MONITORAMENTO ONLINE DA DENGUE: USANDO O GOOGLE PARA PREDIZER EPIDEMIAS

Letícia Octaviano da CRUZ[1]
Rafael IZBICKI[1]

- RESUMO: Infelizmente, relatórios oficiais sobre o avanço no número de casos de epidemias como a dengue levam muito tempo para serem divulgados, o que faz com que a população viva em estado de desinformação e gestores públicos tenham dificuldades em implementar políticas para combater as proliferações. Neste trabalho, temos como objetivo mostrar como dados coletados gratuitamente em tempo real pelo Google Trends podem ser utilizados para que se preveja o número diário de novos casos de dengue no estado de São Paulo. Para estimar a evolução de tal quantidade, foram utilizados o método dos mínimos quadrados, lasso e florestas aleatórias, com covariáveis criadas a partir da mensuração do volume de buscas diárias de algumas palavras-chave, como "tratamento dengue", "sintomas dengue" e "febre dengue". O modelo estimado através do método dos mínimos quadrados foi o que apresentou melhores predições, se mostrando adequado para fornecer estimativas atualizadas do número de casos de dengue para um período de até oito meses após a divulgação do último relatório oficial. Isso possibilita que os órgãos competentes e o governo adotem medidas necessárias para contenção de epidemias de modo mais eficaz e barato que o tradicional.

- PALAVRAS-CHAVE: Epidemia ; dengue; predição; infoveillance.

1 Introdução

A popularização do acesso à internet por meio de computadores pessoais, celulares e tablets propicia um aumento substancial na quantidade de informação disponível na rede. Em particular, o registro de interações de usuários com a internet através de ferramentas de busca como o Google e redes sociais como o Twitter oferece uma excelente oportunidade para a elaboração de bancos de dados

[1]Universidade Federal de São Carlos - UFSCar, Departamento de Estatística, CEP:13565-905, São Carlos, SP, Brasil. E-mail: *leticia.estat@yahoo.com; rafaelizbicki@gmail.com*

que permitem a investigação de fenômenos sociais, econômicos e políticos, entre outros. Neste trabalho, utilizamos dados do Google com a finalidade de fornecer *insights* sobre a proliferação de uma epidemia no território brasileiro: a dengue.

Tradicionalmente, relatórios oficiais sobre a proliferação de dengue e outras epidemias levam muito tempo para serem divulgados. Isso ocorre pois, além das inúmeras burocracias advindas de processos públicos, estes se baseiam em casos reportados por postos de saúde, os quais precisam ser compilados de modo a se criar uma única base de dados consistente. Assim, as informações disponíveis para a população sobre certa epidemia não refletem o presente, mas sim o status de sua proliferação semanas antes da divulgação oficial. Por outro lado, uma das grandes vantagens do uso de dados provindos da internet é que estes podem ser processados quase que instantaneamente, de modo que informações relativas à proliferação de uma dada epidemia cheguem à população o mais rapidamente possível. Assim, estes recursos podem ser usados para que o governo e outros órgãos competentes adotem as medidas necessárias para a contenção de epidemias de modo mais eficaz que o tradicional. Além disso, o uso de dados da internet possui a vantagem de que o custo para sua extração é muito pequeno quando comparado ao custo de obtenção de dados tradicionais. O uso de dados provindos da internet para monitoramento de dados relativos à saúde pública é chamado de *Infoveillance* (EYSENBACH, 2009).

Diversos são os autores que propõe ferramentas para o monitoramento de epidemias em tempo real com base em dados obtidos na internet. A Google Flu Trends (GOOGLE, 2017b), por exemplo, baseia-se em dados sobre a quantidade de buscas de determinadas palavras-chave no Google. Utilizando essas informações, a Google Flu Trends consegue prever precisamente a quantidade de pessoas com gripe em diversas regiões do mundo (COOK et al., 2011; GINSBERG et al., 2009). Outra ferramente que tem objetivo similar, mas que usa dados provindos do Twitter, é o Flu Detector (LAMPOS; CRISTIANINI, 2012). Contudo, as previsões fornecidas por esta são feitas somente para a região do Reino Unido. Para algumas referências adicionais sobre monitoramento online de epidemias de gripe (DREDZE et al., 2014; LAMPOS; CRISTIANINI, 2010; LAMPOS; BIE; CRISTIANINI, 2010; LEE; AGRAWAL; CHOUDHARY, 2013; STILO et al., 2014).

