POLITECNICO DI MILANO

Corso di Laurea Magistrale in Ingegneria per l'Ambiente e il Territorio - Environmental and Land Planning Engineering

Dipartimento di Ingegneria Civile e Ambientale

Geolocated Twitter data as a proxy for the analysis of natural disasters: the Hurricane Florence case study

Supervisor: Prof. Federica Migliaccio

Co-supervisor: Katarina Spasenovic, Ph.D. candidate

Master Graduation Thesis by:

Federica Gaspari

Matr. 899057

Academic Year 2019/2020

Abstract

The growing availability of geolocated social network contents – so-called Social Media Geographic Information – brought new questions regarding their potential in different situations, from urban mobility planning to crisis scenario. The needs of first-hand feedbacks from affected areas and the social media response during natural disasters have drawn attention of researchers, especially the ones involved in exploiting and interpreting the complex use of SMGI for event detecting and monitoring in time and space. In particular, Twitter, thanks to the possibility of accessing its data through official Application Programming Interfaces, has become the subject of a significant number of studies focused on a great variety of natural disasters (earthquake, hurricanes, floods, fires etc.).

This work presents possible geo-statistical and temporal analysis on Twitter posts published during the hurricane Florence emergency that occurred in the United States of America in 2018. The spatial and temporal distribution of geolocated posts has been analysed at different scales, exploring the composition of the SMGI dataset, and identifying at a global scale the areas characterized by a higher level of social media activity. More detailed geostatistical analyses have been performed in order to evaluate the significance of the given geolocated tweets with respect to the natural disaster. The distribution has been explored calculating Nearest Neighbour Indexes. The identification of the most affected areas has been analysed with hot spot analyses and its Getis Ord Gi* index. The results of these analyses and the visual representation of the Kernel Density estimation have then been compared with National Oceanic and Atmospheric Agency (NOAA) and Federal Emergency Management Agency (FEMA) hurricane reports, highlighting the potential for identification of the landfall site through the Twitter dataset and the main issue associated to the influence of densely populated areas on the calculations.

Keywords: natural disaster, hurricane, spatial data analysis, spatio-temporal analysis, Twitter, social media, SMGI

Sommario

La crescente disponibilità di contenuti geolocalizzati sui social networks – i cosiddetti Social Media Geographic Information – ha posto nuove domande in merito al loro potenziale in diverse situazioni, dalla pianificazione della mobilità urbana fino agli scenari di crisi legati a disastri naturali. La necessità di informazioni immediate relative alle aree direttamente colpite da un'emergenza e il livello di attività registrato sui social media in occasione di disastri naturali hanno richiamato l'attenzione dei ricercatori, interessati a utilizzare e a interpretare la complessità di utilizzo dei SMGI per il rilevamento e il monitoraggio degli eventi nel tempo e nello spazio. In particolare, Twitter, grazie alla possibilità di accedere ai suoi dati attraverso le Application Programming Interface ufficiali, è diventato rapidamente il soggetto di un rilevante numero di studi focalizzati su ogni tipo di disastro naturale (terremoti, uragani, alluvioni, incendi etc.).

Questo lavoro presenta alcune possibili analisi geostatistiche e temporali sui post di Twitter pubblicati nel corso dell'emergenza legata all'uragano Florence che ha colpito gli Stati Uniti nel 2018. La distribuzione spaziale e temporale dei post geolocalizzati è stata valutata a scale differenti, in modo da esplorare la composizione del dataset di SMGI e identificare a livello globale le aree caratterizzate da maggiore attività sulla piattaforma di Twitter. Successivamente sono state eseguite analisi geostatistiche più dettagliate per valutare la rappresentatività dei tweets geolocalizzati in rapporto all'evento considerato in una finestra temporale di cinque giorni. Inizialmente, la loro distribuzione è stata valutata attraverso il calcolo del Nearest Neighbour Index. In seguito, è stato calcolato l'indice locale di Getis Ord Gi* per l'analisi degli hot-spot per identificare le aree con attività social più rilevanti. I risultati di queste analisi e le rappresentazioni cartografiche della stima della densità di Kernel sono state in seguito confrontate con i report relativi all'uragano Florence di National Oceanic and Atmospheric Agency (NOAA) e Federal Emergency Management Agency (FEMA), evidenziando il potenziale di identificazione del sito di arrivo sulla terraferma di Florence e le maggiori problematiche associate all'influenza sui calcoli delle aree densamente popolate.

Parole chiave: disastro naturale, uragano, analisi dei dati spaziali, analisi spazio-temporale, Twitter, social media, SMGI

Table of Contents

List of Figures

List of Tables

Acronyms

AGI	Authoritative Geographic Information
API	Application Programming Interface
DRR	Disaster Risk Reduction
FEMA	Federal Emergency Management Agency
GIS	Geographic Information System
GPS	Global Positioning System
ISO	International Organization for Standardization
JSON	JavaScript Object Notation
LBSN	Location Based Social Network
NHC	National Hurricane Centre
NOAA	National Oceanic and Atmospheric Agency
NWS	National Weather Service
OSM	OpenStreetMap
PoI	Point of Interest
SMGI	Social Media Geographic Information
USCB	United States Census Bureau
USGS	United States Geological Survey
VGI	Volunteered Geographic Information
WGS84	World Geodetic System 1984

Introduction

Modern web technologies and tools have completely changed the idea of community and society, leading to the new methodologies that will help to understand the world and its evolution in time and space. During the last two decades, the growing accessibility to broadband connection and the pervasive diffusion of handheld and mobile devices have led to the so-called *Web 2.0* era, implying a different approach to information in term of both its nature and applications.

Previously, web users were part of a passive consumer audience, not included in the process of data creation and elaboration. Thanks to Web 2.0 tools and the rise of collaborative and crowdsourced projects, the traditional dynamics linked to the communication and sharing of information have dramatically evolved in more inclusive ways. The role of individual users has become more complex over the time, enabling content creation and interaction with each others.

The amount of data produced and shared in real time on the web is constantly bringing new questions and challenging tasks that should be solved. Volume, velocity, and variety are key factors that should be considered when dealing with *Big Data* and, more specifically, with geographic information. The concept of web users as human sensors (Goodchild, 2007) experiencing events and disseminating information requires the introduction of new definitions and approaches to handle and interpret derived data. *Volunteered Geographic Information* (VGI) are the natural products of this evolution in geospatial data science, containing details about the user activities in time and space but also other thematic information.

The contexts of information contribution and analysis should be carefully considered in order not to misunderstand the motivation and the meaning of the user activities (Roick & Heuser, 2013). This aspect requires even more attention when the subjects of a study are *Social Media Geographic Information* (SMGI), a specific category of new geographic data derived from social networks has common points with VGI but also many important differences that require alternative methodologies.

The growing popularity of *Location Based Social Networks* (LBSN) provides an important amount of up-to-date user shared data that could contain valuable spatial, temporal, and thematic information. Recent studies have demonstrated that SMGI in specific cases represent significant proxies of user behaviours and observations on events with strong connection to their spatial component.

Geolocated posts published on social media platforms like Twitter, are instant information and possibly direct link to the people involved in certain event, having this on mind geolocated posts could be game-changing elements in the effect analysis of social events, marketing evaluation, mobility or urban plans but also natural disasters. For this last event type, SMGI potentials are still under studies because the nature of the data and its availability could be crucial for damage detection as much as for emergency response and reporting. These social media products, integrating official data provided by authorities, could make the difference in the cost reduction of data collection that usually needs expensive surveys. However, when dealing with SMGI, it is very important to consider the several issues related to the use of these data that intrinsically have quality and representativeness problems.

The case study: Twitter posts published during the Hurricane Florence

Considering the previously explained concepts, in recent years evaluating the accuracy and the representativeness of SMGI is one of the main objectives of studies in the field of emergency analysis and evaluation. For this reason, the case study of this thesis is represented by the event Hurricane Florence, a natural disaster that occurred in September 2018 and hit the West coast of the United States of America (US). This was the first major event of the 2018 Atlantic hurricane season that caused 22 direct deaths and 30 additional indirect fatalities in North Carolina, South Carolina and Virginia as reported by the official tropical cyclone documentation (Stewart & Berg, 2019), published by the National Hurricane Center (NHC), the division of the US National Oceanic and Atmospheric Administration (NOAA) and National Weather Service (NWS).

The goal of this thesis is to perform some possible geo-statistical and temporal analysis on Twitter posts published during the emergency to evaluate the significance of this SMGI with respect to the natural disaster. The complete Twitter dataset has been collected by Harvard University through Twitter API filtering the shared contents by the hashtags and keywords associated to the hurricane (e.g. (#)hurricane, (#)Florence etc.).

The Hurricane Florence Twitter dataset has been pre-processed in Python environment with Hydrator application developed by DocNow. The data preliminary preparation and analysis with descriptive statistics calculations have been computed with the joint use of MATLAB by MathWorks and Microsoft Excel. Finally, the core spatial and temporal analyses have been performed in a Geographic Information System (GIS) environment with ArcGIS software produced by ESRI, with the aim to identify the distribution patterns and hot spots. The obtained results should be compared with official

reference data provided by US national authorities (NOAA, NHC and Federal Emergency Management Agency (FEMA)) in order to evaluate accuracy, coherence and representativeness of the process results.

Thesis outline

The thesis work has been organized as follow:

- The introduction provides the information about the main idea and the goals of the executed work, introducing the case study and the tools used for the analysis.
- Chapter 1 includes an overview of the main concepts associated to VGI and SMGI, focusing on their enabling factors and definitions, helping to understand similarities and differences between these two types of geographic information based on the nature of data, user types, application contexts and possible issues.
- In Chapter 2 the Twitter data structure is presented with methodologies used to retrieve data through Application Programming Interface (API). The peculiarities and potentials of tweet contents are then illustrated with an overview of some representative case studies focused on different types of natural disaster events applied during different emergency phases.
- Chapter 3 introduces the Hurricane Florence case study and a complete description of the Twitter dataset analysed, explaining the data preparation, the filtering and pre-processing.
- Chapter 4 presents the results of the spatial elaborations applied to identify the most active/affected US states, the calculation of the most common spatial-autocorrelation indexes and the computation of the Kernel density map.
- Finally, in Chapter 5 the conclusions and the final observations are highlighted and compared with reference data. Following discussion about the strengths and weaknesses of this type of data were presented with ideas about the other possible analyses to be performed or alternative approaches to SMGI dataset.

Chapter 1

Volunteered Geographic Information and Social Media Geographic Information

The growing availability of new types of data such as Volunteered Geographic Information and Social Media Geographic Information require different analysis methods. This new procedure needs different approaches from the ones applied in the past for traditional cartographic data and products. In order to better understand the nature of this geographic information, it is essential to explore the context and the reason of their wide spreading, highlighting their main characteristics and potentials in relation to the typical user profiles without ignoring their data structure. This chapter gives a presentation of VGI and SMGI main characteristics under a geospatial perspective, also considering the fundamental technological and sociological aspects which contributed to these phenomena redefining the concept of geography.

1.1 The context and the reasons of this phenomena

The last two decades of technology have recorded a stable internet connection and a significant diffusion of new tools and applications with the aim to enable social interaction between users. This has been possible thanks to the combination of science, technology and sociology whose complex dynamics affect the derived geographic information.

Enabling factors

The rise of forums, blogs and social networks is the evidence of the interaction development between users on the internet in technological and social sense. The constant need for connection and the feeling of being part of a global community are the foundations of the collective intelligence (Levy, 1994) concept. For this concept, the web represents an ideal place where everyone is able to convey ideas, knowledge and efforts as a part of a coordinated and collaborative cognitive project driven by technological innovations.

This aspect represents a complex element that should be analysed since it implies the knowledge of a wide variety of concepts from different subjects. Nonetheless, the continuous and ubiquitous

interactions between individual users that are part of a network can solve the problem of information requesting. In this way, the interconnection of many users provides a constant exchange of complementary skills in many fields. This new structure in which every single user is involved in the contribution and thinking process resulted in the crowdsourced model (Howe, 2006), a system that is strictly connected to the Web 2.0 era based on the growing availability of interaction and data management tools (O'Reilly, 2007).

The direct effect of this model is indeed the bustling dissemination of information and data inside the web network, with a significative amount of content produced by a mass community that involves also not-professional users in the data creation process. This huge data production and availability, also called *Big Data*, constantly offer new challenges in the context of content management.

Geospatial science is one of the fields most impacted by this evolution that led to the so-called *Web Cartography* (Plewe, 2007) and *Neogeography* (Haklay M. et al., 2008), implying a new definition of the user/spatial data consumer role and a different type of geographic information that previously were exclusive duties of official cartographic agencies and companies. Every single user, through Citizen Science applications or Location Based Social Networks (Cohn, 2008), can now disseminate and retrieve valuable spatial data. Thanks to the internet access as well as the mobile devices equipped with Global Positioning System (GPS) receivers that allow to both track activities and sense events and behaviours. Smartphones have indeed become of common use because of the considerably reduced costs.

Term origins and definitions

The constant growth in the amount of this new type of spatial data required the introduction of different and dedicated definitions to be addressed. Comparing the Neogeography implications and peculiarities with the traditional approaches, in 2007 Goodchild termed *Volunteered Geographic Information*. This is a special case of web user-generated geospatial contents that every person, including non-professionals, can voluntarily create and share with peers. The essential characteristic of VGI is the presence of a geographic component that can have the form of a couple of coordinates (latitude and longitude) or a *geotag*, a label representing a topographic element or a Point of Interest (PoI) associated to a bounding box whose extreme vertices have known coordinates values. The VGI could then be considered as the result of a democratization of geo-science that became more accessible for interested amateur users and triggered new bottom-up dynamics in the content making and management process (Sui et al., 2013).

However, the inclusion of the spatial component as an additional media in the mechanism of social networks has led to the need of a specific classification that only partially overlaps the previously presented VGI definition. *Social Media Geographic Information* (SMGI) can then be considered as any piece or collection of multimedia data or information with an explicit or implicit geographic reference collected through the social networking web or mobile applications (Campagna, 2016). This type of geographic information requires also more detailed consideration of the type of users and multimedia contents. For this reason, it should be observed in a different way from the one adopted for current VGI researches and projects.

1.2 Data and user characteristics

VGI and SMGI can be found in everyday aspects of daily routines. Smartphone applications of activity tracking such as Runtastic and Waze allow their user to create a profile, interact with other users and share information about running workout tracks or mobility and road traffic. Other platforms, instead, like OpenStreetMap (OSM) (whose web interface is shown in Figure 1) or Mapillary are devoted to mapping with innovative techniques which engage the data collectors and contributors. Eventually social networks like Twitter (Figure 2), Facebook, Instagram, Flickr and instant messaging services (e.g. WhatsApp, Telegram) enables their user to share location information with their contacts or network.

