Detecção e previsão de eventos de alto impacto utilizando dados de redes sociais online

Denise E. F. Brito, Wagner Meira Jr.

[1]Departamento de Ciência da Computação – Universidade Federal de Minas Gerais
Belo Horizonte – MG – Brazil

{denise.brit, meira}@dcc.ufmg.br

Abstract. *This work aims to design, evaluate and apply regression methods to early detection of events using public data available in online social networks. The process comprises four phases, which consists of verifying the viability of predicting the event through informal data, identifying the models to be used, calculating the parameters of the prediction function and evaluating the model in a case study. The context is the dengue outbreak in Brazil, integrating the Dengue Observatory project, where the resulting models were able to correctly predict the severity of the surges, in a per week basis and for the largest Brazilian cities, for 99.12% of the disease incidence values.*

Resumo. *Este trabalho tem como objetivo projetar, avaliar e aplicar modelos de regressão para detecção precoce de eventos utilizando dados públicos disponíveis em redes sociais online. O processo compreende quatro fases, que consistem em verificar a viabilidade de prever o evento através de dados informais, identificar os modelos a serem utilizados, calcular os parâmetros da função de previsão e avaliar o modelo num estudo de caso. O contexto é a epidemia de dengue no Brasil, integrando o projeto do Observatório da Dengue, onde os modelos resultantes foram capazes de prever a severidade dos surtos, numa escala semanal e para as maiores cidades brasileiras, para 99.12% dos valores de incidência da doença.*

1. Introdução

Eventos de alto impacto são processos ou fenômenos que podem causar mortes ou danos variados, de problemas de saúde a danos a propriedades (incluindo rebanhos), passando por serviços, meio ambiente e equilíbrio social ou econômico. A severidade do evento depende da sua amplitude e da latência para a adoção de medidas que minimizem ou remediem os vários tipos de impacto, o que mostra a relevância de mecanismos eficazes para detectar precocemente ou prever tais eventos.

A detecção e previsão desses eventos é uma área de pesquisa abrangente, que vem se mostrando útil em diversos contextos, como economia, segurança pública, meio ambiente, vigilância epidemiológica e outros. Diversas tecnologias têm sido utilizadas para realizar essas tarefas, incluindo tecnologias da informação e comunicação. A popularização de ferramentas como máquinas de busca e redes sociais trouxe a possibilidade de utilizá-las como fonte de dados para a detecção e previsão desses eventos.

Com o crescimento das redes sociais online, usuários do mundo todo compartilham experiências e informações sobre suas vidas e sobre o local onde vivem. Algumas

dessas redes permitem que o usuário escreva mensagens de acesso público e disponibilizam uma forma de coletá-las, como o Twitter. Os usuários do Twitter podem colocar em seu perfil sua localização e podem postar mensagens públicas curtas, de até 140 caracteres, que por sua vez podem ser coletadas utilizando-se uma API padronizada [1]. Essas características do modo de interação inerentes ao Twitter, assim como o comportamento dos seus usuários de relatar fatos cotidianos, tornam essa ferramenta uma importante fonte de dados para detectar e prever eventos de alto impacto.

A realização dessas tarefas compreende diversas etapas. O processo se inicia com a coleta dos dados e várias atividades de engenharia de dados, como limpeza, identificação de entidades relevantes e georeferenciamento dos dados, além da determinação das mensagens relevantes para fins de detecção e previsão. Por outro lado, devemos ter dados reais sobre os eventos que queremos detectar ou prever, cuja correlação com os dados de redes sociais tem de ser verificada e cujo potencial para prever os eventos tem de ser avaliado.

