UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

RELATIONSHIP BETWEEN THE BEHAVIORS OF SOCIAL MEDIA AND PHYSICAL INFRASTRUCTURE AFTER DISRUPTION

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE

By

ENORA MAZE

Norman, Oklahoma

2017

RELATIONSHIP BETWEEN THE BEHAVIORS OF SOCIAL MEDIA AND PHYSICAL INFRASTRUCTURE AFTER DISRUPTION

A THESIS APPROVED FOR THE

SCHOOL OF INDUSTRIAL AND SYSTEMS ENGINEERING

BY

Dr. Kash Barker, Chair

Dr. Charles Nicholson

Dr. Ziho Kang

Acknowledgments

Firstly, I would like to thank my advisor Dr. Kash Barker, for his help along this project, his constant good mood and his precious touristic tips.

I am also deeply grateful to my family for allowing me to live such a blissful experience, and for their unconditional support. None of this would have been possible without the help of my friends Alison Jalanti Cuchet, Jérémy Pfeifer, Cyril Beyney, Olivia Perret, Ange Umwali, Rafeef Al-Sammarraie, Sunayana Samantaray and many others who made this experience unforgettable.

I would also thank Dr. Charles Nicholson and Dr. Ziho Kang for accepting to be part of my committee. Finally, I would acknowledge Kirsten Perry for her precious work on Twitter analysis.

Contents

List of Tables

List of Figures

Abstract

Social media have developed quickly over the years, on a worldwide scale. It has become a major tool for expressing ideas, sharing political opinions, publicity and market trending, assistance, etc., used on a daily basis by millions of people, in several languages, and continuously expanding. This evolution has caught the attention of researchers, as much more data became accessible. One research area concerns the study of social media behavior during natural disasters. These studies try to determine whether social media are social sensors, but only a few focuses on the physical environment. Here, the main objective is to establish whether Twitter is a sensor of the physical environment during a natural disaster.

In order to understand the relationship between Twitter and the physical environment, a data set of tweets is compared to a measurable disruption caused by the natural hazard. The tweets need to be relevant to the disruption, and so are filtered using specific keywords. Then, they are decomposed into a time series, and compared with a time series of the measurable disruption with a cross-correlation function.

Two examples of disruption are studied here, both during Hurricane Sandy in 2012. The first one compares the behavior of Twitter with the number of power outages, and the second one with the water peak elevations. Both examples do not yield to conclusive results, as no significant correlation is found. However, it doesn't mean that a correlation does not exist at all. The analysis is strongly dependent on the quality of the data set, and unfortunately some values are missing on important time periods on the Twitter data set. Also, the water peak elevations data set do not contain many points, and they are not taken at a regular time interval, which may have biased the analysis.

Chapter 1

Introduction

1.1. Purpose of the study

In a world where communication is faster than ever, social media have become a new tool of expression. Used on a daily basis, they allow people to express opinions, to share ideas and information. This quick development encourages new ways of use, like national agencies who provide real time information on gas station availability during Hurricane Sandy [4], to help organize protests during the Arab Spring [27], or to ensure of your safety to your relatives with Facebook "Safety check". Many researchers around the world have grown an interest in social media, as it makes so much information available. Most of the studies are linked to sentiment analysis, and trying to determine the mood of a population for example, for national elections in Sweden in 2010, or in France and in the United States in 2012, tweets related to the elections were analyzed to detect which sentiments were predominant before and after the elections [10, 16]. Social media studies have often been conducted to determine whether they can be considered as sensors of our society, being during natural disasters [2, 4] or political events [7, 10, 16]. However, only a few concerns the relationship between social media and the physical environment.

1.2. Social Media

The growth of social media has been fast, especially in the United States, as in January 2015 the part of the total active accounts on social media represented 58 percent of the North American population [1]. While Facebook remains the most popular platform counting 42 percent of the total social activity in the United States, Twitter arrives second with 19 percent. Established in 2006, Twitter is a social platform, which allows users to send in real time, public or private messages up to 140 characters, and has developed all around the world. Indeed, Twitter "has become the pulse of a planet-wide news organism, hosting the dialogue about everything from the Arab Spring to celebrity deaths" [24]. This system is very appreciated by researchers who can analyze real time reaction, perform sentiment analysis, etc. These studies are also possible because of all the ways Twitter made available to collect data, such as the streaming API for example.

1.3. Problem statement

Previous studies have been conducted on social media during natural disasters. However, most of these analyzed the event detection [6], the reaction of people thanks to sentiment analysis [2, 16**?**], or try to determine a more efficient way to select tweets relevant to the disasters [2].

Indeed, most of the work related to Twitter, in general and also during natural disasters, is based on sentiment analysis, and attempts to figure out whether Twitter is a sensor of the society. That's why this study is quite new, as it tries to understand the relationship between Twitter and the physical environment during a natural disaster. Determining the nature of this relationship could help in predictive analytics, and develop methods on forecasting community performance from infrastructure

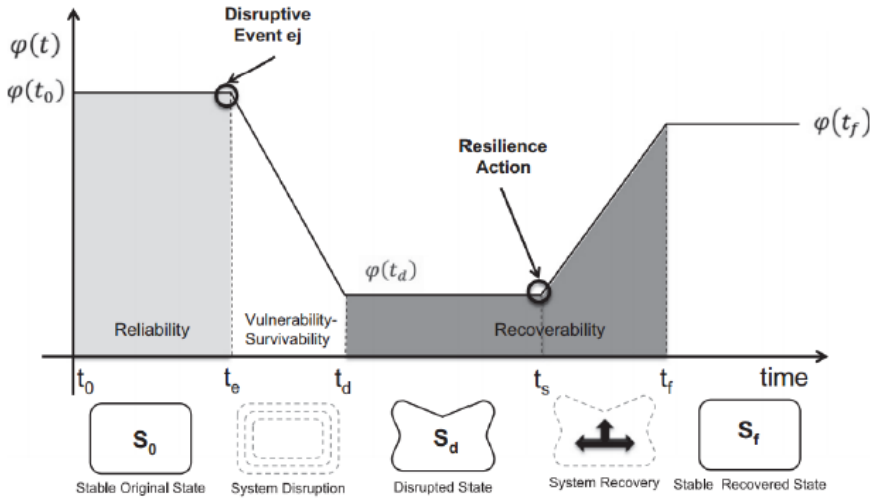

Figure 1.1: Behavior of the service function $\varphi(t)$ across state transitions.

performance. All this work could help improve situational awareness, which is defined by ESRI as a "human mental process that can be enhanced using technology to access, analyze and present information to have a greater understanding of existing conditions and how they will change over time." The research question for this project is *" Does Twitter mimic the physical environment during a disaster ? "*

In order to answer this question, the idea is to compare the behavior of Twitter with a measurable disruption caused by a natural hazard. This study is focused on the infrastructure performance from the disruption event, and aim to compare this performance with the frequency of tweets related to it on a common time period.

A data set of tweets is examined, and then refined thanks to specific keywords to ensure its relevance to the natural disaster and the physical disruption. The idea is then to compare it with data of the physical disruption caused by the hazard, with a cross-correlation analysis.

1.4. Structure of the thesis

The first part of the thesis describes the literature review on analysis of social media during natural disasters. Then, in Chapter 3, the methodology used for the study is explained. The results and analysis are detailed in Chapter 4 with two examples, and finally, the fifth chapter concludes the thesis.

Chapter 2

Literature Review

This chapter analyzes previous works and literature on the use of social media during disasters.

2.1. Use of Social Media during Disasters

Social media have grown exponentially over the years, and whereas their first utility kept focused on expressing opinions, sharing ideas, or discussing with friends, it appears that it can also have new functions. Indeed, the number of users and the format of these media have generated a lot of data. Many researchers have been wondering if these media can be used in case of emergency. Therefore, in 2011 the U.S. Geological Survey studied Twitter as a possible earthquake detection tool [6]. In this case, they use the fact that users tend to send tweets very quickly after feeling the ground shaking. In order to detect earthquakes, they developed a short-term average, long-term average algorithm. With this method, they managed to detect 48 earthquakes, "with two false triggers in five months of data", out of the 5175 reported by the USGS catalog. This number can seem small, but in their study, they argue that most of these earthquakes were not powerful enough to be felt, or

that they stroke in deserted places.

A previous study was led in 2010, on real-time earthquake detection using tweets as sensors [21]. They select the tweets based on specific keywords, the number of words and the situation. Their algorithm also try to find the objective event, for example the epicenter of an earthquake, by applying a "probabilistic spatiotemporal model" which calculates the origin and the trajectory of the disaster location [21].

2.2. Use of Twitter as Sensor

Many studies have been conducted since the creation of Twitter in 2006 on sentiment analysis, to determine whether Twitter can be considered as a social sensor.

One analyzes human behavior on Twitter to see if humans can become the new biggest "sensor network" [26]. Humans are considered as sensors, as they make observations that are either true or false. However, the major issue is to ensure of the reliability of the human perception. Indeed, some people may attribute to themselves observations made by others. The results are nonetheless quite promising as it tends to show that the veracity of human claims can be correctly estimated through three examples on Hurricane Sandy, Hurricane Irene and on the Egyptian president resignation in February 2011.

Another study tries to determine if Twitter is a social sensor of natural disasters, with different level of sensitivity [2]. By analyzing Twitter data, they compare several elements, such as the variation of tweet frequency before, during and after the disaster, the proximity to the center of the disaster, the diversity of expressed feelings, and change in tweet frequency regarding to the social vulnerability [2]. This study is conducted on five examples with different types of disasters, like the Moore Tornado in 2013, or the Black Forest Fire in Colorado in 2013. This analysis

tends to show that Twitter is indeed a social sensor, but presents different level of sensitivity.

2.3. Geo-location

Starting in 2009, the geo-location is an optional function developed by Twitter, providing either the exact GPS coordinates, or a location chosen by the user. However, this system is not automatic as users will have to agree first when setting up their Twitter account. As a result, only a few tweets are actually geotagged. Indeed, according to Leetaru and al. [11] the number of tweets with geographic coordinates represents about 2.02 percent of the total number of tweets posted each day.

Still, some researchers try to bypass that issue to find a global idea of the location of the tweets. Indeed, by text mining the tweets directly to find location words they managed to obtain geographic information [13]. Then they used a recognition technique to ensure the locations correspond with the place, or even the street.

Getting the geographic data can be really valuable, especially in case of emergencies. In their study, MacEachren and al. [12] uses a "geovisual analytics approach to support situational awareness for crisis event", which shows that it can be very useful for assistance supplies.

However, some of the data can be biased, as users can manually put a location that differs from their current position, or because in case of emergencies, some people may wait to be in a safe place to tweet [4].

2.4. Cross-Correlation

The cross-correlation function is a method used in statistics to measure the covariance of two vectors, or also in signal processing to measure the similarity of two signals. After a look at the literature, it appears this method has never been used on Twitter analysis. However, it can be found on different fields.

In 1992, Keane and Adrian use the cross-correlation analysis on particle image velocimetry to “measure the separation of pairs of particle images between successive frames” [9]. It can also be applied in case of “double- or multiple-exposure single frame images”, where the cross-correlation analysis is conducted on two different areas of the same frame.

The cross-correlation function is also often used in time series analysis, to compare two time series and try to determine if they are linked across time, with a certain lag value. This method is commonly employed in finance, for stock price analysis, to quantify "‘the risk of a given stock portfolio [19].

Chapter 3

Methodology

This chapter focuses on the methodology used for the cross-correlation analysis on data set of tweets. The first part details the data collection, the second part, the mathematical analysis.