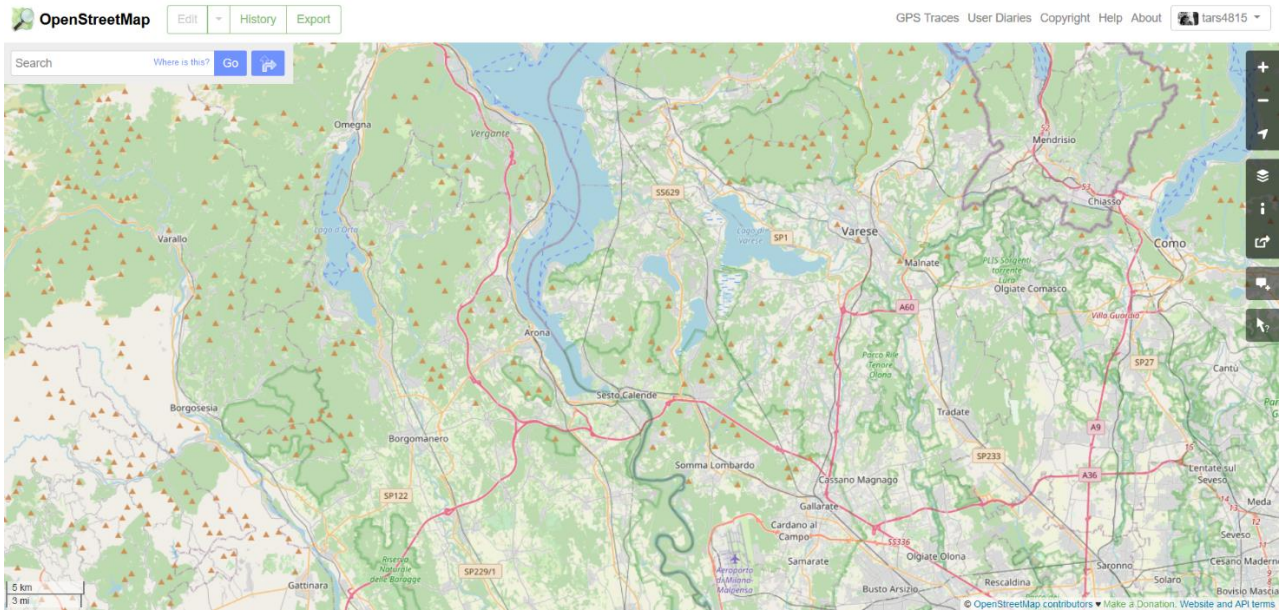

Figure 1 OpenStreetMap web interface

Figure 2 Example of a Twitter post with a geographic reference

The causes and origins of the VGI and SMGI diffusion clearly imply, as previously suggested, that these types of data have peculiarities that differ from the ones of traditional products. As part of Big Data and participative phenomena, non-canonical geographical information are generally characterised by complex, heterogeneous and scalable elements that influences datasets at different levels during the analysis, modelling and visualization especially when it comes to predictive application and decision making process. For this reason, VGI and SMGI can generally be described by the 4Vs model – *Volume*, *Variety*, *Velocity* and *Value* - that highlights the main peculiarities to be considered when dealing with Big Data (Chen et al., 2014).

The continuous and rapid generation of geographic information lead to big constantly growing amount of data, that requires new approach in terms of data collection and analysis. The amount of data referring to complex real-world phenomena is indeed constantly increasing requiring different approaches at different scales. Additionally, the heterogeneity of data is connected to a variety of content to be analysed. The data can come in form of single structured tables or data model, like OpenStreetMap or WikiMapia, to unstructured multiple objects deriving notably from mobile enabled social networks (e.g. Twitter, Flickr or Instagram). The fourth V included in the model refers to the huge significant value – technological but also economical – associated to these geospatial data that are usually hidden in a very low-density distribution.

Despite the very high variability of the data structuration level and content, it is possible to identify 3 key components of both VGI and SMGI (Capineri et al., 2016):

- A geographic reference should be included as the *spatial component*. Its accuracy should then be evaluated with specific methodologies and studies.
- A *thematic content* could be expressed in various forms (text, image, audio, video etc.) and generally depends on the objectives and motivations of a collaborative project or of a social network.
- A *user association tag* including information about profile settings and experience in order to ensure the recognition of intellectual property as part of a community project or network. Sensible information is usually restricted due to privacy issues.

These fundamental components are strictly linked to the characteristics and the profile of the single user, a human sensor that contributes to a specific project or community generating and sharing new information that could be re-used for different purposes. This double nature of the individual in the web environment requires the introduction of a new definition made possible by a community-based technology evolution and by the dissolution of boundaries between passive and active user approach. From the overlapping of the information *producer* and *user* roles derives the definition of *produser* (Coleman et al., 2009).

It is crucial to identify the main characteristics of a produser contribution. One key factor of individual information generation is naturally the personal motivation associated to the contribution (Budhathoki & Haythornthwaite, 2012). These can vary from altruism or special situation needs and communication to pride of place or protection of personal investments or interests too. Additionally, the evolving real-time nature of VGI and SMGI requires also to take into account the fact that contribution, reasons and goals of the same single produser could change in time and space.

The diffusion of automation processes also asks for a consideration of the probability of not-human nature activities linked to the presence of web bot [1] that are developed for automatically run scripts and tasks in the Internet and so possibly inside crowdsourced platforms and social networks too. This aspect should be carefully evaluated because it strongly affects the thematic content, introducing bias in common behaviour analysis. Usage frequency and reputation of the produser are other important

[1] A computer program that works automatically, especially one that search for, find and repost information on the web.

factors that are linked to the reliability and, consequently, to the representativeness of the shared geographic information.

The final key element of the user profile description is the level of awareness of his/her contribution. Indeed, VGI are usually part of volunteered project whose principal aim is the generation and the sharing of geographic information associated to specific content. This consciousness could not always be granted in the case of social media where the geographical coordinates or geotag sharing is not a core component of the process and sometimes can be applied without a complete knowledge of its dynamics.

1.3 VGI and SMGI research fields and applications

In the last years, the complexity and the potential of VGI and SMGI contents deriving from a collective technological, cultural, and scientific innovation increased the research interest. Current researches focused on VGI and SMGI mining and interpretation can be grouped by three areas of interest: event detection and monitoring, social and public event studies, and urban and mobility planning. Before giving an overview of these categories, it is necessary to point out that, in the majority of cases, this classification could be affected by overlapping of different groups, especially when a specific situation or phenomenon implies knowledge and skills belonging to separated study field.

Event detection and monitoring

As previously explained, one of the most valuable characteristics of VGI and SMGI is their immediate real-time component. This, especially in extreme scenarios like natural disasters or critical diseases and pandemics, provides useful data information about the position of affected people or the entity of damages and outbreaks.

One of the first studies analysing the Twitter messages related to natural disaster had focus to detect and monitor the 2008 Sichuan Earthquake (Li & Rao, 2010). The official media and agency coverage and the tweet activity have been compared emphasizing that the seism was detected by Twitter posts within seconds after the occurrence, almost 24 hours before it was reported by expert authorities like USGS. Considering disaster response and recovery, Haiti Earthquake in 2010 represented a turning point for collaborative project focusing in emergency scenarios. The community efforts along the entire rescue coordination and reconstruction journey have been reported and documented by OSM users' activities (Soden & Palen, 2014).

According to the researches, social media users are also more likely to share personal information in case of international pandemic. For example, Lampos et al. (2010) defined a flu score based on Twitter activity that was able to detect and predict on a spatial level disease breakout for selected cities in the UK. More recently, the behaviour of Dutch Twitter users has been monitored in order to evaluate, under a temporal perspective, the rate of activity in comparison with COVID-19 severity level, identifying increasing trends with respect to official emergency communication as shown in Figure 3 (Wang et al., 2020). The considered keywords were associated to the common name for COVID-19, the communication made by National Institute for Public Health and the Environment (whose acronym is "rivm") and the frequency of the term "mondkapje", Dutch word for "face mask".

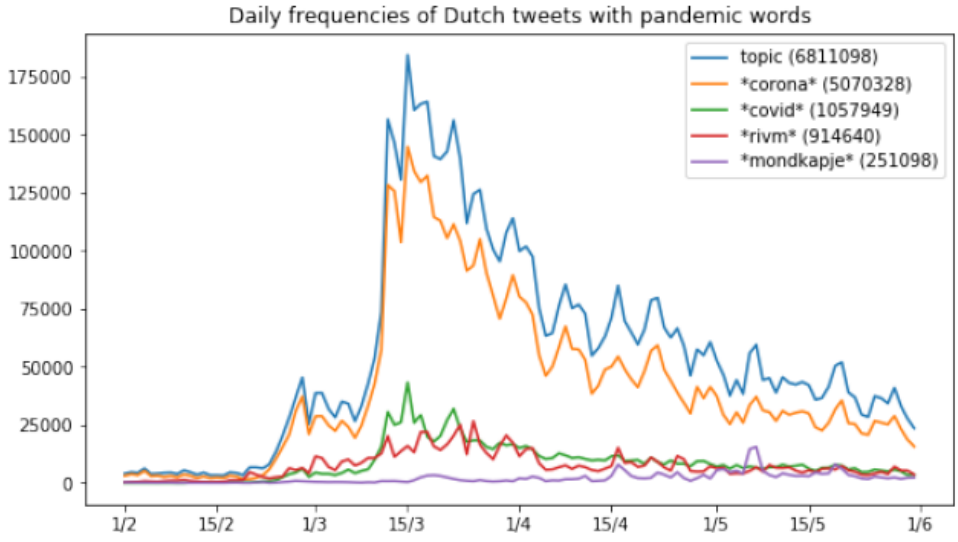

Figure 3 Frequencies of Dutch tweets associated to COVID-19 related topics (Wang et al., 2020)

Social and public event studies

SMGI are a valuable source of data regarding human behaviour and response in space and time. Considering disseminated tweets, many researches made studies about the proximity of users belonging to the same contact network and the spatial dimensions of the social communities (Roick & Heuser, 2013). Another important methodology applied in the case of social studies is the sentiment analysis approach that, combining skills and knowledge ranging from text processing techniques to computational linguistics, identify and interpret the population feelings and behaviours in time and space. In this way, VGI and social posts become proxies by which it is possible to measure the impact and success of a public events or manifestations. The text sentiment interpretation and the position could also be crucial in demographic predictive studies associated to political campaigns.

Considering only the spatial component, it is however possible to obtain a picture of the audience response to a public manifestation. For example, the movements inside the Expo 2015 event that took place in Milan (Italy) have been analysed using geo-statistical index calculations on Instagram data (Migliaccio et al., 2018), identifying distribution pattern and points of interest inside the exhibition area as shown in Figure 4. Additional work on this social media dataset has been done with a spatial-semantic approach for validation (Migliaccio et al., 2019) in which the semantic component has been used to validate the post position and its accuracy.

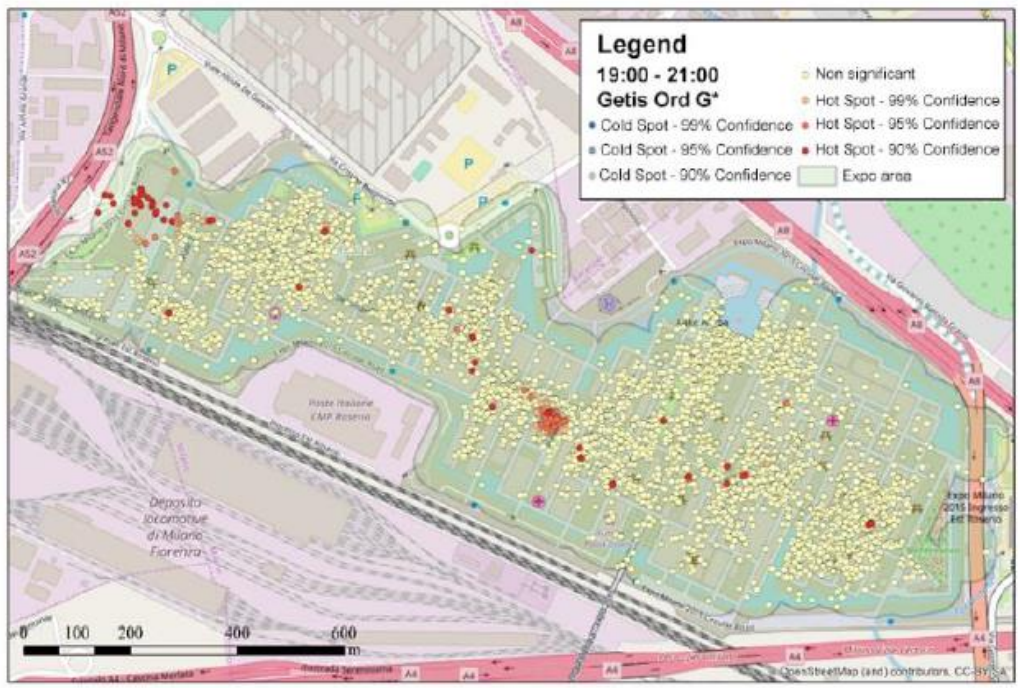

Figure 4 Hot-spots computed with the Getis Ord Gi index in the Expo 2015 Milano area (Migliaccio et al., 2018)*

Urban planning and mobility

Previously introduced applications and projects like Mapillary and Waze help in the dissemination of data but also in the collection of information about the evolving urban landscapes and dynamics. Multimedia social data can be key factor for investigating the structure of cities and the mobility models by analysing the distribution of human activities.

For example, through image sequence published on photographic LSBN like Flickr and Instagram it is possible to retrieve valuable information about citizens' movements inside the city, distinguishing also touristic pedestrian routes or weekdays/weekend differentiated patterns (Steiger et al., 2016). Social media contributions and VGI produced in the context of urban participatory projects could particularly emphasise specific Points of Interest or strategic places that require to be taken into

account when dealing with decision planning and making processes. This approach could make the difference especially in the context of Smart City, whose strategy is strongly based in the constant interconnection and information exchange about citizen experiences, feelings, and needs.

1.4 Possible issues

The variety of the produser community associated to VGI and SMGI dissemination implies valuable potentials but also causes significant possible bias and problems in many analyses.

Quality standards

The main issue to be considered is surely connected to the data quality. Official cartographic data are subjected to high quality standard requirements. The new geographic information, usually generated by non-experts or created as collateral products, sometimes do not even have metadata containing clear and validated information about the quality (Elwood et al., 2012). Considering the standard parameters defined by the International Standardization Organization (ISO) (accuracy, logical consistency, and completeness for all the spatial, temporal, and thematic components), accuracy is the most important issue for both VGI and SMGI. This is related to the fact that generally the geo-location is based on a manual feature positioning (by clicking on a map) or on GPS positioning through a mobile device whose accuracy depends on many technological characteristics. Additionally, the use of geotag as referencing elements implies some toponymy ambiguities: different places can have the same name, as different names could be attached to the same place (Capineri et al., 2016).

Representativeness

The rate and the frequency of VGI and SMGI production is strongly linked to the availability and the accessibility of an internet access. Urban areas, where wi-fi services can be considered highly stable and accessible, can generally be considered as hot-spots for data generation. Therefore, there is a spatial inequality in the geographic information generating activity, with a process unbalanced towards urban agglomeration in contrast with rural zones. Also, the presence of a more heterogeneous user population in big cities implies the combination of a wider variety of human behaviour and routines. The need of thinking of a sociological and anthropological approach to VGI and SMGI leads to another important bias to be identified inside all the different user categories. The so-called digital divide affects not only the spatial distribution of the users but also their age variety. LSBN users, indeed, are not equally distributed in all demographic classes and social backgrounds. Lastly, it is important to remember that both volunteered and social media geographic information models rely

on a basic assumption for which the individual contributions converge to a single unanimous agreement that often coincides with the truth (Elwood et al., 2012). However, the reputation of the individual produser is still a complex element that needs more consideration and further researches.