Este trabalho de iniciação científica consiste em projetar, implementar e testar modelos de regressão que permitam detectar e prever, a partir de dados de redes sociais, a intensidade de ocorrência de eventos de alto impacto. Distinguimos pelo menos 4 fases na execução do trabalho. A primeira fase é a avaliação da existência da correlação entre os dados reais e os dados oriundos de redes sociais. A segunda fase é a determinação dos modelos mais apropriados para os dados, seguido da determinação dos seus parâmetros. A última fase é a aplicação do modelo a dados oriundos de redes sociais para detectar ou prever eventos de alto impacto. Resultados preliminares deste trabalho foram publicados em [Brito et al. 2012].

O nosso cenário de aplicação são as epidemias de dengue no Brasil e este trabalho faz parte do projeto do Observatório da Dengue[2], cujo objetivo é prover uma ferramenta de vigilância epidemiológica de dengue, a partir de mensagens postadas no Twitter, complementar aos métodos tradicionais. A dengue é uma doença transmitida pela picada do mosquito Aedes aegypti, característico de regiões tropicais e subtropicais[3], que afeta milhares de pessoas no Brasil a cada ano. Em trabalhos anteriores do projeto [Gomide 2012], foi comprovado que existe uma correlação alta entre o número de tweets relatando casos de dengue e o número de casos de dengue notificados. Nesses trabalhos, os dados utilizados para análise foram os tweets que possuíam a cidade onde se localizava o usuário e que foram classificados como experiência pessoal com a doença, utilizando um algoritmo classificador [Veloso et al. 2006], e para validação, o número de casos reais de dengue fornecidos pelo Ministério da Saúde, para 2010 e 2011.

2. Metodologia

Nesta seção, descrevemos a metodologia desenvolvida no trabalho, que, como mencionado, compreende quatro fases.

2.1. Correlação

Primeiramente, é preciso testar se os dados oriundos das redes sociais possuem alta correlação com os dados oficiais. Para tal, calcula-se o coeficiente de correlação cruzada

[1] http://apiwiki.twitter.com/

[2] http://http://observatorio.inweb.org.br/dengue

[3] http://http://www.who.int/topics/dengue/en/

entre duas séries [Bourke 1996]:

$$r = \frac{\sum_{i=1}^{n}(t_i - \bar{t})(o_i - \bar{o})}{\sqrt{\sum_{i=1}^{n}(t_i - \bar{t})^2}\sqrt{\sum_{i=1}^{n}(o_i - \bar{o})^2}}$$

onde $t(i)$ é a série dos dados de redes sociais, $o(i)$ é a série dos dados oficiais, $i = 0, 1, 2, ...N-1$ e N é o número de semanas avaliadas. O coeficiente de correlação cruzada pode variar entre -1 e 1. Quanto mais próximo de 1, melhor a correlação entre as séries. Se próximo de 0, indica que as séries não estão correlacionadas, e se próximo de -1, significa correlação inversa. É importante notar que, se o coeficiente calculado com algum atraso na série dos dados oficiais indica melhor correlação do que séries perfeitamente alinhadas temporalmente, isso indica a possibilidade de realizar detecção precoce.

2.2. Modelos

A próxima etapa do trabalho consiste em avaliar diferentes modelos para descobrir quais são os que melhor descrevem os dados. Para escolher os modelos a serem testados, é necessário observar a natureza dos dados, que são discretos, além da restrição de que a função de previsão deve gerar apenas resultados não-negativos. O problema de detectar eventos através de tweets constitui um problema de contagem. Por isso, foram aplicadas as regressões de Poisson, quasi-Poisson e binomial negativa, que fazem parte dos Modelos Lineares Generalizados (GLM) [Zeileis et al. 2008]. Além dessas, foi testada a regressão ortogonal linear [Markovsky and Van Huffel 2007].

A razão pela qual a regressão ortogonal foi usada é que pode haver erros não só na variável aleatória Y (dados oficiais), mas pelo fato de que o X, que representa os tweets, é um dado informal, coletado de uma rede social e, portanto, de várias fontes diferentes (usuários da rede), que ainda passa por um processo de classificação automática e é associado ao município pela informação fornecida pelo próprio usuário. Logo, há uma incerteza inerente ao X.

