Análise geográfica entre mensagens georreferenciadas de redes sociais e dados oficiais para suporte à tomada de decisões de agências de emergência

Thiago H. Poiani[1], **Flávio E. A. Horita**[1], **João Porto de Albuquerque**[1,2]

[1]Instituto de Ciências Matemáticas e de Computação (ICMC)
Universidade de São Paulo (USP) – São Carlos/SP – Brasil

[2]GIScience Research Group
Heidelberg University – Heidelberg – Germany

thpoiani@usp.br, {horita,jporto}@icmc.usp.br

Abstract. *The recent damages caused by floods have called for better preparation from vulnerable communities. New data sources like in-situ sensors and social media have opened different perspectives for supporting data collection, and then improving decision-making of the emergency agencies. Therefore, this paper presents a geographical analysis of the relationship between authoritative data and georeferenced social media messages with the aim of understanding their contributions to decision-making in case of floods. The results showed a straight relationship between georeferenced social media messages and authoritative data. Furthermore, it was revealed that these messages are useful to provide information about the situation at the affected area.*

Resumo. *Os recentes danos causados pelas inundações chamam a atenção para uma melhor preparação das comunidades vulneráveis. Novas fontes de dados como sensores estáticos e mídia social abriram diferentes perspectivas para auxiliar na coleta de dados e, assim, melhorar a tomada de decisões das agências de emergência. Este artigo apresenta uma análise geográfica do relacionamento entre dados oficiais e mensagens georreferenciadas de mídias sociais com o objetivo de entender suas contribuições para a tomada de decisões em inundações. Os resultados mostraram uma forte relação entre mensagens georreferenciadas de mídias sociais e dados oficiais. Além disso, tais mensagens também podem prover informações úteis sobre a situação na área afetada.*

1. Introdução

Inundações são perigos naturais hidrológicos recorrentes em diversas regiões do Brasil e que mais afetaram pessoas e causaram mortes entre o período de 2004 e 2014 [Guha-Sapir et al. 2015]. Para servir como suporte aos desastres naturais no país, o Ministério da Ciência, Tecnologia e Inovação criou, em 2011, o Centro Nacional de Monitoramento e Alertas de Desastres Naturais (CEMADEN)[1]. Pluviômetros instalados em áreas de risco de inundação são monitorados por essa agência de emergência, coletando dados climáticos que dão suporte para tomada de decisões.

[1]http://www.cemaden.gov.br/

O acúmulo de informações de fontes de dados distintas auxilia a formulação de estratégias de gestão de risco de inundações. Mídia social é uma fonte com potencial de uso devido à grande quantidade de informações geográficas voluntárias (VGI) distribuída em um tempo curto por *sensores humanos* [Goodchild 2007].

O objetivo desse artigo é apresentar uma análise geográfica da relação entre dados oficiais e mensagens georreferenciadas de mídias sociais. Para isso, são utilizados dados de sensores pluviométricos do CEMADEN e mensagens coletadas no Twitter. A partir disso, espera-se, além de entender as contribuições das mensagens de mídias sociais, identificar novos locais relatados por sensores humanos que não são monitorados para auxiliar na tomada de decisões das agências de emergências no caso de inundações.

O restante desse artigo está organizado da seguinte forma: na Seção 2 descreve-se a fundamentação teórica e alguns trabalhos relacionados. Na Seção 3 estão as técnicas e metodologias utilizadas nessa pesquisa. Na Seção 4 são apresentados os resultados. Por fim, a Seção 5 apresenta a conclusão e sugere trabalhos futuros.

2. Gestão de Risco de Inundações e Mídias Sociais para Desastres

No Brasil, os problemas de inundações são recorrentes. No período de 2004 a 2014, esses perigos naturais causaram mais dano do que outros tipos de eventos, como secas e escorregamentos de terra [Guha-Sapir et al. 2015]. Nesse contexto, a gestão de risco de inundações se mostra uma importante solução para minimizar os impactos sociais, financeiros e ambientais. Suas atividades podem ser agrupadas em três fases [Ahmad and Simonovic 2006]: (1) Planejamento pré-inundação; (2) Gestão de emergência; e, (3) Recuperação pós-inundação. Em todas estas fases, a coleta de informações é fundamental no suporte às atividades dos tomadores de decisão [Ahmad and Simonovic 2006].