Data availability and privacy restrictions

VGI and SMGI usually include sensitive data about user preferences and habits. For this reason, usually the data retrieval and collection are partially restricted so that people can feel safer also on a legal basis. The disclosure of personalized geographic information on the web is often perceived as a threat to everyone's privacy. The possibility of exploiting personal spatial information is generally considered as a loss of control over data disseminated on the web. These threats can generally be summarized in three privacy categories (Roick & Heuser, 2013): location (associated to the knowledge of the user position), absence (knowing that a user is not on a specific place) and co-location (deriving the position of a user from his/her contact).

VGI systems and social networks are still investigating the more appropriate methods to ensure privacy protection based on user profile settings. Regarding this issue and possible solution, current cloaking and restriction of data collection applied by LSBN could reduce and limit the availability of valuable and representative information needed by future researches and studies.

1.5 Key elements in the comparison between VGI and SMGI

Considering all the factors and elements previously explained, it is now important to highlight the similarities and the differences between VGI and SMGI. Both types of data are characterized by the presence of a spatial, thematic and user components but generally SMGI can take advantage of a not null temporal field (usually associated to the content publishing date) and, in particular, of a so-called *interaction score* (Campagna et al., 2016). Connections, opinions and endorsement dynamics are core mechanisms for social networks and are represented by functionalities that could vary from platform to platform (e.g. *comments*, *likes* and *shares* for Facebook; *likes*, *reply*, *quotes* and *retweet* for Twitter etc.) depending on their targets and objectives. These fields, after detailed statistical explorations, could represent significative enriching elements in research contexts more focused on audience perceptions and reactions.

Another aspect that plays a crucial role in the comparison between VGI and SMGI is the structuration level of the considered participative project or social network. Generally, volunteered geographic information are indeed results of well-defined Citizen Science models whose objective and context

are approved and endorsed by all the contributors that, as part of an organised community, could also define guidelines and quality standard for availability and usability of data. On the other hand, social media users' primary motivations do not include the information creation and sharing for a single common objective. Additionally, SMGI – in other words, social media posts with a geographic reference – represent only a small portion of the total social contents. For example, in the case of Twitter, on average only about 3% of the posted tweets is directly georeferenced (Leetaru et al., 2013).

In the end, it is needed to highlight the main differences between VGI and SMGI contributors. For the first case, produsers are integrated in a fluid role system, in which amateur contributors through hands-on experience could improve their reputation inside the community and gain recognition as reference expert. Also, VGI projects like OpenStreetMap could identify expert members inside their community and define some validation teams whose responsibility is to ensure and validate data quality and coherence. Instead, SMGI users are always subjected to the ambiguous distinction between active and passive contributors (only upgraded developer accounts can directly access APIs after an often-complex request procedure). Geospatial quality check, is not a main priority for social communities whose primary goal, as previously explained, is simple user interactions and information exchange.

These key differences eventually identify the most important elements to be controlled and considered and the main complexities and issues that could affect the elaborations and the analysis of a SMGI case study as the one presented in this thesis work.

Chapter 2

Twitter data and studies related to natural disasters

Over the past years, many web users affected by the occurrence of certain natural disaster rely on the peculiarities of Twitter communication dynamics. It has been used as a valuable support in reporting damages or fatalities and to organize rescue operations, providing relevant information of the event during the first hours. Lately many emergency agencies are realizing the advantage of this social network to have preliminary evaluation of the impact of an event based on the users' activity (Hossman et al., 2011). However, the primary purpose of Twitter, created by Jack Dorsey, Noah Glass, Biz Stone and Evan Williams, was not to support disaster response operations. Born on March 21st, 2006, Twitter is a microblogging and social-networking platform that enables its users to communicate with each other through status update posts called *tweets* whose text is restricted to 280 characters. Nowadays, in every second around 6000 tweets are published. For the first quarter of 2019, Twitter recorded about 330 million active users per month (Statista, 2019). These are the key numbers and statistics of a social media network that, due to its straightforward nature, represents a valuable source that should be exploited when immediate and fast data are required.

2.1 Twitter objects and data description

To understand the usage flexibility of Twitter, it is necessary to explore the basic structure of a tweet and its linked elements inside a relational model as documented by the Twitter Developer official website. Inside the Twitter system, data are encoded using JavaScript Notation (JSON) based on key-value pairs describing the content attributes. The fundamental object is the *tweet,* containing all the information about the status update and the user who published it on Twitter. Four additional children objects are further defined: user, entities (referring to the tweet multimedia content), and places (defined by the post geolocation procedure). All the values of these last three objects are then included in the extended form of the parent tweet. In the end, all the key-value pairs are compressed into a single object whose graphical interface on the Twitter platform is depicted in Figure 5.

Figure 5 Example of the graphical interface of a tweet object

Twitter object

The simple tweet summarizes the four main components of this SMGI type: temporal reference, thematic content, user characteristics, and spatial element. Within its JSON structure, it embeds five main attributes:

- *ID* – included also as *ID_str* in the string representation – that is the unique integer value and primary key that identifies a specific tweet.
- *created_at*, a string field that includes the UTC datetime the tweet was created.
- *text* field representing the textual content of the status update in the Unicode Transformation Format 8 bit (UTF-8).
- *user* which includes all the key-value pairs defined on the dictionary object belonging to the user who published the tweet.
- *entities* that, referring to a specific Twitter entities object, lists all the multimedia and references associated to the tweet. In case of a tweet containing only a simple text, this element is null.
- a Geo object that could be either a *coordinates* field or a *place* geotag. In both cases, this object is nullable.

In addition to these attributes, the tweet object counts other elements associated to the nature of the post published. In the case of a retweet – user interaction for which the broadcast of a tweet can be amplified by another not-author user that shares the original tweet – or a quoted tweet – a retweet with an additional text comment, the tweet indeed includes other fields like *retweeted_status* (containing the text of the original retweeted post) or the boolean *is_quoted_status* (referring to the quoted nature of the content). Eventually, these components, the detected language of the tweet (*lang*) and the interaction rating attributes (*retweet_count*, *reply_count* and *favorite_count*) represent game-

changing thematic elements for studies that focus on the interactions between the users and the possible reliability or significance of a content.

Tweet user characteristics and thematic component

In a similar way, every registered Twitter user is then identified by an object with a total of 42 fields whose essential unalterable attributes are: the integer unique value *ID* (and its correspondent *ID_str*) and the *created_at* temporal attribute that the refers to the UTC time the user account was originally created.

All the other elements associated to the user object could be changed over the time and they depend on users' preferences. Considering these differences, other two attributes to be observed when identifying the user are *screen_name* and *name*. The first one is the nickname chosen by the user that can be recalled or tagged by other contacts inside tweet texts by typing it after "@". *Screen_name* should be unique inside the Twitter user database and should not include spaces. It could be changed at any time. *Name,* instead, corresponds to the name with which the user would like to be addressed on the platform. However, it is not necessarily the true name since legal person can be regularly associated to Twitter profile.

Additionally, other important user object attributes provide significant information about the rate of activities and interactions of the single profile, including the number of users followed by the user in order to keep track of friends or contacts and contents of interest (*friends_count*), the total number of people that keep track of the single user activity (*followers_count*) and the level of interaction with other users (*statuses_count* that is an updated field of the total number of post published and *favorites_count* which counts the number of contents with which the user interacted with a "favorite" reaction). The remaining fields of the user objects are then related to privacy preferences and profile appearances settings.

The *entities* object, instead, represents a valuable tweet component on a thematic level. Indeed, it includes in a JSON key-value format potentially significative elements such as *hashtags* (identified by the presence of a word preceded without space by the "#" symbol), *media* (reporting whether it is a photo, a video or an animated Graphics Interchange Format (GIF) file and indicating its size and url source), website *urls*, *user_mentions* (including the name and the ID of the user mentioned in the tweet) and *poll* (if a poll with multiple answer options is attached to the original tweet).

Twitter spatial component and methods for its geolocation

The last and, under many points of view, most complex tweet component to deal with is the spatial one. The official Twitter documentation refers to it as the Geo object, a Twitter element that has not a generally standardized definition. The geographic component, indeed, can refer to both a point and a polygonal feature. In some cases, both types of Geo object can be assigned to the tweet.

In the case of a point feature, a *coordinates* attribute is defined as a collection of float numbers with a couple of values referring to longitude and latitude with a World Geodetic System 1984 (WGS84) World Mercator projection. For this reason, the position linked to the tweet is always the user one retrieved by the GPS tracker of the device logged into Twitter when the content has been published.

On the opposite side, the *place* object is associated to a polygon feature and, in particular, to a bounding box already defined inside the Twitter places archive. This element is identified by a geotag, whose name can be either a geographical location (structured hierarchically: city, admin, country etc.) or a point of interest (for instance a shop, a restaurant etc.), associated to 4 couples of coordinates corresponding to the extreme values of the rectangular-shaped bounding box. This geotag can generally be manually selected within a list of options proposed to the user at the moment of the posting. In this case, the position of the user can be then inferred only approximating it with an average of the extreme longitude and latitude values.

The two possible geolocation methodologies finally suggest some crucial differences to be weighted in the context of spatial analysis. It can be generally said that to a *coordinates*-geolocated object corresponds a more accurate and significative spatial content while for a *place*-tagged spatial content the information reliability should be treated carefully because sometimes the indicated location could differ from its true position at the moment of publication. Also, if the point position of a *place*-located tweet is approximated with an average method, the accuracy and the representativeness of the element location could vary significantly with respect to the extension and the level of details of the geotag used.

Eventually, another useful element for spatial analyses could be represented by the *location* attribute associated to the user profile that, compared with the location associated to the single tweet, could give additional information about the user activity (local or tourist etc.). Anyway, it should be taken into account that the location indicated in the profile settings can be arbitrarily defined by the user itself and could include reliability and representativeness biases. All these observations are crucial elements for defining queries inside the Twitter archive.

2.2 How to access Twitter data

Twitter, through its Developer platform, allows interested users to access and search data. However, the official documentation and tools are available only for selected users who have completed a mandatory application procedure for a Twitter Developer account. This process is indeed required to prevent abuse of the social network platform, ensuring the protection of sensitive data with the application guidelines, and to help Twitter company to understand the needs of the developer community. During the process, the applicant is asked to declare the motivation (professional, hobbyist or academic) for the use of the developer tools. The user should describe with as much detail and accuracy as possible how Twitter data and users' details will be analysed and how tweet, *retweet*, *like* and *follow* functionalities will be integrated in a possible future project. The submitted application is reviewed and finalized positively if all the declarations follow the Twitter guidelines and restrictions.

Once the individual user account is upgraded to Developer one, it is possible to access Twitter tools as long as the searched and downloaded data are not entirely aggregated and published outside the social network environment. Additionally, every activity concerning sensitive user information dissemination and Off-Twitter matching (i.e. associating Twitter content with a natural personal if not explicitly expressed on the user's profile) do not comply with applicable laws and all parts of the Developer Agreement and Policy. Consequently, as reported on the restricted use cases documentation, it is possible to redistribute content obtained through Twitter tools with another party only by sharing tweets or user IDs which then the end user can rehydrate (i.e. request the full tweet or user content with all its attributes through specific Twitter application). In this way, interested Developer users should always make reference to Twitter official tools for obtaining complete information.

To begin the tweets searching and downloading procedure after the account upgrade it is necessary to create a client application, defining its name and declaring its aim. This step allows the registered user to obtain the 4 parameters (*consumer_key*, *consumer_secret*, *access_key* and *access_secret*) required to access Twitter private account information through OAuth 1.0a [2]authentication method. To each authorized user may correspond many different applications.

After all these procedures it is then possible to access to Twitter Application Programming Interfaces (APIs) tools handling significative amount of both historical and real-time data. However, Twitter

[2] OAuth 1.0a is an open standard for access delegation, commonly used as a way for Internet users to grant websites or applications access to their information on other websites but without giving them the passwords.

enables different levels of searchability to its data offering three different groups of APIs whose potentials vary from limited to complete information access, implying also different costs: Standard, Premium and Enterprise APIs (Figure 6).

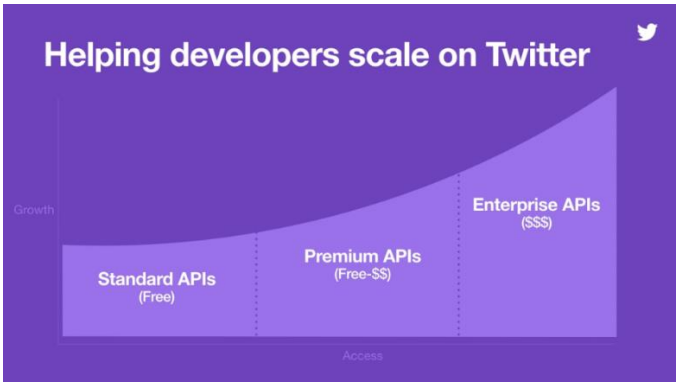

Figure 6 Twitter API different characteristics.

Only the first type of APIs is completely free but with important restrictions that should be considered especially when dealing with SMGI representativeness. Hence, Standard APIs are affected by restriction for the two groups of possible data request:

- POST group that consists in a submit, change, or destroy data method applied in the Twitter environment and the user or the application is writing into the system. This request group, for example, includes the possibility of automatic interaction by posting tweets, retweeting, following a specific user or reacting with a "favorite" to a specific content.
- GET group that consists in retrieving data and reading information from Twitter. Some possible GET operations are obtaining the entire status timeline of a specific user, retrieving the complete following list of a user, and searching tweets by filtering Twitter activity with specific keywords.

To each group correspond a different request rate limit. In the case of free Standard APIs POST procedure limits are defined by the generic Twitter user activities, for example every user/application can make a maximum of 300 status posting request every 3 hours (2400 daily tweets). The limits of the GET request differ for user and for application and are restricted for the time of 15 minutes. Additionally, every specific GET request has a different limit. For example, the GET search/tweets operation allows a total of 450 requests per application whilst for each single user it is possible to make a maximum of 180 requests. Also, each request, if maximized, corresponds to 100 tweets as possible results of a desired query. This means that with an application every day it is possible to

obtain a maximum of 4320000 tweets[3]. Nonetheless, it is important to notice that this is just a permitted amount of a query filtering result which includes a maximum tweet number that could be greater or lesser than the allowed one.

The GET search is also strongly affected by the used APIs that influence both the Twitter archive supported history and data fidelity. Indeed, Standard APIs search permits a sampling of recent tweets published in the past 7 days and the collected sample is not necessarily complete because some tweets may be randomly skipped from the search. As mentioned in Twitter Developer documentation, standard search API is "focused on relevance and not completeness". Extensions to these limitations can be obtained only through paid Premium or Enterprise Twitter APIs that can ensure a full data fidelity and extend supported searchable history from the standard 7 days to 30 days or to full archive (tweets from as early as 2006, Twitter launch datetime). All these restrictions imply data availability and completeness issues.

2.3 Retrieving Twitter geolocated data

Filtering tweets by location is a crucial step to obtain an SMGI dataset for geo-statistical analyses and studies. This procedure is made possible by some libraries (both official and community-supported) that cover the Twitter API across several programming languages and platform (JavaScript, Node.js, Python, R and Ruby are the most common tools). Tweepy is the most popular and used Python community-supported library because it easily integrates the use of Standard APIs, enabling large groups of developers and interested users to access Twitter data. Particularly, it helps defining the two main location queries used to obtain geolocated tweets in a Python environment.

