Statistical relations between flood days and Twitter activity

Vitor Y. Hossaki[1,2], Wilson Ceron[3], Jeferson F. Mendes[2], Rogério G. Negri[2], Lívia R. Tomás[1], Luciana R. Londe[1], Roberta B. Bacelar[4], Leonardo B. L. Santos[1]

[1]Centro Nacional de Monitoramento e Alertas de Desastres Naturais (CEMADEN), 12247-016 – São José dos Campos – SP – Brazil

[2]Instituto de Ciência e Tecnologia (ICT) – Universidade Estadual Paulista (UNESP) 12247-016 – São José dos Campos – SP – Brazil

[3]Universidade Federal de São Paulo (UNIFESP) 12231-280 – São José dos Campos – SP – Brazil

[4]Curso de Comunicação Social – Faculdade Anhanguera 12236-660 – São José dos Campos – SP – Brazil

{vitor.yuichi, rogerio.negri, jeferson.feitosa}@unesp.br

luciana.londe@cemaden.gov.br, wilson.seron@unifesp.br

{santoslbl, liviatomas, roberta.baldo}@gmail.com

Abstract. *Flooding is an increasingly frequent phenomenon and has been causing economic losses and even deaths in several cities worldwide, such as in the city of São Paulo, Brazil - the case study of this paper. Research shows that the use of voluntary geographic information is a helpful complementary tool for monitoring and analyzing a myriad of hydrological and meteorological events. This work analyzes and demonstrates that the number of tweets, about a specif set of keywords, is statistically higher on flooding days than on non-flooding ones.*

1. Introduction

Flooding is a frequent phenomenon in urban regions due to population growth and the characteristics of the urbanization process, causing an increase of impervious surfaces and poor drainage. This hydrological phenomenon is among the natural hazard associated with the most significant impact in the world. Over the years, there has been an increase in its global incidence. Global factors, such as Global warming, and local ones, such as the lack of urban planning [Tingsanchali 2012] are some of the variables causing this increase.

In the city of São Paulo, Brazil, flooding has been recurrent since the beginning of its occupation. The urban structure combined with the hydrographic and morphological characteristics helps trigger this phenomenon [Hirata et al. 2013]. Santos *et al.* (2013) estimated that the macroeconomic effects of flooding are 172.3 million reais per year.

Because of the material, economic and human impacts caused by flooding, measures to mitigate and anticipate this phenomenon are fundamental. In this context, social networks show a growing trend in their incorporation in research for the monitoring and analysis of a multitude of events. According to [De Albuquerque et al. 2015], and

[Hossaki 2021], the use of voluntary geographic information, especially the social network Twitter, are fundamental components for greater awareness of the events that occur. It consolidates the perception of elements in the environment and enables a greater understanding of the possible consequences.

In the meantime, some studies such as [Horita et al. 2015], [Hirata et al. 2013] and demonstrate that the use of voluntary geographic information such as Social Networks is essential for decision-making and flood management.

This work aims to verify quantitatively the information posted by users on a social network in a spatial and temporal radius, looking for statistical relations with hydrological events.

2. Materials and methods

2.1. Study area

The study area is in the city of São Paulo, Brazil, most precisely in the Tamanduateí River basin (Figure 1). This basin has an area of $323\,\mathrm{km}^2$ and extends to the hydrographic basins of Rio Pinheiro, Rio Guaió, Rio Aricanduva and Córrego de Tapuapé.

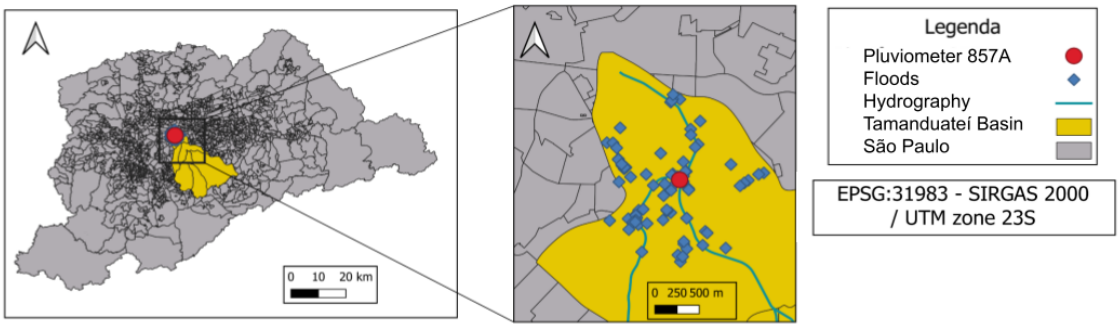


Figure 1. Study area location.