A princípio, foi considerada a regressão de Poisson, que assume que a variável Y possua distribuição de Poisson. Uma característica da distribuição de Poisson é que a esperança da variável aleatória é igual à variância, o que no mundo real muitas vezes não se aplica. A superdispersão ocorre quando a variância é consideravelmente maior do que a esperança e, nesse caso, a distribuição de Poisson não é recomendada. As regressões quasi-Poisson e binomial negativa são semelhantes à de Poisson e são comumente usadas em seu lugar, por comportarem bem a superdispersão. Como cada cidade pode ter um comportamento diferente, tendo em vista as características de difusão epidemiológica e do uso da internet, a avaliação de vários modelos que correlacionem esses dados se faz necessária.

2.3. Cálculo dos parâmetros

Em seguida, foram calculados os parâmetros das diferentes regressões. Os modelos GLM foram testados com o software R [R Development Core Team 2011], com a utilização do pacote MASS [Venables and Ripley 2002] para a binomial negativa, e a regressão ortogonal linear foi aplicada com a implementação encontrada em [Petras and Bednarova 2010].

Nos modelos GLM, o valor esperado de Y (dados reais) condicionado a X (tweets) é dado por: $E[y_i|x_i] = \mu_i$ e $g(\mu_i) = ax_i + b$ onde x_i é a amostra de tweets na semana i, $g()$ é a função de ligação e a e b são os parâmetros a serem encontrados na regressão.

Na regressão quasi-Poisson, a relação entre esperança e variância é linear e se dá por [Ver Hoef and Boveng 2007]: $VAR(Y) = \theta E(Y)$ em que θ é o parâmetro de superdispersão. Na binomial negativa, a relação é quadrática: $VAR(Y) = \kappa E(Y)^2 + E(Y)$ e o parâmetro de superdispersão é: $\kappa E(Y) + 1$. Nas duas regressões, a função $g()$ de ligação mais comumente utilizada é a logarítmica, pois ela faz com que $E[y_i|x_i] = \exp(ax_i + b)$ gerando apenas valores não-negativos.

Na regressão ortogonal linear, assim como no método de mínimos quadrados, o valor esperado de Y se dá por: $E[y_i|x_i] = ax_i + b$ sendo que os dois métodos diferem no cálculo dos parâmetros a e b, pois a regressão ortogonal linear trata ambas as variáveis de forma simétrica, enquanto que no método de mínimos quadrados, o X é considerado correto e o Y é passível de ter ruído.

2.4. Avaliação dos modelos

É preciso verificar se os dados para análise satisfazem as premissas dos modelos escolhidos. Para utilizar a regressão de Poisson, testamos se há superdispersão. Se houver, descarta-se o modelo. A seguir, calculamos, para cada regressão, o erro médio ponderado pelo número de habitantes das cidades com mensagens suficientes: $\sum_{i=1}^{c} \frac{P_i * \sum_{j=1}^{s} |y_j - \hat{y}_j|}{\sum_{k=1}^{c} P_k}$ onde P é o tamanho da população, y é o número de casos de dengue reais, $\hat{y}$ é o número previsto de casos de dengue, c é o número de cidades e s o total de semanas. Escolhemos aquelas com menores erros para realizar um teste estatístico e determinar qual é o modelo que melhor se aplica aos dados analisados.

3. Estudo de caso: Observatório da Dengue

As funções de previsão utilizadas no Observatório da Dengue foram calculadas com base nos dados de 2011, sendo assim, a primeira fase deste trabalho consistiu em aplicar as mesmas funções aos dados correspondentes ao período da 1ª a 30ª semana de 2012, para verificar se a relação entre tweets e casos de dengue permaneceu constante para todas as cidades com mais de cem mil habitantes e se o alarme de incidência da doença continua com boa precisão.