Neste sentido, plataformas de mídia social como Twitter, Facebook e Instagram, permitem aos usuários o compartilhamento de suas informações com outras pessoas através da rede social. Por meio destas plataformas, torna-se possível analisar atividades diárias e, com isso, prever possíveis movimentações sociais. Alguns exemplos de pesquisas voltadas para o campo de desastres visam apoiar à tomada de decisões [Vieweg et al. 2014], auxiliar na predição de eventos [MacEachren et al. 2011] e aumentar o conhecimento situacional [Starbird et al. 2010]. Outro grupo de pesquisa busca analisar as contribuições para a integração de informações de mídias socias e dados oficiais. [Croitoru et al. 2013] revelam a existência de uma relação entre o espaço, rede social e eventos, que pode render na compreensão do comportamento de uma comunidade. [Albuquerque et al. 2015] demonstram que mensagens de redes sociais mais próximas ao evento natural podem possuir mais informações úteis sobre o desastre.

Apesar de tratar da integração de dados oficiais e mensagens de mídias sociais, muitas das pesquisas anteriores falham em utilizar esses dados como forma de filtrar mensagens de mídias sociais. Esta combinação poderia auxiliar na descoberta de conhecimento relevante e, assim, prover mais informações para melhorar a tomada de decisões na gestão de risco de inundações.

3. Metodologia

Esta pesquisa tem como objetivo analisar a relação geográfica entre dados oficiais e mensagens georreferenciadas de mídias sociais. Dessa forma, ela busca responder a seguinte pergunta de pesquisa: *PP) Dados oficiais podem auxiliar na identificação de novas áreas de inundação por meio da análise de mensagens georreferenciadas de mídias sociais?*

Para isso, essa Seção descreve os passos realizados para o desenvolvimento das análises qualitativas e quantitativas, tendo como estudo de caso o estado de São Paulo por possuir uma grande densidade populacional, com 166,25 habitantes por quilômetro quadrado [Instituto Brasileiro de Geografia e Estatística 2010], e 367 sensores pluviométricos monitorados pelo CEMADEN.

3.1. Análise qualitativa

A análise qualitativa é responsável pela classificação de mensagens publicadas na rede social Twitter no período de 7 a 31 de maio de 2015.

Para a coleta de mensagens, foi usado o serviço Twitter Streaming API[2] que permite uma coleta contínua utilizando filtragem por localização feita por um *bounding box*, uma área limite definida por um polígono através das posições geográficas de seus vértices. Um *bounding box* que abrange todo o estado de São Paulo foi determinado como: -53.11 (longitude mínima), -25.48 (latitude mínima), -44.16 (longitude máxima), -19.78 (latitude máxima). A partir disso, as mensagens recebidas foram armazenadas em uma base de dados não relacional orientada a documentos.

A análise dos dados necessitou que os *tweets* fossem normalizados, mantendo assim apenas as propriedades essenciais para a análise de conteúdo: identificador, hora de criação, texto e dados geográficos. Para os *tweets* que não possuíam geolocalização, a propriedade ”dados geográficos”foi definida com valor nulo.

Para a extração dos dados, foram considerados apenas *tweets* que possuíam georreferência do local de envio e mensagens com determinados termos relevantes para a pesquisa. Foram determinadas palavras-chave para evitar que conteúdo irrelevante fosse retornado. Após alguns testes pilotos para definir quais seriam os termos mais relevantes, os seguintes termos foram escolhidos: *chuva*, *chuvisco*, *água*, *garoa*, *nuvem*, *tempestade*, *temporal*, *dilúvio*, *alagamento*, *inundação*, *enchente*. Dessa forma, os *tweets* foram extraídos da base de dados a partir do mecanismo de consulta *full-text search*, que permite o retorno de mensagens que possuem as palavras-chave determinadas e termos similares.

Por fim, essas mensagens foram lidas e classificadas em categorias de acordo com o seu conteúdo. Mensagens sem relação com a proposta do estudo foram classificadas como ”fora do contexto”. Publicações com relação foram classificadas como ”dentro do contexto”, porém as mensagens mais relevantes, que possuíam informações temporais e geográficas, foram classificadas também como ”relevante”. Vale ressaltar também que foi realizado um processamento adicional com base nas coordenadas dos limites de São Paulo para garantir a inclusão de *tweets* apenas do estado.

[2]https://dev.twitter.com/streaming

3.2. Análise quantitativa

A análise quantitativa é responsável por identificar novas áreas de riscos de inundação através da combinação da análise dos *tweets* e dos locais das estações pluviométricas do Centro Nacional de Monitoramento e Alertas de Desastres Naturais.

As medições das estações pluviométricas estão disponíveis através da área de download do Mapa Interativo da Rede Observacional para Monitoramento de Risco de Desastres Naturais[3].