Real-time location filter

If the aim of a study is to obtain real-time tweets and information about social user activities and behaviours, a Tweepy streaming module is a recommended tool. The Twitter streaming API is indeed used to download tweets in real time and to obtain a high volume of data, ensuring a significant value in term of immediacy of event response but also regarding event tracking in time and space. This is possible by building a *StreamListener* class and consequently a *Stream* object that establishes a session through APIs, receives Twitter data and processes them according to the function or filter applied. Even if the Tweepy Stream structure is easy to be defined block by block, it is important to

[3] This is the results of the following calculation: $24 * 4 * 450$ $(requests\ per\ 15 - minutes\ window) *$ 100 $(maximum\ searchable\ tweets\ per\ request)$

remember that Standard APIs do not allow to concatenate a location and a keyword filter in a single command. For this reason, it is necessary to build a double filter as the following example:

```
class StdOutListener(StreamListener):
    def on_data(self, data):
        if 'keyword' in data
            with open('tweetsfile.json', 'a') as tf:
                tf.write(data)
            return True
    def on_error(self, status):
        print(status)
if __name__ == '__main__':
    l = StdOutListener()
    auth = OAuthHandler(consumer_key, consumer_secret)
    auth.set_access_token(access_token, access_token_secret)
    stream = Stream(auth, l)
    stream.filter(locations=[‘lon_01’,’lat_01’,’lon_02’,’lat_02’])
```

In the stream filter it is defined the bounding box (whose first two longitude and latitude coordinates correspond to the bottom left corner whilst the others identify the top right corner) inside which the desired tweets should be searched. Hence, this means that Tweepy does not support multi-sided or disjointed geographical rectangular areas. In this way the stream object will look only for tweets whose *coordinates* or *place* field values are within the input GPS coordinates rectangle. The *keyword* filter is instead defined inside the *StreamListener* class in order to write only desired tweets on the output file that in this case is a JSON format one. If any error (syntax, rate limit exceedance etc.) is encountered, the called listener class will print the error code.

When dealing with bounding boxes and keyword definitions, it is fundamental to understand the desired level of details. A wider bounding box, indeed, implies a bigger amount of Twitter data, but it probably includes more meaningless SMGI that may not be directly affected by the event that a specific study is investigating. These issues could be solved by the application of a second filter selecting tweets that only contain chosen keywords that could be both simple but representative words and hashtags for a given research scenario. So, a list of words could be also used as a filter defined inside the if statement with multiple condition through boolean operators. Clearly, the choice of the filtering words should also consider the influence of the word frequency and the filtering time-window. Considering all these possible issues, Figure 7 shows the filtering results for different query parameters (keyword, Bounding Box extension, streaming time, and streaming daytime) obtained on August 12th and 13th, 2020.

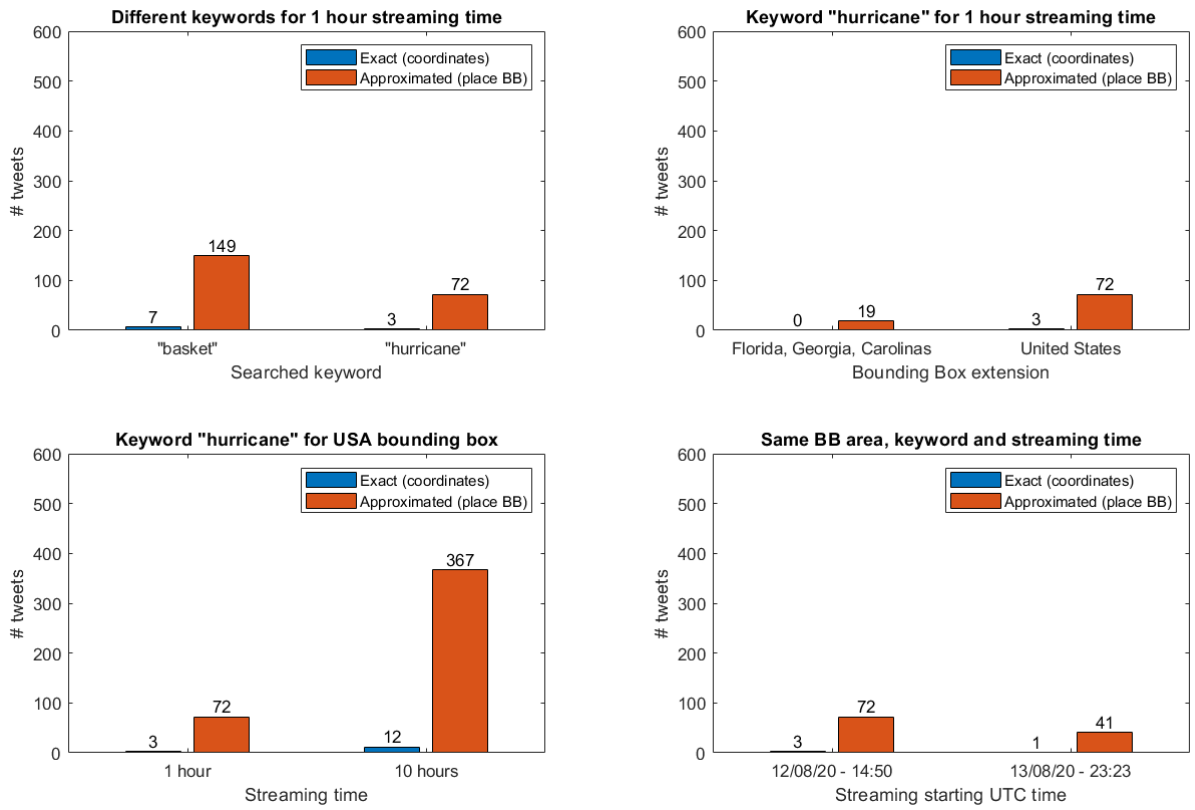

Figure 7 Different composition of real-time tweets depending on keyword, bounding box, streaming time and starting time for a query.

All the resulting tweets are geolocated through the previously described Geo objects (*coordinates* or *place* attribute). For all four examples, the percentage of geolocated Twitter posts obtained with the exact method is always less than 5%. The significative fluctuation of the filtered tweets is an evidence of the crucial keyword, space, and time dependence of the real-time query procedure. For instance, the word ‘basket’, considered as a common noun representing an everyday activity or popular hobby, is associated to a greater number of filtered tweets in comparison with the ‘hurricane’ term, associated to rare severe weather events. However, this keyword trend could change at different temporal or spatial scales, also for the same word filter. In order to prove this, a 1-hour ‘hurricane’ filter has been applied to different geographical bounding boxes: the first one includes the United States territory (except Alaska and Hawaii states) and the second one contains Florida, Georgia, South and North Carolina, US states that are usually impacted during the Atlantic hurricane season. The results showed a less intense activity in the second smaller area: the filtered tweets were one quarter of the ones filtered within the United States bounding box.

On a temporal level, different streaming time and datetime highlighted that, on average, the social activity rate is not constant over the time. This is clearly visible in the comparison of the 1-hour filter

(with a total of 75 tweets) with the 10-hours one (approximately 38 tweets/hour), both performed on the same day (August 12th, 2020) with the same keyword. Moreover, the comparison between 1-hour filtering for different days (August 12th and 13th, 2020) has shown a significant difference on activity rate.

In conclusion, it is important to mention that a combined keyword and position filter does not guarantee the significance or the representativeness of some tweets that could include the searched keyword in different contexts from the scenario of interest. For example, the word 'hurricane' could be associated to popular idioms, around Miami, to the Miami Hurricanes, a Floridian football team. The keyword misleading issue could then be avoided adopting a relevant word list that may help to focus even more the real-time query on the desired scenario. Additionally, multiple word translations should be considered when studying areas with people speaking different languages (e.g. border areas). All these issues remark the complexity of real-time filtering but also enhance the potentials of this method for event detection and live tracking through Twitter activity fluctuations or trends.

Historical data filtering by location

The previous observations about the geographical area definition and the keywords choice are valid also when filtering for historical data. Nonetheless, in this case there are even more significative issues due to the Twitter Standard APIs restrictions. A developer account, indeed, with a basic free authentication permission could navigate through data published only during the last 7 days. Consequently, the moment of filtering strongly affects the procedure, limiting in time the available SMGI dataset that could be even more incomplete.

The historical data filtering in Tweepy is supported by a *Cursor* object that iterates through timelines, user lists and tweets based on specific input filtering parameters. An example of Tweepy Cursor is the following:

```
public_tweets = tweepy.Cursor(api.search, count=100, q="hurricane -filter:retweets",
geocode="25.761681,-80.191788, 1000km", since="2020-08-05").items()
```

This example of a Tweepy *Cursor* module shows some key parameters for filtering by location and keyword with the maximum of 100 tweets per request. The *q* element defines the search query that includes the keyword to look for in the tweets and has an additional filter that ignores retweeted posts. The location filter, instead, is represented by the *geocode* parameter which returns tweets located within a given radius (in kilometres or miles) calculated from a point with given latitude and longitude ("latitude, longitude, radius").

However, the main issue derived from this geo-filtering method is that, when activated, the Standard search API will first attempt to find posts that have exact *coordinates* within the queried *geocode*, then it will look for tweets with a not null *place* attribute. If none of the two attributes is found, the Tweepy *Cursor*, as mentioned both on Tweepy and Twitter Developer documentations, will search for tweets posted by users whose *profile location* can be geocoded reversely into a couple of coordinates within the queried geocode. Hence, this priority searching procedure may result in a downloaded SMGI dataset that includes tweets without latitude and longitude information and with a detected profile location that could have serious liability issues because it can be arbitrarily assigned. Eventually, the temporal component of the query is included in the *since* parameter that define the date time (located maximum 7 days before the search day) after which the tweets should be searched. A narrow search interval could also be defined using *since_id* (starting tweet ID for the query) and *max_id* (equal to the last and most recent post ID to be searched): in this way it is possible to identify in a more detailed way – using implicitly a daytime filter - the time window.

The most sensitive element for the query is the search radius that defines the width of the circular *geocode* area. For example, for a fixed point of coordinates 25.761681, -80.191788 (corresponding to the city of Miami, Florida) changing values of radius (10, 100 and 1000 km) have been adopted in order to evaluate how the total number of resulting tweets changes. In the Figures 8 and 9 are illustrated the results of the filtering by location and keyword 'hurricane' for tweets published between August 5th and 12th, 2020, a period of time that followed the occurrence of the Isaias hurricane in the US gulf coast.

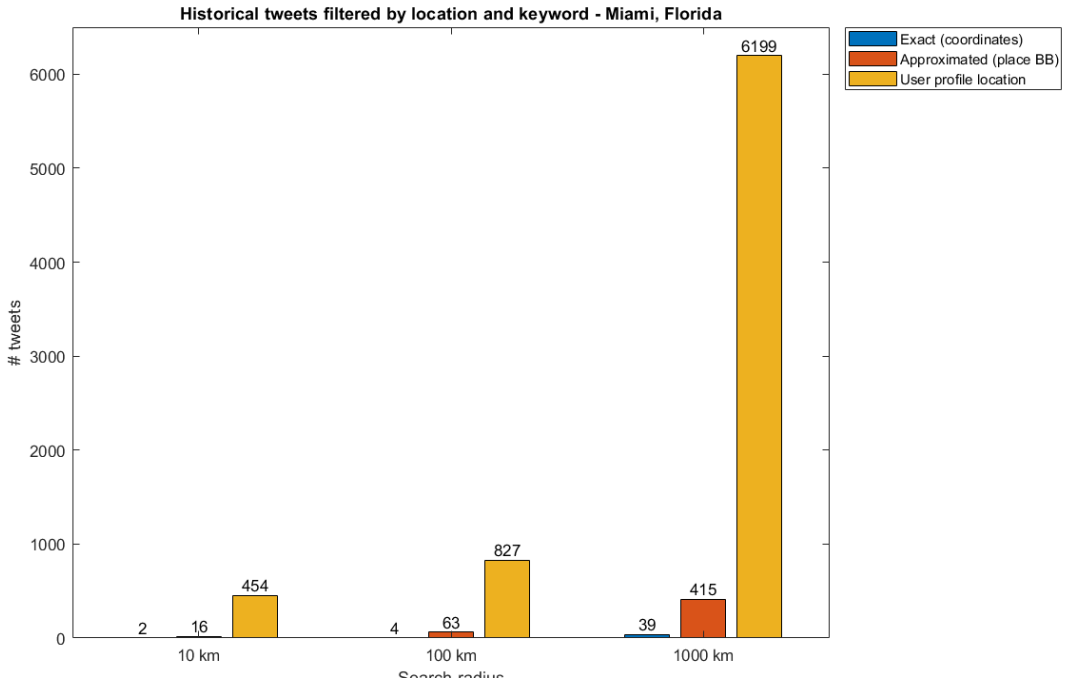

Figure 8 Historical tweet typology depending on search radius

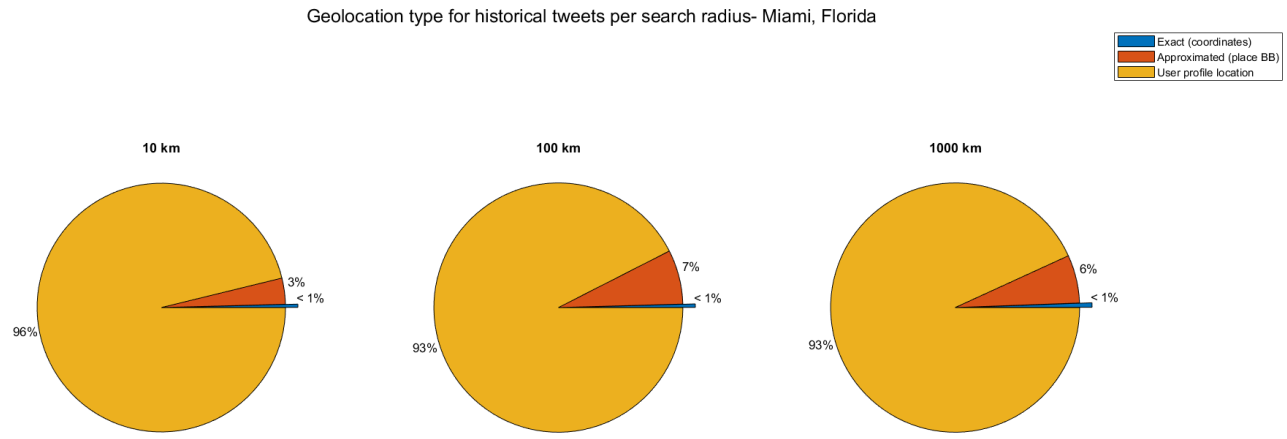

Figure 9 Composition percentage for historical tweets depending on search radius

The geolocation type percentages suggest that the proportion between geolocated posts and tweets whose position is inferred by the user profile location can be considered approximately constant with search radius, with geolocated contents that are always lower than 10%. Hence, these observations reflect the main issues associated to historical data retrieval with Standard API: the majority of the filtered posts include approximated spatial information inherited from a user profile attribute that could be ambiguous and not reliable. Additionally, it has been investigated the type of users who published the filtered SMGI. In Figure 10 the percentages of geo-enabled users (profile who activated the location sharing functionalities) are illustrated depending on the search radius.

