TWITTER PLUVIOMETER

Student: Vitor Yuichi Hossaki
Supervisor: Prof. Dr. Leonardo Bacelar Lima Santos

Report to CNPq/Cemaden Scientific Initiation Scholarship Program

December of 2021

SUMÁRIO

Pág.

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 (2013) estimated that the macroeconomic effects of flooding are 172.3 million reais per year. Based on the impacts mentioned, there is a growing trend in the literature incorporating tools such as the social network to predict flooding.

In this context, some researchers, such as those presented by Horita et al. (2015) and Hirata et al. (2013), demonstrate that the use of social networks provided with voluntary geographic information can be used as an effective instrument in the development of flood monitoring and warning systems.

Similarly, Albuquerque et al. (2015) demonstrates the potential of spatiotemporal data obtained through the social network Twitter for disaster management. In the cited search, posts are filtered through keywords and then binary sorted among those "related" or "not related" to the event of interest, which in turn are aligned to a time series of outbreaks of flooding in a given region. The results obtained indicate the statistical relevance of the information pointed out by the Tweets in relation to the flooded regions. In a similar way Hossaki et al. (2021) also demonstrates that the words used in the Twitter social network linked to the meteorological context on days of flooding are greater than in relation to days that do not occur.

2 OBJECTIVES

The primary purpose of this report is to perform an exploratory and statistical analysis on rainfall, radar, Twitter and flooding data, determining correlations and

statistical distribution of attributes.

As more specific objectives, it consisted of data processing and generation of a single file integrating all databases used in the research. Analyzing the result of filtering tweets regarding the general context of the posts.

3 THEORY FUNDAMENTALS

3.1 Floods

Flooding is a complex phenomenon, as its cause is interrelated to a range of parameters such as climate, urban structural faults, inadequate drainage systems, hydrographic basins, proximity to water bodies, inappropriate use and occupation of the soil, among others (DOOCY et al., 2013).

The absence of urban planning and the rapid modification of the space culminate in soil sealing, contributing to a reduction in concentration time and an increase in the volume of surface runoff, thus amplifying the peak flow and consequently saturating the site's rainwater drainage (HANSMANN, 2013).

The local topography is also a preponderant factor for the occurrence of flooding, since it has already been verified that the places with the highest frequency of flooding have flat morphometric characteristics, depressions or valley bottoms, hindering the local runoff process (BRAGA, 2016).

Also, due to the population's lack of environmental education, the inadequate disposal of solid waste appears as another factor causing the obstruction of the local drainage system, once again leading to flooding.

3.2 Statistics

Statistics is a fundamental discipline to provide method and tools to find deeper insight into data. Data science and Machine Learning need the concepts of statistics to build more sophisticated models and analyses.

One of the fundamental structures of statistics is the hypothesis test which establishes a direct relationship between theory and statistics, since hypotheses can be tested with data (WEIHS; ICKSTADT, 2018).

3.3 Correlation

Correlation is a relationship between two variables represented by ordered pairs (x, y) with x the independent variable and y the dependent variable. In Data Science, correlations are widely used to establish statistical relationships and inferences with greater precision.

However, the correlation does not imply causality, it must be analyzed whether a variable x causes y, in addition to verifying if there is a cause between the two variables through a third, thus discarding that there is only a correlation by sheer coincidence (LARSON et al., 2004).

The most used test for parametric tests is the Pearson coefficient. Given a sample of X and Y, the Pearson coefficient is calculated by the equation 3.1, where ρ is correlation degree ranging from 0 to 1.

$$\rho = \frac{cov(X, Y)}{\sqrt{var(X) \cdot var(Y)}} \tag{3.1}$$

For non-parametric tests, descriptive statistics uses Spearman's coefficient. Pearson's correlation assesses linear relations, Spearman's correlation assesses monotonous relationships, whether linear or not. This correlation is described by the equation

3.2, where n is the number of samples, d is the differences between ranks and is the correlation.

$$\rho = 1 - \frac{6 \sum d^2}{n^2(n^2 - 1)} \tag{3.2}$$

3.4 Normality test

To apply the correct test, it is necessary to determine the distribution using the Shapiro Wilk test, which consists of a normality test. Assuming that the null hypothesis assumes that the sample has a normal distribution. The W test statistic for normality is defined by the 3.3 equation. Where y is the variable in the sample and a the tabulated coefficient.

$$W = \frac{b^2}{S^2} = \frac{(\sum_{i=1}^{n} a_i y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y_i}))^2} \tag{3.3}$$

3.5 Mann Whitney Test

This non-parametric test is used to determine whether there is a difference between two groups of data (PERME; 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; 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.