Foi aplicado um atraso de 0, 1 e 2 semanas na série de casos reais, para verificar se é possível detectar precocemente a incidência de casos de dengue em certa cidade. No entanto, os tweets foram agrupados de forma que as semanas começam nas quintas-feiras, enquanto que os casos reais obedecem à semana epidemiológica, que começa aos domingos. Para compensar essa diferença de meia semana de avanço da série do Twitter, neste trabalho será considerado que houve detecção precoce apenas quando o coeficiente de correlação com atraso de duas semanas for o melhor.

Estabeleceu-se para o Observatório da Dengue que a quantidade suficiente mínima de tweets é de 7 por semana. No entanto, nenhuma cidade obteve tweets suficientes em todas as semanas de 2012 analisadas. Então, a análise foi feita para todas aquelas que obtiveram pelo menos 20 semanas das 30 totais com dados suficientes.

O alerta de volume categoriza o número de casos previstos por cem mil habitantes em três faixas, de acordo com a escala do Ministério da Saúde: abaixo de 100, a incidência é baixa; entre 100 e 300, média e, acima de 300, alta. Para calcular os casos previstos, originalmente foi utilizada a função de previsão, encontrada por regressão linear, baseada

nos dados dos anos anteriores. A função é da forma $y_i = ax_i + b$, onde a e b são os parâmetros; y_i é o número de casos previstos, x_i é o número de tweets e i é o índice da semana. Todos esses valores são específicos para cada cidade, visto que a análise compreende todas as regiões brasileiras e os municípios diferem bastante uns dos outros em características como clima, tamanho da população e acesso à internet. Para validação do alerta de volume, comparou-se a faixa de incidência encontrada pelos casos previstos com a dos casos reais. Neste trabalho, verificamos a pertinência da utilização de outros modelos.

O alerta de tendência utiliza a métrica Z-score para determinar qual foi o comportamento do número de tweets da semana corrente em relação às duas semanas anteriores. O número de semanas anteriores foi estabelecido de acordo com o ciclo de duração da doença. Esse alerta se baseia na hipótese de que, quando os casos reais aumentam, deve haver um aumento também na quantidade de tweets relatando experiência com dengue. Analogamente, quando os casos diminuem, deve ser possível observar uma queda no número de tweets. O Z-score é calculado pela seguinte fórmula: $(x - \mu)/\sigma$ onde μ é a média diária de tweets das duas semanas anteriores e σ é o desvio padrão do período. O alerta de tendência também se divide em três faixas, a saber: abaixo de -1, a tendência é de diminuição; entre -1 e 2, estabilização, e acima de 2, aumento.

Para encontrar a função de previsão, os dados foram divididos da seguinte forma: as 50 semanas de 2011 (todo o ano, exceto pelas semanas 1 e 22, por falta de dados) foram usadas para calcular os parâmetros dos modelos e as 30 semanas de 2012 foram reservadas para teste.

A Figura 1 contém o histograma do valor do parâmetro de superdispersão da regressão quasi-Poisson por número de cidades, retornado pela função que recebe como parâmetro o resultado da regressão e é calculado da seguinte forma:

$$\sum_{i=1}^{n} \frac{model\$weights \times model\$residuals^2}{model\$df.residual}$$

Para alguns municípios, o valor do parâmetro é muito superior a 1, indicando que a variância é muito maior do que a esperança e, portanto, descarta-se a possibilidade do modelo de Poisson.

Figura 1. Número de cidades por parâmetro de superdispersão.

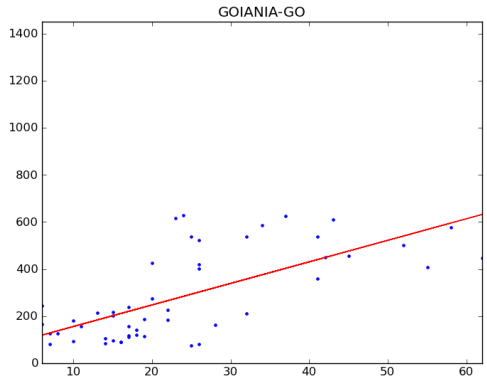

Figura 2. Dados de treino para Goiânia. A reta corresponde à regressão linear feita e os pontos são os valores de tweets (x) por casos reais (y).