Os pluviômetros da área realizam medições a cada 10 minutos quando ocorre chuva contínua, caso contrário, de hora em hora. O arquivo transferido é uma planilha composta por dados dos pluviômetros, com identificador, coordenadas geográficas, hora da medição e volume de chuva. Para esta pesquisa, o documento do mês de maio e do estado de São Paulo foi utilizado.

A maior medição de chuva registrada no período analisado ocorreu em Campos do Jordão, atingindo um valor de 55,4 no dia 13/05 às 02h30. Contudo, o segundo maior valor é 28,4, registrado em Caieiras no dia 10/05 às 20h30. Portanto, essa medição de Campos do Jordão será considerada como um *outlier*, sendo removida da análise.

4. Resultados

No período estudado, foram coletados 1.589.549 *tweets* apenas com o filtro de *bounding box*. Adicionando os filtros de palavras-chave e georreferência, foram retornados 4.171 *tweets*. Com a remoção das mensagens que estavam fora dos limites do estado de São Paulo, foram totalizados 3.037 *tweets* para a análise. A partir da extração, os *tweets* foram classificados, atingindo uma quantidade de 1.614 mensagens fora do contexto, 1.423 dentro do contexto e, dentre estas, 1.181 relevantes para a pesquisa.

Com base na análise dos *tweets*, foi possível identificar dias com picos de publicações, em que a quantidade de mensagens dentro do contexto da pesquisa foi maior que as mensagens fora do assunto (Figura 1). Para investigar se o aumento da quantidade de mensagens relevantes está relacionado aos dias que ocorreram precipitações ou chuvas, foi necessário a análise das medições das estações pluviométricas.

Durante o período analisado, foram realizadas 403.046 medições nas estações pluviométricas. Para uma análise mais consistente dos dias e locais que registraram precipitações, os dados foram filtrados com volume de chuva maior que 0, chegando a uma quantidade de 56.032 medições. Na Figura 2 está representada a quantidade total de medições e as medições com volume de chuva por dia.

Para determinar se é possível identificar novas áreas de risco de inundação a partir da análise de mídia social combinada com pluviômetros, foi realizada uma análise dos dias 10 e 31, por representarem os maiores picos de atividades em ambos os gráficos.

Na Figura 3 está representada a disposição entre os locais de envio de *tweets* relevantes (pontos vermelhos) e as estações pluviométricas (clusters e marcadores azuis) que realizaram medições em 10 de maio. Com essa sobreposição, é possível identificar que a maioria dos locais que os *tweets* foram enviados relatando chuvas possuem pluviômetros próximos, como as regiões de Santos, São Paulo e Ribeirão Preto. Contudo, ainda assim

[3]http://www.cemaden.gov.br/mapainterativo

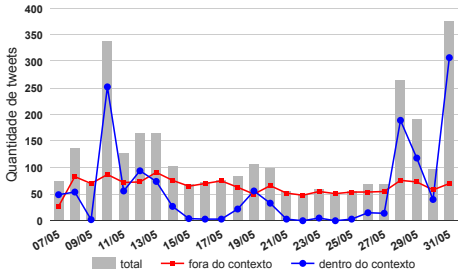

Figura 1. Quantidade de tweets classificados no período analisado

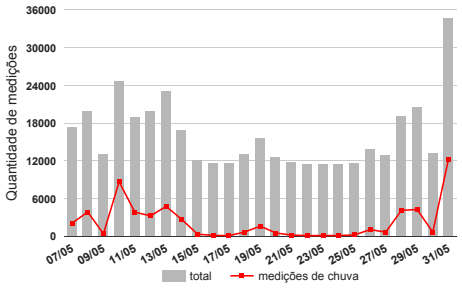

Figura 2. Quantidade de medições dos pluviômetros no período analisado

existem locais sem pluviômetros em que humanos agiram como sensores, sendo possível identificar possíveis novas áreas de risco, como na região de Ibitinga, Araçatuba e Birigui.

Na Figura 4 está apresentado a disposição entre sensores humanos e pluviômetros que realizaram medições no dia 31 de maio. Com essa sobreposição, é possível identificar que os principais *tweets* georreferenciados estão próximos de estações pluviométricas, com poucas exceções, como Presidente Prudente, Assis e Poços de Caldas.

