

Mining Rainfall Spatio-Temporal Patterns in Twitter: A Temporal Approach

Sidgley Camargo de Andrade, Camilo Restrepo-Estrada,
Alexandre C. B. Delbem, Eduardo Mario Mendiando
and João Porto de Albuquerque

Abstract Social networks are a valuable source of information to support the detection and monitoring of targeted events, such as rainfall episodes. Since the emergence of Web 2.0, several studies have explored the relationship between social network messages and authoritative data in the context of disaster management. However, these studies fail to address the problem of the temporal validity of social network data. This problem is important for establishing the correlation between social network activity and the different phases of rainfall events in real-time, which thus can be useful for detecting and monitoring extreme rainfall events. In light of this, this paper adopts a temporal approach for analyzing the cross-correlation between rainfall gauge data and rainfall-related Twitter messages by means of temporal units and their lag-time. This approach was evaluated by conducting a case study in the city of São Paulo, Brazil, using a dataset of rainfall data provided by the Brazilian National Disaster Monitoring and Early Warning Center. The results provided evidence that the rainfall gauge time-series and the rainfall-related tweets are

S.C. de Andrade (✉)

Federal University of Technology – Paraná, Curitiba, Toledo, Brazil
e-mail: sidgleyandrade@utfpr.edu.br; sidgleyandrade@usp.br

S.C. de Andrade

University of São Paulo, São Carlos, Brazil

C. Restrepo-Estrada

São Carlos School of Engineering, University of São Paulo, São Carlos, Brazil
e-mail: camilo.restrepo@udea.edu.co

A.C.B. Delbem

Institute of Mathematical and Computing Sciences, University of São Paulo,
São Carlos, Brazil
e-mail: acbd@icmc.usp.br

E.M. Mendiando

Brazilian National Center of Monitoring and Early Warning of Natural Disasters,
São José dos Campos, Brazil
e-mail: emm@cemaden.gov.br

J.P. de Albuquerque

Centre for Interdisciplinary Methodologies, University of Warwick, Coventry, UK
e-mail: J.Porto@warwick.ac.uk

not synchronized, but they are linked to a lag-time that ranges from -10 to $+10$ min. Furthermore, our temporal approach is thus able to pave the way for detecting patterns of rainfall in real-time based on social network messages.

Keywords Social network · Twitter · Rainfall · Temporal analysis · Time-series correlation

1 Introduction

Social networks are playing an increasingly important role in the fields of information fusion and data mining because of the assistance they give in detecting and monitoring targeted events that attract the attention of users, such as rainfall episodes. In the last few years, there has been a growing interest in analyzing social network messages for disaster management (Steiger et al. 2015). In this context, the focus of recent research has been on analyses of social network messages to detect events such as earthquakes (Crooks et al. 2013; Earle et al. 2012; Sakaki et al. 2010), hurricanes (Kryvasheyeyu et al. 2016), and forest fires (Spinsanti and Ostermann 2013). Another group of studies has adopted an approach for extracting and classifying relevant information about the target-event, e.g. situational updates about an ongoing event (Albuquerque et al. 2015; Herfort et al. 2014; Starbird et al. 2012; Imran et al. 2013).

However, the previous studies fail to address the problem of the temporal validity of the social network data in real-time. This validity is important for establishing the correlation between social network activity and the different phases of targeted events. For instance, in real situations a decision-maker might want to understand the relation between the social network messages and the phase of the targeted event. In view of this, there is a need to fill a gap in these studies since they only examine social network messages after the event and show how they spread from the moment the targeted event starts. In other words, previous studies concentrated on conducting a post-hoc analysis based on the whole dataset instead of treating the social network stream as a time-series.

Furthermore, another limitation is that some of studies used a large time-scale for exploring the spatio-temporal correlation. Others assigned a temporal resolution that is based on social network messages alone, rather than supporting them with authoritative data to improve credibility. For example, Albuquerque et al. (2015) explored the correlation between the significance and proximity of the flood-related tweets with water levels using a one-day time-scale. Kryvasheyeyu et al. (2016) investigated the Twitter messages aggregate hourly to predict the location and severity of hurricane damage. Earle et al. (2012) and Sakaki et al. (2010) did not include the temporal resolution of the authoritative data for detecting earthquakes. For example, Sakaki and Matsuo (2012) “detect an earthquake when five positive tweets arrived in 5 min”. Assis et al. (2015) prioritized flood-related tweets only if a sensor reported a high water level within of catchment. However, rainfall detection in real-time using

rainfall gauge data with social network messages in urban areas requires a flexible and a finer time-scale than other events such as earthquakes, hurricanes, fires and floods. It is because extreme rainfall episodes, such as convective rainfall, are events that occur in few hours or minutes, with a higher degree of intensity for rainfall durations of less than 1 h (Llasat 2001; Marchi et al. 2010).

The problem of temporal validity can be understood in terms of the existing timing differences between two or more time-series. It is closely linked to event detection in real-time or near real-time with the aid of social network messages. For example, we know that social network messages can occur at the beginning, middle or end of the targeted events, but *how can the appropriate time for using the social network messages in a real-time setting be detected?* This question can be answered through a temporal validity of data using a lag-time of the time-series. It allowed us to take note of the existing correlation between the data in different phases of the rainfall events, and thus could be useful for detecting, monitoring or disaster management in real-time. Hence, the time-series analysis must include the lag-time variable to ensure it is more effective. Thus, it can serve to support the extraction of key information and improve the use of social network messages in practical solutions, such as monitoring and early warning systems.

With regard to the lag-time and its potential use, the following research question can be raised:

- Is there a temporal relationship between the rainfall gauge time-series and the rainfall-related social network stream?

Our answer to this question is based on the assumption that the temporal relationship between authoritative data and event-related social network messages can be explained by duration, frequency and the lag-time of the time-series. For this reason, the cross-correlation between rainfall gauge data and rainfall-related tweets is analyzed by means of the temporal units and their lags, rather than only the duration and frequency of the time-series. In addition, an attempt is made to describe the lag-time as a means of showing how it can be applied (together with the existing approaches) to detect events and support the retrieval of appropriate information in rainfall events.