O sucesso da Google Flu Trends levou o Google a criar a Google Dengue Trends (GOOGLE, 2017a), usada para monitorar epidemias de dengue. Outros autores (EL-METWALLY, 2015; GLUSKIN et al., 2014) também criaram ferramentas com o objetivo de monitorar esta doença. Infelizmente, estes recursos consideram o Brasil como uma unidade só, isto é, não há previsões separadas para cada estado, algo de extremo interesse tanto para a população brasileira, quanto para gestores de políticas públicas. Ainda que no Brasil exista o *Info Dengue Rio* (INFODENGUE, 2017), este é restrito ao estado do Rio de Janeiro. Além disso, não há informações detalhadas de como os modelos são construídos, de modo que não é possível reproduzi-los e melhorá-los. Neste trabalho, temos como objetivo mostrar como o Google Trends pode ser utilizado para estimar o número diário de novos casos de dengue no estado de São Paulo, no sentido de predizer epidemias que possam estar acontecendo antes mesmo de estatísticas oficiais serem compiladas.

Na Seção 2 descrevemos os dados utilizados neste trabalho, bem como a metodologia que foi aplicada para a criação de modelos preditivos. A Seção 3 apresenta os resultados obtidos. Finalmente, na Seção 4 apresentamos as conclusões deste trabalho, assim como perspectivas de trabalhos futuros.

2 Metodologia

2.1 Descrição dos dados

Utilizamos dados provindos da interação de usuários com a internet para monitorar a epidemia de dengue no estado de São Paulo. Em particular, foram utilizados dados provindos de três fontes: Portal da Saúde (SAÚDE, 2017), Ministério da Saúde (através da Lei de Acesso à Informação) e Google Trends, ferramenta com a qual é possível verificar (gratuitamente) a frequência com que um termo é procurado ao longo de um período em cada região do planeta.

Para os anos de 2004 à 2014 foram utilizados dados recebidos por e-mail através de contato direto com o Ministério da Saúde, contendo o número de novos casos de Dengue no estado de São Paulo por semana epidemiológica – um padrão utilizado por convenção internacional[1]. Para a quantificação dos números de novos casos de dengue no ano de 2015 foram utilizados os boletins epidemiológicos disponibilizados no site do Portal da Saúde, em que os dados apresentam o número de casos acumulados por semana epidemiológica.

Finalmente, foram feitas buscas no Google Trends por diversas palavras chaves relacionadas à epidemia estudada. Mais especificamente, buscamos por "manchas dengue", "sintomas dengue", "febre dengue", "tratamento dengue", "dengue plaquetas"e "chikungunya". Os dados do Google Trends não representam os números referentes ao volume absoluto de pesquisa, mas apenas uma taxa que varia de 0 a 100, que representa a relação entre o número de pesquisas efetuadas para cada termo e o número total de pesquisas efetuadas no Google ao longo do tempo. Isto é, os dados são normalizados e essas taxas são calculadas de 7 em 7 dias. Restringimos as buscas ao estado de São Paulo, isto é, apenas buscas feitas nesta região foram contabilizadas.

Como os dias de observação dos novos casos e da taxa de buscas de um termo não seguem o mesmo padrão (enquanto o número de novos casos esta por semana epidemiológica, a taxa de buscas das palavras estão por semana), estimamos o número de novos casos para todos os dias através do ajuste de uma curva suave via *Smoothing Spline*; veja exemplo na Figura 1. Para o cálculo da taxa de buscas diárias foi utilizada uma média ponderada, em que foi dado maior peso para a taxa mais próxima da observação a ser estimada.

[1] As semanas são contadas de domingo a sábado, sendo a primeira semana do ano aquela que contém o maior número de dias de janeiro e a última a que contém o maior número de dias de dezembro.

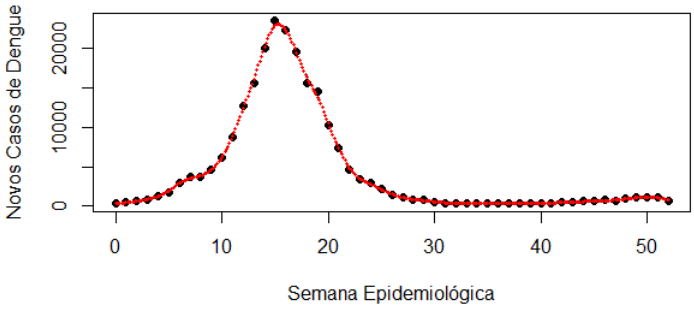


Figura 1 – Exemplo de curva ajustada através do método *Smoothing Spline* para o número de novos casos de Dengue.

Com a finalidade de facilitar a comparação do volume de casos de dengue e o volume de pesquisa das palavras relacionadas a epidemia, normalizamos o número de novos casos de dengue, de modo que o mesmo varie de 0 a 100, como os dados do Google Trends. Com isso, obtivemos um banco de dados composto de 4382 linhas – correspondentes à quantidade de dias contemplados – e 8 colunas – correspondentes às taxas de busca das palavras-chaves, à taxa de novos casos e ao intervalo de tempo calculado tendo como ponto inicial o primeiro dia de 2004. A Tabela 1 apresenta as dez primeiras e as dez últimas linhas do banco de dados utilizado para a realização das análises.