3.1. Data Collection

A data set of tweets related to Hurricane Sandy was collected. The tweets were selected with their hashtags, for example #HurricaneSandy, or #Sandy. This data set was then transmitted to me, thanks to Dr. Barker. It contains tweets from October $26th$ to November $5th$, 2012. The data set is also composed of the tweets' ID, the users' ID and the time stamps. As mentioned previously, Twitter made geodata available since August 2009. The geo-data consists either of a place or GPS coordinates. However, the geo-localization is not systematic since the user needs to set it up manually. According to Leetaru and al. [11], only 2.02 percent of tweets includes geographic data on a regular day. In order to retrieve the tweets' coordinates, I used the Twitter API services in Python with the *tweepy* package. The API tool allows registered application users to get information such as tweets,

user names, coordinates, etc. Each request has a limited number of information allowed. The tweets' ID are then used to retrieve the corresponding coordinates. Unfortunately, no coordinates were found in this search, for this data set. One reason could be that I wasn't allowed to get the few existing coordinates because of my basic registration status on the Twitter API service.

The first step of the analysis is to determine the relevance of the data set. For this problem, the need is to filter the data set in order to keep tweets related to both Hurricane Sandy and the physical disruption. In order to do so, a text mining technique is used to find some keywords. The analysis is made in R, with the platform RStudio. The keywords need to be relevant to the problem. Therefore, the word "hurricane" is directly chosen. Also, every tweet is limited to 140 characters, users will have a tendency to employ shorter words when tweeting. That's why the word "storm" is also picked. The other keywords have to be related to the physical disruption, and preferably short too. Tweets containing the specific keywords are then selected, using the function *grep* [6].

The data set is then refined to keep only the tweets containing the keywords. Then, these tweets are decomposed into time series of fifteen minute intervals, taken from October $26th$ to November $5th$, 2012. Unfortunately, some values are missing on this data set. Four missing time periods can be identified :

- From $15:15$ October $29th$ to $8:15$ October $30th$,
- From $19:30$ October $30th$ to $23:45$ October $30th$,
- From $15:30$ October $31st$ to $23:45$ October $31st$,
- From $15:15$ November $4th$ to $23:45$ November $4th$.

The first missing period is quite long, which probably makes it the most disturbing one. These holes in the data set can have several origins : a problem during the data

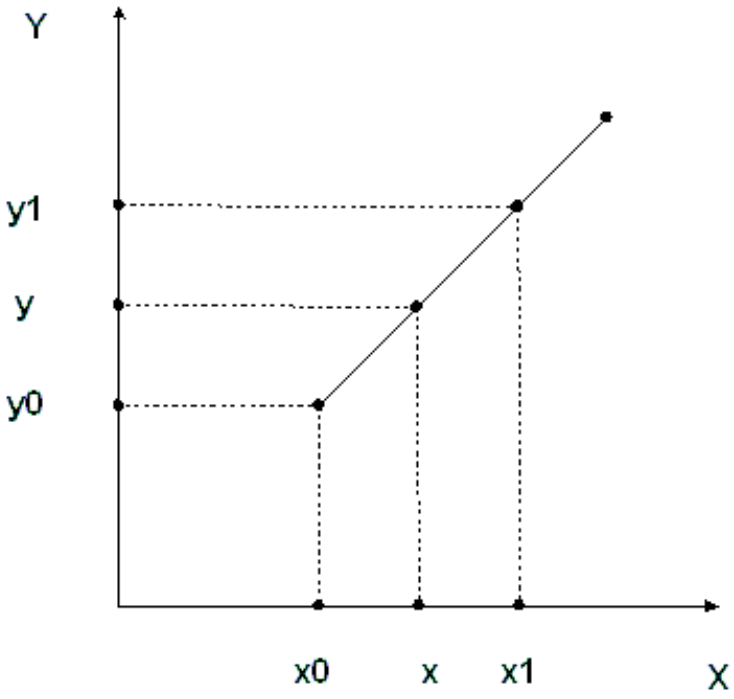

Figure 3.1: Example of a linear interpolation between two points (x_0, y_0) and (x_1, y_1).

collection due to the network, or an error on the code, or maybe human related. Indeed it is possible that people didn't use Twitter during a certain amount of time to get into a safe place, or perhaps because of their battery running low due to power outages. In order to deal with this incomplete data set, a linear interpolation is performed. Linear interpolation is a technique used to estimate the value of a continuous function, between a discrete set of known points. For two known points (x_0, y_0) and (x_1, y_1), it will determine the slope of the straight line between them.

The equation of linear interpolation is [17]:

$$\frac{y - y_0}{x - x_0} = \frac{y_1 - y_0}{x_1 - x_0}. \tag{3.1}$$

This equation can be derived, to express y in term of x:

$$y = y_0 + (y_1 - y_0)\,\frac{x - x_0}{x_1 - x_0}. \tag{3.2}$$

To perform the linear interpolation, the function *na.approx()* in R is used. A spline interpolation has also been considered. However, the results are not conclusive as it sometimes approximates a negative number of tweets for certain time period.

Every time series can be decomposed into three elements [8]: a seasonal element (daily here), a trend element and a remainder element, which can be described by the following equation:

$$y_t = S_t + T_t + E_t, \tag{3.3}$$

where y_t represents the data at period t, S_t is the daily or seasonal element at period t, T_t is the trend element at period t and E_t the remainder. A general decomposition of a time series into its three elements is shown in Figure 3.2.

We can see the general trend obtained, which is the main interest of this analysis since we want to determine if the trend of the tweets time series matches accurately the reality. However, it is impossible to remove completely randomness and the seasonal effect.

3.2. Cross-Correlation Analysis

The idea is to compare both time series to determine whether a correlation exists between them. In order to do so, the cross-correlation function is used. It "measures how closely two different observables are related to each other at the same or differing times" [22]. However, to apply this method, some work is required beforehand.

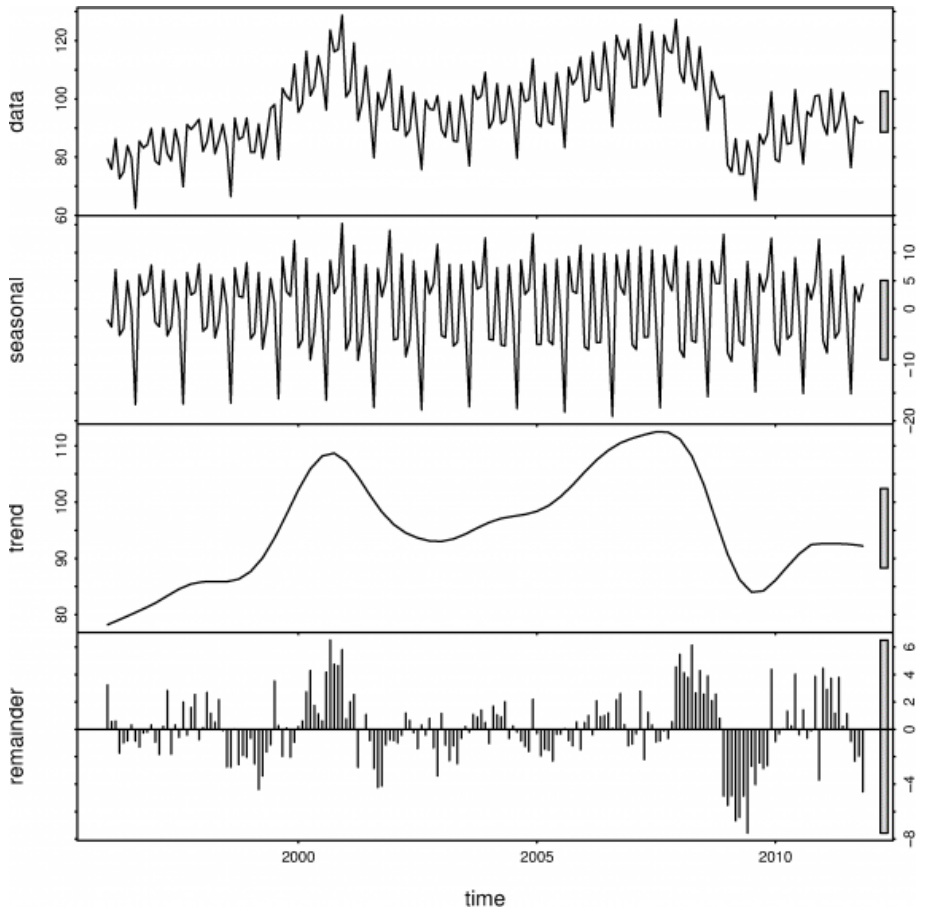

Figure 3.2: Decomposition of a time series into its seasonal, trend and remainder elements.

Indeed, the cross-correlation function needs both time series to be stationary in order to work properly. A process is said to be stationary if the mean, the variance and the auto-correlation process do not vary across time [20]. So, the first step is to stabilize the variance of our data, and ensures that the homoscedasticity condition is met. Therefore, a Box Cox transformation is performed. It also normalizes the distribution of the data. This method was formulated by two statisticians George Box and Sir David Roxbee Cox in 1964 [18]. Based on the previous work of Tukey (1957), the Box Cox transformation can be written as follow:

$$y^{'}(\lambda) = \begin{cases} \frac{y^{\lambda}-1}{\lambda}, \text{if } \lambda \neq 0 \\ \log y, \text{if } \lambda = 0, \end{cases} \tag{3.4}$$

where $y^{'}$ represents the transformed data, and λ the exponent, range from -5 to 5. To determine the best transformation i.e. the best approximation of a normal curve, all the values of λ are tested. The function *Box.Cox.lambda* in R helps finding the optimal value of λ. Then the function *Box.Cox* is used to perform the transformation with the most accurate value of λ. This method only works for positive values, however, by adding a constant the equation can be modified to also fit for negative values.

The second step is to ensure there is no auto-correlation. Therefore, each time series is differentiated once. "The first difference of a time series is the series of changes from one period to the next" [15], which can be described by the following equation [22]:

$$Y_t^{'} = Y_t - Y_{t-1}. \tag{3.5}$$

Differencing allows to remove the trends due to the accumulation of randomness [23]. If, after differencing once, the time series are not stationary then one can continue

taking the successive differences.

In order to test the stationarity of both time series, a unit root test is performed. Here, the Augmented Dickey Fuller test is used. It is a bit more powerful than the regular Dickey Fuller test, since it also takes the lagged values into account [14]. The first equation is an example of the Dickey Fuller test for an auto-regressive model, and the second one the model for the Augmented Dickey Fuller test.

$$y_t = \rho y_{t-1} + u_t, \tag{3.6}$$

where y_t is the variable, ρ a coefficient and u_t the error term.

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + ... + \delta_{p-1} \Delta y_{t-p+1} + \epsilon_t, \tag{3.7}$$

with α a constant, β a time trend coefficient and p the lag order. The hypotheses for this test are :

$$\begin{cases} H_0 : \gamma = 0 \\ H_1 : \gamma < 0 \end{cases} \tag{3.8}$$

The H_0 hypothesis being that there is a unit root, therefore the alternative hypothesis is that the time serie is stationary.

Then finally, after ensuring the stationarity of both time series, the cross-correlation function is performed. It manages to compare the resemblance of the series as a function of lag of one set relative to the other. Both time series do not need to be evenly spaced, nor to even overlap [22]. Indeed, the cross-correlation coefficient will be calculated "only for lags which shift the two sample intervals so that they overlap significantly" [22]. This function is also defined as the normalized cross-covariance function :

$$\rho_{xy}(\tau) = \frac{E\left[(X_t - \mu_X)\left(Y_{t+\tau} - \mu_Y\right)\right]}{\sigma_X \sigma_Y}, \tag{3.9}$$

with μ_X and μ_Y the means of X_t and $Y_{t+\tau}$, and σ_X, σ_Y their standard deviation.