The aim of this paper is thus to adopt a temporal approach to assess the correlation between rainfall data and social network messages at different lag-times. This assessment is based on the spatio-temporal features of a single time-series obtained from the social network messages and rainfall measurements. In light of this, this work seeks to provide a novel approach by means of lag-times rather than only the duration and frequency of the time-series.

This paper is structured as follows. Section 2 introduces the problem statement and hypothesis of this research. Section 3 gives a detailed account of the case study used for evaluating the approach, while Sect. 4 examines the methodology employed. Following this, Sect. 5 shows the results. Finally, Sect. 6 discusses the results obtained, draws some conclusions and makes recommendations for future work.

2 Problem Statement and Hypothesis

Figures 1 and 2 depict, to a limited extent, the increase in the rainfall and frequency of rainfall-related tweets in two periods of January 2016, in São Paulo, Brazil, during the peak of the rainy season. As can be seen, there is a similarity (based on the peaks) between the rainfall time-series and rainfall-related tweets time-series. However, there is not an exact correspondence between both the time-series. In ordinary circumstances, social network users often report messages related to rainfall episodes. That is, they post texts, photos and videos, for example of gray clouds, drizzle, rainfall, and floods, and on rare occasions, give a forecast of rain. These posts appear at different times, i.e. the temporal scale and frequency of rainfall-related social network messages are not known. They may occur before or after the rain, whilst other occur during a period of rainfall (Table 1). Moreover, the social network users can forward older information instead of new information (Earle et al. 2012).

On the other hand, there is a well-known temporal scale for rainfall data. The information usually, comes from specialist equipment (e.g. rainfall gauges) installed in urban areas. In Brazil, for instance, a widely-used temporal scale in urban areas is 10 min (Figs. 1 and 2). This means that it is a challenge to explore the correlation between the rainfall time-series and rainfall-related social network messages.

On the basis of this problem statement, the hypothesis raised here is that *there is a lag-time between rainfall and rainfall-related social network messages*. In formal terms, our hypothesis can be expressed as follows:

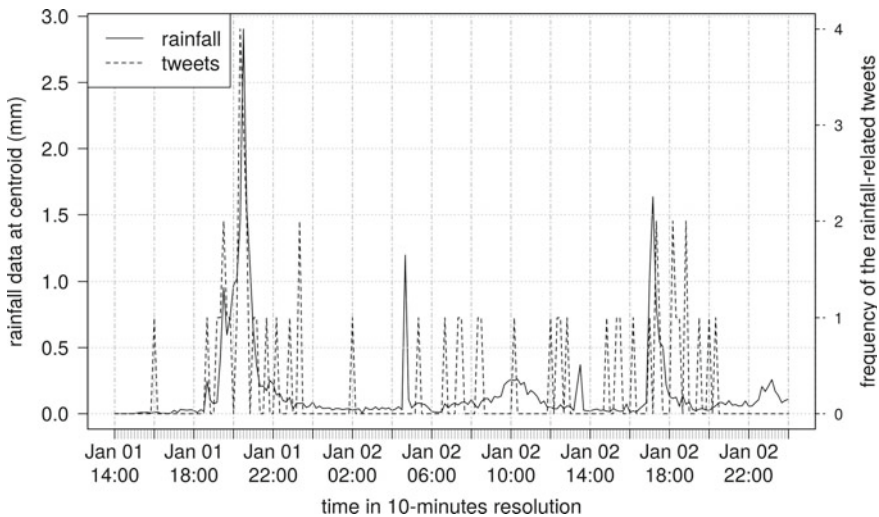


Fig. 1 Increasing rainfall and frequency of rainfall-related tweets from January 1st 14:00 BRST to 3rd 00:00 BRST, São Paulo, Brazil (with 10-min temporal resolution)

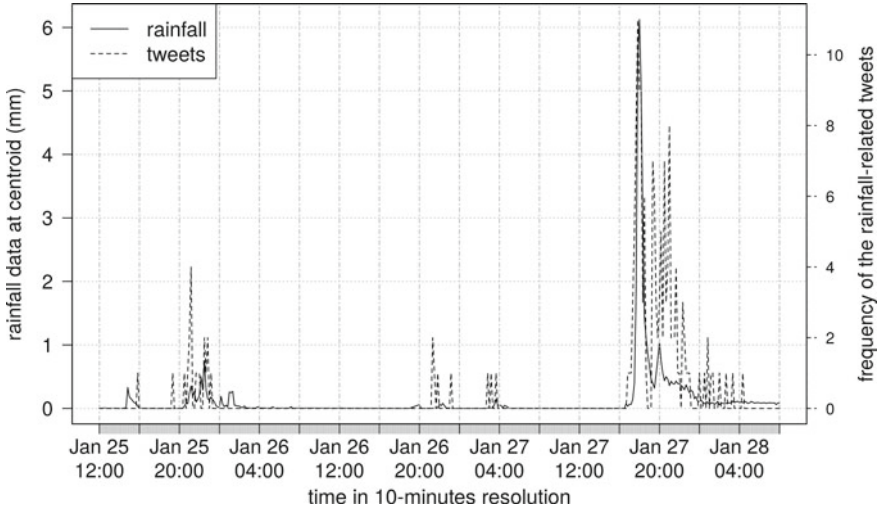


Fig. 2 Increasing rainfall and frequency of rainfall-related tweets from January 25th 12:00 BRST to 28th 08:00 BRST, São Paulo, Brazil (with 10-min temporal resolution)

Table 1 Examples of rainfall-related tweets classified per time (before, after or during the rainfall)