Tabela 1 – Primeiras e últimas observações do banco de dados contendo as taxas de busca das palavras chaves, a taxa de novos casos de dengue e o intervalo de tempo contado a partir do primeiro dia do ano de 2004

Dia	taxa_FebreDengue	taxa_Chikungunya	taxa_ManchasDengue	taxa_DenguePlaquetas	taxa_SintomasDengue	taxa_TratamentoDengue	taxa_Observada
0	0.00	0.00	0.00	0.00	0.00	0.00	0.71
1	0.00	0.00	0.00	0.00	0.00	0.00	0.71
2	0.00	0.00	0.00	0.00	0.00	0.00	0.70
3	0.00	0.00	0.00	0.00	0.00	0.00	0.69
4	0.00	0.00	0.00	0.00	0.00	0.00	0.68
5	0.00	0.00	0.00	0.00	0.00	0.00	0.68
6	0.00	0.00	0.00	0.00	0.00	0.00	0.67
7	0.00	0.00	0.00	0.00	0.00	0.00	0.66
8	0.00	0.00	0.00	0.00	0.00	0.00	0.65
9	0.00	0.00	0.00	0.00	0.00	0.00	0.65
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
4371	13.57	75.57	21.86	0.00	5.00	13.86	8.64
4372	14.43	79.43	23.14	0.00	5.00	14.14	8.65
4373	15.29	83.29	24.43	0.00	5.00	14.43	8.66
4374	16.14	87.14	25.71	0.00	5.00	14.71	8.67
4375	17.00	91.00	27.00	0.00	5.00	15.00	8.68
4376	16.33	88.00	26.00	0.00	5.00	14.78	8.70
4377	15.91	86.09	25.36	0.00	5.00	14.64	8.71
4378	15.62	84.77	24.92	0.00	5.00	14.54	8.72
4379	15.40	83.80	24.60	0.00	5.00	14.47	8.74
4380	15.24	83.06	24.35	0.00	5.00	14.41	8.75
4381	15.11	82.47	24.16	0.00	5.00	14.37	8.76

Notamos que, como os métodos propostos neste trabalho só utilizam dados

públicos para sua construção, não foi necessária a aprovação do estudo em um comitê de ética. Também notamos que o sucesso dos métodos preditivos construídos nesse trabalho depende fortemente de fatores que podem mudar com o tempo. Por exemplo, um termo que é importante hoje para predizer epidemiais de dengue pode vir a se tornar inútil em no futuro por questões alheias ao problema da dengue. Por exemplo, se algum vídeo no Youtube intitulado "febre dengue" vier a fazer muito sucesso, o número de buscas por esse termo pode vir a aumentar sem haver qualquer relação com a dengue.

2.2 Modelo de regressão

Utilizamos o banco de dados criado com o objetivo de estimar a taxa de novos casos de dengue no tempo t (variável resposta, y_t) com base na taxa de buscas de expressões relacionadas com a dengue no Google Trends neste mesmo tempo (covariáveis, $\mathbf{x}_t$).

Para estimar tal quantidade, ajustamos um método de regressão, $r(\mathbf{x})$, em $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_{t-1}, y_{t-1})$. Com base na função estimada, podemos então prever $y_t, y_{t+1}, \ldots$ utilizando $\widehat{r}(\mathbf{x}_t), \widehat{r}(\mathbf{x}_{t+1}), \ldots$. Os métodos que foram utilizados para estimar a função de regressão $r(\mathbf{x})$ foram:

- [**Método dos mínimos quadrados**] Assumimos que $r(\mathbf{x})$ é linear
- [**Lasso**] A escolha do parâmetro de penalização λ é feita através da validação cruzada, isto é, estimamos o risco preditivo em cada λ de interesse e selecionamos o que apresenta o melhor modelo.
- [**Florestas aleatórias**] Como recomendado por (JAMES et al., 2013), o número de preditores considerados em cada divisão foi de $m \approx \sqrt{p}$, em que p é o número de palavras-chave. Além disso, o número de árvores utilizadas foi de 500.

3 Resultados

3.1 Análise descritiva

Nesta seção são apresentados resumidamente os dados do número de novos casos de dengue no Estado de São Paulo e a taxa de buscas das palavras relacionadas à dengue no Google.

A Figura 2, que apresenta o número de novos casos de dengue no Estado de São Paulo no decorrer dos anos de 2004 e 2015, indica que não havia um volume substancial de novos casos de dengue até o ano de 2009. A partir deste ano, houve um aumento do número de novos casos, porém este diminuiu gradativamente até 2013, quando voltou a crescer. No ano de 2015, este número atingiu seu ponto máximo. Pode-se também notar uma periodicidade no número de novos casos de dengue ao longo do ano: no início de cada ano sempre há mais casos, possivelmente devido à chuva e ao calor.