3.3. Windowed Cross-Correlation

Another cross-correlation method has been developed in 2002, by Boker, Xu, Rottondo and King, called the windowed cross-correlation [3]. It calculates the cross-correlation function on a limited time segment, where both time series are supposed to be stationary. This process is then repeated by moving the calculation window from one value, until the whole time series is treated. This analysis gives a set of local cross-correlation function. Then a pick peaking algorithm is applied to determine the function maxima, and then to establish the series of successive delays corresponding to these maxima [3].

Let's define X and Y two time series of length N, $X = \{x_1, x_2, ..., x_N\}$ and $Y = \{y_1, y_2, ..., y_N\}$. The length of the window is n, and d is the maximum lag value for the analysis, with $d > 0$. For a first segment of length n, the serie X can be written as $X_{(a,n)} = \{x_a, x_{a+1}, x_{a+2}, ..., x_{a+n-1}\}$. Then we calculate the local cross-correlation coefficient, for every lag k between $-d$ and d, defined by :

$$r(X_{(a,n)}, Y_{(a+k,n)}) = \frac{1}{n} \sum_{i=a}^{a+n} \frac{(x_i - \overline{X}_{(a,n)})(y_{i+k} - \overline{Y}_{(a+k,n)})}{\sigma_{X_{(a,n)}} \sigma_{Y_{(a+k,n)}}}, \tag{3.10}$$

with $X_{(a,n)}$, $Y_{(a+k,n)}$ the means of the segment of both time series X and Y of length n, respectively starting by x_a and y_{a+k}, and $\sigma_{X_{(a,n)}}$ and $\sigma_{Y_{(a+k,n)}}$ their standard deviation. The result is a serie of $2d + 1$ coefficients, defining the local cross-correlation function, indexed on the central value of the segment $X_{(a,n)}$.

Chapter 4

Analysis and Results

This chapter details the analysis conducted on two cases to determine the relationship between Twitter and physical disruption during a disaster.

4.1. Examples

In order to establish if a correlation exists between the behavior of Twitter during a disaster and a physical disruption caused by the disaster, two cases related to Hurricane Sandy are studied.

As mentioned previously, Hurricane Sandy has been devastating and caused a lot of damaged in the United States, especially on the Eastern seaboard, which cost is estimated around $50 billion. The first example studies the relationship between the number of power outages due to Hurricane Sandy and the power- and electric-related tweets published at that time. For the second example, the analysis is conducted on the water peak elevations on the Eastern seaboard caused by the heavy rainfall during Hurricane Sandy. These values are then compared with the flood-related tweets.

Keyword Combination	Total Frequency of the Keywords Combination
hurricane + electric	7682
storm + electric	8010
hurricane + power	40040
storm + power	69322

Table 4.1: Frequency analysis of the keywords combination on Hurricane Sandy data set.

4.2. Case 1 : Power outages

4.2.1 Tweets Selection

For this example, keywords need to be related to power outages in general, then associated with the previous keywords picked in Chapter 3. The word "power" is directly chosen. The second keyword is "electric", as it is also related to power outages. Moreover, it ensures that the word "electricity" is also picked.

Table 4.1 records the frequency of the keywords combinations in the data set. As shown, the words "storm" and "power" are the most commonly employed, before the combination of "hurricane" and "power". These numbers are not surprising. As mentioned previously, tweets are limited to 140 characters and users seem to rather use shorter words than longer one. The same principle applies for the combination of "storm" and "electric".

The data set is then refined to keep only the tweets containing the combinations of keywords given in Table 4.1, and decompose into a time series of fifteen-minute intervals. The Figure 4.1 represents the time series of the total number of tweets, from Hurricane Sandy, in which the words "power" and "electric" appear. As mentioned in Chapter 3, several time periods are missing. Indeed, the four black lines that we observe on the graph represents the missing points.

The Figure 4.2 represents the time series of the total number of power- and electric-related tweets after linear interpolation.

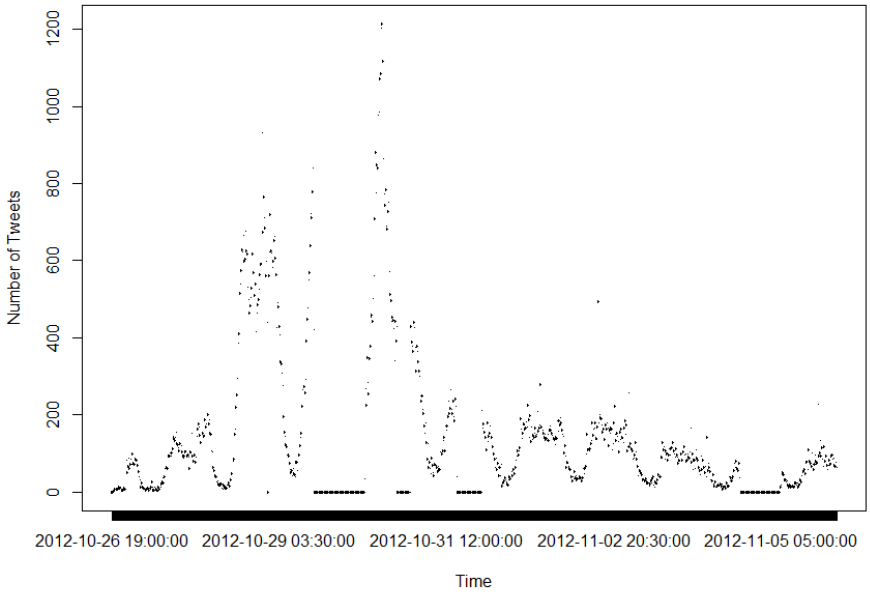

Figure 4.1: Time series of the total number of power- and electric- related tweets from Hurricane Sandy.

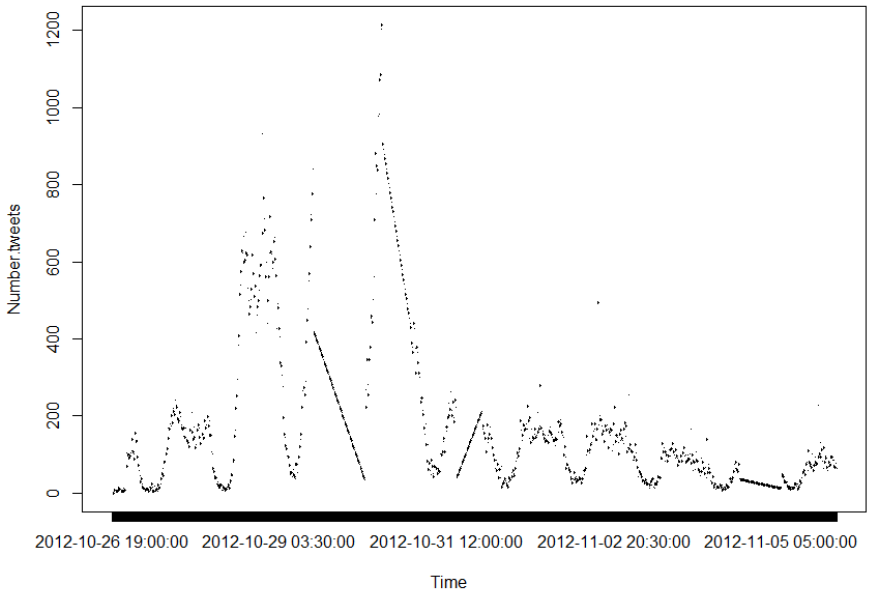

Figure 4.2: Time series of the total number of power- and electric- related tweets from Hurricane Sandy, after linear interpolation.

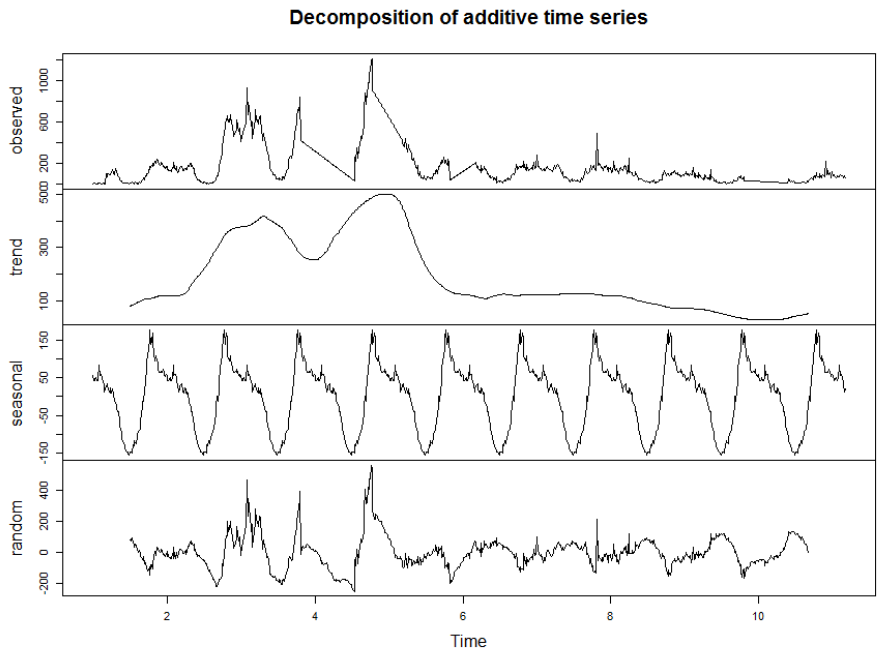

Figure 4.3: Decomposition of the tweets time series into its seasonal, trend and remainder elements.

As for Figure 4.1, the x-axis doesn't display all the date time by lack of space. On Figure 4.2, we can observe cyclic or daily elements which could imply that the number of tweets posted on this data set is depending on the time they were posted. Figure 4.3 displays the decomposition of the tweets time series into its trend, seasonal and random elements.

4.2.2 Data Collection on Power Outages

The data set of power outages consists of the number of outages for different states of the United States, touched by Hurricane Sandy. It encompasses the following states: Pennsylvania, New Jersey, Connecticut, Maryland, Delaware, Maine, Ohio, Rhode Island, Vermont, North Carolina, Massachusetts, Virginia and New Hampshire, and the city of New York. However, the detailed number of power outages by state through time is not available, so this data set has been analyzed as whole, over

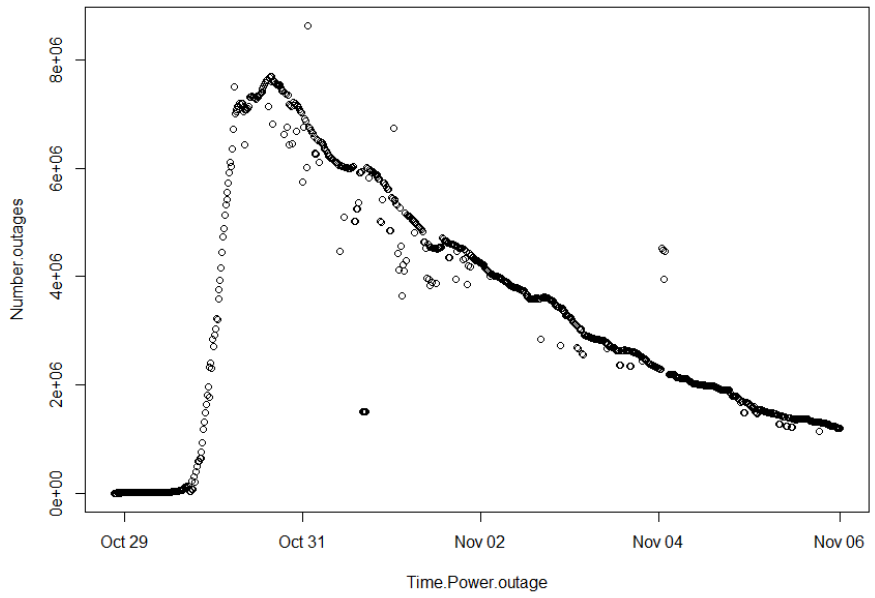

Figure 4.4: Time series of the total number of power outages taken over fifteen minute intervals.

fifteen minute intervals. It includes values from October $28th$ at $9:15$ pm to the end of November 2012. Figure 4.4 represents the raw time series of power outages. From the graph, it looks like there is a peak of power outages during Hurricane Sandy on October $30th$ around $10-11$ pm.

