erros entre cidades de diferentes padrões demográficos e, conseqüentemente, diferentes padrões de incidência de arboviroses.

A partir dos gráficos da figura 9, poderíamos concluir que a performance do modelo Lasso é mais do que satisfatória, tendo em vista a alta concentração de erros próximos a zero. No entanto, analisando a figura 10, temos outra perspectiva.

Os gráficos presentes na figura 10 consistem em *scatterplots* em que cada ponto representa uma semana em uma cidade. A diferença é que os erros estão representados em escala quantílica, levando em consideração a distribuição acumulada dos erros para cada semana epidemiológica do ano. Assim, podemos observar que, embora tenhamos pequenos erros em termos de magnitude, se considerarmos a distribuição histórica de casos para as cidades, vemos que também temos alta concentração de erros quantílicos próximos a 1.

Um baixo desvio absoluto não é sinônimo de boa performance do modelo, visto que para cidades menos populosas é natural que os desvios sejam baixos. O que

nos interessa é saber se os desvios, de fato, são consistentes com os dados históricos, bem como com o contexto espaço-temporal em questão.

Ao compararmos a figura 13 com a figura 10, podemos ver que a *Random Forest* nos fornece erros quantílicos com menor dispersão em relação à população, quando comparada ao Lasso. Isso nos dá uma intuição de que o modelo *Random Forest* performe melhor e de maneira mais consistente.

Quando comparamos as figuras 13 e 10, vemos que quanto mais baixa for a variância, melhor será a performance do modelo. De certa forma, isso nos ajuda na criação de uma função de perda para os modelos: ao final, queremos que um modelo de baixa variância residual seja pouco penalizado.

Comparando as figuras 15 e 13, é nítido que a *Quantile Forest*, quando comparada à *Random Forest*, apresenta erros de menor dispersão em relação à população das cidades, tanto dentro como fora da amostra. O fato de trabalharmos com cidades muito pequenas (i.e.: que nos fornecem muitas observações iguais a zero) não pareceu afetar a boa performance do modelo.

Percebemos que, mesmo comparado a um modelo de redes neurais, o modelo *Quantile Forest* parece superá-lo, em termos de dispersão dos erros quantílicos (figuras 15 e 17). Um aspecto que, talvez, deva ser explorado mais a fundo é o fato de que o LSTM parece manter sua performance *in-sample* no conjunto *out-of-sample*. O modelo *Quantile Forest* tem pior performance *out-of-sample* quando comparada ao seu próprio desempenho *in-sample*.

4.2 Semanas epidemiológicas

Ao analisarmos a figura 11 juntamente com a figura 12, é possível notar algumas relações importantes em termos de performance ao longo do ano. Da figura 11, podemos dizer que os erros do modelo, em termos quantílicos, são menores nas semanas iniciais do ano, mais especificamente até a semana 30. Da semana 30 em diante, a performance cai consideravelmente, tanto para o Rio de Janeiro quanto

para o Ceará ¹.

Comparando com o que nos mostra a figura 12, vemos que a melhor performance do modelo se dá justamente nos períodos epidêmicos, ou seja, de maior incidência de dengue.

Percebemos que, no geral, o modelo Lasso apresenta uma tendência de superestimação do número de casos, visto que os erros quantílicos são majoritariamente positivos.

Para o modelo *Random Forest*, a performance parece um pouco mais consistente durante o ano, principalmente para o estado do Ceará (figura 14). Ainda assim, percebemos uma leve queda de performance nas semanas finais do ano, que consiste no período inter-epidêmico. Também observamos esse aspecto no modelo Lasso, porém com mais intensidade (figura 11).

Vimos que o modelo *Quantile Forest* performa surpreendentemente bem ao longo do ano, com baixíssima dispersão em relação ao erro quantílico igual a zero.

Um aspecto que vale ser ressaltado sobre este conjunto de dados específico é o fato de que, nos anos de 2017 e 2018, foi observada uma incidência de dengue atipicamente menor, tanto em períodos epidêmicos como em inter-epidêmicos. Estes dois anos formam os conjuntos *out-of-sample* e, provavelmente, isso afetou a performance dos modelos por meio de uma superestimação de número de casos, principalmente Lasso e *Random Forest* (figuras 10 e 14).

4.3 *Quantile Regression Forest*

4.3.1 Rede de cidades

Os modelos desenvolvidos em [7] constroem regressores diferentes para cada uma das cidades. Isto é, a *Random Forest*, por exemplo, nos fornece um modelo diferente

¹Focamos as análises comparativas nesses dois estados por apresentarem aspectos geográficos e climáticos bem diferentes.

para cada cidade de cada um dos três estados explorados. O mesmo acontece para os outros modelos.