3.6 Materials and method

3.6.1 Study Area

This project admits as a study area the region of São Paulo inserted in the hydrographic basin of the River Tamanduateí (Figure 3.1). This basin has an area of 323 km^2 and extends to the hydrographic basins of the Pinheiro, Guaió, Aricanduva and Córrego de Tapuapé rivers. This area was defined from the vicinity of a rain gauge, according to a spatial radius of 2000 m, which encompasses different flooding regions, available Tweets and a meteorological radar cell.

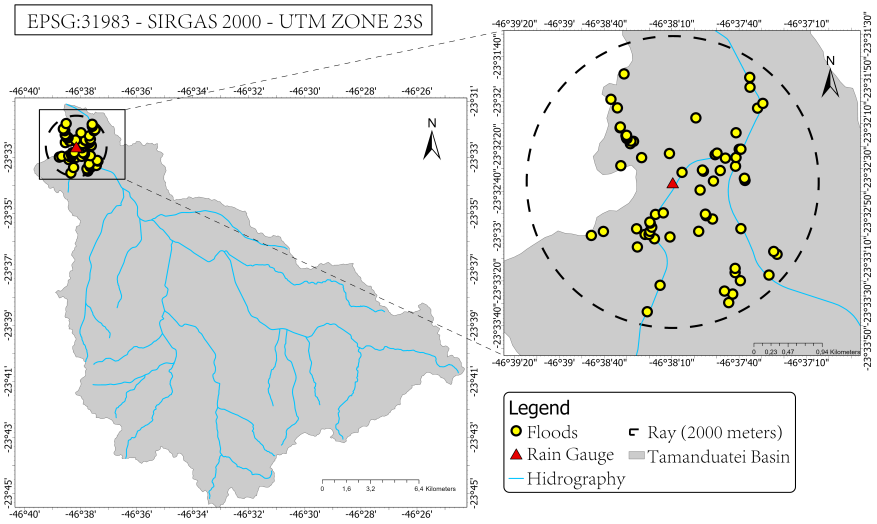


Figura 3.1 - Study Area

3.6.2 Data and tools

Twitter data were extracted through an API (*Application Programming Interface*) provided by the social network itself. The pluviometric data are collected from the pluviometer #833A, belonging to the National Center for Alerts and Natural Di-

sasters (CEMADEN), which are made publicly available by the institution. The historical series of flooding in the study area is available through CEMADEN.

The data from the meteorological radar were extracted by a station located in the city of São Roque. This equipment, maintained by the Department of Airspace Control (DECEA), it monitors displacement, action of clouds and instability nuclei, measuring the volume of precipitation in a given location. Furthermore, this radar has a range of 250 km, covering the entire metropolitan region of São Paulo. The radar product used is called CAPPI (*Constant Altitude Plan Position Indicator*), which has a spatial resolution of approximately 1 km and a temporal resolution of 10 minutes. For the conversion of reflectivity (dBZ) into separation rate (mm/h) the Marshall-Palmer ratio (ANDW, 1948) will be used, and then represented as "daily accumulated".

The development of the project will be guided by programming via the *Python* language. The manipulation, filtering and processing of data will be supported by the *Pandas* (VANDERPLAS, 2016) and *Numpy* (MCKINNEY, 2012) libraries.