As shown previously, the time series can be decomposed into its trend, seasonal and random components. The Figure 4.5 represents its decomposition. By quickly comparing the trend from both time series, we can see that they differ a bit. Indeed, it looks like there is a recess on the tweet time series decomposition where there should be a peak. This might be explained by the missing values on the tweets data set.

As observed on the graph, it seems that there is a lot of noise on certain time period, especially when many tweets were posted.

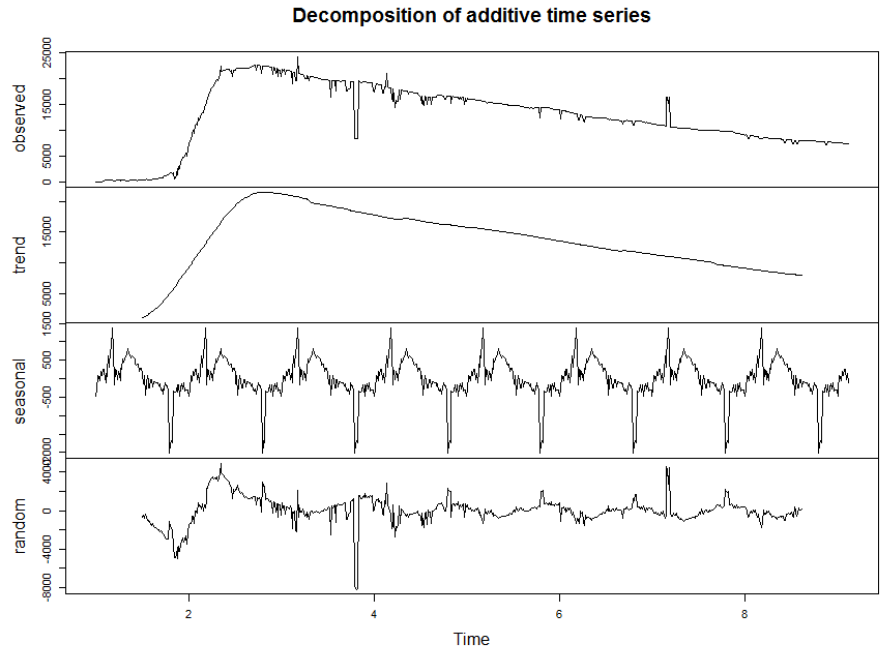

Figure 4.5: Decomposition of the power outages time series into its seasonal, trend and remainder elements.

4.2.3 One Time Differentiation

In this part, the results of the cross-correlation analysis for the power outages case, with a one time differentiation are provided.

The Figure 4.6 represents the cross-correlation coefficients between the tweet time series and the power outages time series. For this analysis, the time interval or frequency of both data sets needs to be corresponding. That's why the tweet time series interval is limited to $21:15$ on October $28th$ to midnight on November $6th$, 2012. The x-axis represents the lag time, which can be seen as the number of 15 minute intervals the tweet time series lags the power outages time series. The y-axis produces the cross-correlation number, in the range of -1 to 1. As shown in Figure 4.6, the maximal absolute value for the ACF is obtained for a lag time of $h = 6$, and is slightly below 0.15. The negative coefficient means that an under-average number of power outages is related to a above-average number of tweets 90 minutes

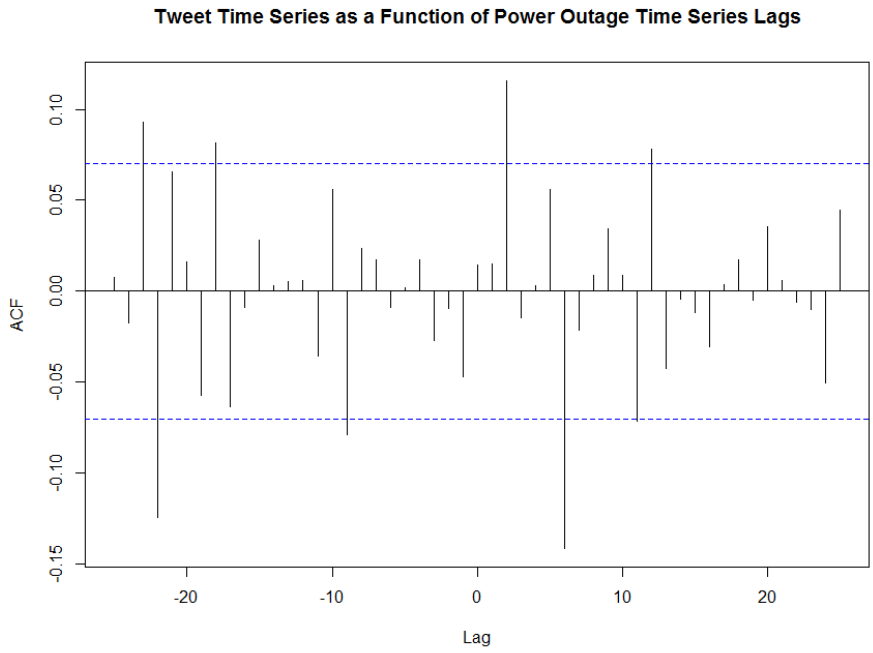

Figure 4.6: Cross-correlation graph between the conditioned time series of power outages and the time series of power or electricity related tweets from Hurricane Sandy.

later. Two other points, at $h = 2$ and $h = -22$ presents about the same absolute value for their ACF. However, these coefficients are supposed to lie between -1 and 1, therefore 0.15 or -0.15 do not seem to be really significant.

The Figure 4.7 provides nine different scatter plots of the power outages time series versus the tweet time series, given for a 0 to 9 interval lag times. In each graph, the correlation number is given on the right box. By observing these nine plots and their correlation number, it looks impossible to detect a relationship between the number of power outages and the number of tweets. Therefore, it looks like in this case, the analysis is inconclusive.

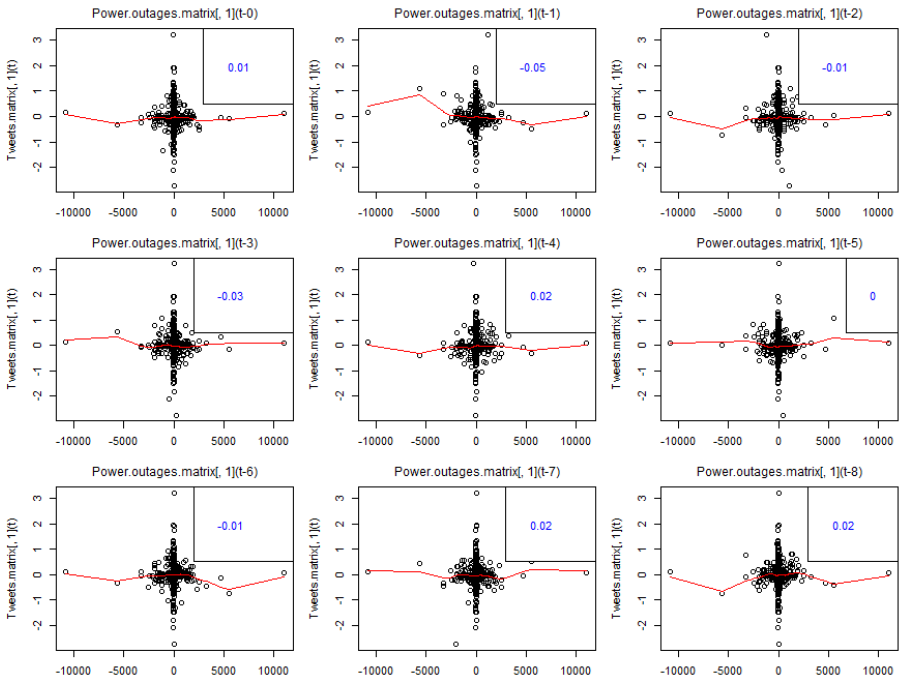

Figure 4.7: Scatterplots of number of power outages versus number of tweets, where the tweet time series is graphed at increasing lags of the power outages time series.

4.2.4 Two Time Differentiation

In this section we test the same example, except that both time series are differentiated twice.

As previously, the Figure 4.8 represents the cross-correlation coefficient between the tweet time series and the power outages time series. As shown in the graph, the biggest ACF in absolute value is just a bit greater than 0.15, and occurs at $h = -23$. Again, this coefficient is negative. This can be interpreted as a relation between an above-average number of power outages and an under-average number of tweets $3h45min$ minutes later. However, we cannot talk about existing correlation as this number is pretty low for an ACF coefficient. The same principle applies for the absolute value of the ACF at $h = 5$ or $h = -24$.

The Figure 4.9 gives the nine scatterplots of the number of power outages versus the number of tweets, taken from a 0 to a 9 interval lags times. Again, the correlation co-

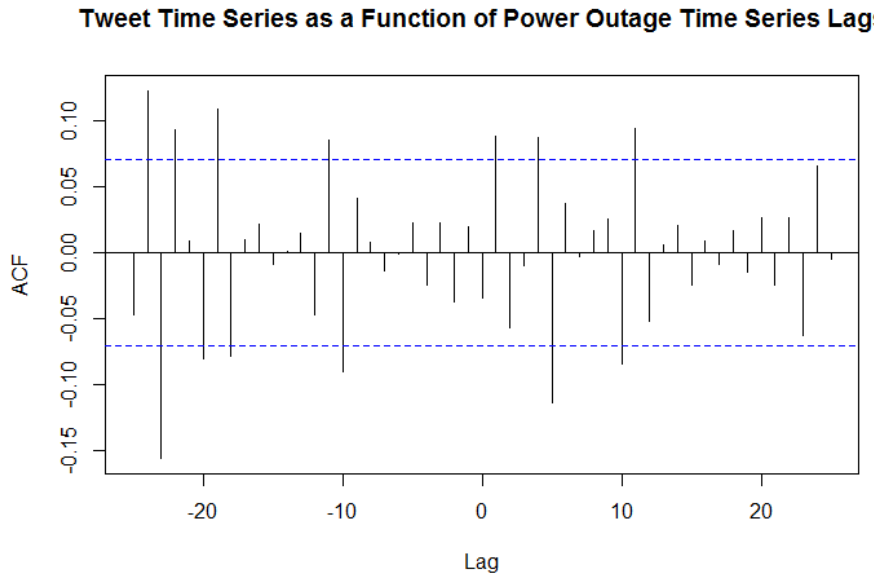

Figure 4.8: Cross-correlation graph between the conditioned time series of power outages and the time series of power or electricity related tweets from Hurricane Sandy, after a two time differentiation.

efficients are pretty low, and it looks like there is no detectable relationship between the number of power outages and the number of tweets. Therefore, differentiating twice do not seem to be more conclusive in this case.

4.2.5 Windowed Cross-Correlation

In this part, the results of the windowed cross-correlation are presented, for different window sizes. Table 4.2 gives the cross-correlation coefficient for the one time differentiation case, and Table 4.3, for the two time differentiation case. As we can see in Table 4.2, the values are quite low, as expected from the previous analysis. The highest coefficient is found for the smallest window size, and is on the same range of values than the one found with the regular cross-correlation function. These results tend to confirm that there is no significant correlation found here.