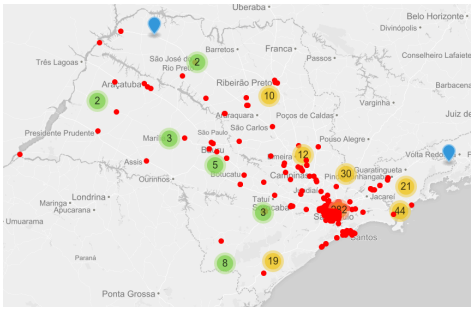

Figura 3. Disposição entre *tweets* e estações pluviométricas no dia 10 de maio

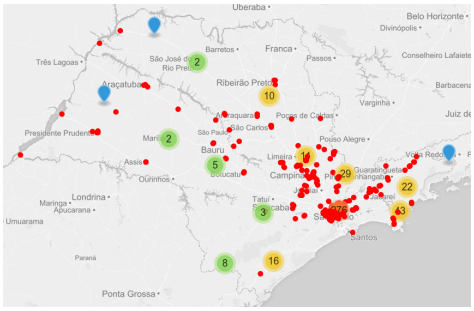

Figura 4. Disposição entre *tweets* e estações pluviométricas no dia 31 de maio

5. Conclusão

Nesse trabalho foi realizada uma análise quali-quantitativa para investigar se mensagens georreferenciadas de mídias sociais contêm informação útil para identificar novas áreas de risco de inundação. No período analisado, foram detectados picos de atividades com alta concentração de publicações de mensagens e medições de chuvas pelas estações pluviométricas. Com a análise do conteúdo de mensagens georreferenciadas relevantes, identificamos que os autores escrevem informações climáticas da região na qual se encontram, além de informar possíveis áreas de risco de alagamento. Dessa forma, pode-se afirmar que dados oficiais podem ser utilizados para auxiliar na filtragem de mensagens de mídias sociais e, assim, permitir a descoberta de informação relevante. Essa análise também serviria como uma etapa de preparação na gestão do risco de inundações, pois com uma grande concentração de mensagens sobre um mesmo evento, torna-se possível localizar novas áreas de risco.

Como trabalhos futuros, recomenda-se a elaboração de mapas de vulnerabilidade de inundação baseados em informações de redes sociais, comparando e avaliando com o mapa de vulnerabilidade da Agência Nacional das Águas (ANA)[4]. Uma análise geoestatística dos dados coletas (por exemplo, indicadores locais de associação espacial) mostrou-se necessária, sendo então adicionada nos próximos artigos. Além disso, tanto a criação de modelos para identificação de mudanças climáticas a partir de análise de redes sociais, quanto a automação das etapas de coleta e categorização de mensagens de mídias sociais são áreas promissas para trabalhos futuros.

Agradecimentos

THP agradece ao CNPq (130153/2015-0) e FAPESP (2015/05929-3) pelo apoio financeiro. FEAH e JPA agradecem a CAPES (Edital Pró-alertas 24/2014). FEAH agradece o suporte financeiro do CNPq (202453/2014-6). JPA agradece a CAPES (88887.091744/2014-01) e Heidelberg University (Excellence Initiative II / Action 7) por apoiar a sua contribuição à essa pesquisa.

Referências

Ahmad, S. and Simonovic, S. P. (2006). An Intelligent Decision Support System for Management of Floods. *Water Resources Management*, 20(3):391–410.

Albuquerque, J. P., Herfort, B., Brenning, A., and Zipf, A. (2015). A Geographic Approach for Combining Social Media and Authoritative Data Towards Identifying Useful Information for Disaster Management. *International Journal of Geographical Information Science*, pages 1–23.

Croitoru, A., Crooks, A., Radzikowski, J., and Stefanidis, A. (2013). Geosocial Gauge: a System Prototype for Knowledge Discovery from Social Media. *International Journal of Geographical Information Science*, 27(12):2483–2508.

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4):211–221.

Guha-Sapir, D., Below, R., and Hoyois, P. (2015). EM-DAT: International Disaster Database. Université catholique de Louvain.

Instituto Brasileiro de Geografia e Estatística (2010). Censo Demográfico. Disponível em: http://www.censo2010.ibge.gov.br/sinopse/index.php?dados=10&uf=00. Acesso em: 22 out.

MacEachren, A. M., Robinson, A. C., Jaiswal, A., Pezanowski, S., Savelyev, A., Blanford, J., and Mitra, P. (2011). Geo-twitter analytics: Applications in crisis management. In *25th International Cartographic Conference*, pages 3–8.

Starbird, K., Palen, L., Hughtes, A. L., and Vieweg, S. (2010). Chatter on The Red: What Hazards Threat Reveals about the Social Life of Microblogged Information. *Proceedings of the 2010 ACM conference on Computer Supported Cooperative Work (CSCW)*, pages 241–250.

Vieweg, S., Castillo, C., and Imran, M. (2014). Integrating Social Media Communications into the Rapid Assessment of Sudden Onset Disasters. *Social Informatics*, 8851:444–461.

[4]http://www2.snirh.gov.br/home/item.html?id=cf201bd9b2c540fa951b0619006eb2af