For the application of statistical tests, the *Scipy* (VIRTANEN et al., 2020) library will be used. Finally, necessary database operations will be performed with the support of the Geographic Information System *QGIS* (SAMELA et al., 2018)

3.6.3 Method in the first project

Firstly, all geolocated tweets from January to March 2019 within a radius of 2000 meters were extracted through API. The centroid of this ray is located in pluviometer 857A in São Paulo and the coordinates of this place can be found on the website of the institution CEMADEN. The data extracted through the Twitter API, uses the UTC format to inform the day and time of posts. Therefore, in the pre-processing of the tweets data, the date-time was converted from UTC for the América/São Paulo region. Also, to facilitate visualization and optimize the data, some unnecessary

columns were removed. Finally, the date was delimited for the analyzed time window.

With the data from the social network, the filtering of tweets was carried out from a list of words separated as associated phenomena such as meteorological (METEO) and hydrological (HIDRO) 3.1.

Tabela 3.1 - List of keywords.

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

From the analysis of the time of occurrence of the floods using Microsoft Excel, a large portion of the floods occur between 4 pm and 8 pm. Therefore, the tweets were filtered from the list of words (Table 3.1), and in the time frame from 4:00 pm to 8:00 pm. The floods were processed through and filtered only to those belonging to the analyzed time interval.

The frequency of words in the METEO and HYDRO list was also generated, separating it into two data series, number of times of occurrence in days with flooding and days without flooding. This binary classification allowed the generation of Boxplots.

From the two series mentioned above, the MannWhitney non-parametric statistical test was applied to verify whether there is statistical relevance in the difference between the two series.

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

3.6.4 Method in the second project

Based on the results of the Mann Whitney test, it was possible to use new data sources by integrating with the tweets. So similarly, tweets were collected via the API. To filter the posts for the context of flooding and rain, the words 'chuva', 'chove', 'chuvoso' and 'chuvosa' were used. These terms, according to (ANDRADE et al., 2021), are less temporally and spatially volatile and are commonly used in the city of São Paulo to refer to the phenomena of rain and flooding. Any tweets that contain one of the above words are selected and counted for the day the post was sent by the user. Then, these posts are aggregated for each of the days of the analyzed time frame.

Data from the 833A pluviometer were extracted from the website of the CEMADEN institution. The files containing the equipment information are found separated by month and measurements of all other rain gauges. In the primary treatment of the data, the files were concatenated and unified in a single DataFrame file, later only the measurements of the 833A rain gauge were filtered. This equipment, in times of absence of rain, records the information once in an hour, otherwise, in the detection of precipitation it starts measuring the rain every 10 minutes. Therefore, to analyze the rainfall data, the accumulated rainfall was calculated every 10 minutes for each day.

Radar data measurements are already processed and the information is found with the daily rainfall accumulation in a given cell. These cells comprise each of the rain gauge points. Therefore, values were extracted only from the point coinciding with the 833A pluviometer.

The flooding data was thoroughly reviewed by a member of the CEMADEN team. This information comes from meteorological and hydrological monitoring institutions. The outbreak of a flood is recorded from the moment there is an obstruction

of a road due to accumulation of water and precipitation. Data are primarily found records of flooding throughout the entire Tamanduateí Basin and for the entire year of 2019. The day, time, duration, street name, latitude, longitude, start and end time of the flood are recorded in the file.

For the processing of the flooding data, geoprocessing tools in QGIS were used. First, a circumference of radius 2000 m around the 833A rain gauge was created using the Buffer algorithm. Soon after, through the intersection algorithm between layers, only the floods belonging to the circle created by the Buffer were selected. The data were then delimited for the time window analyzed and the occurrences of flooding per day were counted.

With the processed data in hand, each of the attributes was integrated into a single DataFrame through the Merge function. In this table, all attributes are organized by the index that corresponds to the date, and each of the corresponding columns accumulated rainfall in the pluviometer and radar, frequency of tweets and number of occurrences of flooding.

To explore the data and verify the relationships between the attributes, the data were plotted as a function of the time window in a single graph. less frequent flooding. For more accurate inferences, Pearson and Spearman correlations were calculated. In selecting the appropriate correlation algorithm, the Shapiro Wilk test was used to determine whether the attributes followed a normal distribution, that is, parametric or non-parametric.

After determining the type of statistical distribution, the appropriate correlation algorithm was defined. Thus, inferences were made relating the attributes with the occurrence of flooding.