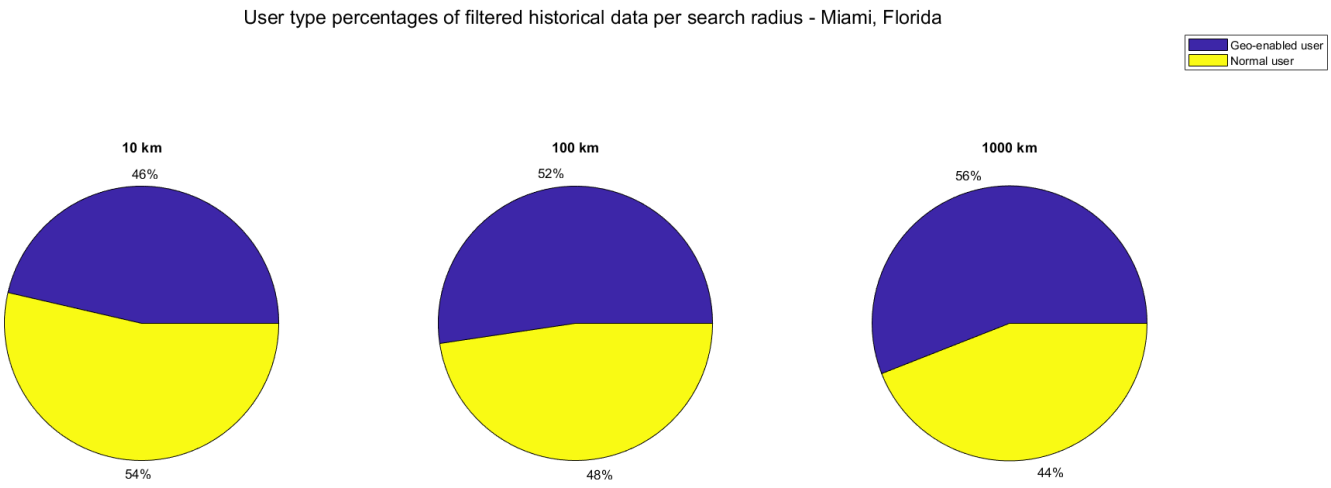

Figure 10 Type of users associated to historical tweet activities depending on search radius

The pie charts highlighted the evidence that an increase of the searched area corresponds to a greater number of tweets published by geo-enabled users. However, for this example, the change does not imply an increase of more precise geolocated tweets. This evidence could be motivated by the fact that users who have activated the positioning functionality were not interested in sharing explicitly their position because they were not directly affected by the searched topic.

On the other hand, a simple keyword filtering on historical data have shown other significative results consistent with the Twitter documentation in particular only a small part of tweets is geolocated (less than 5%). The performed query was defined by a list of searchable keywords ('terremoto', 'temblor', 'sismo', 'Oaxaca', 'Alerta Sismica', '#TenemosSismo' and '#TemblorCDMX') considered representative of the occurrence of an earthquake in a Spanish-speaking area. The tested case is represented by the magnitude 7.4 event recorded in the Oaxaca state (Mexico) on June 23rd, 2020. Additionally, the search time-window has been defined with an initial daytime (through the *max_id* parameter) corresponding to the publishing time of the first tweet alert written by AlertaSismica SASMEX, the Twitter account for official Mexican alerting system, and a window-length equal to 16 hours. The *Cursor* searched and downloaded 1617 tweets without any location query parameter. The pie chart in Figure 11 shows the proportions between geolocated tweets (with *coordinates* or *place*) and un-referenced posts returned by the Tweepy Cursor.

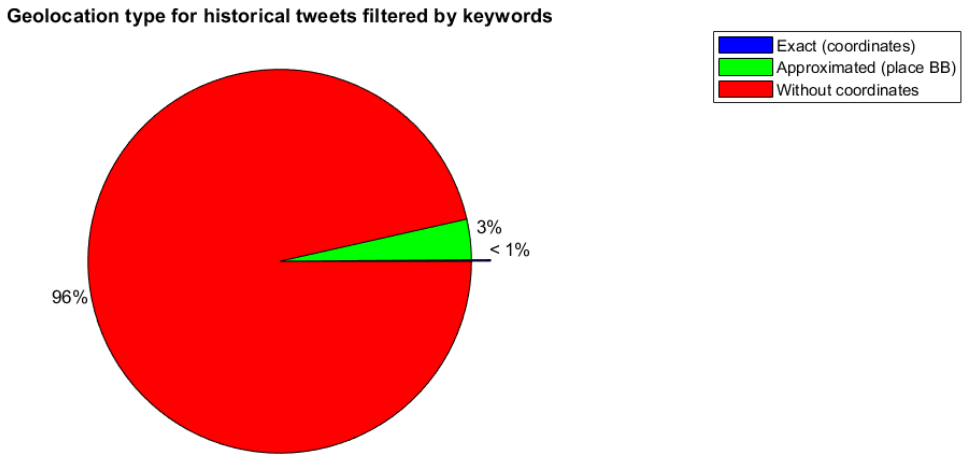

Figure 11 Dataset composition for Oaxaca earthquake test for historical tweets

The obtained dataset is mainly composed by tweets without any geographical information, resulting in a geolocated portion that is less than 5% of the total. Additionally, the georeferenced tweet is not necessarily located in the area affected by the seismic event. Indeed, for this case, in the Tweepy Cursor the spatial filter module (a fixed bounding box) was not applied. For this reason, the dataset could include SMGI from other regions of the world. Instead, regarding the keywords list

effectiveness, it could be said that the chosen words determined a strong filter for the query, considering that Spanish language has been detected in the text of 1599 out of 1617 tweets.

Finally, it is needed to remark that historical data, even with all the restriction associated to the use of Standard APIs, embed valuable additional information that could be missed in the case of real-time data. Indeed, old tweets have also recorded social interactions (number of 'favorite' reaction or retweet associated to a specific content) that would not be available if the posts were collected in real time.

Text filtering and geocoding

The third and last option for retrieving tweets with location information relies on the potential of semantic and textual analysis of SMGI content. The most common procedure suggested also in the official Twitter documentation, is to choose specific keywords for a query and then identify inside the textual component the presence of addresses or toponyms. Once the location names are detected, a geocoder uses them as input text and returns the corresponding coordinate. Geocoding algorithms are often time-expensive and require more sophisticated computing and programming abilities. What so ever, this method could solve the problem of small insignificant amount of geotagged tweets (as previously mentioned and supported also by Twitter documentation, the percentage of geolocated posts usually varies from 1 to 2% of the total social contents).

2.4 Twitter SMGI research in natural disaster

Thanks to its characteristics, Twitter has been widely used as a valuable data source in crisis scenario. As mentioned before when introducing SMGI, the information and the observations shared by users could provide important insights about the impact of a natural disaster on the population and on the surrounding damaged environment. Indeed, Twitter can be considered as a platform where events, technologies and emotions move and develop, and no net division exist between news, rational reporting, and personal reactions. Consequently, this complex background requires methodologies and workflows able to understand whether data are representative or simple noises. Analyses should then consider the tremendous potentials and the sophisticated challenges offered by a specific crisis scenario, combining them with the issues associated to the user population variety.

As suggested by Harrison & Johnson, 2016, researches of SMGI in natural disasters must be classified clearly, understanding the event typology (earthquakes, floods, hurricane etc.), its duration and frequency, and the concerned emergency management phase (mitigation, preparedness, response,

recovery as shown in Figure 12). All this is necessary to integrate SMGI instruments and studies in future Disaster Risk Reduction (DRR) frameworks (Kankanamge et al., 2019). Therefore, these operations have to consider that social media data usually provide valuable information on the first instances of a crisis response, while the impacts of an event often requires years, sometimes decades, to be fully evaluated on the human ecosystem.

Figure 12 The Disaster Management Cycle and its four phases (Harrison & Johnson, 2016)

Every analysis needs also to observe carefully social network information, attempting to define and continuously improve workflow, preventing ambiguous data from influencing results but also looking for new ways to deal with ethical and privacy issues. For instance, a crucial challenge regarding data representativeness is related inevitably to geographical and demographic biases, because Twitter use is still strictly linked to younger and populated areas' user groups that can take advantage of stable internet connections and significative smartphone and handheld devices penetration. So, the oldest and generally most vulnerable communities are often the least likely to be represented by big crisis data (Crawford & Finn, 2015). Additionally, it is important to remark that the Twitter activity is not necessarily the result of human user population: in fact, a large number of tweets is produced and disseminated automatically by bots, non-human developed agents that algorithmically and autonomously manage every action and interaction associated to a Twitter account. Crisis datasets, implying usually significative mass media coverage, could be affected by this automatic behaviour that recirculates and amplifies the most dramatic images and updates.

Recent studies focused on different types of natural disasters have tried to solve the presented issues, suggesting specific workflows and operations to ensure data quality, accuracy and representativeness often through the comparison of SMGI with Authoritative Geographic Information (AGI) – provided by topographic surveys – or with remote sensing open data (e.g. Sentinel, NASA Earth Data etc.) in the context of multidisciplinary studies (Klonner et al., 2016)

Earthquakes detection and early warning through Twitter activity

The unpredictability and immediacy of seismic processes gained the attention of the first researches on SMGI in the context of natural disasters. Inside crisis Twitter datasets, indeed, a spike in posts or an increase of particular hashtags or keyword trends could suggest the occurrence of the event. For example, the potential of Twitter data for earthquake detection have been assessed by exploring the tweets generated after the 2009 Morgan Hill seismic event in California (Earle et al., 2011). The dataset, filtered by 'earthquake' keyword and composed only by geocoded posts, suggested a potential detection in less than 30 seconds, implying that Twitter is potentially faster than the usual USGS notification lag (1,5 to 20 minutes). Also, the study found evidences of a rough coincidence between the area affected by the highest values of ground motion and the spatial distribution of tweets (Figure 13). However, the lack of a good percentage of geolocated content was considered as a strong limitation for any representativeness evaluation.

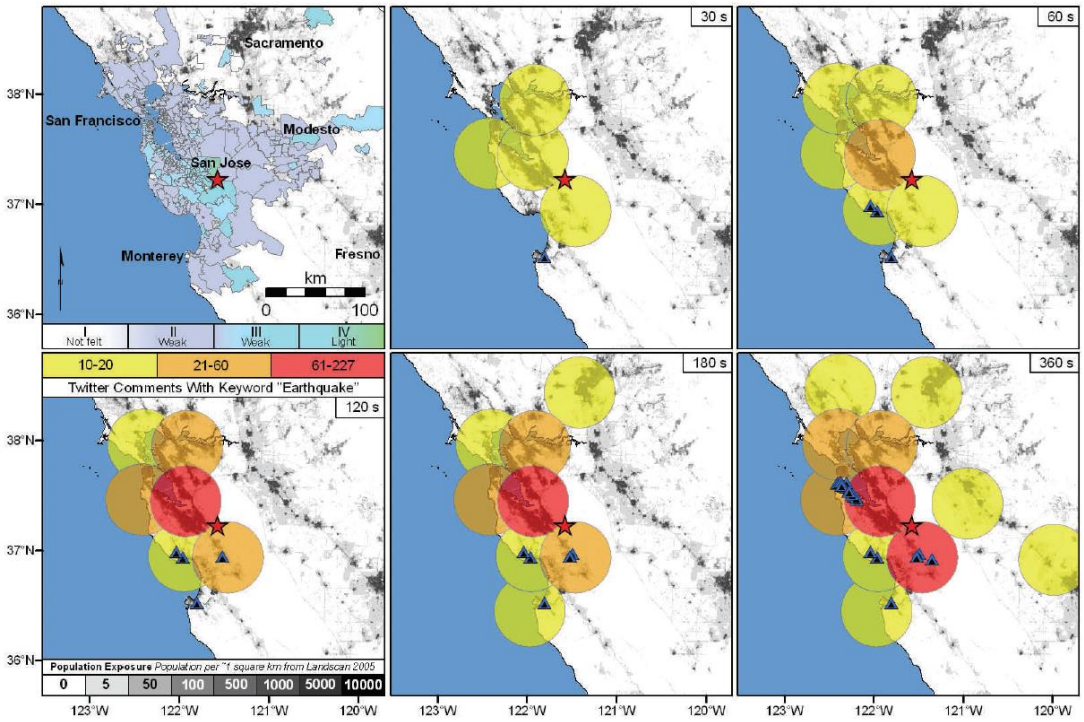

Figure 13 Spatio-temporal comparison of Twitter post and USGS earthquake data with different panel referring to discrete times after the earthquake as indicated in the upper right corner of the map. The red star represents the epicenter and the tweets with exact latitude and longitude geo-references are shown as black triangles with blue outlines (Earle et al., 2011)

Another study focused on the 2011 Great Tohoku earthquake and tsunami (Acar & Muraki, 2011) highlighted the presence of high number of retweeted posts causing misleading data dissemination and interpretation. Additionally, this case study highlighted the involvement of social media users located in both directly and indirectly hit areas. The researchers noted after a textual analysis that the first group of users shared tweets mainly about safety updates and damage reports whilst the indirectly

involved users were more interested on opinion-related comments about earthquake secondary effects (transportation, nuclear plant risks etc.).

The definition of automatic processes that filter and analyses tweets with streaming techniques could be a massive revolution for the traditional alerting procedures. For instance, a real-time Web system for the detection and monitoring of emergency situations structured as illustrated in Figure 14. based also on keyword filtering and machine learning techniques, have been applied to a 70-days window in 2013 in Italy, detecting the 75% of seismic events with magnitude greater than 3,5 reported by Istituto Nazionale di Geofisica e Vulcanologia (INGV) official seismographs network (Avvenuti et al., 2015). Despite the promising results, potential Natural Language Processing tools for enriching the small fraction of geolocated tweets still have to be investigated in order to obtain a finer-grain SMGI crisis mapping able to further helps decision makers during emergency response and recovery phases, especially for early warning procedures.

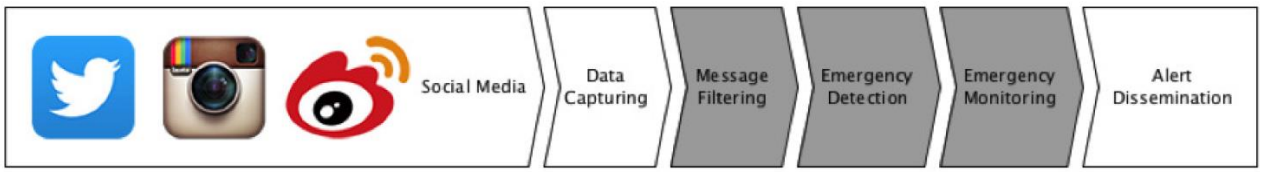

Figure 14 Emergency system architecture (Avvenuti et al., 2015)

Twitter data patterns for hurricane preparedness and response

Considering the huge economic impact of disasters occurring every year during the Atlantic and Pacific hurricane seasons, hurricane events represent another significative study and research field for SMGI. One of the major severe weather events that attracted the attention of many researchers was the hurricane Sandy, occurred during autumn 2012. It made landfall in the USA near Brigantine, New Jersey, causing significant flood in the New York metropolitan areas and gaining lots of attention on social media. Indeed, Twitter via Reuters reported a new record for event media activity with a total of about 20 million tweets about the storm. This great amount of social posts about the event (either geotagged or not) represented a valuable source for researches focused mainly on pre-event preparation and on disaster response.