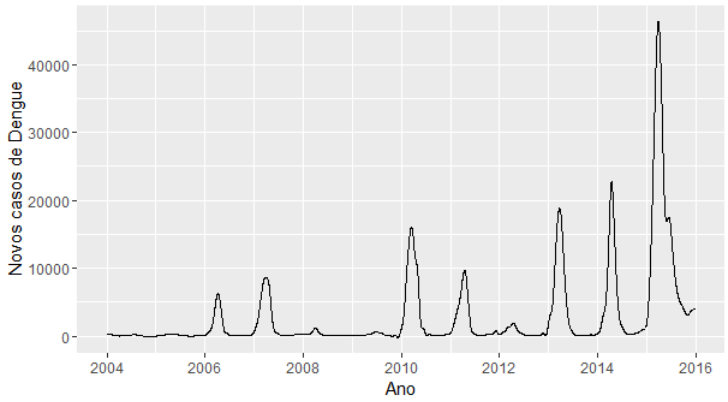
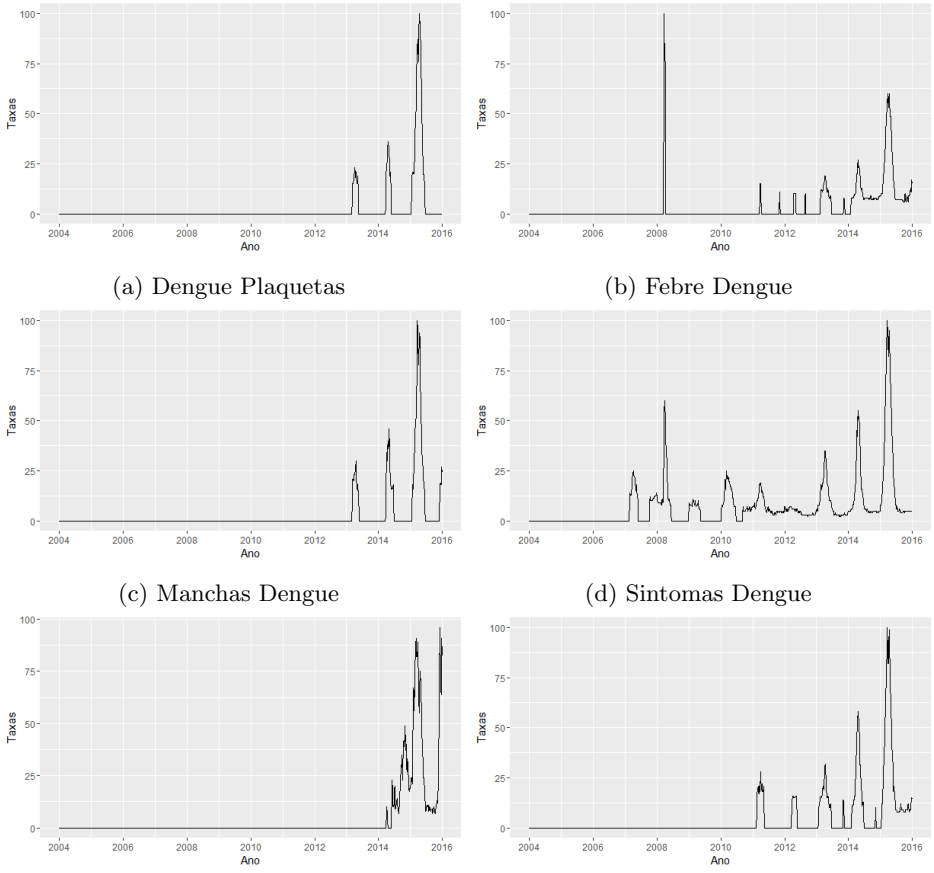


Figura 2 – Número de novos casos por dia no decorrer dos anos de 2004 e 2015.

A Figura 3 mostra a frequência diária de busca de cada palavra no Google no estado de São Paulo entre os anos de 2004 e 2015. Via de regra, observa-se que o uso do Google para a busca pela maioria dessas palavras intensificou-se após 2011. "Chikungunya" é uma excessão, pois só começou a ser observada em 2014. "Sintomas Dengue", por sua vez, vem sendo buscada de forma substancial desde 2007. Nota-se também que "Febre Dengue" apresentou um pico em 2008, com padrão bastante destoante das demais observações. Em todos os casos, nota-se também uma periodicidade das frequências, que aparentam se correlacionar bastante com o número de novos casos de dengue apresentados na Figura 2 (i.e., no início de cada ano sempre há mais casos).

Pode-se observar também que o comportamento do número de buscas pela palavra-chave "Manchas Dengue" foi bastante parecido ao da palavra-chave "Dengue Plaquetas". Finalmente, observa-se que o comportamento das três ultimas elevações das taxas relativas à palavra-chave "Sintomas Dengue" é bem parecido àquele detectado para "Manchas Dengue" e "Dengue Plaquetas", assim como ao número de casos reais observados (Figura 3).



(a) Dengue Plaquetas

(b) Febre Dengue

(c) Manchas Dengue

(d) Sintomas Dengue

(e) Chikungunya

(f) Tratamento Dengue

Figura 3 – Taxa de buscas por dia de diversas expressões no Google no estado de São Paulo do ano de 2004 a 2015.