4 RESULTS AND DISCUSSION

4.1 Results in the first project

For all time windows analyzed, METEO-type words were predominant in all cases. Users' posts recurrently use in their sentences words such as rain, storm, among others. In some cases, users use the word classes METEO and HYDRO metaphorically, not referring directly to the analyzed flooding contexts.

In this context, the words were also temporally separated, and a Boxplot was plotted from the data obtained (Figure 4.1).

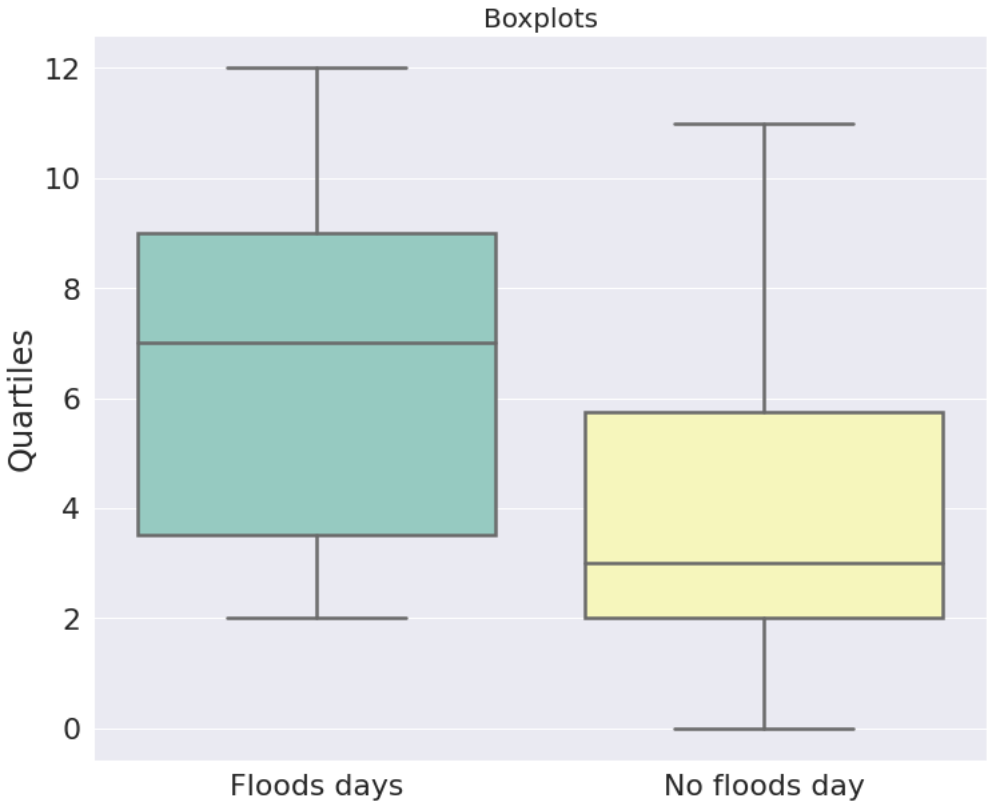


Figura 4.1 - Box plot, fourth time window

Boxplot visually demonstrates the relevance of METEO class words on flood days compared to non-flood days. To prove this statistical relevance, the Mann Whitnney