2.2. Materials

The data from the social network Twitter were extracted through the API (Application Programming Interface) [Krishnamurthy et al. 2008]. Flood registries were obtained from the Emergency Management Center of the city of São Paulo - CGE website [**?**]. Those flood registers come from the moment there is obstruction of the road caused by water accumulation.

For data manipulation, filtering and treatment, several Python libraries are used: Pandas [McKinney et al. 2011], Matplotlib [Nelli 2015], and Seaborn [Waskom 2021]. For the application of statistical tests was used Scipy [Virtanen et al. 2020] and Numpy [Van Der Walt et al. 2011]. Finally, some filtering in the flooding database was performed with geoprocessing tools from QGIS software [Uchoa and Ferreira 2004].

2.3. Method

To process the tweets data for the context linked to the floods, a word list (Table 1) was used to filter only the relevant posts. This list has words related to meteorological contexts

Table 1. List of keywords.

METEO	chuva	rain	temporal	lightning
	rainbow	raio	precipitacao	trovão
HIDRO	alagado	alagamento	enchente	
	inundacao	enxente		

(METEO) and hydrological phenomena (HIDRO). This filtering method is similarly used in international surveys that relate Twitter data to disaster phenomena [Wang et al. 2018].

A point in the lower portion of the Tamanduateí basin is the center of a circle with spatial radius SI 2000 m, where the daily social activity is quantified.

With the processed data, a time-series of words occurrence is generated (Table 1) for days of flooding and days of non-flooding. A boxplot is generated for visualization through this information, and the non-parametric Mann-Whitney test is applied for statistical significance analysis.

This non-parametric test is used to determine whether there is a difference between two groups of data [Perme and Manevski 2019]. The null H_0 determines that two analyzed series X and Y are equal. if your if p is less than the significance $\alpha = 0.05$, the null hypothesis is rejected [MacFarland and Yates 2016]. The statistical test was divided among the three months studied, applied in the first month, then in the two consecutive months, and finally, the entire temporal window, in which the variable X represents the temporal distribution of words on the days that flooding occurred and Y the distribution on days without flooding.

Complementary, Google Maps' photos of the study area show the topography of the places.

3. Results and Discussion

In general, most tweets presented the context associated with flooding or phenomena such as rain, despite some tweets in a metaphorical way. In addition, there is a very high prevalence of the words of the METEO class concerning the HIDRO class (Table 1).

From the generation of the time series, a boxplot is generated. The graph visually demonstrates the predominance of words on flooded days compared to non-flooded days.

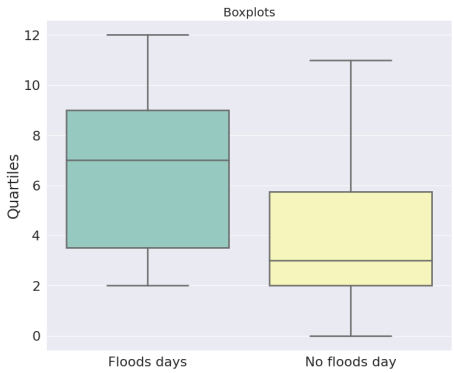


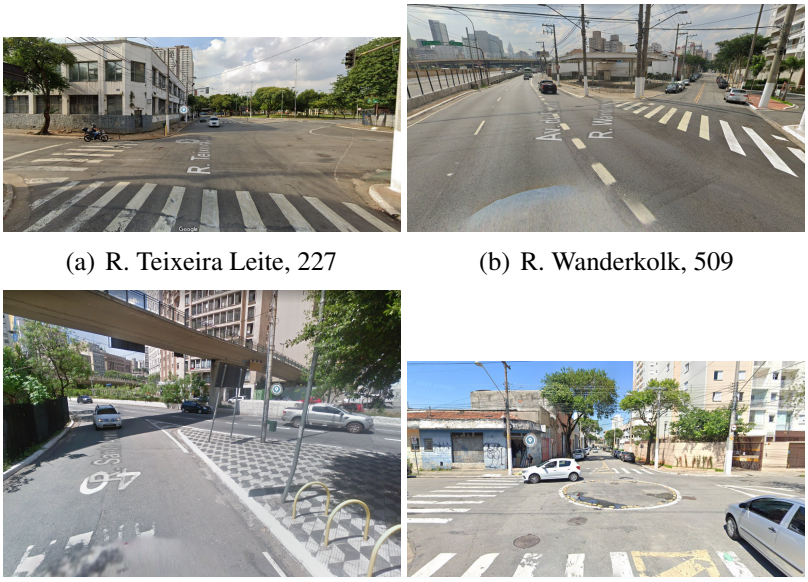
Figure 2. Box-plot for flood and non-flood distributions.