As we can see in Table 4.3, for every window size the maximum cross-correlation

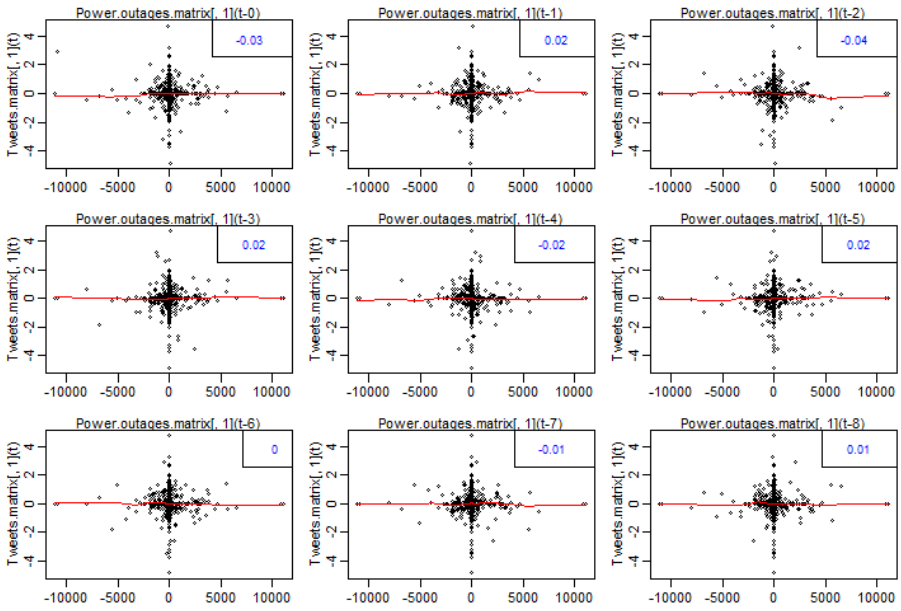

Figure 4.9: Scatterplots of number of power outages versus number of tweets, where the tweet time series is graphed at increasing lags of the power outages time series.

Window Size	Maximum Correlation Coefficient
2	0.1666667
5	0.0666667
10	0.0333333
15	0.00215538
30	0.001075269
60	0.0005376344
90	0.0003584229

Table 4.2: Maximum correlation coefficient for different window sizes, for a one time differentiation example.

Window Size	Maximum Correlation Coefficient
2	0.00
5	0.00
10	0.00
15	0.00
30	0.00
60	0.00
90	0.00

Table 4.3: Maximum correlation coefficient for different window sizes, for a one two differentiation example.

Keyword Combination	Total Frequency of the Keywords Combination
hurricane + flood	30416
hurricane + water	30814
hurricane + rain	71132
storm + flood	23323
storm + water	18987
storm + rain	67212

Table 4.4: Frequency analysis of the keywords combination on Hurricane Sandy data set.

coefficient is always null. These results are conformed with the previous analysis as no correlation were found. Therefore, it tends to confirm that there is no correlation between both time series in this example. However, it seems surprising that no coefficient is found at all. It could be explained by the fact that maybe, the small coefficients found previously occur for a very large delay, and so the window sizes are too small.

4.3. Case 2 : Floods

This part provides the results of the cross-correlation analysis for the flooding example.

4.3.1 Tweets Selection

As for the first example, several keywords are selected regarding to their relevance to the physical disruption caused. The first keyword is flood, which also allows to select words like flooding or flooded during searches. The second one is water, and the last one rain. Table 4.4 presents the results of the frequency analysis of the keywords combinations in the Hurricane Sandy data set. As we can see, the combinations including the keyword rain are the most frequently used.

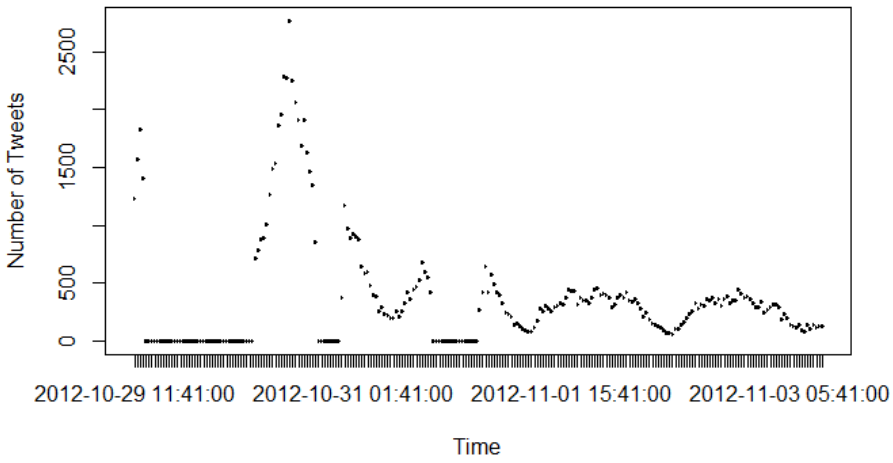

Figure 4.10: Time series of the total number of flood- water- and rain-related tweets from Hurricane Sandy.

All the tweets containing the keywords are then selected and decomposed into a time series taken over thirty minute intervals. Figure 4.10 represents the tweets time series, and as previously some values are missing. These four time periods are represented by the black lines. Figure 4.11 shows the tweets time series after linear interpolation.

4.3.2 Water Peak Elevation Data Collection

This data set is composed of water peak elevations due to Hurricane Sandy, on the Eastern seaboard, involving the following states : Virginia, Maryland, Delaware, New Jersey, New York, Connecticut, Rhode Island, Massachusetts, and the city of New York. This data are public and provided by the United States Geological Survey (USGS).

Unfortunately, this data set is not as complete as the power outages data set. Indeed, there aren't as many points, and this flood data do not occur on regular time interval,

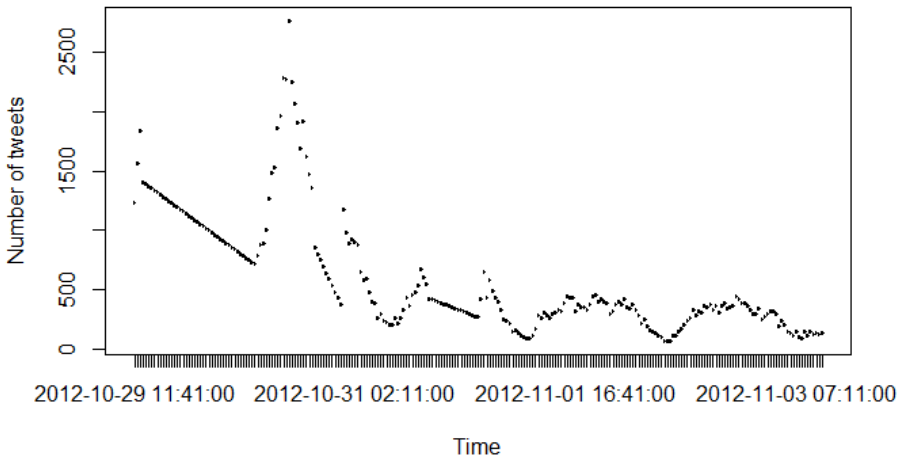

Figure 4.11: Time series of the total number of flood- water- and rain-related tweets from Hurricane Sandy, after linear interpolation.

which could possibly damaged the quality of the cross-correlation analysis.

Figure 4.12 represents the water peak elevations time series. As we can observe on the graph, the majority of the floods seem to occur on the $29th$ of October 2012.

As previously, the time series can be decomposed into its trend, seasonal and random elements. As we can observe on the Figure 4.13, it seems difficult to detect a trend.

4.3.3 One Time Differentiation

In this part, the results of the cross-correlation analysis for the flood example, with only one time differentiation is presented.

The Figure 4.14 represents the correlation coefficients between the tweet time series and the water peak elevations time series. We can observe on the graph that the highest ACF in absolute value appear on $h = -6$ and $h = 1$. As for previously these values are quite low, and we can't conclude that a correlation exists between these two time series here.

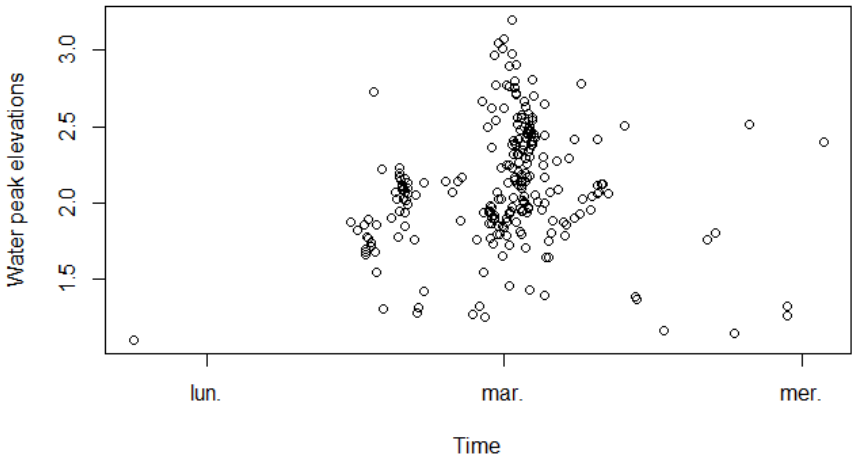

Figure 4.12: Time series of the water peak elevations from Hurricane Sandy.

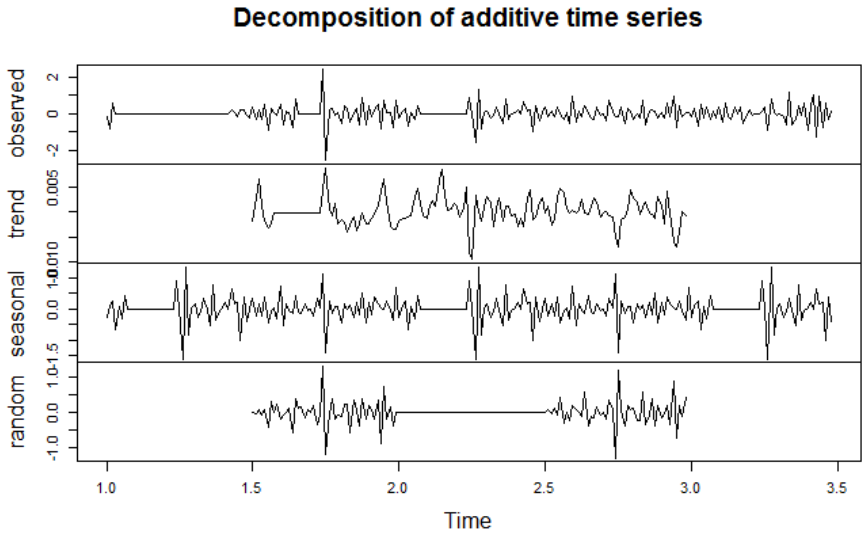

Figure 4.13: Decomposition of the water peak elevations time series into its trend, seasonal and random elements.