Uma característica importante da modelagem das *features* foi a *clusterização* de cidades, que consiste numa técnica estatística para análise de dados extremamente útil na análise exploratória [16]. Parece razoável dizer que, não apenas aspectos da própria cidade explicam sua incidência de dengue, mas também características de municípios próximos à cidade em questão. Sendo assim, para cada estado, foram construídos *clusters* hierárquicos de cidades com base em suas características intrínsecas. Todos os três modelos em [7] obtiveram melhores resultados após a criação dos *clusters*.

Portanto, para a predição do número de casos de dengue em uma cidade, foram consideradas também informações sobre cidades próximas a ela.²

Para os modelos construídos a partir de árvores de decisão (em nosso caso, *Random Forest* e *Quantile Regression Forest*), conseguimos facilmente quantificar o poder preditivo de cada *feature* para os modelos. Esta métrica é chamada de *feature importance* e nos trouxe algumas informações interessantes.

A partir das *feature importances* do modelo em cada cidade, construímos uma medida de influência entre cidades. Seja n o número de *features* de uma cidade i . Temos que a influência da cidade j na cidade i é definida como:

$$\mu_{ij} := \sum_{k=1}^n f_{ij}^{(k)}, \quad (7)$$

onde $f_{ij}^{(k)}$ representa a importância de uma *feature* k da cidade j utilizada no modelo da cidade i . Note que não necessariamente temos $\mu_{ij} = \mu_{ji}$, visto que a cidade i pode ser muito relevante para o modelo de j , mas a cidade j pode ter baixo poder preditivo para a cidade i .

²Ao falarmos em distância, não necessariamente englobamos o aspecto geográfico. Podemos encontrar cidades com padrões de incidência de dengue muito parecidos e que, no entanto, não são geograficamente próximas.

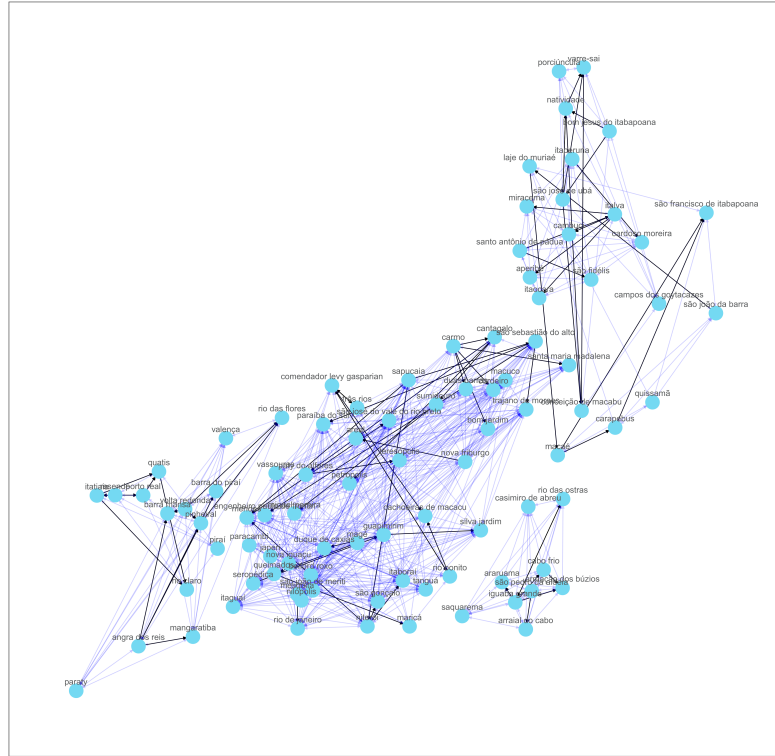


Figura 19: Grafo direcionado de influências entre cidades do Rio de Janeiro. Cada aresta ligando uma cidade j a outra cidade i representa a influência definida em (7). As arestas em azul claro representam influências menores ou iguais a 0,2 e as arestas em preto, nos mostram influências maiores que este valor.

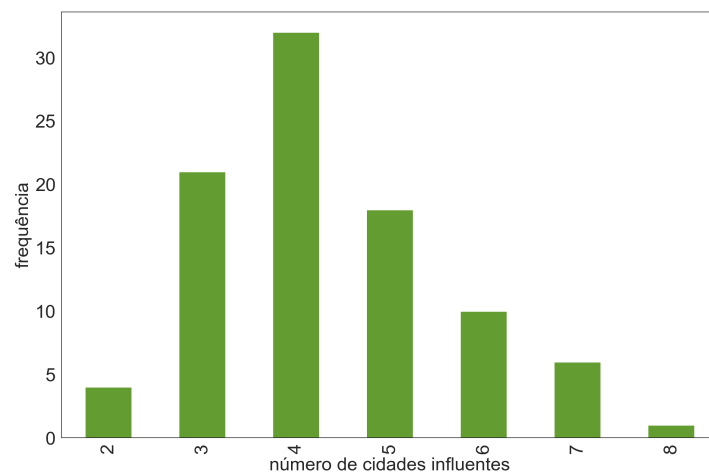


Figura 20: Histograma da frequência da quantidade de cidades do Rio de Janeiro que aparecem nas 10 *features* mais importantes de cada uma delas.