Além da superdispersão, ao traçar o gráfico de tweets (X) por casos reais (Y), em que cada ponto corresponde a uma amostra de uma semana, é possível observar (Figura 2) outra dificuldade para encontrar um modelo adequado: há muitas semanas em que ocorrem valores bastante diversos de casos, para o mesmo número de tweets, o que é uma evidência de que a regressão ortogonal linear pode se aplicar.

Ao calcular o erro médio ponderado pelo número de habitantes de cada cidade, para cada modelo, considerando as 50 semanas de treino, a regressão binomial negativa obteve 25220.54414; a quasi-Poisson, 9739.89995; a ortogonal, 6532.08542 e a linear, 5883.06794.

4. Análise comparativa dos modelos

A Tabela 1 mostra as cidades que obtiveram dados suficientes em pelo menos 20 das 30 semanas analisadas, ordenadas de forma decrescente pelo número de casos de dengue. O cálculo dos coeficientes da correlação cruzada mostra que para as dezesseis cidades da tabela, dez possuem ao menos um dos coeficientes superior a 0.6, indicando que os tweets podem ser utilizados para detecção de epidemias de dengue, sendo que para Rio de Janeiro, Fortaleza, Recife e Natal, o valor é superior a 0.8, justamente as cidades com maior número de casos notificados dentre todas com tweets suficientes durante pelo menos 20 semanas das 30 analisadas.

Em Natal e São Paulo, o coeficiente de maior valor é o correspondente ao calculado com atraso de duas semanas da série de dados oficiais em relação ao Twitter, mostrando que é possível haver detecção precoce de epidemias. Esse resultado é ilustrado por tweets em que os usuários descrevem seus sintomas e acreditam que estão com dengue, mas ainda não procuraram assistência médica ou ainda não foram diagnosticados.

A Tabela 2 mostra o alerta de volume e de tendência para a cidade do Rio de Janeiro. O volume de casos de dengue e a proporção em relação ao número de tweets sofreram alterações de 2011 para 2012, considerando o mesmo período e município.

Os casos previstos foram calculados com a função de previsão derivada anteriormente, calibrada com os dados de 2010 e 2011. Por esses motivos, o número de casos previstos difere bastante dos casos reais, mesmo quando a correlação é alta, como no

Tabela 1. Correlação entre tweets e casos de dengue com atraso de 0, 1 e 2 semanas

Cidade	Casos de dengue	Corr 0	Corr 1	Corr 2
RIO DE JANEIRO – RJ	122722	0.901898	0.937201	0.914303
FORTALEZA – CE	35253	0.982306	0.892424	0.677947
RECIFE – PE	8684	0.928756	0.901308	0.876676
NATAL – RN	8577	0.714485	0.792293	0.833409
GOIÂNIA – GO	7045	0.366342	0.348550	0.221251
SALVADOR – BA	4834	0.431263	0.397027	0.197968
JOÃO PESSOA – PB	2726	0.752976	0.680269	0.556181
MANAUS – AM	2175	0.604449	0.618463	0.504184
ARACAJU – SE	2165	0.161315	0.294573	−0.222900
SÃO PAULO – SP	1986	0.429890	0.576528	0.605924
BELÉM – PA	1839	0.707398	0.740384	0.714267
BRASÍLIA – DF	1143	0.610918	0.613832	0.479530
SANTOS – SP	927	0.673293	0.520498	0.366419
BELO HORIZONTE – MG	747	−0.165831	−0.334045	−0.322071
CURITIBA – PR	37	0.438474	0.038517	0.012693
PORTO ALEGRE – RS	24	0.370767	0.450461	0.324811

Rio de Janeiro. No entanto, o alerta de volume continua adequado, pois este segue a classificação do Ministério da Saúde, que divide a incidência em três faixas. Sendo assim, o alerta de volume obteve 99.12% de acerto, comparando a faixa dos casos previstos à faixa em que recaem os casos reais, nas 285 cidades com população superior a cem mil habitantes, durante as semanas 1 a 30 de 2012.