Date/Time	Rainfall-related tweets	Translation	Time
2016-01-12 06:11:39	“Escuro e quente, a tarde a chuva vem, não se enganem, é verão!!!! (...)”	“Dark and warm, in the afternoon there will be rain; make no mistake, i, it’s summer !!!! (...)”	Before
2016-01-28 13:47:57	“A chuva de ontem @ Em MOEMA https://t.co/3kggiLckGZ ”	“Yesterday’s rain @ In MOEMA city https://t.co/3kggiLckGZ ”	After
2016-01-27 17:37:39	“Chuva chuva e mais chuva... https://t.co/vzC9w01qQ8 ”	“Rain rain and more rain... https://t.co/vzC9w01qQ8 ”	During

Hypothesis. Let $Q_t = \{q_1, q_2, \dots, q_i, \dots, q_m\}$ and $P_t = \{p_1, p_2, \dots, p_j, \dots, p_n\}$ be defined by two time-series, the frequency of rainfall-related social network messages and rainfall data, respectively. The elements q_i and p_j are indexed in time T . For all $t \in T$ there is a constant $k \in \mathbb{Z}$ that makes the relationship function ρ at time t other than zero.

$$\forall_{t \in T}, \exists_{k \in \mathbb{Z}} : \rho(P_t, Q_{t+k}) \neq 0 \quad (1)$$

where P_t and Q_{t+k} are observations (values) of the variables at time t and $t + k$, and k is the lagged k -periods. Here, the function ρ is a measure of correlation between both time series. We have checked ρ through the cross-correlation of time-series, as will be seen further on.

Furthermore, an attempt is made answer the two following questions: (Q1) *What happens when the k constant is other than zero?* (Q2) *How many lagged k -periods are needed to provide a suitable description of the social networks messages about rainfall gauge data?* These questions are necessary to show how our temporal approach can be applied, together with the existing approaches, to detect patterns and support the extraction of key information about rainfall events.

3 Case Study

3.1 Context of the Study—Rainfall in Brazil

The case study was carried out in São Paulo, Brazil, where heavy rainfall often affects the city's infrastructure and citizens. Figure 3 shows the scenario with the spatial distribution of active rainfall gauges and rainfall-related tweets in January 2016. It can be seen that the rainfall-related tweets are closer to the regions with rainfall gauges data than the regions without rainfall gauges data.

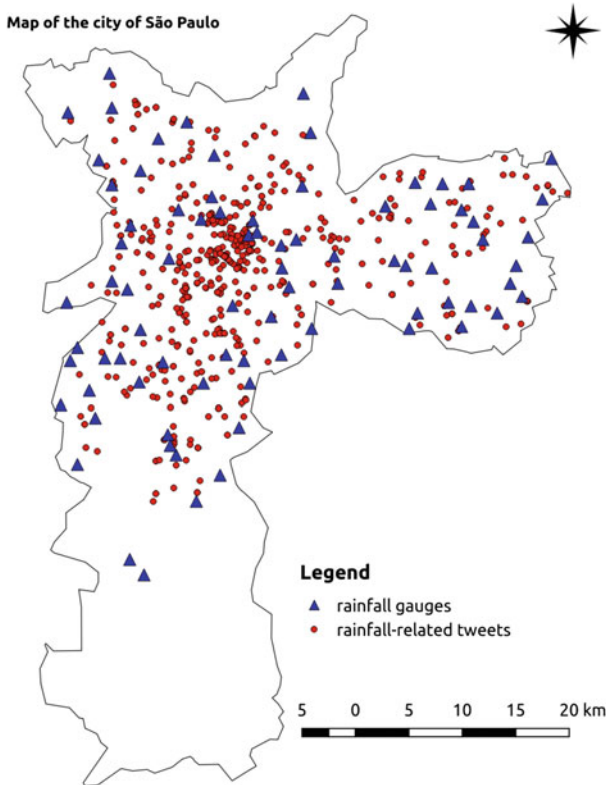


Fig. 3 Map of the distribution of active rainfall gauges and rainfall-related tweets in January 2016, in São Paulo, Brazil

3.2 Description of the Dataset and its Time-Series

Social network messages

Our social network dataset contains 243,333 georeferenced tweets retrieved by Twitter Streaming API from January 1st to January 30th 2016. All the tweets selected were geotagged within the administrative borders¹ of the city of São Paulo.

At first, the tweets collected via Twitter Streaming API were filtered with the aid of keywords and synonyms (including stem words and wildcard). The keywords in Brazilian-Portuguese were “chuv*” (chuva, chuveiro, chuvarada, etc.), “garoa*” (garoando, etc.), “temp*” (temporal, tempestade, tempo ruim, etc.), “alag*” (alagamento, alagado, etc.), and “inund*” (inundação, inundado, etc.). These keywords were chosen and extended from a list of previous studies (Assis et al. 2015). After this, we carried out a manual search for the subset of rainfall-related tweets to remove the false-positives (2,916 tweets were removed). In this context, the false-positive are tweets that contains one or more keywords, but they are not inside the context, i.e., rainfall-unrelated tweets. For example, there are several rainfall-unrelated tweets containing the keyword “garoa” because the city of São Paulo is known as the land of drizzle. Furthermore, we removed all retweets messages (228 retweets), i.e., messages concerning of another Twitter user and that contains the text “RT”. With regard to messages with Twitter Emoji (e.g. 🌧️) and grammar mistakes (e.g. “Xuva” instead of “CHuva”), they were classified along with unrelated messages (239,569 tweets were classified as unrelated). Table 2 shows some examples of rainfall-related tweets. In our case study only 0.25% (620 tweets) were marked by us as rainfall-related tweets.

Table 2 Examples of rainfall-related tweets that have been classified manually

Date/Time	Rainfall-related tweets	Translation
2016-01-09 01:19:57	“CHUVAAAAAAAAA”	“heavy rain”
2016-01-09 15:20:40	“Que chuvinha mais gostosa”	“what wonderful rain”
2016-01-11 07:43:41	“ta garoando aqui”	“it’s drizzling here”
2016-01-27 15:58:40	“Preparando o barco, chove muito em sampa.”	“Get the boats ready, it’s pouring with rain in São Paulo”
2016-01-27 18:56:30	“Tempestade! @ Avenida Ipanema Sorocaba https://t.co/HkvrBIBgeW ”	“A Storm! @ Ipanema Sorocaba Avenue https://t.co/HkvrBIBgeW ”

¹We took the geometry of the city from the Global Administrative Areas (GADM).

