Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 9, n. 1, 2022.

Trabalho apresentado no XLI CNMAC, Unicamp - Campinas - SP, 2022.

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics
Preprint

Statistical analysis of data from Twitter and weather sensors

Vitor Y. Hossaki[1]
UNESP-ICT, São José dos Campos, SP
Wilson Ceron[2]
UNIFESP, São José dos Campos, SP
Rogério G. Negri[3]
UNESP-ICT, São José dos Campos, SP
Leonardo B. L. Santos[4]
CEMADEN, São José dos Campos, SP

VGI (*Volunteered Geographic Information*) systems allow their users to voluntarily share georeferenced data via social networks. Due to the generation of momentary geoinformation in the occurrence of events, several studies have carried out analyzes on the behavior patterns of the social network Twitter in relation to different natural and anthropic events.

However, the complexity of collecting relevant data in the face of the large volume of information, added to the verification of the reliability of the filtered content, pose challenges to researchers [2]. The integration of VGI data and *in situ* sensors, combined with an architecture that processes data and calculates statistics, have shown positive results for the monitoring of adverse events of natural origin.

In this context, this work aims to analyze statistical correlations between data from the Twitter social network with data from a rainfall station and weather radar.

Rainfall data were recorded by the 833A rain gauge, belonging to CEMADEN (National Center for Monitoring and Alerts for Natural Disasters), which are publicly available by that institution. The radar data made available by the Department of Airspace Control (DECEA), with spatial and temporal resolutions of 1 km and 10 min, respectively, expressed in terms of reflectivity (dB) into separation rate (mm/h), were converted according to the Marshall-Palmer relationship and then used in the calculation of accumulated values per day over the study area.

Twitter data extraction was performed through the API (*Application Programming Interface*), which allowed the selection of georeferenced tweets that were published within a radius of 2000 m in relation to the rain gauge considered. Subsequently, the tweets were filtered and aggregated by day, based on the following rain-related keywords: 'chuva', 'chove', 'chuvoso', 'chuvosa' e 'chuvarada'. According to [1], such keywords are less spatially and temporally volatile than local and idiosyncratic terms. The information covers the first three months of 2019 and data processing was performed using the *Python 3* language and the SciPy library.

After processing the data, in order to verify the adherence of the data to the Gaussian distribution, the Shapiro-Wilk test was applied according to a significance $\alpha = 0.05$. The results of this initial analysis showed that the data do not follow a Gaussian distribution, justifying in turn the use of Spearman's (non-parametric) correlation measure. This measure expresses the strength and

[1] vitor.yuichi@unesp.com
[2] wilsonseron@gmail.com
[3] rogerio.negri@unesp.br
[4] santoslbl@gmail.com

direction of the association between variables that do not have a Gaussian distribution [3]. As this measure approaches ± 1, the greater the correlation (direct/indirect) between these variables.

Table 1 displays the Spearman correlation values, thus showing a significant correlation between the variables. The occurrence of precipitation and the detection of the event by meteorological sensors (i.e., rain gauge and radar) indicate a correlation with Twitter posts.

Table 1: Spearman's Correlation | p-value

	Pluviometer	Tweets	Radar
Pluviometer	$1 \mid 0$	$0.48 \mid 1.31 \cdot 10^{-6}$	$0.41 \mid 4.82 \cdot 10^{-5}$
Tweets	$0.48 \mid 10^{-6}$	$1 \mid 0$	$0.57 \mid 4.51 \cdot 10^{-9}$
Radar	$0.41 \mid 4.8 \cdot 10^{-5}$	$0.57 \mid 4.51 \cdot 10^{-9}$	$1 \mid 0$

The results indicate that data from weather radar and rain gauge present statistical correlation with information shared by users of the Twitter network, when they mention the occurrence of rain. However, when the keywords considered are used in different contexts, the filtering/selection process of tweets may consider posts unrelated to the meteorological phenomenon in question.

Thus, due to the statistical relationship found between the analyzed databases, we can conclude that the social network Twitter is an alternative tool in the analysis of atmospheric phenomena. As a future perspective, we intend to investigate machine learning models capable of relating rainfall data, meteorological data and tweets for the purpose of flood forecasting.

Acknowledgments

We thank the National Council for Scientific and Technological Development (CNPq) and the Foundation for Research Support of the State of São Paulo (FAPESP - Proc: 2021/01305-6) for promoting this research. Thanks to Lívia Tomás, Luciana R. Londe and Roberta B. Bacelar for supporting the research.

References

[1] Sidgley Camargo de Andrade et al. "The effect of intra-urban mobility flows on the spatial heterogeneity of social media activity: investigating the response to rainfall events". In: **International Journal of Geographical Information Science** (2021), pp. 1–26.

[2] Joao Porto De Albuquerque et al. "A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management". In: **International journal of geographical information science** 29.4 (2015), pp. 667–689.

[3] Yadolah Dodge. **The concise encyclopedia of statistics**. Springer Science & Business Media, 2008.