Deviations from normal behaviours in Twitter rates combined with other social network sources (e.g. Foursquare) could be analysed in order to expose patterns of human activities during normal and emergency situations, detecting significant changes in pattern over the time. At the time of hurricane Sandy in the New York area a peak has been recorded for the tweets associated to grocery and shopping activities right before the forecasted arrival (Grinberg et al., 2013) (Figure 15). The relation between Twitter users and the area affected by a disaster could be further explored with combined

approach, exploiting both the spatial and the emotional component of SMGI. For example, this is possible by implementing geo-statistical analyses (mainly based on spatial centrographic measures) with textual recognition and classification (sentiment analysis). This procedure called sentiment geo-mapping has been applied to hurricane Sandy geo-located dataset, detecting the affected areas with sufficient accuracy (Caragea et al., 2014).

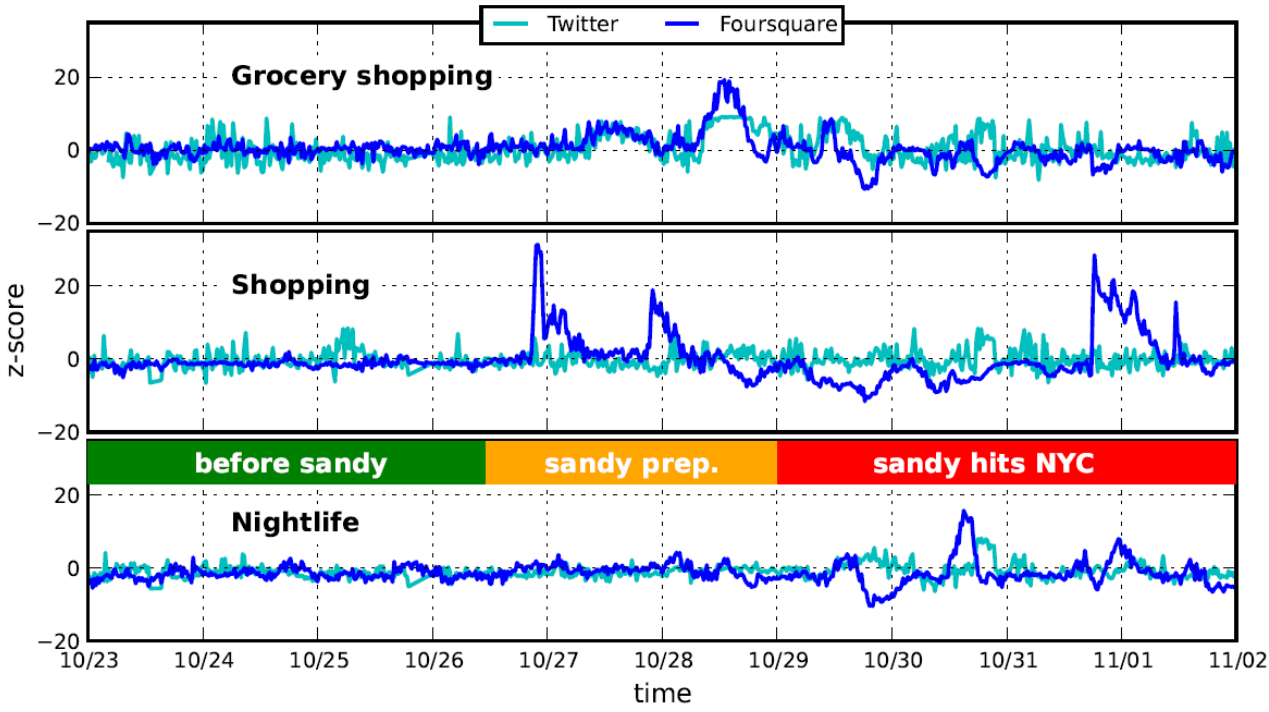

Figure 15 Pattern disruption on Foursquare and Twitter during Hurricane Sandy (Grinberg et al., 2013)

On the other hand, the spatial and multiscale dimensions of geo-social Twitter datasets could be a key element in damaged detection. For the hurricane Sandy event, the most affected areas were detected by Twitter activities and compared with the official source Federal Emergency Management Agency (FEMA). The results have shown good level of mapped damage in regional scale/level. However, at the urban level, reported damage demonstrated to be hardly detectable due to the strong influences of socio-demographic aspects (e.g. wealthier neighbourhoods could be overrepresented) (Shelton et al., 2014). A spatiotemporal approach on a Twitter dataset derived from a content-relevance analysis, has been taken for the case study of hurricane Michael that hit Florida, USA, in 2018 (Spasenovic et al., 2019). Geo-statistical tools such as Kernel density map and hot-spot analysis have been adopted for comparing relevant tweet distributions with the hurricane path documented by NOAA. Obtained results have shown the good convergency of social media data with the recorded landfall area both in space and time.

Floods and wildfires spatiotemporal evolution for situational awareness

The last two types of natural disasters that need to be treated individually are wildfires and floods. In comparison with earthquakes and hurricanes, these crisis events are generally characterized by consequences and damages at a smaller scale with a significative evolution in time. Researches on wildfires and floods again refers to the assumption that – even without severe filtering procedures - SMGI, posted during crisis situation and located near burnt or flooded areas, are more likely to be related to the events themselves as an extension of Tobler's first law of geography (Tobler, 1970 and Albuquerque et al., 2015).

In the context of 2014 wildfire, SMGI researchers highlighted the temporal coherence between official alerting timing and Twitter activity peak in the San Diego county (Wang et al., 2016). However, spatial analyses on tweets distribution and density identified important influence pattern of "gatekeepers" account (popular users or opinion leaders from whom the public acquires information e.g. local media or news reporters) whose activity could make more significant some social content – even if located away from the impacted area – only because of a strong social interacting rates (retweets or favourite reactions). The effects of this bias could be reduced by integrating a network analysis on connections between retweeting users.

Studies on floods have been conducted mainly comparing the results of flood peak propagation models with the spatio-temporal distribution of tweets classified by their textual content, combining both hydrological, geographical and sociological approaches. Referring to the 2013 River Elbe flood in Germany, researchers found out that classifying SMGI by their textual content (Flood level, volunteer actions for disaster response, media documentation, traffic conditions and other), as shown in Figure 16 the correlation between the position of flooded area along the river path and the distribution of disaster-related geolocated tweets is significative while media-related contents follow a different pattern driven by major cities with higher distances from the damaged regions (Herfort et al., 2014).

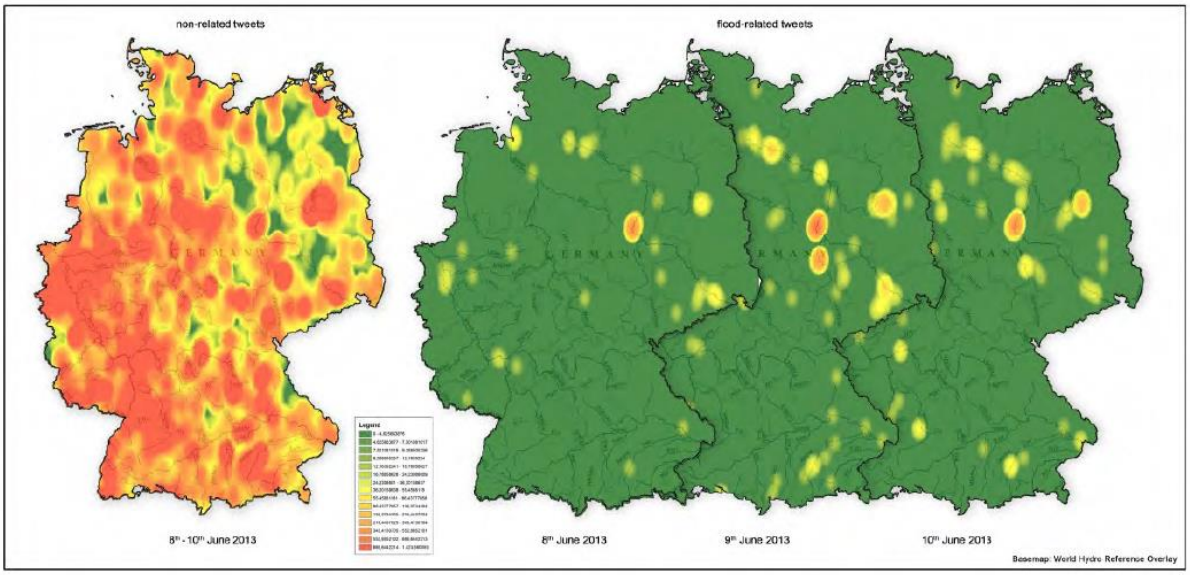

Figure 16 Spatio-temporal distribution of tweets for the River Elbe case study (Herfort et al., 2014)

In conclusion, SMGI studies and possible future automatic applications for DRR should take care about the main issues associated to the use and integration of this data. Spatial scale could represent a crucial aspect for research findings as much as the workflow definition (data filtering and statistical approaches). Additionally, the human component of the dataset and its network dedication should be considered with attention. It is valuable source of society dynamics evolution during emergency situations and possible source of misleading effect in data interpretation and relevance detection.

Chapter 3

Hurricane Florence case study: tweets retrieval and pre-processing

Current researches have demonstrated the potential and limitations of the use of Twitter SMGI for natural disasters. The constantly growing amount of social media data brings new challenges on the extension and applicability of tools used for specific case studies. With focus to explore more this topic, the thesis work analyses Twitter dataset associated to the Hurricane Florence of 2018. The case study has been chosen because of the great availability of hurricane-related tweets and because of the completeness and open access of report, data and information published by US agencies (NOAA and NWS). This chapter describes the main characteristics of the event and Twitter dataset used for the analyses. Additionally, observation and significative insights of user behaviours are already illustrated with the first filtering and processing procedures in Chapter 3.3.

3.1 Hurricane Florence: case study presentation and tools used

As mentioned in the NOAA-NWS report (Stewart & Berg, 2019), Florence was the second major hurricane of the 2018 Atlantic season – started on May 25th and ended on October 31st - for both damage costs and deaths. Originated on August 30th from a convective tropical wave south-east of Cabo Verde Islands, it was a long-lived, category 4 hurricane according to the Saffir -Simpson Wind Scale (Table 1), causing 22 direct deaths and 30 indirect fatalities in the United States.

Table 1 Saffir-Simpson Wind Scale (Saffir, 1978)

Category	Sustained Winds	Types of Damage
1	119-153 km/h	Very dangerous winds will produce some damage
2	154-177 km/h	Extremely dangerous winds will cause extensive damage
3 (major)	178-208 km/h	Devastating damage will occur
4 (major)	209-251 km/h	Catastrophic damage will occur
5 (major)	252 km/h or higher	Catastrophic damage will occur

The NOAA "best track" chart of Florence path is given in Figure 17 with the changes on its classification. It firstly was classified as a hurricane on September 4th until it was downgraded to tropical storm on September 7th. Then Florence was again graded as a category 4 hurricane on September 9th, reaching its peak wind velocity of 241 km/h (130 kt) and its lowest pressure value of 937 mb on September 11th (Figure 18) when it was located about 1300 km east-southeast of Cape Fear, North Carolina. Florence finally made its landfall as a category 1 hurricane near Wrightsville Beach, North Carolina, around 11:15 UTC on September 14th.

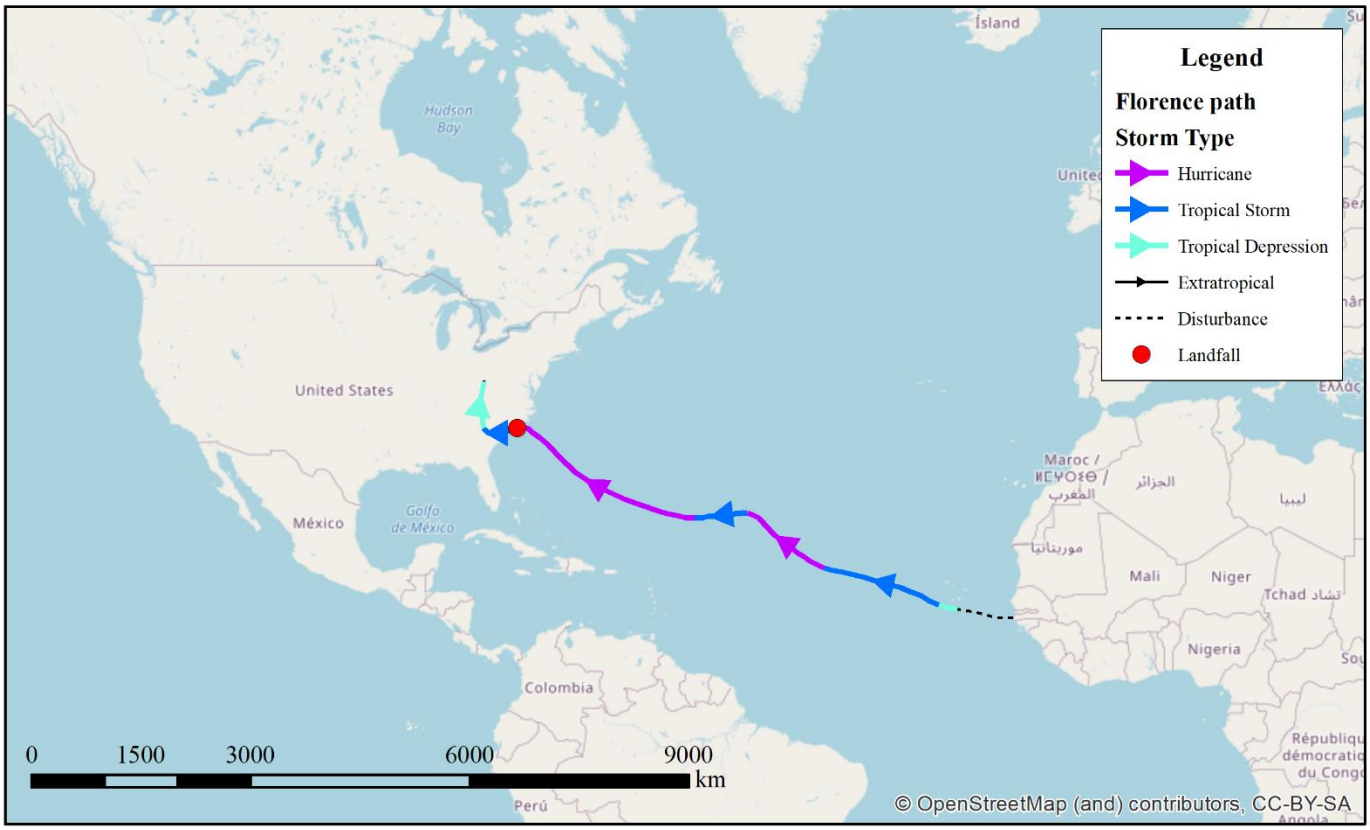

Figure 17 Florence path and landfall

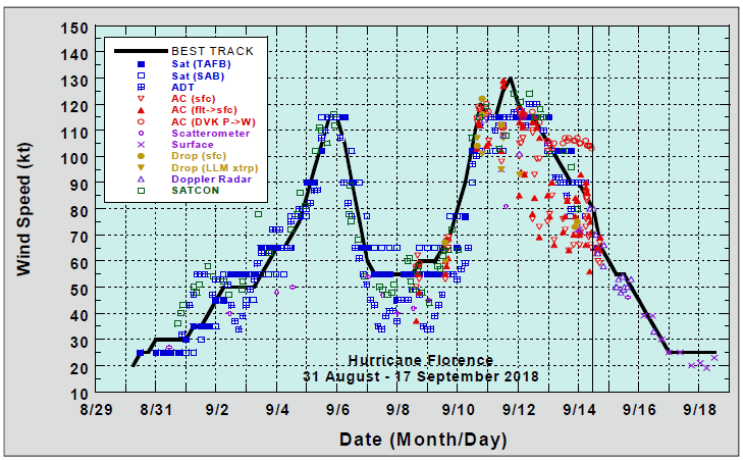

Figure 18 Best track for Florence wind speed observations (Stewart & Berg, 2019)

On September 11th at 9:00 UTC a hurricane watch was issued for the coastal area, from Edisto Beach (South Carolina) until the North Carolina/Virginia border. Twelve hours later, it was changed to hurricane warning – for which hurricane conditions are expected, not only possible – for the area between South Soutee (South Carolina) and Duck (North Carolina). As mentioned on the NOAA official website, on this day local authorities and officials of the counties concerned by the hurricane arrival started the emergency procedure, for the most threatened locations and evacuation. The warning was officially discontinued on September 14th at 21:00 UTC.