3.2 Modelos ajustados

Os ajustes dos modelos foram realizados utilizando a linguagem (R CORE TEAM, 2017). Para a construção dos modelos, foram utilizados dados a partir de 2010 pois, como visto na Seção 3.1, não há um volume substancial de buscas no Google antes deste período. Com a finalidade de avaliar o poder preditivo de cada método, todos os modelos foram ajustados em quatro períodos diferentes, todos tendo como início o ano de 2010, mas com término, respecticamente:

- o fim de 2012,
- o meio de 2013,
- o início de 2014 e
- o início de 2015.

O poder preditivo foi então verificado através da predição diária do número de casos de dengue em um período de até 240 dias (8 meses) após a última data usada para estimar os modelos.

3.2.1 Mínimos quadrados

Inicialmente o modelo foi ajustado utilizando como covariáveis todas as palavras-chave investigadas. Como somente as palavras-chave "sintomas dengue", "febre dengue", "tratamento dengue" foram significativas, tal modelo foi então reajustado utilizando apenas tais covariáveis. Os resultados das estimativas dos parâmetros são apresentados na Tabela 2.

Tabela 2 – Estimativa dos parâmetros do modelo por mínimos quadrados

	Período I	**Período II**	**Período III**	**Período IV**
Intercepto	-5.373	-4.991	-4.914	-3.219
taxa_FebreDengue	-0.167	-0.151	-0.151	-0.198
taxa_SintomasDengue	1.552	1.507	1.494	1.273
taxa_TratamentoDengue	-0.039	-0.004	0.002	-0.045

A Figura 4 indica que as predições para 240 dias à frente dos modelos ajustados via mínimos quadrados (em vermelho) são bastante razoáveis, se assemelhando bastante aos níveis reais de casos de dengue observados. A única excessão ocorre no período final de 2015, em que as predições não parecem ser tão precisas. Possivelmente isto não ocorreu devido a problemas não das predições, mas sim na mensuração do número de casos reais naquele período final: como descrito na Seção 2.1, os dados de 2015 foram obtido utilizando-se outra fonte daqueles até 2014. Assim, possivelmente existe uma diferença na metodologia de coleta destes.

Finalmente, nota-se também que há uma leve superestimação do número de casos de dengue nos picos de 2014 e 2015.

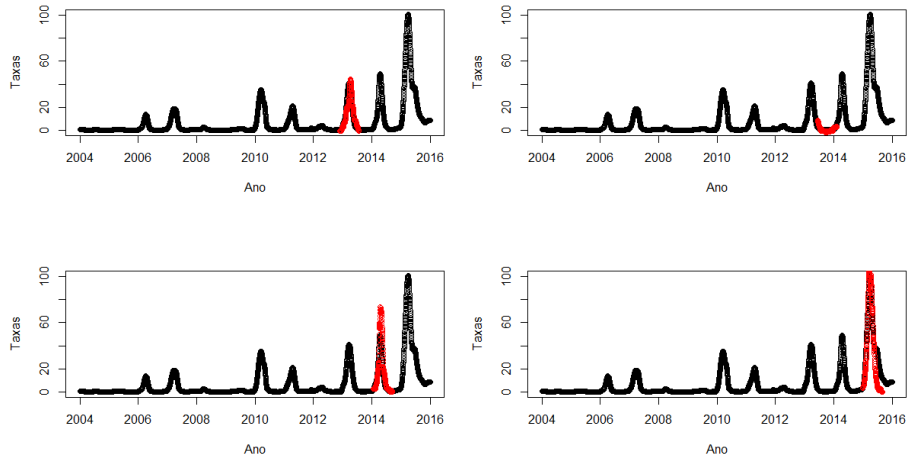


Figura 4 – Gráficos da taxa de novos casos de dengue com predição de 240 dias dada pelo modelo ajustado através do método dos mínimos quadrados.

3.2.2 Lasso

Para o ajuste do lasso, foram utilizadas as taxas de pesquisa de todas as palavras-chave investigadas. Os resultados das estimativas dos parâmetros para os quatro modelos relativos aos quatro períodos investigados são apresentados na Tabela 3. Nesta tabela, apresentamos somente covariáveis que apresentaram coeficientes diferentes de zero em ao menos um dos modelos. Observa-se que, em geral, as covariáveis selecionadas são as mesmas que aquelas significativas no método de mínimos quadrados.