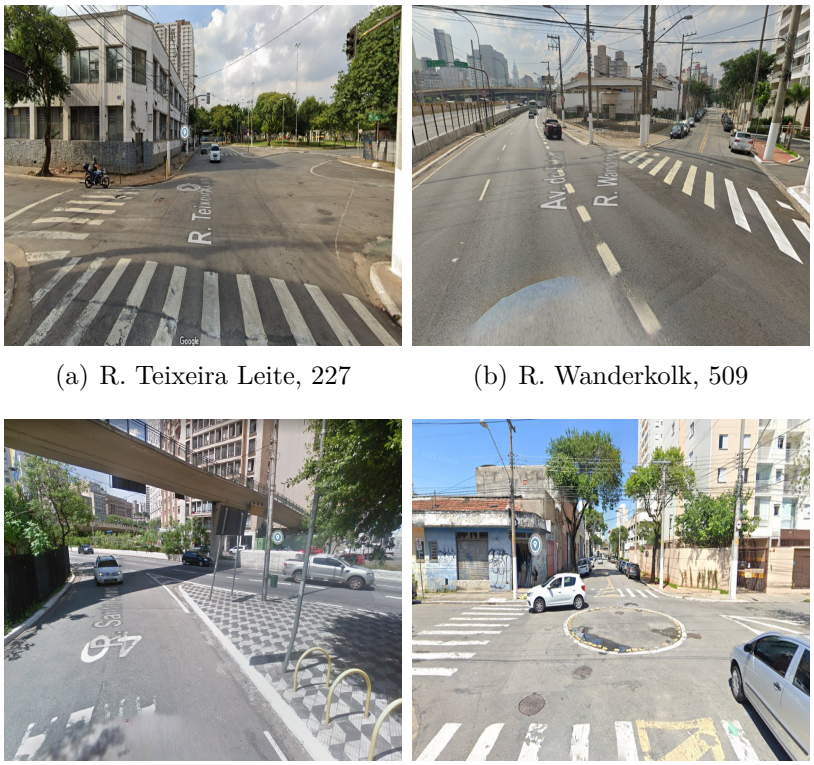
test was applied.

The table 4.1 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.

Tabela 4.1 - 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 4.2) 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. 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. Ãscar Horta, 187

Figura 4.2 - Regions where more floods in the study area (Photos from Google maps)

4.2 Preliminary Results in the second project

The tweets filtered based on the words presented in the 3.6.4 methodology, presented a large part of the tweets related to the context of rain and flooding. The use of other words such as 'raio' and 'tempestade', result in tweets using these words to designate a metaphorical context or with a sense displaced from the phenomena of rain and flooding. Therefore, the application of the word 'chuva' and its variations present a more stable meaning to refer to meteorological phenomena.

However, it can be observed that the filtering algorithm detects tweets using the word rain in a more poetic sense (Table 4.2).

Tabela 4.2 - Related and unrelated Tweets

Examples of Tweets	
Related tweets	Tweets out of context
quem aqui gosta de pokemon?\nvideo de dias atras porque a chuva estragou meus planos hoje	minha forÃ§a esta na solidao. nÃ£o tenho medo nem de chuvas tempestivas nem de grandes ventanias soltas.
sabado com chuvas e minhas aluna vieram fazer um alongamento para tira toda preguica	a ordem e seguir em frente romper a tempestade e nao se ater aos ventos, raios e chuvas.

After the unification of all attributes in a single DataFrame, a graph was plotted relating these variables to the time window studied 4.3.