Rainfall data

The rainfall data were obtained from the National Center for Monitoring and Early Warning of Natural Disasters (CEMADEN)² through a REST API. There were 81 active rainfall gauges in São Paulo in the data collection period. The rainfall measurements were provided in linear depth (millimeters) and with two sizes of temporal window: 10 min when it was raining and 60 min otherwise.

We only used reliable rainfall gauges and removed those that had no data throughout the month and those which registered negative values. In total, we removed 21 rainfall gauges.

4 Methodology

As can be seen from Fig. 4, we designed a temporal approach for analyzing the cross-correlation between rainfall gauge data and rainfall-related Twitter messages. We evaluated this approach through the case study that was undertaken.

Our temporal approach comprised three steps, which are as follows:

- Step 1** to gather sets of rainfall-related tweets and authoritative rainfall data;
- Step 2** to generate a time-series of the rainfall data and rainfall-related tweets; and
- Step 3** to analyse the cross-correlation between both the rainfall time-series and rainfall-related tweets time-series.

In Step 1, we selected a set of tweets that were published with the same spatio-temporal rainfall data obtained from the rainfall gauges and that contained content related to rain events. These filtered tweets are called rainfall-related tweets. After this, we generated two time-series, i.e., one for rainfall gauges data and another for rainfall-related tweets (Step 2). Both time-series were generated with time-scales of 10, 20 and 30 min. The rainfall-related tweets time-series corresponds the frequency of rainfall-related tweets during the period of analysis, whereas the rainfall time-series are rainfall gauges data interpolated at the centroid using the Inverse Distance Weighting method. Finally, we calculated the correlation between them (Step 3) for each time-scale (10, 20 and 30 min). Step 1 involves manual tasks, whereas Steps 2 and 3 are both automatic. The task types and process flow are drawn up in the Business Process Model and Notation (BPMN) to allow a visualization of the whole process used in our methodology.

²CEMADEN website is available at www.cemaden.gov.br.

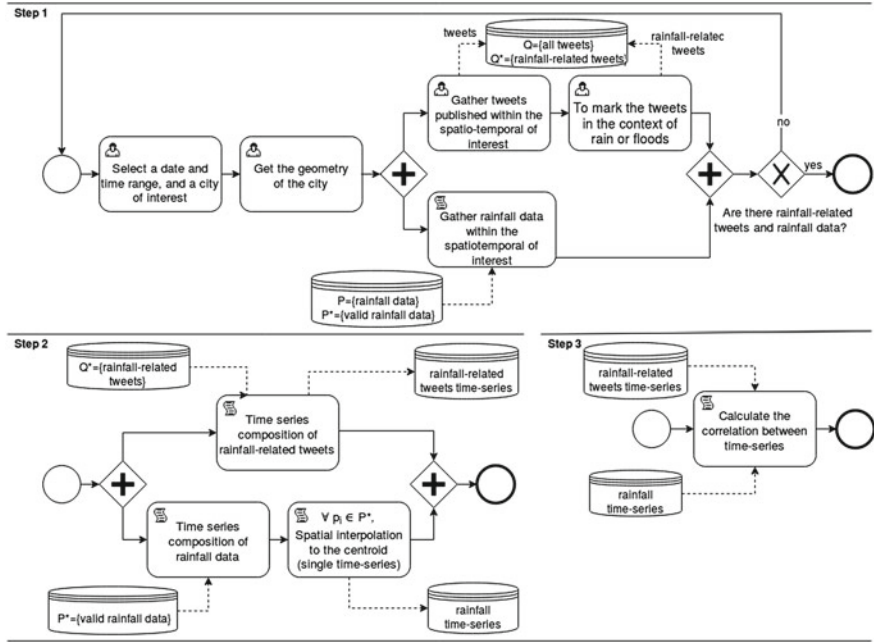


Fig. 4 Methodological approach for designing the temporal approach

4.1 Methods

The next sections describe the methods related to spatial interpolation of a rainfall time-series (Step 2), cross-correlation between rainfall and rainfall-related tweet time-series (Step 3) and how they are applied to our case study.

4.1.1 Inverse Distance Weighting

Inverse Distance Weighting (IDW) is a technique that is commonly used for estimating the values of rainfall measurements at the centroid of a catchment by calculating the weighted averages of the available rainfall gauges (Bartier and Keller 1996; Ahrens 2006; Yang et al. 2015; Mair and Fares 2011). The rainfall gauges nearest to the centroid will have a greater weight for calculating the average rainfall. The rainfall measurements at the centroid can be calculated as follows:

$$p_c^t = \frac{\sum_{i=1}^N (\frac{1}{d_i^r} p_c^t)}{\sum_{i=1}^N (\frac{1}{d_i^r})} \text{ if } d_i \neq 0, \text{ for all } i \quad (2)$$