Tabela 3 – Estimativa dos parâmetros do modelo pelo método do Lasso

	Período I	**Período II**	**Período III**	**Período IV**
Intercepto	-5.328	-4.979	-4.826	-4.078
taxa_FebreDengue	-0.159	-0.107	-0.105	0
taxa_Chikungunya	0	0	0	-0.039
taxa_ManchasDengue	0	-0.010	0	-0.348
taxa_DenguePlaquetas	0	0.036	-0.017	0
taxa_SintomasDengue	1.542	1.500	1.475	1.378
taxa_TratamentoDengue	-0.032	0	0	-0.001

A Figura 5 indica que as predições para 240 dias à frente dos modelos ajustados via lasso (em vermelho) não são tão razoáveis quanto aqueles obtidos via mínimos quadrados (Figura 4). Em particular, nota-se que há uma tendência em subestimar o valor real do número de casos observados. Apesar de o modelo conseguir

acompanhar a tendência de crescimento/decrescimento da taxa, as predições não são tão precisas quanto as obtidas via mínimos quadrados.

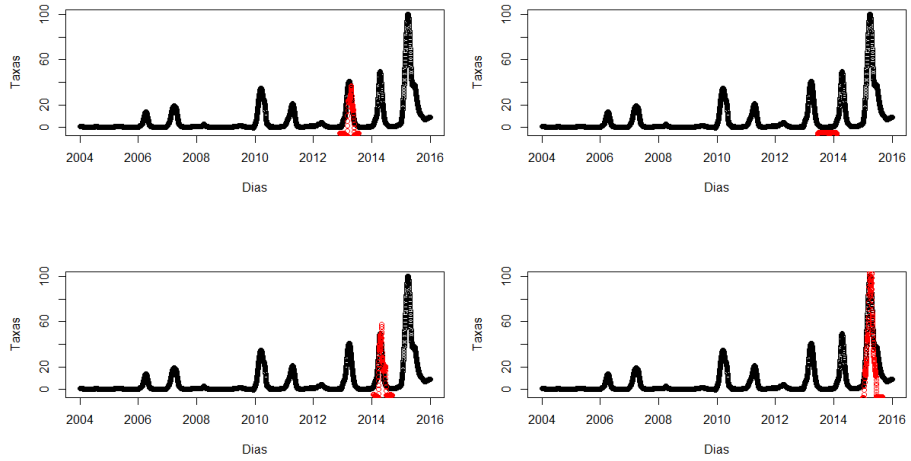
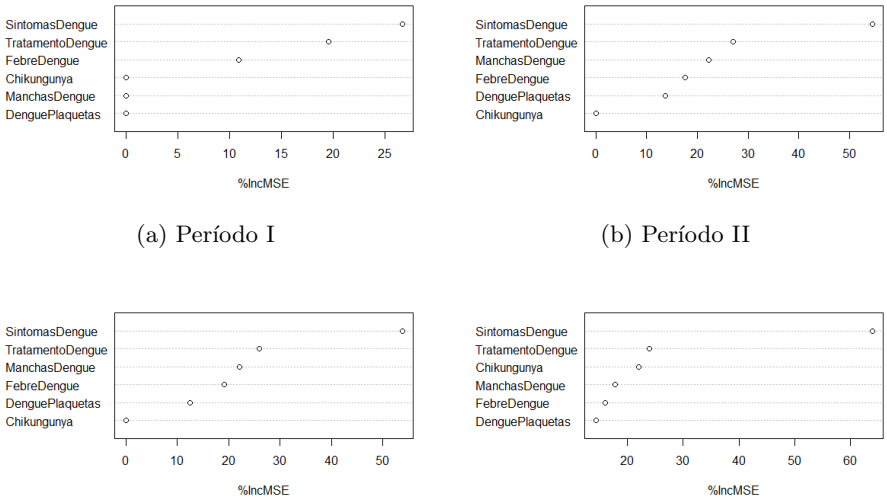


Figura 5 – Gráficos da taxa de novos casos de dengue com predição de 240 dias dada pelo modelo ajustado através do Lasso.

3.2.3 Florestas aleatórias

O método de florestas aleatórias traz uma medida de importância para cada covariável no modelo de predição, o %IncMSE, que mede o quanto adicionar uma covariável na árvore diminui (em média) o erro quadrático médio, veja (JAMES et al., 2013). A importância das covariáveis nos modelos ajustados para a predição do número de novos casos de dengue são apresentadas na Figura 6. As palavras chave "sintomas dengue" e "tratamento dengue" estiveram entre as variáveis mais importantes nos quatro períodos estudados.



(a) Período I

(b) Período II

(c) Período III

(d) Período IV

Figura 6 – A importância das covariáveis nos modelos ajustados para a predição do número de novos casos de dengue.

A Figura 7 indica que as predições para 240 dias à frente dos modelos ajustados via florestas aleatórias (em vermelho) não são tão razoáveis quanto aqueles obtidos via mínimos quadrados (Figura 4). Em particular, nota-se que há uma tendência em subestimar o valor real do número de casos observados. Ao contrário do que ocorria no lasso (Figura 5), tal subestimação ocorre nos picos, e não em regiões de baixa ocorrência de casos novos. Isso possivelmente ocorre pois as predições dadas por florestas aleatórias consistem em médias das variáveis respostas do conjunto de treinamento. Assim, nunca é possível estimar um valor maior que os valores observados. O número de novos casos de dengue, contudo, cresceu ano a ano entre 2013 e 2015, de modo que o número de casos estimados máximo foi bastante inferior ao observado nestes três anos. Em outras palavras, apesar de o modelo conseguir acompanhar a tendência de crescimento/decrescimento da taxa, florestas aleatórias não permitem que se extrapole as predições para além dos valores já observados das variáveis respostas.