Florence caused devastating freshwater flooding across most of the south-eastern United States and significant storm surge flooding in portions of eastern North Carolina. These were the consequences of rainfall exceeding 0,25 meters across much of eastern and central North and South Carolina. A maximum total rainfall equal to 0,91 meters was measured in the area of Elizabethtown, North Carolina, setting a new state record for tropical cyclone rainfall. Also, the hurricane circulation caused a total of 44 tornadoes across three states (North Carolina (27), Virginia (11) and South Carolina (6)).

On the economic side, the NOAA National Center for Environmental Information (NCEI) reported that wind and water damage caused by Florence sum up to approximately $24,2 billion (22 billion only in North Carolina), making it the second most destructive hurricane of 2018 (the first one was Michael with a total amount of $25,1 billion). The hurricane was responsible for 22 direct deaths in the United States: 15 in North Carolina (NC), 4 in South Carolina (SC) and 3 in Virginia (VA). The main causes were freshwater floods (submerged vehicles, drownings), wind (falling trees) and tornadoes (Figure 19). Heavy rainfall generally caused many flash and severe flooding events whilst wind spawned tree uprooting, blown off roof incidents, power outages and traffic interruptions.

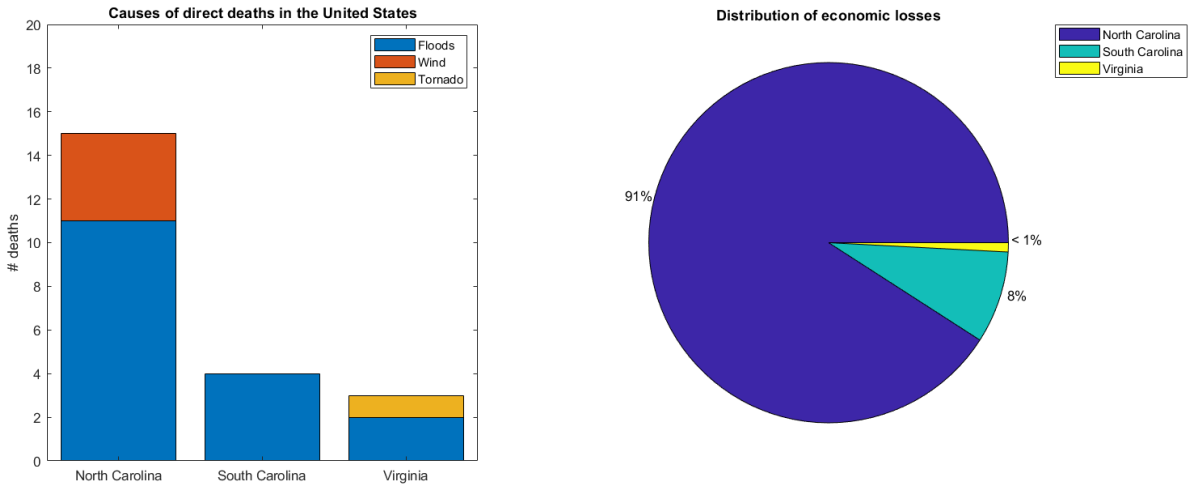

Figure 19 Death causes and economic losses reported in North Carolina, South Carolina and Virginia

3.2 Tweets source, data preparation and statistical composition

As mentioned on paragraph 2.2, the Twitter Developer guidelines do not allow to share entirely tweet objects. The only possible way to publish or share Twitter posts information is through their tweet IDs. The hurricane Florence Twitter dataset was made available by Harvard Dataverse archive (Wrubel, 2019). It contains IDs for 7766964 tweets published between September 11th and October 4th, 2018. The dataset metadata file reports that the Florence-related social media status updates have been collected in real-time using a Twitter stream API with query keywords 'Florence', 'hurricane Florence', '#hurricaneFlorence' and '#Florence'. Figure 20 shows the workflow followed in the preliminary phase preparing the data for the next elaborations and analysis in the ArcGIS environment.

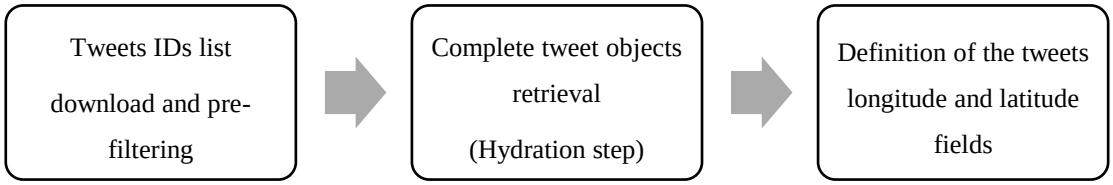

Figure 20 Data preparation workflow

The original dataset can be exported through the George Washington archive web interface where it is possible to select some preliminary filtering parameters for the download. In this case, only georeferenced tweets (original posts, quoted retweets and replies) were downloaded for further analyses. Hence, not-quoted simple retweets were not considered for the next steps with the assumption that, since they are simple sharing objects, they generally do not include first-hand opinions or significant evidences related to a specific event. Figure 21 highlights the significant differences in composition percentages between geolocated tweets and unreferenced Twitter posts.

The downloaded dataset reflects the main positioning issue associated to SMGI. Indeed, only a small portion of the Twitter data – equal to 1% - includes a spatial reference (*coordinates* or *place* attribute). This fact should be taken into account for future considerations about the level of representativeness of geolocated tweets.

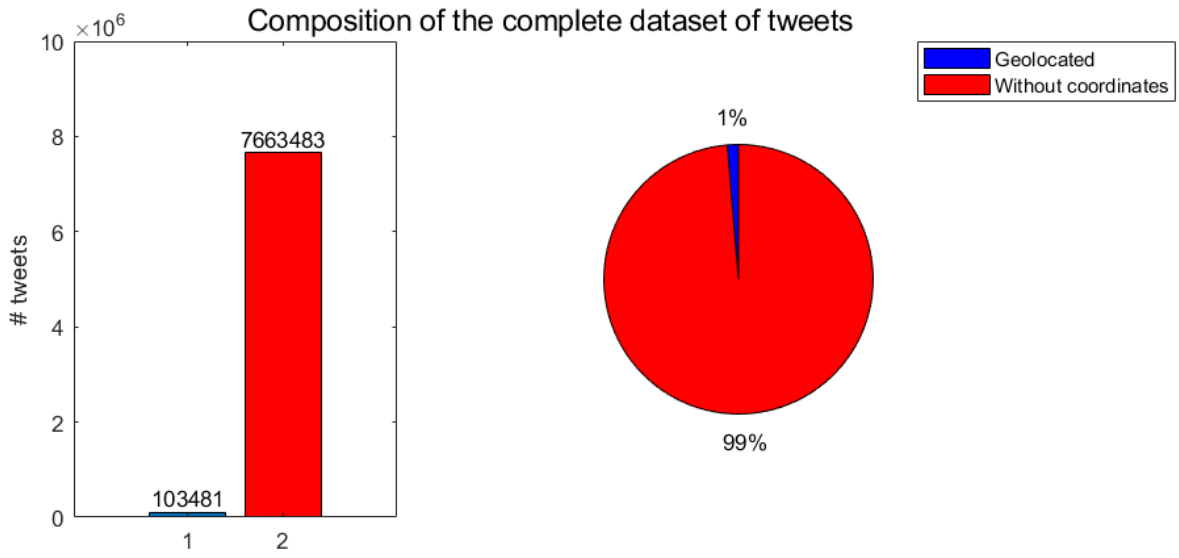

Figure 21 Comparison between geolocated tweets and unreferenced posts

Before the Tweet IDs hydration procedure, it has also been evaluated the composition of geolocated tweets in term of post typology: original tweet (a social media post created and posted by a user) and quoted retweet (a shared retweet with an attached text field containing the retweeting user comment). The ratio between the two typologies is almost equal to 1:1 (Figure 22). However, the higher percentage for original tweets could suggest that Twitter users were more interested in sharing their opinions or observations than quoting and retweeting others' social media contributions.

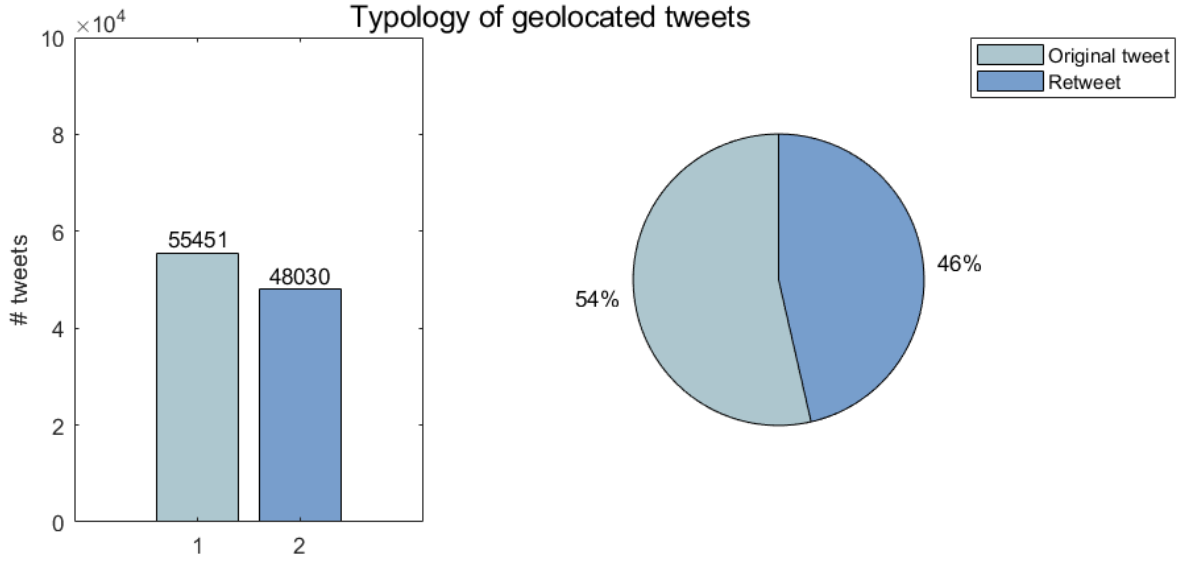

Figure 22 Comparison between original tweets and quoted retweets

The hydration procedure (complete JSON tweet objects retrieval from IDs list) can be performed using specific tools or libraries developed mainly in Python language. For this case study, for simplicity of use and interface, it has been chosen the Hydrator program developed by Documenting the Now (DocNow), a users' community that aims to build tools to help archivists, activists and researchers working with social media data.

This step also helps to have a first evaluation of the level of the dataset availability and completeness. In fact, the output products' metadata gives an overview of the dataset statistics, highlighting, in particular, the percentage of tweets that have been deleted since the collecting procedure took place. Figure 23 shows that the percentage of deleted tweets sums up to 17%, implying that a significant number of posts were deleted by the users after the original publications or some users cancelled their Twitter accounts in the months following the collecting stream API. Also, in Figure 24 it is possible to see how the ratio between original posts and quoted retweets changed after the hydration process.

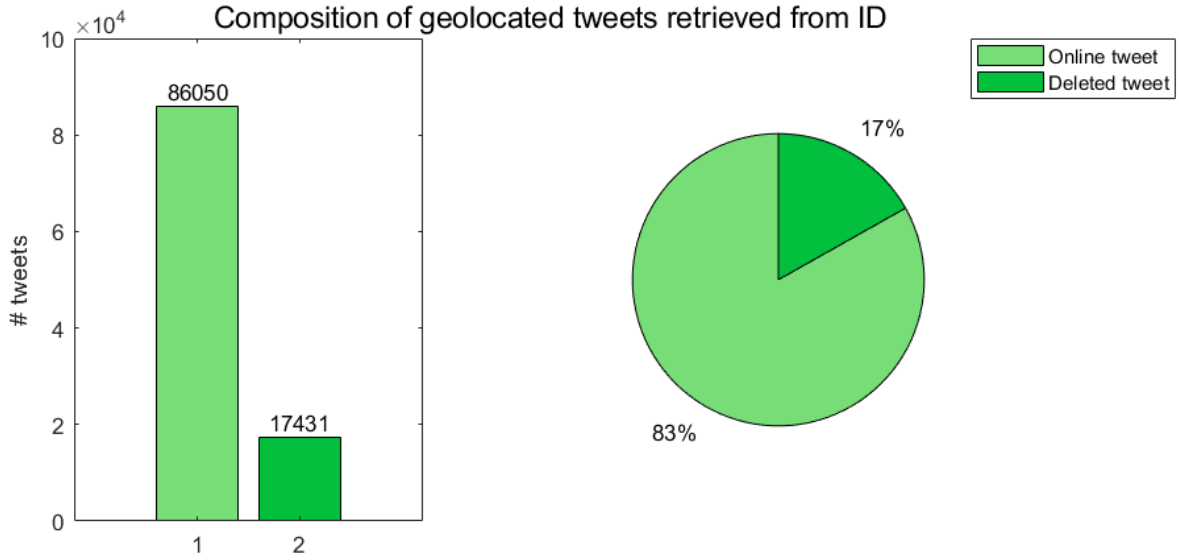

Figure 23 Comparison between available and deleted tweets

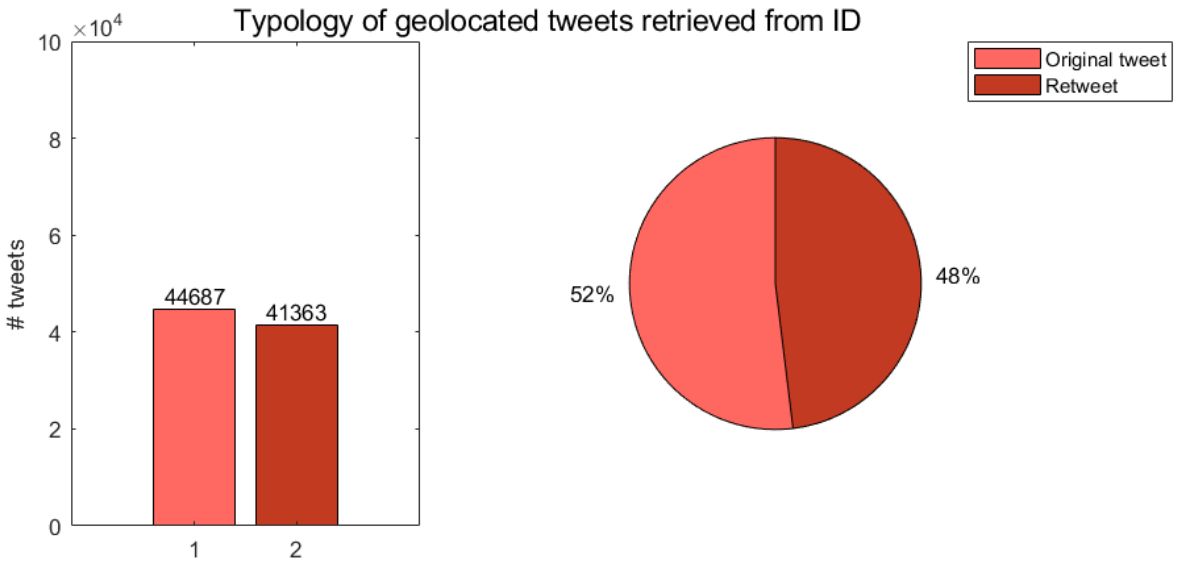

Figure 24 Comparison between original tweets and quoted retweets after the hydration step

Once retrieved the tweet objects that are still online, the output raw Florence-related dataset is characterized by 190 separate attributes for a total of 86050 records. After a first evaluation, only 18 original tweet attributes have been considered relevant for the purpose of this case study, ignoring fields with redundant information or with information linked to the user profile appearance settings (background colour, profile or cover image etc.).