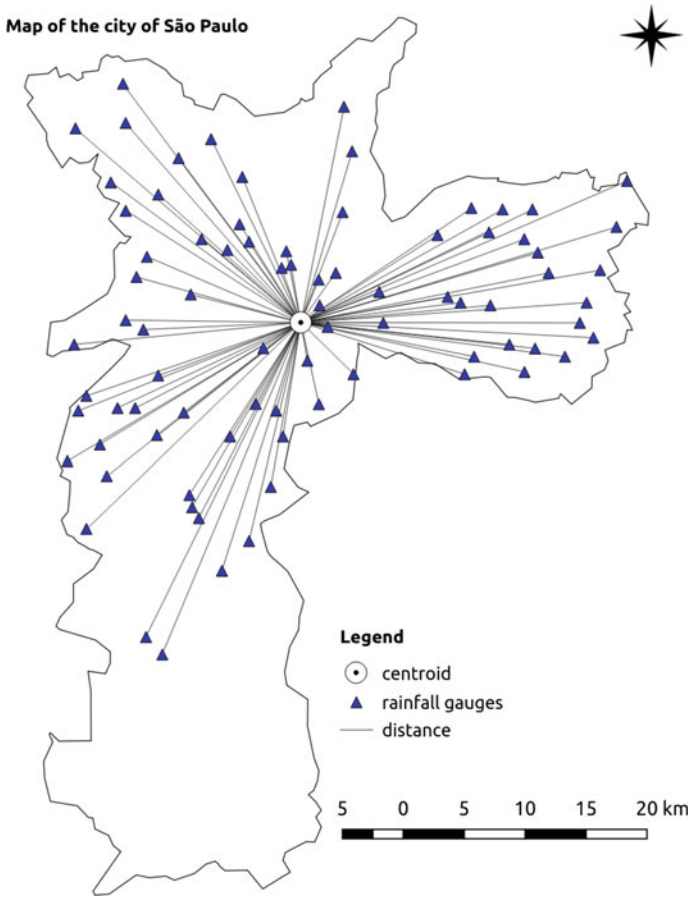


Fig. 5 Spatial interpolation calculated from rainfall for the time period 1st–30th January 2016

where p_c^t is a rainfall at the centroid, p_i^t is a rainfall in each rainfall gauge, d_i is the distance from the rainfall gauge i to the centroid and r is the power parameter. In our case, we considered, for the sake of simplification, that the urban area of the city of São Paulo forms a single catchment. Figure 5 depicts the spatial interpolation that is calculated from rainfall data using the parameter $r = 2$ for the entire period of the case study.

4.1.2 Cross-Correlation

Let q_t and p_t be two stationary processes. The correlation is any statistical relationship, whether causal or otherwise, between two random variables or two sets of data. The cross-correlation is already a function that estimates the correlation between q_t

and p_t at pairs of time points (Bacchi and Kottegoda 1995). The cross-correlation at lag k is given as follows:

$$\rho_{qp}(k) = \frac{\sum_{t=1}^N (q_{t+k} - \mu_q)(p_t - \mu_p)}{\sqrt{\sum_{t=1}^N (q_t - \mu_q)^2 \sum_{t=1}^N (p_t - \mu_p)^2}} \quad (3)$$

where μ_q and μ_p are the expected values of q and p , respectively, and k is a lag of the variable q . The cross-correlation is a method to check if the two series have randomness. This randomness is ascertained by computing autocorrelations for data values at varying lag-times. If they are random, these correlations should be near zero for any and all lag separations. Otherwise, one or more of the correlations will be significantly non-zero. With k negative, they are predictors of p_t , it is sometimes said that q leads to p . When one or more q_{t+k} , with k positive, are predictors of p_t , it is sometimes said that q lags p .

5 Results

5.1 10-min Temporal Resolution

As can be seen from Figs. 1 and 2 (see Sect. 1), the rainfall gauge time-series and the rainfall-related tweets are not synchronized, i.e., there is not an exact correspondence between both the time-series. This evidence appears throughout the case study. Another piece of evidence is that the peaks of both the time-series sometimes shift in time, i.e., the peak of the rainfall-related tweets time-series appears before the peak of the rainfall gauge time-series and vice versa. For example, from 20:00 BRST to 21:00 BRST on 1st January the frequency of rainfall-related tweets came before the rainfall (Fig. 1), but from 17:00 BRST to 18:00 BRST on 2nd January the frequency of rainfall-related tweets came after the rainfall (Fig. 2). A similar kind of behavior can be observed in other periods in January 2016.

Figure 6 depicts the cross-correlation for the period 20:00 BRST–21:00 BRST on 1st January 2016, whereas Table 3 summarizes the cross-correlation values for the same period. The blue dotted line (Fig. 6) represents the confidence interval of 95%. Although the significance correlation is in the range of -7 to 7 lag-times, the greatest correlation is -1 lag-time ($k = -1$). This means most rainfall-related tweets come 10 min before the rainfall phenomenon. The most likely explanation is that rainfall-related tweets can predict future rainfall or detect rainfall activity in real time or near real-time, i.e., they can occur before the rain (see Table 1). This means that this setting can be used for forecasting and monitoring rainfall and thus provide better support for decision-making.

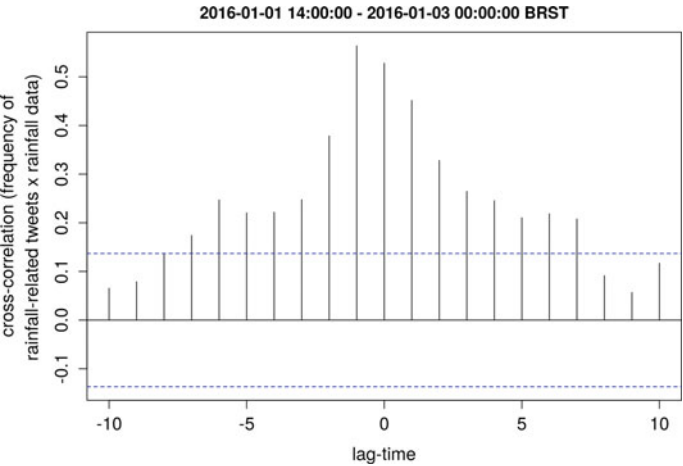


Fig. 6 Cross-correlation and visualization between rainfall data time-series and rainfall-related tweets with a 10-min time-scale from January 1st 14:00 BRST to January 3rd 00:00 BRST 2016

Table 3 Cross-correlation values between rainfall data time-series and rainfall-related tweets with a 10-min time-scale from January 1st 14:00 BRST to January 3rd 00:00 BRST 2016

Lag-time	Cross-correlation	Lag-time	Cross-correlation
−10	0.065	1	0.451
−9	0.078	2	0.328
−8	0.135	3	0.264
−7	0.174	4	0.245
−6	0.246	5	0.210
−5	0.220	6	0.218
−4	0.221	7	0.207
−3	0.247	8	0.091
−2	0.378	9	0.056
−1	0.563	10	0.116
0	0.528		

A similar result is achieved in the period 25th–28th January (Fig. 7), i.e., the greatest correlation is -1 lag-time ($k = -1$). In this case, the confidence interval (Fig. 7) and cross-correlation values (Table 4) are highest.