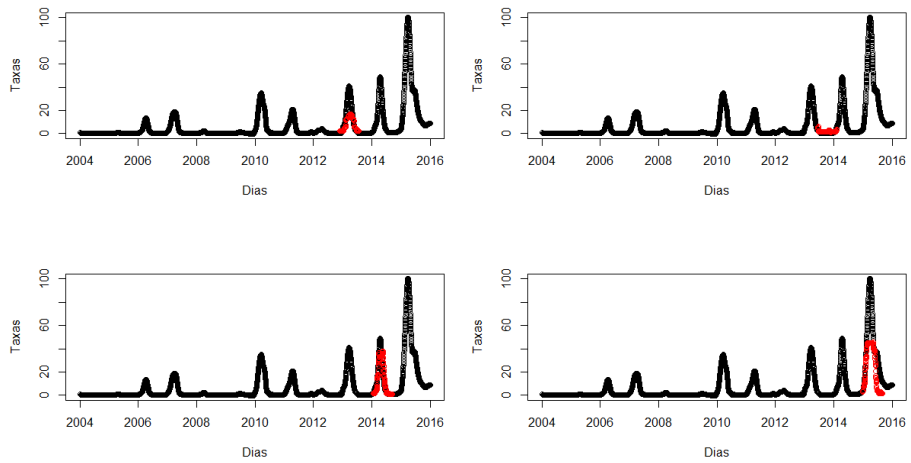


Figura 7 – Gráficos da taxa de novos casos de dengue com predição de 240 dias dada pelo modelo ajustado através do método de Florestas Aleatórias.

4 Conclusões

Neste trabalho apresentamos modelos preditivos que são capazes de estimar como o número de novos casos de dengue variam com o tempo a partir do volume de buscas de certas palavras-chave no Google. Tais modelos são de extrema importância, já que permitem que políticas publica de combate à dengue sejam implementadas com maior eficácia, visto que não é necessário esperar meses para que se tenha dados oficiais sobre onde se localizam os focos de dengue.

Vimos que as palavras-chave "tratamento dengue", "sintomas dengue"e "febre dengue" foram as que apresentaram maior importância na construção dos modelos, estando mais diretamente associadas ao número de casos de dengue observado. Assim, entre os termos monitorados, essas foram as expressões que foram mais preditivas de epidemias. Em particular, uma regressão linear estimada via o método de mínimos quadrados com essas covariáveis levou a predições muito precisas com um alcance de oito meses após a divulgação do último relatório oficial, capturando tendência de aumento e queda com boa acurácia. Assim, o objetivo principal deste trabalho foi alcançado, visto que tal modelo possibilita uma ação mais veloz de políticas públicas para o combate à dengue, já que não é necessário esperar a divulgação de dados oficiais para que se tome as providências necessárias.

Este trabalho sugere diversas direções para pesquisas futuras sob o ponto de vista metodológico. De um lado, a superestimação apresentada pelo método de mínimos quadrados e a subestimação presente em florestas aleatórias sugere que métodos de agregação de métodos de predição (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; IZBICKI; SANTOS, 2017) podem levar a resultados superiores

aos estimadores considerados individualmente. De outro lado, é possível que melhorias individuais possam ser implementadas a cada um dos métodos considerados. Por exemplo, o lasso possivelmente não apresentou bons resultados por não funcionar bem para dados inflados no zero. Assim, adaptar tal métodos para dados com essa característica pode render bons resultados. Também pode-se buscar utilizar estimadores de densidades condicionais (IZBICKI; LEE, 2016, 2017) para se obter não somente uma estimativa pontual para o número de casos de dengue, mas também toda a incerteza associada a essa quantidade, i.e., $f(y|\mathbf{x})$. Finalmente, outra sugestão de trabalho futuro é a utilização de séries temporais com covariáveis para este problema, com o intuito de fazer previsão levando em consideração a ordenação temporal das observações.

De um ponto de vista aplicado, seria interessante aplicar os métodos propostos para a predição de outras epidemias, como a Zika e a Chikungunya, assim como aplicar tais métodos a outros estados do Brasil além de São Paulo. Finalmente, seria de grande utilidade para a população brasileira e para as autoridades sanitárias que fosse feita uma implementação de um *dashboard* mostrando as predições de epidemia desenvolvidas neste trabalho em tempo real para usuários na internet. Isso certamente faria com que houvesse um combate mais efetivo à dengue no Brasil.