Na figura 20, podemos ver que nenhuma cidade do Rio de Janeiro obteve apenas uma cidade influente. Isso significa dizer que, para todas elas, *features* de outras cidades mostraram um alto poder preditivo.

5 Conclusão

Como inicialmente descrito na seção 1, nosso objetivo principal era conseguir, de algum modo, analisar de maneira mais justa e coerente as performances de quatro modelos fundamentalmente diferentes para a predição de casos de dengue em uma série de cidades do Brasil. Tivemos bastante cuidado ao escolher as localidades com as quais trabalharíamos, buscando aquelas cujos aspectos climático-geográficos e urbanos fossem consideravelmente distantes para, com isso, verificar as performances dos modelos nos contextos mais diversificados possíveis.

Procuramos uma maneira alternativa de definir o erro de uma predição. Queríamos, neste processo de avaliação dos modelos, incorporar características específicas conhecidas das séries temporais com as quais trabalhamos, como bem definidas na seções 1 e 2. Para nós, não faz sentido simplesmente nos apoiarmos nos erros absolutos médios, já que isso em nenhum sentido traduz os processos epidemiológicos em questão.

As análises presentes nas últimas seções tornaram mais fácil e natural a comparação entre os modelos propostos. No entanto, acreditamos que testes estatísticos mais profundos devam ser realizados, a partir dos erros quantílicos desenvolvidos. Podemos encontrar em [17] alguns métodos que poderíamos ter como ponto de partida.

É evidente que a métrica de penalidade dos modelos baseada em quantis traduz melhor seus respectivos comportamentos, além de permitir um entendimento mais profundo de suas distribuições. Certamente, uma abordagem futura possível consiste em utilizar a métrica quantílica como função de perda dos modelos, ou seja, alterar a medida que eles buscam otimizar. Nossa hipótese é de que provavelmente as performances irão crescer ainda mais.

Referências

- [1] Henchal, E. A. & Putnak, J. R. The dengue viruses. *Clinical microbiology reviews* **3**, 376–396 (1990).
- [2] Guzman, M. G. *et al.* Dengue: a continuing global threat. *Nature reviews microbiology* **8**, S7–S16 (2010).
- [3] Roth, G. A. *et al.* Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the global burden of disease study 2017. *The Lancet* **392**, 1736–1788 (2018).
- [4] Rodriguez-Roche, R. *et al.* Virus evolution during a severe dengue epidemic in cuba, 1997. *Virology* **334**, 154–159 (2005).
- [5] Codeco, C. *et al.* Infodengue: a nowcasting system for the surveillance of dengue fever transmission. *BioRxiv* 046193 (2016).
- [6] Codeco, C. *et al.* Infodengue: A nowcasting system for the surveillance of arboviruses in brazil. *Revue d'Épidémiologie et de Santé Publique* **66**, S386 (2018).
- [7] Mussumeci, E. *A machine learning approach to dengue forecasting: comparing LSTM, Random Forest and Lasso*. Ph.D. thesis (2018).
- [8] Santos, B. M. C. d., Coelho, F. C., Armstrong, M., Saraceni, V. & Lemos, C. Zika: an ongoing threat to women and infants. *Cadernos de saude publica* **34**, e00038218 (2018).
- [9] TaUIL, P. L. Urbanization and dengue ecology. *Cadernos de saúde pública* **17**, S99–S102 (2001).
- [10] Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58**, 267–288 (1996).
- [11] Abu-Mostafa, Y. S., Magdon-Ismail, M. & Lin, H.-T. *Learning from data*, vol. 4 (AMLLBook New York, NY, USA:, 2012).

- [12] Legendre, A. M. *Recherches sur l'attraction des sphéroïdes homogènes* (De l'Imprimerie Royale, 1785).
- [13] Liaw, A., Wiener, M. *et al.* Classification and regression by randomforest. *R news* **2**, 18–22 (2002).
- [14] Breiman, L. Random forests. *Machine learning* **45**, 5–32 (2001).
- [15] Meinshausen, N. Quantile regression forests. *Journal of Machine Learning Research* **7**, 983–999 (2006).
- [16] Anderberg, M. R. *Cluster analysis for applications: probability and mathematical statistics: a series of monographs and textbooks*, vol. 19 (Academic press, 2014).
- [17] Gneiting, T. & Ranjan, R. Comparing density forecasts using threshold-and quantile-weighted scoring rules. *Journal of Business & Economic Statistics* **29**, 411–422 (2011).