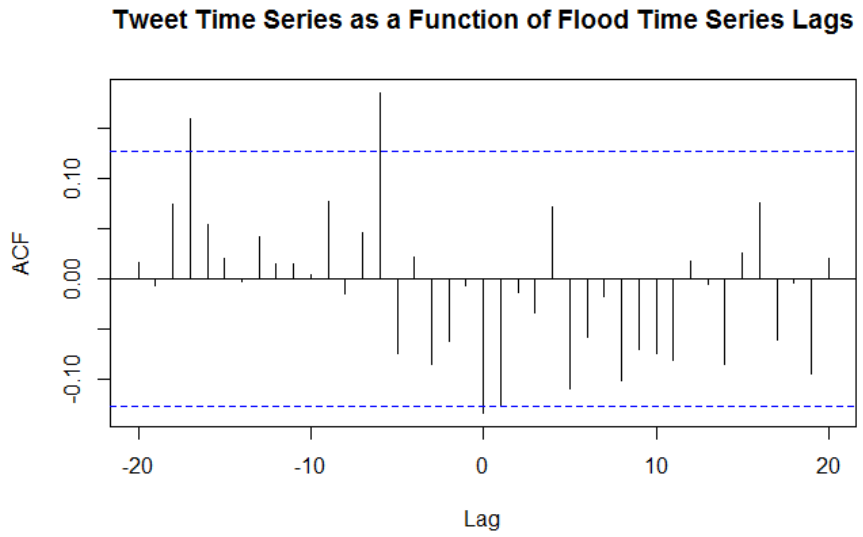

Figure 4.14: Cross-correlation graph between the conditioned time series of water peak elevations and the time series of flood or rain related tweets from Hurricane Sandy.

Again, with Figure 4.15, the scatterplots of the water peak elevations versus the number of tweets, taken from a 0 to a 9 interval lags times are given. When looking at the graph, it seems impossible to determine a relationship between the number of tweets and the water peak elevations.

4.3.4 Two Time Differentiation

This section presents the results of the cross-correlation analysis for the flooding example, with a two time differentiation.

The Figure 4.16 represents the cross-correlation function between the tweet time series and the water peak elevations time series. As we can observe on the graph, it looks like there isn't many high cross-correlation coefficient in absolute value. The greatest ones appear for $h = 3$ and $h = 8$, with a CCF slightly under 0.15, for a negative correlation. Again, these numbers are too small for us to consider that there is an actual correlation between these two time series.

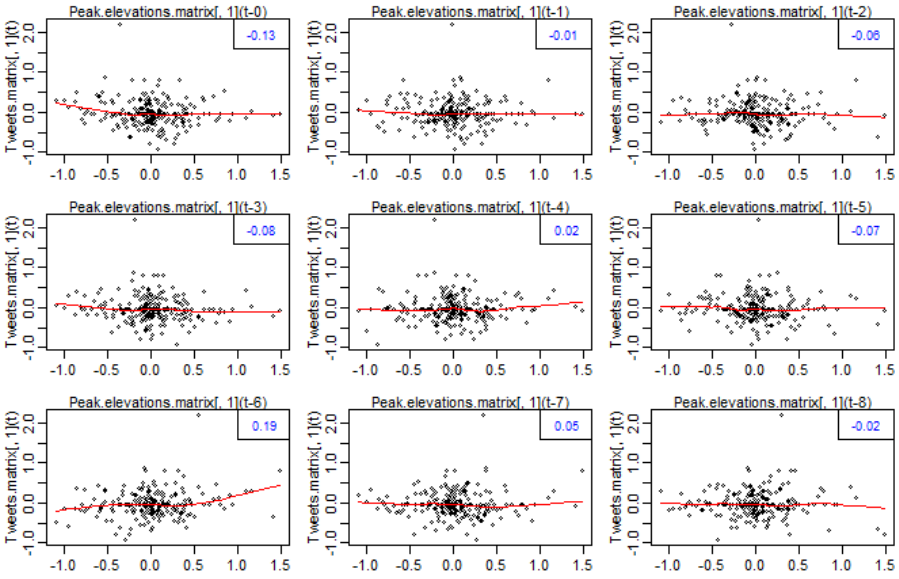

Figure 4.15: Scatterplots of number of water peak elevations versus number of tweets, where the tweet time series is graphed at increasing lags of the flood time series.

On Figure 4.17, the scatterplots of the water peak elevations versus the number of tweets, taken from a 0 to a 9 interval lags times are displayed. By observing the values of the cross-correlation coefficients, we cannot say that a correlation exists here. We can conclude that there is no significant correlation between the two time series here.

4.3.5 Windowed Cross-Correlation

Here are presented the results of the windowed cross-correlation, for different window sizes. The table 4.5 gives the cross-correlation coefficients for the one time differentiation example, and Table 4.6, for the two time differentiation case. As we can observe on both Table 4.5 and Table 4.6, the results are null, for every window size. There is no accuracy improvement compared to the previous analysis with the regular cross-correlation function. Indeed, there is even less information as the cross-correlation coefficients found previously are not detected here. This can be

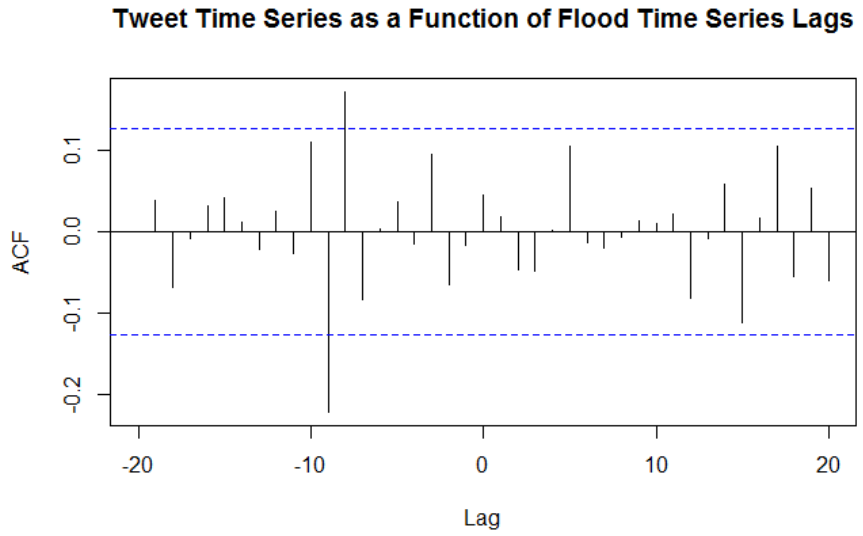

Figure 4.16: Cross-correlation graph between the conditioned time series of water peak elevations and the time series of flood or rain related tweets from Hurricane Sandy, after a two time differentiation.

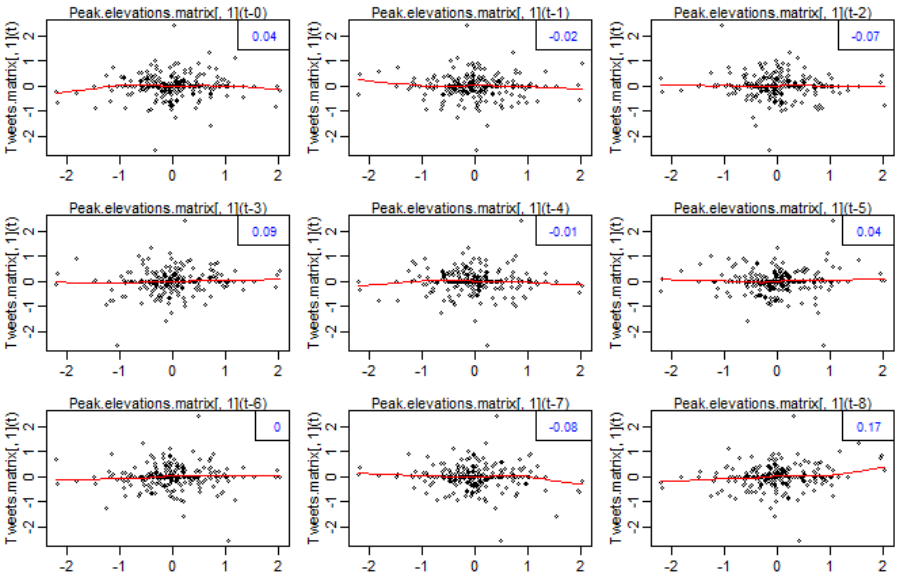

Figure 4.17: Scatterplots of number of power outages versus number of tweets, where the tweet time series is graphed at increasing lags of the power outages time series, after a two time differentiation.

Window Size	Maximum Correlation Coefficient
2	0.00
5	0.00
10	0.00
15	0.00
30	0.00
60	0.00
90	0.00

Table 4.5: Maximum correlation coefficient for different window sizes, for a one time differentiation example.

Window Size	Maximum Correlation Coefficient
2	0.00
5	0.00
10	0.00
15	0.00
30	0.00
60	0.00
90	0.00

Table 4.6: Maximum correlation coefficient for different window sizes, for a one two differentiation example.

explained by the fact that this method may be not as efficient as the previous one [5].

Chapter 5

Conclusion

The main motivation on this project is that since Twitter has been created, many studies have been conducted to exploit its capacity to be a social sensor of our society, especially through sentiment analysis, with less interest on the physical environment and their relationship. That's why, in this study, we compare Twitter's behavior during a natural disaster, with a measurable physical disruption caused by the hazard, with a cross-correlation analysis.

The cross-correlation analysis did not yield to successful conclusive results. Indeed, for both examples studied, the highest absolute values of the CCF coefficients are quite low, below 0.20. Therefore, it is not possible to conclude that a correlation exists between the number of tweets related to the hazard, and the measurable disruption for these examples. Moreover, the results are pretty similar for both examples between the one time differentiation case and the two time differentiation case. No improvement has been observed by differentiating a second time both time series. The second time differentiation is to ensure of the stationarity of the time series, and that trend effect, like the time of the day where most tweets are posted, is eliminated. Since for both cases the number of differentiation do not seem to have an impact on the results of the CCF analysis, it could mean that the effect of the

daily posting trend is quite limited.

The windowed cross-correlation analysis also did not provide successful proof of a relationship between the physical environment and Twitter. Indeed, except for the one time differentiation case on the power outages example during Hurricane Sandy, no correlation is detected at all. The windowed cross-correlation function is an alternative, but in no case a more accurate method than the regular CCF. Therefore, it is not surprising that the coefficients found are not higher in absolute value than the previous ones. However, not having any correlation at all for most cases is questionable.

All these results are similar as no clear correlation is found for any case. However, it is also impossible to conclude for both examples, that the time series are not correlated at all. This study only shows that we didn't find a correlation, not that it didn't exist. Indeed, several factors may have influenced the analysis, and the major one is the quality of the data sets used here. The Twitter data set is common to both example, and even though it is composed of several million tweets related to Hurricane Sandy, important holes may have impacted a lot the quality of the analysis. The missing periods correspond mainly to times when Hurricane Sandy stroke violently the Eastern seaboard, and the use of a linear interpolation to fill the gap may have increase uncertainty. Several factors may have caused these holes as mentioned previously. It can be due to a technical issue during the data collection, or may be human related. It is possible that people would stop using Twitter, or social media in general to get safe during a natural disaster, or that they would use carefully their electronic devices to save battery in case of power outages for example. Also, for the second case, the water peak elevation data set do not contain many points and there are not consistently taken over thirty minute intervals.

Unfortunately, after this study, it remains impossible to answer the research question. The cross-correlation analysis do not allow us here to declare that the behavior

of Twitter during a natural disaster is linked to the physical environment. However, for all the reasons mentioned previously, we also cannot conclude that these two entities are uncorrelated. To pursue this research, studying another natural disaster example with different data sets can be considered, preferably a large scale hazard, where official national agencies provide public data on the physical disruption such as power outages. It would also have been interesting to perform a spatial analysis, and compare the repartition of the tweets with the locations of the disruption. Another research interest could be to highlight the social index of the damaged zones, and determine their use of social media during a natural catastrophe. Kyle Walker, a geography teacher at the Texas Christian University, developed an interactive map on social indexation in several states of the Unites States [25], which could be interesting to compare if spatial coordinates are available on the data sets.