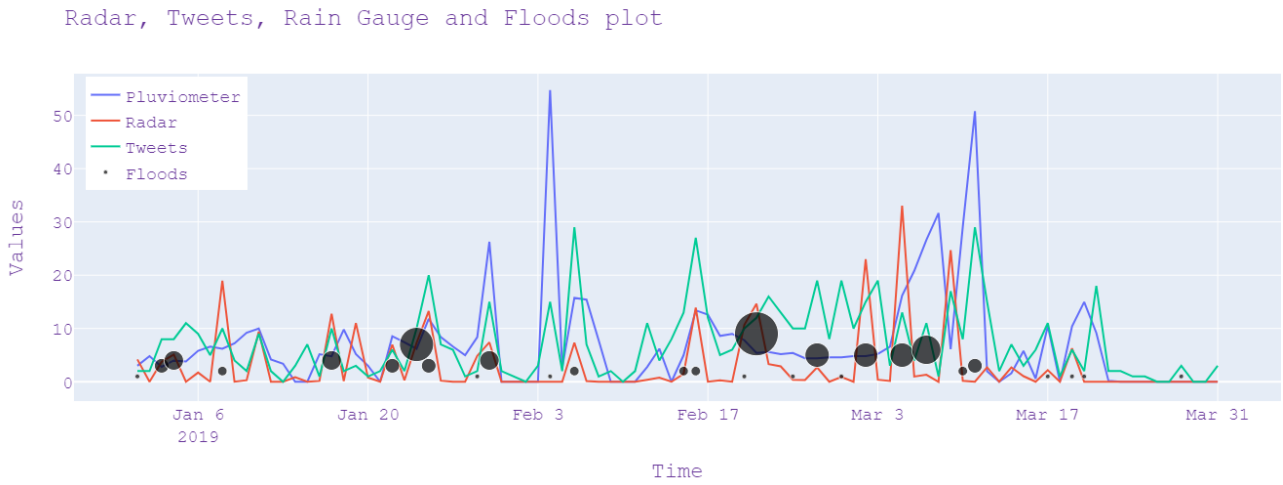


Figura 4.3 - Plot

The plotting of data indicates that on every day that flooding occurred there was an incidence of rain indicated by the rain gauge and radar and rain-related tweets. As can be seen in the figure 4.3, the frequency of flooding on a given day determines the size of the black circle, and that the days with the highest incidence of flooding were not peaks of rain detected by meteorological equipment.

To determine the statistical relationships a series of tests were performed. First, it was necessary to determine if the attributes were a normal distribution, so that in this way parametric or non-parametric statistical tests could be applied. For this verification, the Shapiro Wilk test was used (4.2).

Tabela 4.3 - P-Values results

Shapiro-Wilk test $\alpha = 0.05$	
Attributes	P-value
Rain Gauge	7.175038015984347e-13
Tweet Frequence	2.0394061550632614e-07
Radar	7.194410709808908e-15
Flood	1.445436313501046e-14

In the table above 4.2, the p-values of all attributes were below the limit of α, discarding the null hypothesis that the distribution is normal. Based on the results, to calculate the correlations, Spearman's non-parametric test (4.4) was used for these data.

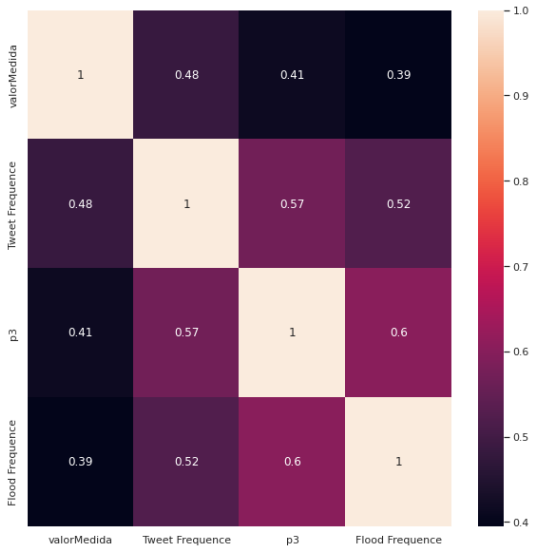


Figura 4.4 - Spearman Correlation

From the correlations, it is clear that the radar (p3), followed by the tweets, have the highest correlation with the frequency of flooding and also the most significant value. According to Correlation... (2021), the cited attributes indicates a strong correlation with floods.

Furthermore, rainfall data indicate the occurrence of precipitation in opposition to Radar. As seen in figure 4.3, the graph indicates the low correlation between these two attributes.

From the results, the most relevant correlations will also exert greater influence on the flooding prediction.

5 CONCLUSION AND PERSPECTIVES

The first project work indicates a statistical relation between the tweets associated with the meteorological and hydrological context with flooding. 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 general, in the second project it can be observed that the value of the attributes with the floods have moderate and strong correlations. Demonstrating that there is potential to submit data to classification algorithms.

In possession of primary analysis, the next step is to prepare the data for Machine Learning algorithms, optimize data quality and detect outliers.