Table 2 shows the statistical analysis based on the nonparametric Mann-Whitney test. For samples with a few days, the test demonstrates that the null hypothesis is not rejected, that is, evidencing that there is no statistical difference between the two series (days of flooding, not flooding). However, as the data period expands, the test shows a considerable difference between the series.

Table 2. Mann-Whitney statistical results.

	January	January, February	January, February and March
Statistic of test	49	199	558
p-value	0.05725	0.02169	0.012948
Reject H_0?	No	Yes	Yes

The regions where there were more floods (Figure 3) share similar characteristics, flat or slightly sloping regions favor the accumulation of water in the region. Large parts of flooding in urban areas occur in flat areas or with depressions and valley bottoms, usually with surface runoff hampered by the topography of the location [Braga 2016]. It can be observed that there is a particular pattern of regions where more flooding occurs within the study area, and those specific topographical characteristics of the place lead to water accumulation.



(a) R. Teixeira Leite, 227 (b) R. Wanderkolk, 509

(c) R. Santos Amaro, 11 (d) R. Óscar Horta, 187

Figure 3. Regions where more floods in the study area (Photos from Google maps)

4. Conclusion and perspectives

The work indicates a statistical relation between the tweets associated with the meteorological and hydrological context with flooding.

It can be seen that there are certain methodological limitations, such as the low amount of georeferenced tweets, the spatial radius of the study area delimiting some elements, such as the more significant amount of tweets and frequency of flood.

Due to the myriad of factors linked to the outbreak of this hydrological phenomenon, other supplementary monitoring mechanisms are needed. In short, Twitter can be a powerful complementary tool. By joining this social network with other monitoring stations, sophisticated monitoring and alerting systems can be developed.

In the future, it is intended to apply artificial intelligence models in order to determine which is the most accurate algorithm to relate rainfall, meteorological data, and tweets for flood forecasting.

References

Braga, J. O. (2016). Alagamentos e inundações em áreas urbanas: estudo de caso na cidade de santa maria-df.

De 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*, 29(4):667–689.

Hirata, E., Giannotti, M. A., Larocca, A. P. C., and Quintanilha, J. A. (2013). Mapeamento dinâmico e colaborativo de alagamentos na cidade de são paulo. *Boletim de Ciências Geodésicas*, 19:602–623.

Horita, F. E., de Albuquerque, J. P., Degrossi, L. C., Mendiondo, E. M., and Ueyama, J. (2015). Development of a spatial decision support system for flood risk management in brazil that combines volunteered geographic information with wireless sensor networks. *Computers & Geosciences*, 80:84–94.

Hossaki, W. C. (2021). A twitter-based meteorological radar. *INIC. SJC*.

Krishnamurthy, B., Gill, P., and Arlitt, M. (2008). A few chirps about twitter. In *Proceedings of the first workshop on Online social networks*, pages 19–24.

MacFarland, T. W. and Yates, J. M. (2016). Mann–whitney u test. In *Introduction to nonparametric statistics for the biological sciences using R*, pages 103–132. Springer.

McKinney, W. et al. (2011). pandas: a foundational python library for data analysis and statistics. *Python for high performance and scientific computing*, 14(9):1–9.

Nelli, F. (2015). Data visualization with matplotlib. In *Python data analytics*, pages 167–235. Springer.

Perme, M. P. and Manevski, D. (2019). Confidence intervals for the mann–whitney test. *Statistical methods in medical research*, 28(12):3755–3768.

Tingsanchali, T. (2012). Urban flood disaster management. *Procedia engineering*, 32:25–37.

Uchoa, H. N. and Ferreira, P. R. (2004). Geoprocessamento com software livre. *Publicação eletrônica. Rio de Janeiro*.

Van Der Walt, S., Colbert, S. C., and Varoquaux, G. (2011). The numpy array: a structure for efficient numerical computation. *Computing in science & engineering*, 13(2):22–30.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272.

Wang, R.-Q., Mao, H., Wang, Y., Rae, C., and Shaw, W. (2018). Hyper-resolution monitoring of urban flooding with social media and crowdsourcing data. *Computers & Geosciences*, 111:139–147.

Waskom, M. L. (2021). Seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021.