Tabela 2. Previsão e validação do número de casos previstos para Rio de Janeiro

Tweets	Casos previstos	Alerta previsão	Casos reais	Alerta real
127	1142.212	baixa incidência	939	baixa incidência
108	1031.518	baixa incidência	1060	baixa incidência
106	1019.866	baixa incidência	1471	baixa incidência
84	891.694	baixa incidência	1633	baixa incidência
81	874.216	baixa incidência	1812	baixa incidência
109	1037.344	baixa incidência	2261	baixa incidência
98	973.258	baixa incidência	2513	baixa incidência
176	1427.686	baixa incidência	3086	baixa incidência
227	1724.812	baixa incidência	4041	baixa incidência
241	1806.376	baixa incidência	3508	baixa incidência
211	1631.596	baixa incidência	3422	baixa incidência
274	1998.634	baixa incidência	2939	baixa incidência
270	1975.33	baixa incidência	834	baixa incidência

Nas regressões quasi-Poisson e binomial negativa, o gráfico de tweets por casos previstos tem um formato exponencial, não correspondendo ao gráfico de tweets por casos reais para todas as cidades, o que é possível notar já no ajuste com os dados de treino. Ao aplicar os dados de teste na função encontrada, esta gera alarmes falsos para muitas

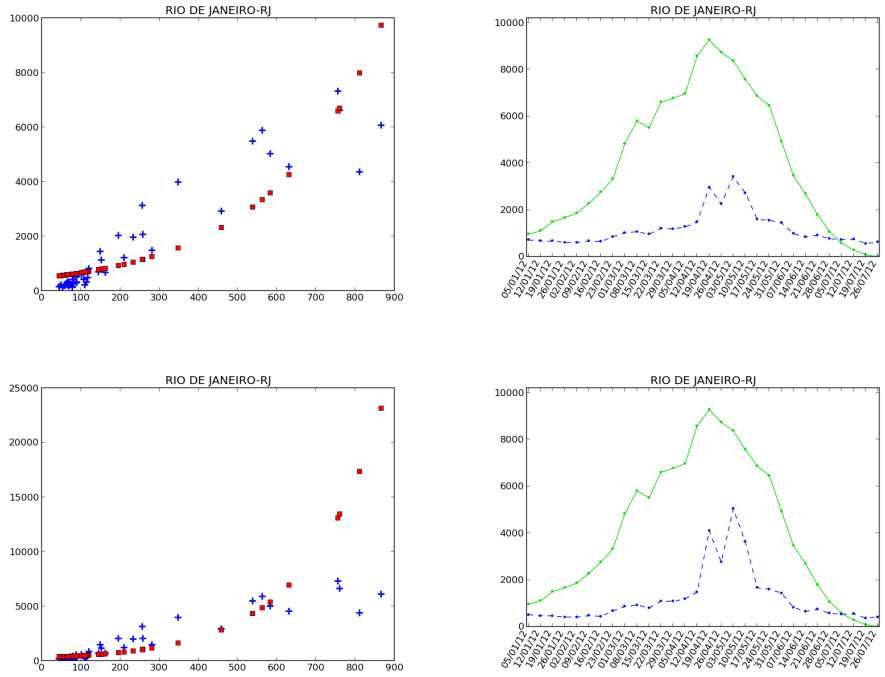

Figura 3. Resultados para Rio de Janeiro. Acima, regressão quasi-Poisson. Abaixo, regressão binomial negativa. Do lado esquerdo, dados de treino. As cruzes azuis são os casos reais e os quadrados vermelhos, casos previstos. À direita, dados de teste. A linha contínua em verde representa os casos reais e a azul, pontilhada, os casos previstos.

cidades, principalmente para aquelas em que um tweet significa vários casos a mais de dengue. Entre as duas regressões, a quasi-Poisson apresentou menores picos de alarmes falsos e por isso parece se adequar melhor do que a binomial negativa.