REFERÊNCIAS BIBLIOGRÁFICAS

ALBUQUERQUE, J. P. D.; HERFORT, B.; BRENNING, A.; ZIPF, A. A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. **International journal of geographical information science**, Taylor & Francis, v. 29, n. 4, p. 667–689, 2015. 2

ANDRADE, S. C. de; ALBUQUERQUE, J. Porto de; RESTREPO-ESTRADA, C.; WESTERHOLT, R.; RODRIGUEZ, C. A. M.; MENDIONDO, E. M.; DELBEM, A. C. B. The effect of intra-urban mobility flows on the spatial heterogeneity of social media activity: investigating the response to rainfall events. **International Journal of Geographical Information Science**, Taylor & Francis, p. 1–26, 2021. 9

ANDW, J. M. Mc. k. palmer,âthe distribution of raindrops with size,â. **J. Meteor**, v. 5, p. 165–166, 1948. 7

BRAGA, J. O. Alagamentos e inundações em áreas urbanas: estudo de caso na cidade de santa maria-df. 2016. 3

CORRELATION (Pearson, Kendall, Spearman). Aug 2021. Disponível em: <https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/correlation-pearson-kendall-spearman/>. 16

DOOCY, S.; DANIELS, A.; MURRAY, S.; KIRSCH, T. D. The human impact of floods: a historical review of events 1980-2009 and systematic literature review. **PLoS currents**, Public Library of Science, v. 5, 2013. 3

HANSMANN, H. Z. Descrição e caracterização das principais enchentes e alagamentos de pelotas-rs. **Universidade Federal de Pelotas, Pelotas-RS**, 2013. 3

HIRATA, E.; GIANNOTTI, M. A.; LAROCCA, A. P. C.; QUINTANILHA, J. A. Mapeamento dinâmico e colaborativo de alagamentos na cidade de são paulo. **Boletim de Ciências Geodésicas**, SciELO Brasil, v. 19, p. 602–623, 2013. 2

HORITA, F. E.; ALBUQUERQUE, J. P. de; DEGROSSI, L. C.; MENDIONDO, E. M.; UEYAMA, J. Development of a spatial decision support system for flood risk management in brazil that combines volunteered geographic information with wireless sensor networks. **Computers & Geosciences**, Elsevier, v. 80, p. 84–94, 2015. 2

HOSSAKI, C.; FREITAS, N. et al. Statistical relations between floods and twitter activity. **GEOINFO**, p. 1–6, oct 2021. 2

LARSON, R.; FARBER, B.; PATARRA, C. tradução técnica. **Estatística aplicada**. [S.l.]: Prentice Hall, 2004. 4

MACFARLAND, T. W.; YATES, J. M. Mann–whitney u test. In: **Introduction to nonparametric statistics for the biological sciences using R**. [S.l.: s.n.], 2016. p. 103–132. 5

MCKINNEY, W. **Python for data analysis: Data wrangling with Pandas, NumPy, and IPython**. [S.l.]: "O'Reilly Media, Inc.", 2012. 7

PERME, M. P.; MANEVSKI, D. Confidence intervals for the mann–whitney test. **Statistical methods in medical research**, SAGE Publications Sage UK: London, England, v. 28, n. 12, p. 3755–3768, 2019. 5

SAMELA, C.; ALBANO, R.; SOLE, A.; MANFREDA, S. A gis tool for cost-effective delineation of flood-prone areas. **Computers, Environment and Urban Systems**, Elsevier, v. 70, p. 43–52, 2018. 7

SANTOS, E. T. d. **Impactos econômicos de desastres naturais em megacidades: o caso dos alagamentos em São Paulo**. Tese (Doutorado) — Universidade de São Paulo, 2013. 2

TINGSANCHALI, T. Urban flood disaster management. **Procedia engineering**, Elsevier, v. 32, p. 25–37, 2012. 2

VANDERPLAS, J. **Python data science handbook: Essential tools for working with data**. [S.l.]: "O'Reilly Media, Inc.", 2016. 7

VIRTANEN, P.; GOMMERS, R.; OLIPHANT, T. E.; HABERLAND, M.; REDDY, T.; COURNAPEAU, D.; BUROVSKI, E.; PETERSON, P.; WECKESSER, W.; BRIGHT, J. et al. Scipy 1.0: fundamental algorithms for scientific computing in python. **Nature methods**, Nature Publishing Group, v. 17, n. 3, p. 261–272, 2020. 7

WEIHS, C.; ICKSTADT, K. Data science: the impact of statistics. **International Journal of Data Science and Analytics**, Springer, v. 6, n. 3, p. 189–194, 2018. 4