References

[1] We are social. `https://wearesocial.com/`. Accessed : 2017-10-06.

[2] C. Beyney. Quantitative analysis of social media sensibility to natural disasters. Master Thesis, University of Oklahoma, 2015.

[3] S. Boker, M. Xu, J. Rotondo, and K. King. Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series. *Psychol Methods*, 7, 2002.

[4] D. Cain. *Twituational awareness : gaining situational awareness via crowd-sourced #disaster epidemiology.* PhD thesis, Naval Postgraduate School, Monterey California, 2013.

[5] M. Coco and R. Dale. Cross-recurrence quantification analysis of categorical and continuous time series : an r package. *Front Psychol*, 5, 2014.

[6] P. Earle, D. Bowden, and M. Guy. Twitter earthquake detection : earthquake monitoring in a social world. *Annals of Geophysics*, 54, 2011.

[7] D. Gayo-Avello. A meta-analysis of state-of-the-art electoral prediction from twitter data. *Social Science Computer Review*, 31:649–679, 2013.

[8] R. Hyndman and G. Athanasopoulos. Time series components in forecasting: principles and practice. `https://www.otexts.org/fpp/6/1`, 2013. Accessed : 2017-03-20.

[9] R. Keane and R. Adrian. Theory of cross-correlation analysis of piv images. *Applied Scientific Research*, 49:191–215, 1992.

[10] A. Larsson. Studying political microblogging : Twitter users in the 2010 swedish election campaign. *New media and society*, 14:729–747, 2011.

[11] K. Leetaru, S. Wang, G. Cao, A. Padmanabhan, and E. Shook. Mapping the global twitter heartbeat : the geography of twitter. *First Monday*, 18, 2013.

[12] A. McEachren and al. Geotwitter analytics support for situational awareness. In *Proceedings of the 2011 IEEE Conference on Visual Analytics Science and Technology, Providence, RI, USA, 23-28 October, 2011*, 2011.

[13] S. Middleton, L. Middleton, and S. Modafferi. Real-time crisis mapping of natural disasters using social media. *Intelligent Systems, IEEE*, 29:9–17, 2014.

[14] R. Mushtaq. Augmented dickey fuller test. `https://ssrn.com/abstract=1911068orhttp://dx.doi.org/10.2139/ssrn.1911068`, 2011. Accessed : 2017-07-13.

[15] R. Nau. Statistical forecasting : notes on regression and time series analysis. `https://people.duke.edu/~rnau/411home.htm`, 2017. Accessed : 2017-06-19.

[16] F. Nooralahzadeh, V. Arunachalam, and C. Chiru. 2012 presidential elections on twitter - an analysis on how the us and french election were reflected in tweets. In *Proceedings of the 19th International Conference on Control Systems and Computer Science, Bucharest, Romania, 29-31 May, 2013*, pages 240–246, 2013.

[17] University of Utah. Interpolation. `https://sutherland.che.utah.edu//2450Notes/Interpolation.pdf`, 2014. Accessed : 2017-04-19.

[18] J. Osborne. Improving your data transformations : applying the box cox transformation. *Practical Assessment, Research and Evaluation*, 15, 2010.

[19] V. Plerou and al. Universal and nonuniversal properties of cross correlations in financial time series. *Physical Review Letters*, 83, 1999.

[20] J. Prins. Nist/sematech e-handbook of statistical method. `http://www.itl.nist.gov/div898/handbook/`, 2003. Accessed : 2017-07-13.

[21] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web, Raleigh, North Carolina, USA, 26-30 April, 2010*, pages 851–860, 2010.

[22] J. Scargle. Studies in astronomical time series analysis. fourier transforms, autocorrelation functions, and cross-correlation functions of evenly spaced data. *The Astrophysic Journal*, 343:874–887, 1989.

[23] C. Shazili. *Advanced data analysis from an elementary point of view*. Cambridge University Press, 2017.

[24] B. Stone. Twitter, the start-up that wouldn't die. `http://www.businessweek.com/articles/2012-03-01/twitter-the-startup-that-wouldnt-die`, 2012. Accessed : 2017-03-24.

[25] K. Walker. Educational attainment in america. `http://personal.tcu.edu/kylewalker/maps/education/#12/37.5610/-122.4539`, 2017. Accessed : 2017-07-13.

[26] D. Wang and al. Using humans as sensors : an estimation-theoretic perspective. In *Proceedings of the 13th International Symposium on Information Processing in Sensor Networks, Berlin, Germany, 15-17 April, 2014*, 2014.

[27] G. Wolfsfeld, E. Segev, and T. Sheafer. Social media and the arab spring : politics comes first. *The International Journal of Press/Politics*, 18:115–137, 2013.

Appendices

Appendix A

Box Cox Transformation

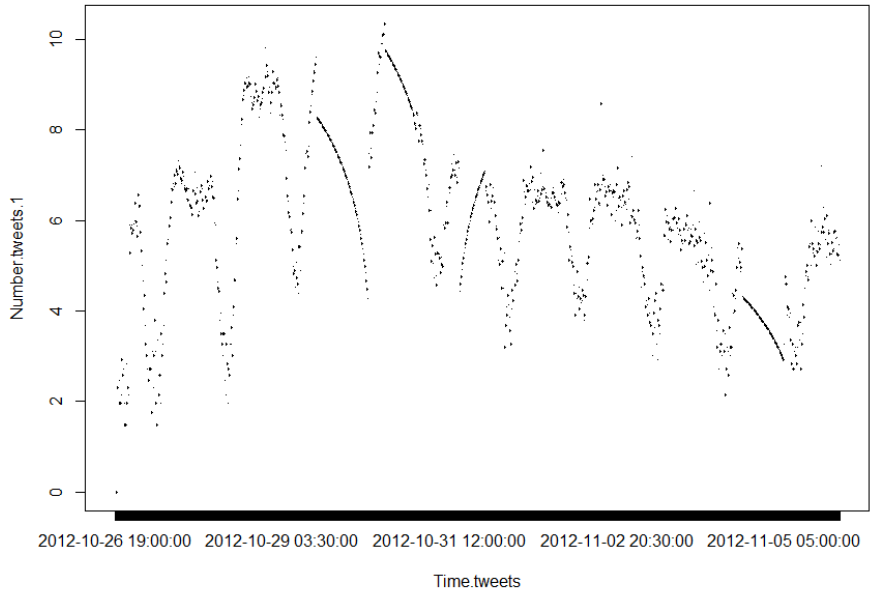

Figure A.1: Time series of the total number of power- and electric-related tweets, after a Box Cox transformation with $\lambda = 0.1$.

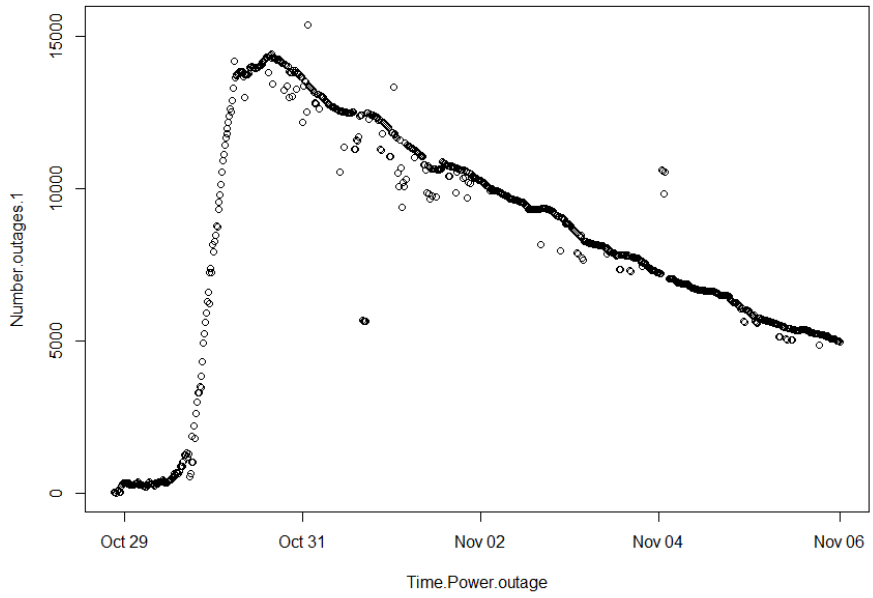

Figure A.2: Time series of the total number of power outages, after a Box Cox transformation with $\lambda = 0.6$.

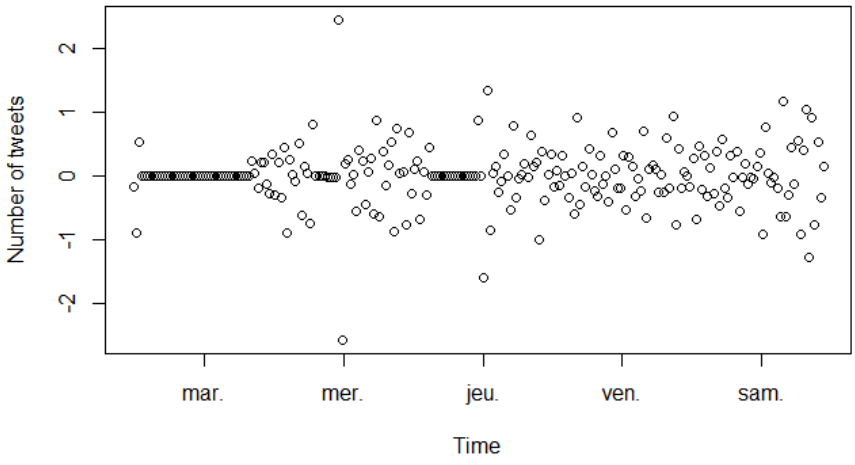

Figure A.3: Time series of the total number flood related tweets, after a Box Cox transformation with $\lambda = 0.1$.

Appendix B

Differentiation

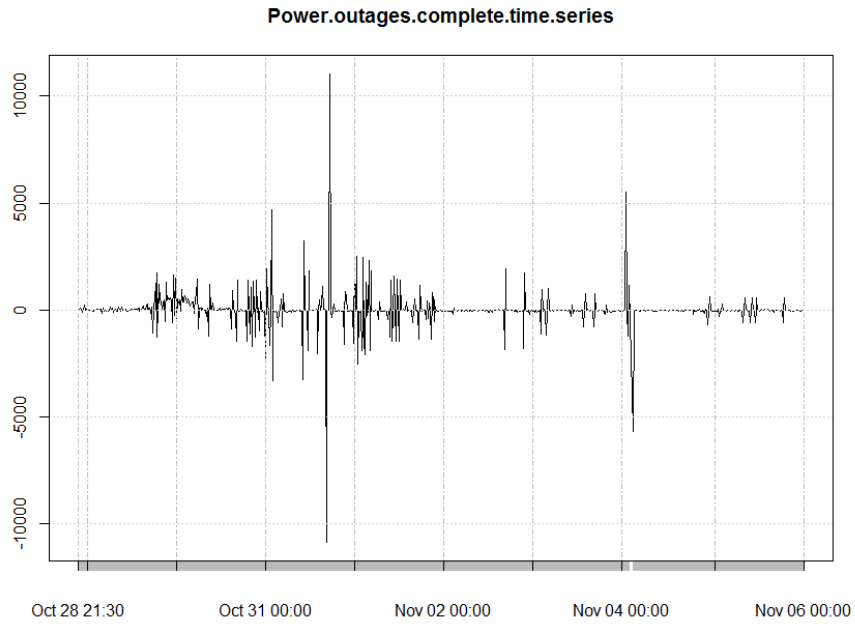

Figure B.1: Time series of the differentiated, transformed power- and electricity-related Twitter data.