In fact, the results show that the highest correlation is negative, but the positive correlation of 1 lag-time ($k = 1$) could be of value for monitoring activities. For example, the decision-maker might be monitoring what the social network users are saying during the rain phenomenon, as well as assessing the consequences and effects of the rain (e.g. situational updates about an ongoing event). This setting is very useful for monitoring extreme events, such as heavy rainfall episodes. Hence, the lag-time can be used to define a threshold time for monitoring the affected area. For

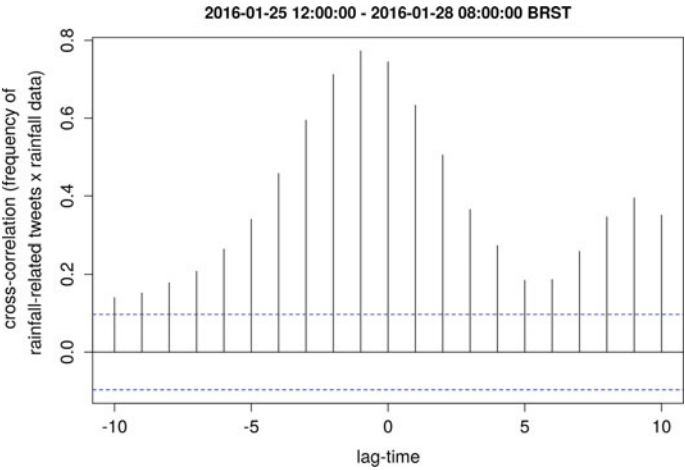


Fig. 7 Cross-correlation and visualization between rainfall data time-series and rainfall-related tweets with a 10-min time-scale resolution from January 25th 12:00 BRST to January 28th 08:00 BRST 2016

Table 4 Cross-correlation values between rainfall data time-series and rainfall-related tweets with a 10-min time-scale from January 25th 12:00 BRST to January 28th 08:00 BRST 2016

Lag-time	Cross-correlation	Lag-time	Cross-correlation
-10	0.140	1	0.633
-9	0.151	2	0.505
-8	0.178	3	0.365
-7	0.207	4	0.273
-6	0.264	5	0.184
-5	0.340	6	0.186
-4	0.457	7	0.259
-3	0.594	8	0.345
-2	0.712	9	0.394
-1	0.772	10	0.350
0	0.744		

example, a maximum size of a temporal window of 20 min can be defined for getting information after the event has been detected, i.e. $0 < k < 2$.

Another possible setting is when the lag-time is zero ($k = 0$). In this case, rainfall data and rainfall-related tweets could be analyzed for monitoring activities, either together or separately. The first alternative is ideal to improve the analysis in real-time. The second alternative could be applied when the rainfall data are insufficient. In this case, it is necessary to assess the credibility of the social network messages.

Figure 8 shows the results for the entire period (from January 1st to 30th January 2016). It can be seen that the cross-correlation is greater than 0.5 in the lag-time from

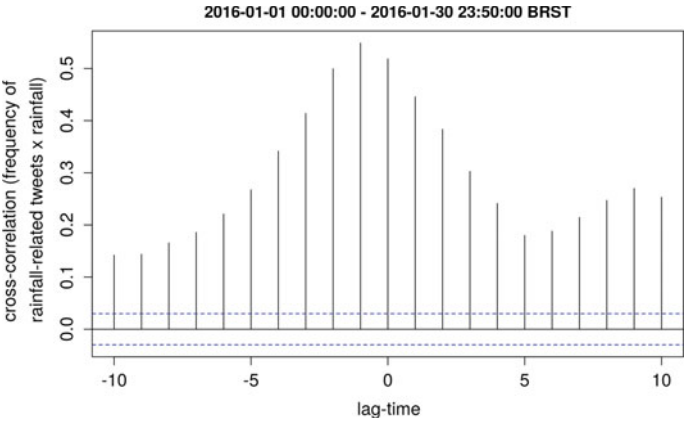


Fig. 8 Cross-correlation and visualization between rainfall data time-series and rainfall-related tweets with a 10-min time-scale in January 2016

Table 5 Cross-correlation values between rainfall data time-series and rainfall-related tweets with a 10-min temporal resolution in January 2016

Lag-time	Cross-correlation	Lag-time	Cross-correlation
−10	0.141	1	0.445
−9	0.143	2	0.383
−8	0.165	3	0.302
−7	0.185	4	0.241
−6	0.220	5	0.179
−5	0.267	6	0.188
−4	0.341	7	0.214
−3	0.413	8	0.247
−2	0.499	9	0.270
−1	0.548	10	0.253
0	0.518		

−1 to 0 (Table 5). In brief, the −1 lag-time setting (−10 min) is enough to be used in forecasting rain, whereas the 0 lag-time (0 min) can be used in a real-time setting.

5.2 20 and 30-min Temporal Resolution

Similar results were obtained for the entire period of the case study with 20 and 30-min time-scales (Figs. 9 and 10). Nevertheless these results reveals additional findings. They show clearly the temporal validity problem for large time-scales. As

can be seen from Tables 5, 6 and 7, the cross-correlation in the 0 lag-time increases as the time-scale increases too. In this way, the results indicate that a large time-scale affect the temporal cross-correlation, by making the rainfall data time-series converge with the rainfall-related time-series in an “exact correspondence”—i.e., it shows a false view of the scenario for decision-making. Figure 11 compares the cross-correlation values for the time-scales of 10, 20 and 30 min. In additional, it shows that large time-scales are not suitable for detecting and monitoring rainfall-events in real-time.