A regressão ortogonal não funciona bem para todas as cidades, também gerando picos de alarme falso. No entanto, para as cidades da Tabela 1, o número de casos previstos ou permaneceu bem semelhante à regressão Linear, ou se aproximou do número de casos notificados. Para determinar se os modelos são significativamente diferentes, foi calculado o Teste F com confiança de 95%.

Para todas as cidades em que o resultado é maior do que 1.61537, que é o equivalente ao quantil 0.95 com graus de variância 48, significa que a regressão ortogonal é estatisticamente diferente da regressão linear simples e que a primeira se ajusta melhor aos dados da cidade em questão. Ou seja, para as cidades de Belém, Recife (Figura 5) e Aracaju, a regressão ortogonal apresentou significativa vantagem em relação à regressão linear. Para as demais cidades, os modelos são considerados equivalentes e ambos podem ser utilizados.

5. Conclusões

Com este trabalho, verificamos que mensagens postadas no Twitter podem ser utilizadas para detecção de eventos de alto impacto, como foi comprovado no contexto do Observatório da Dengue. A condição necessária para que o método funcione é que o local no qual ocorrem os eventos que se quer detectar possuam quantidade suficiente de tweets.

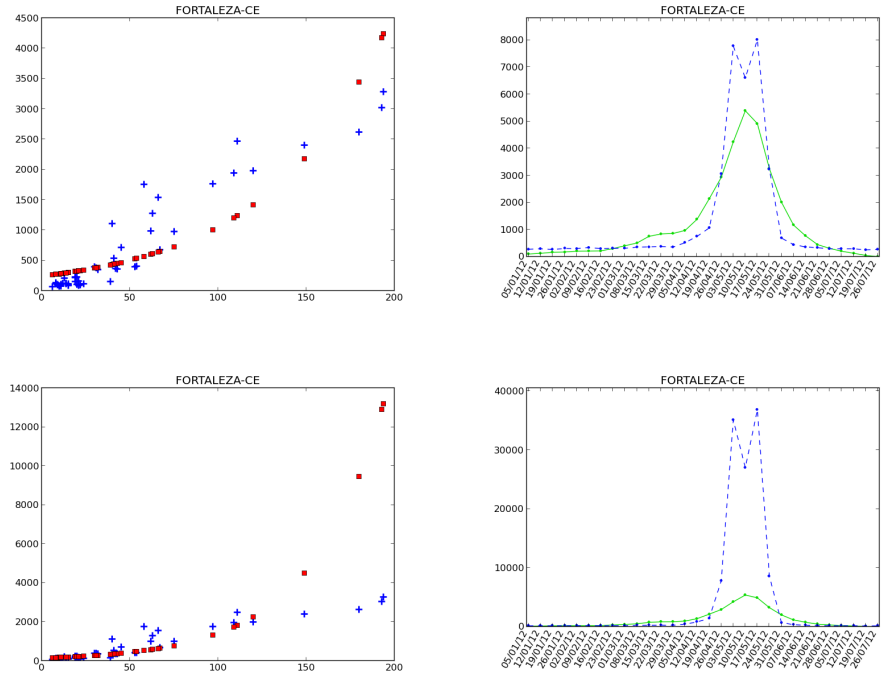

Figura 4. Resultados para Fortaleza. Acima, regressão quasi-Poisson. Abaixo, regressão binomial negativa. Do lado esquerdo, dados de treino. As cruzes azuis são os casos reais e os quadrados vermelhos, casos previstos. À direita, dados de teste. A linha contínua em verde representa os casos reais e a azul, pontilhada, os casos previstos.