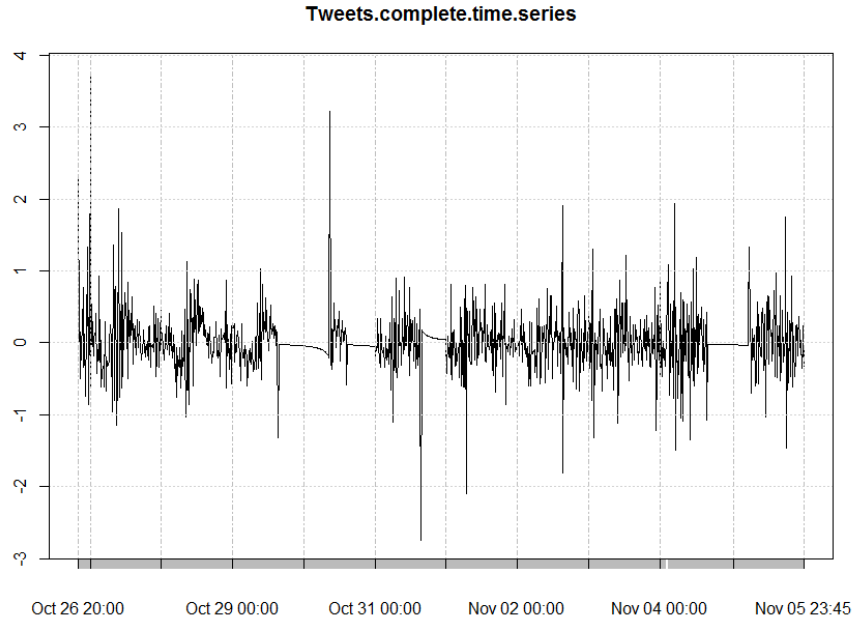

Figure B.2: Time series of the differentiated, transformed power outages data.

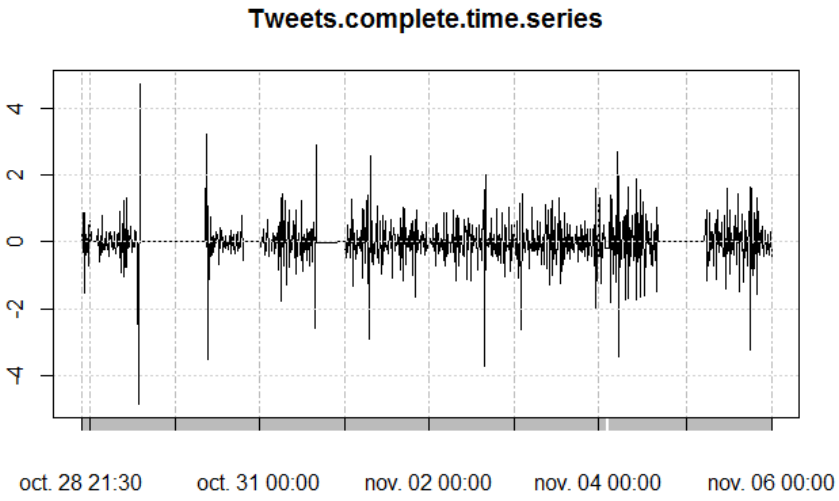

Figure B.3: Time series of the second order differentiated, transformed power- and electricity-related Twitter data.

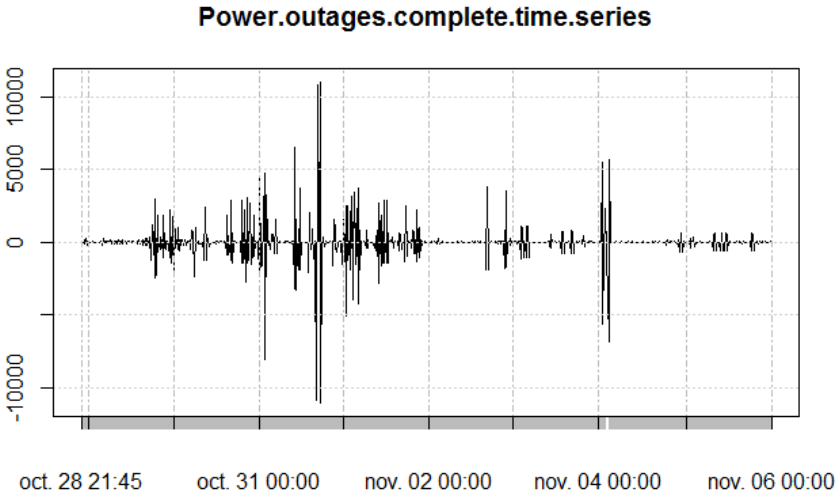

Figure B.4: Time series of the second order differentiated, transformed power outages data.

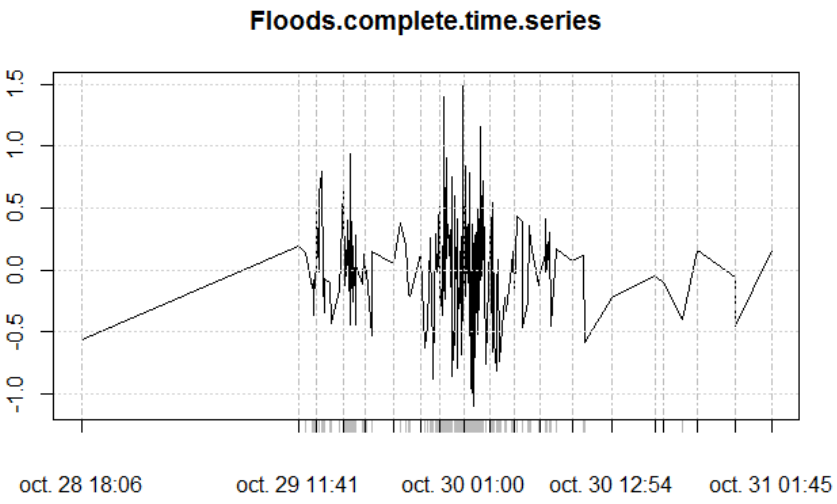

Figure B.5: Time series of the differentiated, transformed flood-related Twitter data.

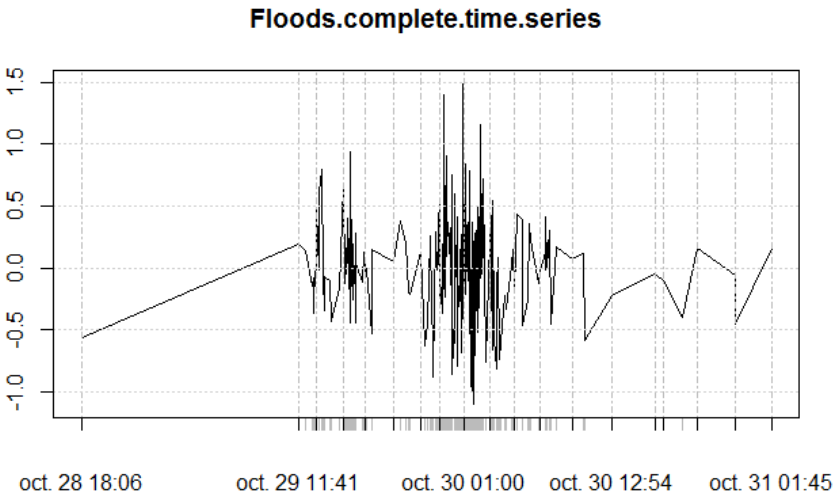

Figure B.6: Time series of the differentiated, transformed water peak elevations data.

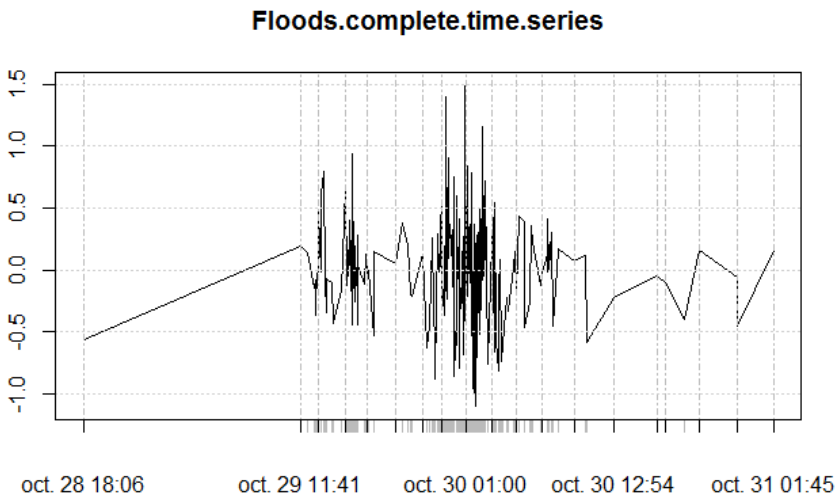

Figure B.7: Time series of the differentiated, transformed flood-related Twitter data.

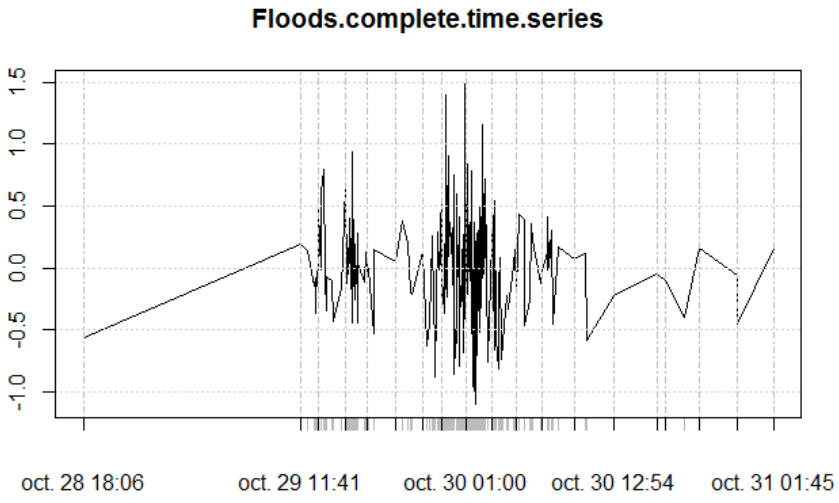

Figure B.8: Time series of the differentiated, transformed water peak elevations data.

Appendix C

Augmented Dickey Fuller test

Results of the Augmented Dickey Fuller test for every time series, in both examples. A low p-value, and a high Dickey Fuller statistic mean that the time series is stationary.

Time series	Differentiation order	Dickey Fuller coeff	Lag order	p-value
Power outages	1	-10.346	9	0.01
Tweets	1	-7.9078	9	0.01

Table C.1: Results of the Augmented Dickey Fuller test for the power outages example with a one round differentiation.

Time series	Differentiation order	Dickey Fuller coeff	Lag order	p-value
Power outages	2	-18.37	9	0.01
Tweets	2	-14.683	9	0.01

Table C.2: Results of the Augmented Dickey Fuller test for the power outages example with a two time differentiation.

Time series	Differentiation order	Dickey Fuller coeff	Lag order	p-value
Water peak elevations	1	-5.5078	6	0.01
Tweets	1	-8.5456	6	0.01

Table C.3: Results of the Augmented Dickey Fuller test for the flood example with a one round differentiation.

Time series	Differentiation order	Dickey Fuller coeff	Lag order	p-value
Water peak elevations	2	-5.5078	6	0.01
Tweets	2	-8.5456	6	0.01

Table C.4: Results of the Augmented Dickey Fuller test for the flood example with a two time differentiation.