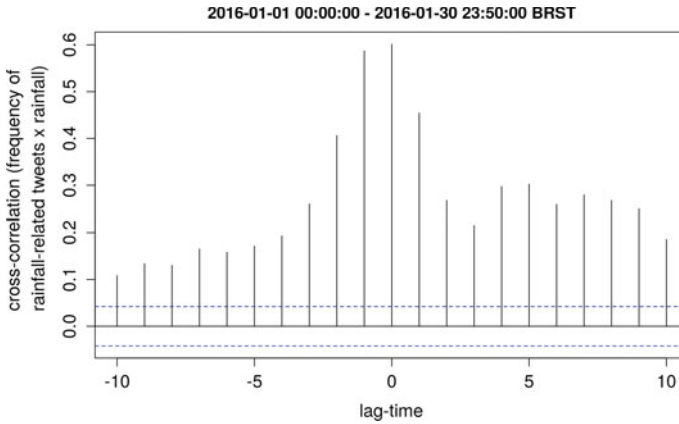


Fig. 9 Cross-correlation and visualization between rainfall data time-series and rainfall-related tweets with a 20-min time-scale in January 2016

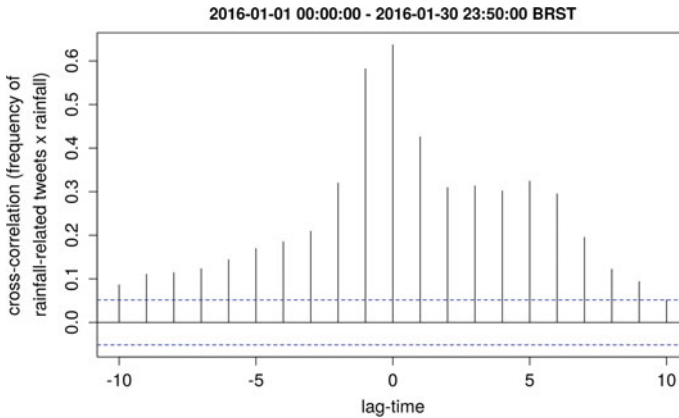


Fig. 10 Cross-correlation and visualization between rainfall data time-series and rainfall-related tweets with a 30-min time-scale in January 2016

Table 6 Cross-correlation values between rainfall data time-series and rainfall-related tweets with a 20-min time-scale in January 2016

Lag-time	Cross-correlation	Lag-time	Cross-correlation
−10	0.108	1	0.454
−9	0.133	2	0.267
−8	0.130	3	0.215
−7	0.165	4	0.297
−6	0.158	5	0.302
−5	0.171	6	0.259
−4	0.193	7	0.279
−3	0.260	8	0.267
−2	0.406	9	0.249
−1	0.586	10	0.185
0	0.601		

Table 7 Cross-correlation values between rainfall data time-series and rainfall-related tweets with a 30-min time-scale in January 2016

Lag-time	Cross-correlation	Lag-time	Cross-correlation
−10	0.085	1	0.425
−9	0.110	2	0.309
−8	0.114	3	0.313
−7	0.123	4	0.301
−6	0.143	5	0.324
−5	0.169	6	0.294
−4	0.184	7	0.195
−3	0.209	8	0.122
−2	0.319	9	0.093
−1	0.581	10	0.051
0	0.637		

6 Discussion and Conclusion

This paper outlines a temporal approach to explore the cross-correlation between social network messages and rainfall data from a meteorological source. A case study was undertaken in São Paulo, Brazil, and we identified lag-time between a time-series containing data from rainfall and a time-series of rainfall-related tweets. The case study shows clearly the temporal validity problem, as well as the need for a cross-correlation analysis to investigate the use of social network messages in practical solutions, such as monitoring and early warning systems.

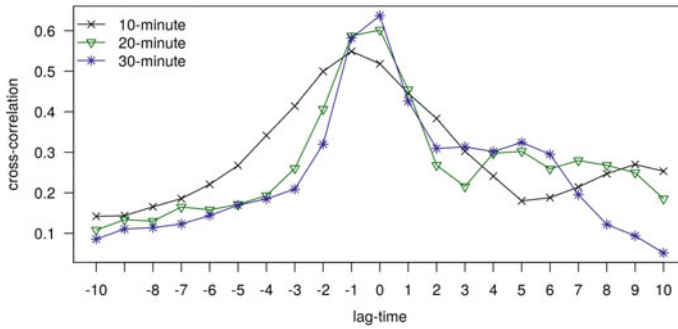


Fig. 11 Comparison of cross-correlation between rainfall data time-series and rainfall-related tweets with 10, 20 and 30-min time-scales in January 2016

The results provide evidence that the rainfall gauge time-series and the rainfall-related tweets are associated with different lag-times (Figs. 6, 7 and 8), but they are highly associated with a lag-time that ranges from -10 to $+10$ min (Tables 3, 4 and 5). Statistically, the lag-time can vary from negative to positive (see Tables 3, 4 and 5). This temporal characteristic can be hard to detect with the human eye without a temporal correlation. For example, when the analyst chooses a small time-scale, the peaks of rainfall data and rainfall-related social network activity are relatively close. Thus, it is not clear to a decision-maker whether the social network activity is or is not useful for decision-making. However, our results show that the rainfall time-series patterns can be approximated by social network messages when proper account is taken of the duration, frequency, and lag-time. In view of this, our approach can be applied to detect patterns of rainfall events in real-time using authoritative data and social network messages. In addition, the findings suggest that this approach can be useful to fit or approximate the best temporal scale between two time-series in any setting (e.g. real-time, near real-time and post-hoc analysis) since the individual data sources can be represented from a single time-series. This is very useful for monitoring activities, where the scale and frequency of the data can change over time.