5 Agradecimentos

Este projeto foi parcialmente financiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (104788/2016-0 e 306943/2017-4) e pela Fundação de Amparo à Pesquisa (2014/25302-2 e 2017/03363-8). Os autores agradecem os revisores e editores pelas sugestões.

CRUZ, L. O.; IZBICKI, R. Dengue online monitoring: using Google to predict epidemics. *Rev. Bras. Biom.,* Lavras, v.36, n.3, p.512-526, 2018.

- *ABSTRACT: Unfortunately, official reports on the development of dengue in Brazil take a long time to be released. This creates an environment where policy makers do not have enough accurate information they can use to improve how they act in order to prevent dengue to spread. In this paper we show how data collected in real time from Google Trends can be used to predict the current number of dengue cases in the state of São Paulo. In order to estimate the number of new cases as a function of time, we use the least square method, lasso and random forests, having the search volume of several keywords as covariates such as "tratamento dengue", "sintomas dengue", and "febre dengue" in the models. The least square method presented better predictions and gave reasonable estimates for up to eight months after the last release of an official report. This allows authorities to take necessary actions to prevent dengue from spreading in a much cheaper and effective way.*

- *KEYWORDS: Epidemic; dengue; prediction; infoveillance.*

Referências

COOK, S. et al. Assessing Google flu trends performance in the United States during the 2009 influenza virus a (H1N1) pandemic. *PloS one*, Public Library of Science, v. 6, n. 8, p. e23610, 2011.

DREDZE, M. et al. Healthtweets. org: A platform for public health surveillance using twitter. In: CITESEER. *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*. [S.l.], 2014.

EL-METWALLY, A. A. Google search trend of dengue fever in developing countries in 2013-2014: An internet-based analysis. *Journal of Health Informatics in Developing Countries*, v. 9, n. 1, 2015.

EYSENBACH, G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the internet. *Journal of medical Internet research*, Internet Healthcare Coalition, v. 11, n. 1, 2009.

GINSBERG, J. et al. Detecting influenza epidemics using search engine query data. *Nature*, Nature Publishing Group, v. 457, n. 7232, p. 1012–1014, 2009.

GLUSKIN, R. T. et al. Evaluation of internet-based dengue query data: Google dengue trends. *PLoS neglected tropical diseases*, Public Library of Science, v. 8, n. 2, p. e2713, 2014.

GOOGLE. *Google Dengue Trends*. 2017. Disponível em: ⟨https://www.google.org/denguetrends/⟩.

GOOGLE. *Google Flu Trends*. 2017. Disponível em: ⟨https://www.google.org/flutrends/⟩.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning*. Second. Berlin: Springer, 2009.

INFODENGUE. *Info Dengue Rio*. 2017. Disponível em: ⟨http://alerta.dengue.mat.br/informacoes/⟩.

IZBICKI, R.; LEE, A. B. Nonparametric conditional density estimation in a high-dimensional regression setting. *Journal of Computational and Graphical Statistics*, Taylor & Francis, v. 25, n. 4, p. 1297–1316, 2016.

IZBICKI, R.; LEE, A. B. Converting high-dimensional regression to high-dimensional conditional density estimation. *Electron. J. Statist.*, The Institute of Mathematical Statistics and the Bernoulli Society, v. 11, n. 2, p. 2800–2831, 2017. Disponível em: ⟨https://doi.org/10.1214/17-EJS1302⟩.

IZBICKI, R.; SANTOS, T. M. d. *Machine Learning sob a ótica estatística*. [s.n.], 2017. Disponível em: ⟨https://rizbicki.wordpress.com/teaching/⟩.

JAMES, G. et al. *An introduction to statistical learning*. Berlin: Springer, 2013.

LAMPOS, V.; BIE, T. D.; CRISTIANINI, N. Flu detector-tracking epidemics on twitter. In: *Machine Learning and Knowledge Discovery in Databases*. [S.l.]: Springer, 2010. p. 599–602.

LAMPOS, V.; CRISTIANINI, N. Tracking the flu pandemic by monitoring the social web. In: IEEE. *Cognitive Information Processing (CIP), 2010 2nd International Workshop on*. [S.l.], 2010. p. 411–416.

LAMPOS, V.; CRISTIANINI, N. Nowcasting events from the social web with statistical learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, ACM, v. 3, n. 4, p. 72, 2012.

LEE, K.; AGRAWAL, A.; CHOUDHARY, A. Real-time disease surveillance using twitter data: demonstration on flu and cancer. In: ACM. *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.], 2013. p. 1474–1477.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2016. Disponível em: ⟨https://www.R-project.org/⟩.

SAÚDE, M. da. *Portal da Saúde*. 2017. Disponível em: ⟨http://portalsaude.saude.gov.br/index.php/situacao-epidemiologica-dados-dengue⟩.

STILO, G. et al. Predicting flu epidemics using twitter and historical data. In: *Brain Informatics and Health*. [S.l.]: Springer, 2014. p. 164–177.

Recebido em 08.11.2016.

Aprovado após revisão em 07.12.2017.