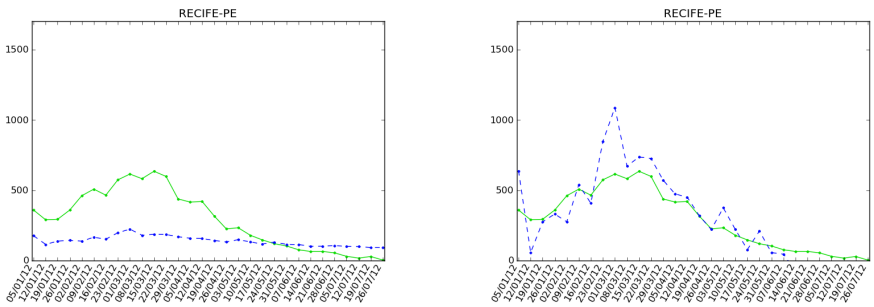

Figura 5. Resultados para Recife. À esquerda, regressão linear simples e à direita, regressão ortogonal linear. A linha contínua em verde representa os casos reais e a azul, pontilhada, os casos previstos.

Particularmente, no caso da dengue, as funções de previsão calculadas pelas regressões lineares, tanto a de mínimos quadrados quanto a ortogonal, apresentam melhores resultados do que as regressões quasi-Poisson e binomial negativa, sendo que para algumas cidades, a ortogonal modelou melhor o problema, enquanto que para as demais, as duas regressões se mostraram estatisticamente equivalentes.

Tabela 3. Valores do Teste F

Cidade	Estado	Valor do Teste F
RIO DE JANEIRO	RJ	1.007859
FORTALEZA	CE	1.003830
CUIABÁ	MT	1.204502
BELÉM	PA	1.698209
MANAUS	AM	1.004843
RECIFE	PE	3.443550
CURITIBA	PR	1.000772
BRASÍLIA	DF	1.172616
PORTO ALEGRE	RS	1.001448
SALVADOR	BA	1.104516
SÃO PAULO	SP	1.080026
JOÃO PESSOA	PB	1.102533
GOIÂNIA	GO	1.090568
SANTOS	SP	1.052270
ARACAJU	SE	2.083052
NATAL	RN	1.034813
BELO HORIZONTE	MG	1.615080
MOSSORÓ	RN	1.097748

Referências

Bourke, P. (1996). Cross correlation. http://paulbourke.net/miscellaneous/correlate/.

Brito, D., Gomide, J., Santos, W., Meira Jr., W., Veloso, A., and Almeida, V. (2012). Um sistema de alarme para vigilância epidemiológica de rumores utilizando redes sociais. In *Proceedings of the 27th Brazilian Symposium on Databases*, pages 225–232.

Gomide, J. S. (2012). Mineração de Redes Sociais para Detecção e Previsão de Eventos Reais. Master's thesis, Universidade Federal de Minas Gerais, BR.

Markovsky, I. and Van Huffel, S. (2007). Overview of total least-squares methods. *Signal Process.*, 87(10):2283–2302.

Petras, I. and Bednarova, D. (2010). Total least squares method. http://www.mathworks.com/matlabcentral/fileexchange/31109.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Veloso, A., Meira Jr., W., and Zaki, M. J. (2006). Lazy associative classification. In *International Conference on Data Mining*, pages 645–654. IEEE Computer Society.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

Ver Hoef, J. and Boveng, P. (2007). Quasi-poisson vs. negative binomial regression: how should we model overdispersed count data? *Ecology*, 88(11):2766–72.

Zeileis, A., Kleiber, C., and Jackman, S. (2008). Regression models for count data in r. *Journal of Statistical Software*, 27(8):1–25.