Thus, the main value of this work is that we have put forward a novel approach to describe the temporal relationship between authoritative data and rainfall-related social network messages using lag-times rather than only relying on the duration and frequency of the time-series. This opens up a new way of exploring and improving the temporal models and approaches with the aim of creating functional relations to enhance hydrological modelling for monitoring rainfall in real-time. Moreover, it is possible to improve existing detection systems, such as PrioritizeSN (Assis et al. 2015), Toretter (Sakaki and Matsuo 2012), GeoCONAVI (Spinsanti and Ostermann 2013), to name just a few.

Future work should further extend this approach by incorporating other social network platforms (e.g. Instagram and Flickr) and case study scenarios (e.g. other cities and countries) to be able to obtain a generalization. Furthermore, the centroid

of the rainfall-related tweets might be explored to understand the extent to which they can be correlated with a rainfall time-series, although a spatial analysis is needed for this. Finally, methods for handling factors of uncertainty can be employed to correct the rainfall gauge data instead of removing them for spatial interpolation.

7 Data Access Statement

All data created during this research are openly available from the University of Warwick data archive at <http://wrap.warwick.ac.uk/87173>.

Acknowledgements This research was partially funded by the Engineering and Physical Sciences Research Council (EPSRC) through the Global Challenges Research Fund. The authors would like to express their thanks for the financial support provided by the Coordination for higher Education Staff Development (CAPES, Grant No. 88887.091744/2014- 01). S. C. Andrade would like to thank the Araucária Foundation of Supports Scientific and Technological Development in the State of Paraná (FAPPR) and State Secretariat of Science, Technology and Higher Education of Paraná (SETI) for their financial support. C. Restrepo-Estrada is grateful for the financial support from CAPES-PROEX.

References

- Ahrens B (2006) Distance in spatial interpolation of daily rain gauge data. *Hydrol Earth Syst Sci* 10(2):197–208
- de Albuquerque JP, Herfort B, Brenning A, Zipf A (2015) A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. *Int J Geogr Inf Sci* 29(4):667–689
- Assis LFG, Herfort B, Steiger E, Horita FEA, de Albuquerque JP (2015) Geographical prioritization of social network messages in near real-time using sensor data streams: an application to floods. In: *Proceedings of the XVI Brazilian symposium on geoinformatics*, pp 26–37
- Bacchi B, Kottegoda NT (1995) Identification and calibration of spatial correlation patterns of rainfall. *J Hydrol* 165(1):311–348
- Bartier PM, Keller C (1996) Multivariate interpolation to incorporate thematic surface data using inverse distance weighting (idw). *Comput Geosci* 22(7):795–799
- Crooks A, Croitoru A, Stefanidis A, Radzikowski J (2013) #earthquake: twitter as a distributed sensor system. *Trans GIS* 17(1):124–147
- Earle P, Bowden D, Guy M (2012) Twitter earthquake detection: earthquake monitoring in a social world. *Ann Geophys* 54(6):708–715
- Herfort B, de Albuquerque JP, Schelhorn SJ, Zipf A (2014) Exploring the geographical relations between social media and flood phenomena to improve situational awareness. *Springer International Publishing, Cham*, pp 55–71. doi:[10.1007/978-3-319-03611-3_4](https://doi.org/10.1007/978-3-319-03611-3_4)
- Imran M, Elbassuoni S, Castillo C, Diaz F, Meier P (2013) Extracting information nuggets from disaster-related messages in social media. In: *Proceedings of the ISCRAM 2013—10th international conference on information systems for crisis response and management*, pp 791–801
- Kryvasheyey Y, Chen H, Obradovich N, Moro E, Van Hentenryck P, Fowler J, Cebrian M (2016) Rapid assessment of disaster damage using social media activity. *Sci Adv* 2(3). doi:[10.1126/sciadv.1500779](https://doi.org/10.1126/sciadv.1500779)

- Llasat MC (2001) An objective classification of rainfall events on the basis of their convective features: application to rainfall intensity in the northeast of Spain. *Int J Climatol* 21(11):1385–1400
- Mair A, Fares A (2011) Comparison of rainfall interpolation methods in a mountainous region of a tropical island. *J Hydrol Eng* 16(4):371–383
- Marchi L, Borga M, Preciso E, Gaume E (2010) Characterisation of selected extreme flash floods in Europe and implications for flood risk management. *J Hydrol* 394(1–2):118–133 (flash floods: observations and analysis of hydrometeorological controls)
- Sakaki T, Matsuo Y (2012) Earthquake observation by social sensors. InTech. doi:[10.5772/29629](https://doi.org/10.5772/29629)
- Sakaki T, Okazaki M, Matsuo Y (2010) Earthquake shakes twitter users: real-time event detection by social sensors. In: *Proceedings of the 19th international conference on world wide web*, pp 851–860
- Spinsanti L, Ostermann F (2013) Automated geographic context analysis for volunteered information. *Appl Geogr* 43:36–44
- Starbird K, Muzny G, Palen L (2012) Learning from the crowd: Collaborative filtering techniques for identifying on-the-ground twitterers during mass disruptions. In: *Proceedings of the ISCRAM 2012—9th international conference on information systems for crisis response and management*, BC, Simon Fraser University
- Steiger E, de Albuquerque JP, Zipf A (2015) An advanced systematic literature review on spatiotemporal analyses of twitter data. *Trans GIS* 19(6):809–834
- Yang X, Xie X, Liu DL, Ji F, Wang L (2015) Spatial interpolation of daily rainfall data for local climate impact assessment over greater Sydney region. *Adv Meteorol* 2015:1–12

Societal Geo-innovation

Selected papers of the 20th AGILE conference on
Geographic Information Science

Bregt, A.; Sarjakoski, T.; van Lammeren, R.; Rip, F.
(Eds.)

2017, XII, 367 p. 136 illus., 105 illus. in color.,

Hardcover

ISBN: 978-3-319-56758-7