Regarding the tweets spatial component, the coordinates associated to tweets with a not null place attribute have been processed in order to obtain a single couple of values. Since it is necessary to represent each SMGI record as a point feature in a GIS environment, it has been applied a correction to the place attribute, defining two separate fields of longitude and latitude associated to the averages of the extreme longitude and latitude values of the *place* bounding box. Then, based on the assumption that the *coordinates* attribute generally gives a more accurate spatial information, for tweets with both place and coordinates, only the second field has been considered and used to define the position of social media records.

In order to easily evaluate the nature of the geolocation method of a tweet, it has also been added a categorical field (*Coor_type*) that is equal to 'Point' if the tweet is associated to a *coordinates* attribute or to 'Approximate' if the position is derived from a *place* geotag. This field could represent an easy-to-use field for the next analysis (e.g. a possible to weight used to give more relevance to one of the methods during an elaboration).

The resulting pre-processed dataset – inside which every tweet is identified by a unique *ID* - is then characterized by these data components to which correspond specific tweet attributes:

- Temporal component represented by the *created_at* field.
- Spatial reference associated to a couple of coordinates. The longitude and latitude values are defined on the basis of the most precise not null geolocating attribute attached to the tweet. Other useful information about the position of a tweet are given by *place.country* (the country inside which the tweet is located with the place geotag) and *place.full_name* (the geotag official name listed inside the Twitter places database).
- Thematic component recording the social activity rate of a single tweet: the status text (*full_text*), number of retweet (*retweet_count*) and favorite (*favorite_count*) reactions, tweet language (*lang*), hashtags included on the text (*entities.hashtags*) and *is_quote_status*, a boolean attribute used to understand whether a tweet is a quoted retweet or not.
- User component associated to the main characteristics of a tweeting profile: number of profiles that the tweet user follow (*user.friends_count*), number of profiles that follow the tweet user (*user.follower_count*), the total amount of tweets published on the user timeline (*user.status_count*), the *user.id* and *name* to identify the profile and the *user.location*.

Figure 25 is showing the comparison between the Florence-related tweets with exact coordinates and ones whose position is approximated through the *place* bounding box (place BB). Additionally, the two different groups of geolocated tweets are compared with the entire dataset that includes also social media posts without a geographic reference (Figure 26).

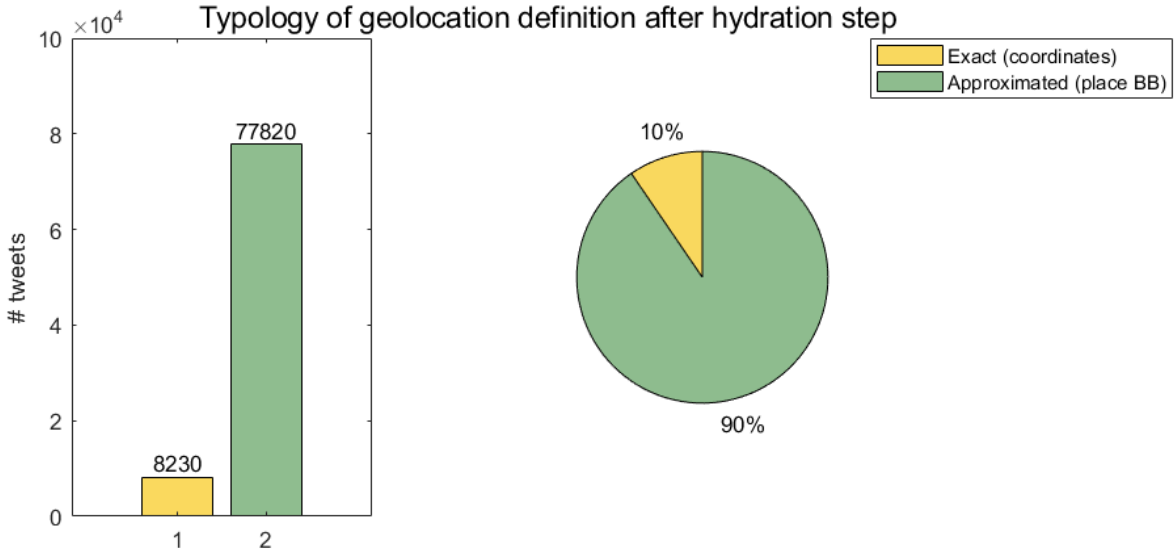

Figure 25 Comparison between tweets geolocated with their exact position (coordinates field) and the ones with an approximate position (place BB)

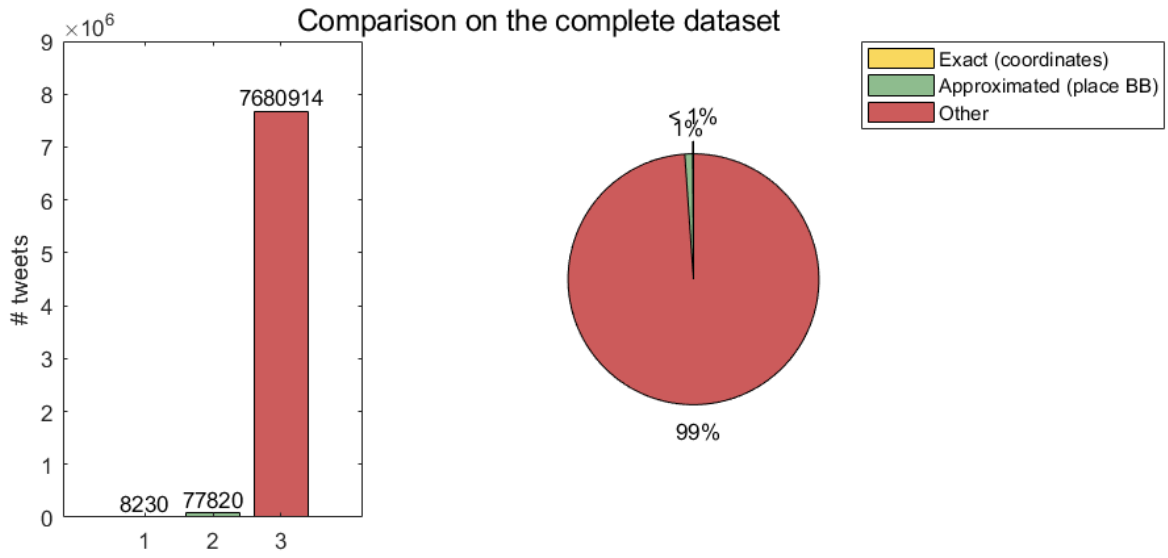

Figure 26 Comparison between the two groups of geolocated tweets and the remaining unreferenced tweets of the entire dataset.

As predictable, the geolocated component represent only a small portion of the entire downloaded dataset. For this case study, it is smaller than the 2% of the total, consistent with the general Twitter documentation observations referring to the geolocated component as usually less than 5% of a complete dataset. The proportion between the unreferenced tweets is even more impressive when comparing it only with the posts associated to an exact position, the ones with a *coordinates* attribute. This means that the most accountable type of geolocated tweets beforehand can not be considered representative of the behaviour of the Twitter user population, underestimating the main activities and reactions happening during the occurrence of a hurricane. However, this consideration emphasises some crucial questions for this case study, asking whether the georeferenced data, even in smaller percentages, are enough to identify pattern comparable with reference data.

3.3 Filtering and pre-processing procedures

Once defined the geographic reference for the pre-filtered Twitter dataset, it is possible to have a first visual evaluation of the global distribution of the social media posts. This step executed on the ArcGIS environment is necessary to understand the scale of the raw dataset, identifying areas affected by a relevant social media activity and defining spatial filters able to isolate only tweets strictly geographically correlated to the occurrence of the hurricane Florence. Figure 27 shows the workflow that guides this filtering and pre-processing step.

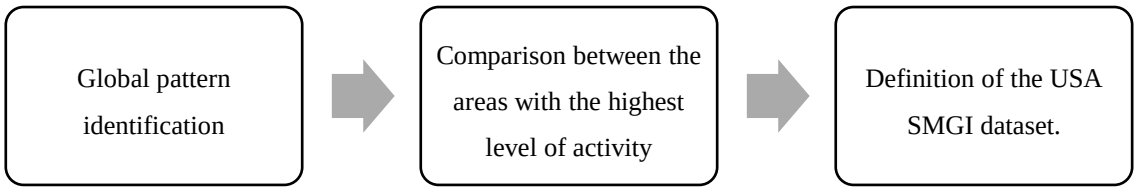

Figure 27 Workflow for the pre-processing of the collected Twitter dataset

For all the cartographic representations it has been adopted a WGS 1984 World Mercator projected coordinate system. Fig. 28 illustrates the distribution at a global scale of the input Twitter georeferenced dataset composed by 86050 tweets published on the social network between September 11th (when the hurricane Florence watch was issued) and October 4th, 2018. The countries' national borders are defined by the data downloaded as shapefile (geospatial vector data format) from the OpenStreetMap database through the Geofabrik server application.

With this first simple unprocessed visualization, it is already possible to observe that a strong concentration of tweets is detected in North America and Europe. United States are the country affected by the occurrence of Florence and then logically the social activity was directly influenced by the hazardous event. The tweets distribution in Europe, instead, need to be explored and motivated with further considerations. A widespread consolidated Internet accessibility could be a reason for this high number of tweets. Additionally, flight connections and worries for cancellations could have enhanced the Twitter activities in the main cities with international airports. Particularly, this aspect could influence the number of tweets published in the area around London (United Kingdom), due to the presence of the Heathrow airport, the busiest European flight facility by passenger traffic and by connections with the United States as reported in September 2018 by the British Civil Aviation Authority. It also important to notice that, even if the hurricane Florence had an international media coverage, two of the most populated countries in the world (China and India was the first and second countries with largest population according to the National Bureau of Statistics of China and to the Open Government Data Platform India) registered low activities on the social network platform. The reasons for this could be Twitter access restrictions (in China, Twitter is banned since 2012 and can be used only through VPN officially approved by the government (Bamman et al., 2012)) or the use of other social networks (in India, as reported by Statista, the leading social platform are Facebook, Instagram and Youtube).

In Figure 29 it is possible to understand which are the countries characterised by the highest numbers of tweets published within their boundaries and compare them with official NOAA hurricane track.

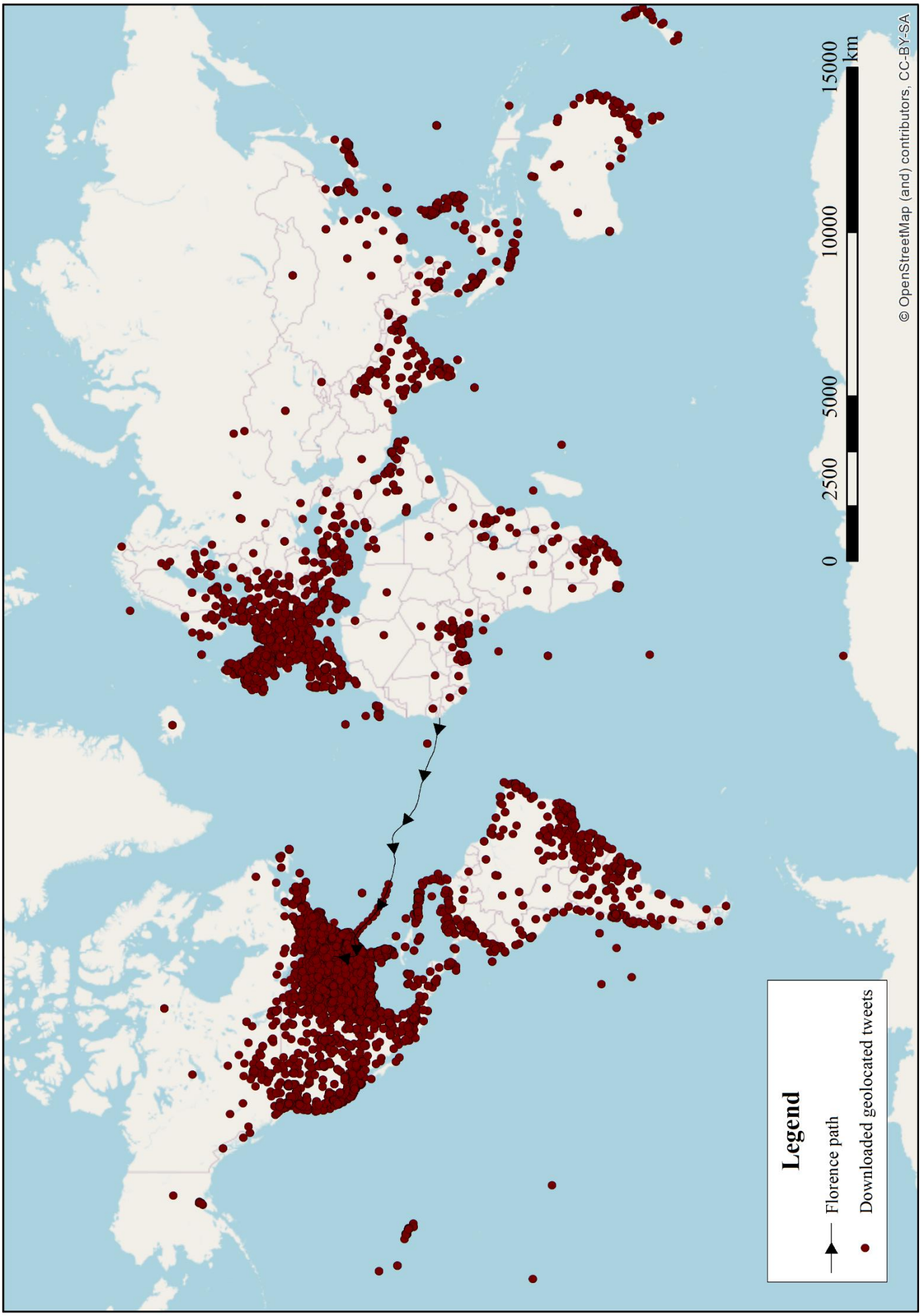

Figure 28 Global distribution of the downloaded geolocated tweets

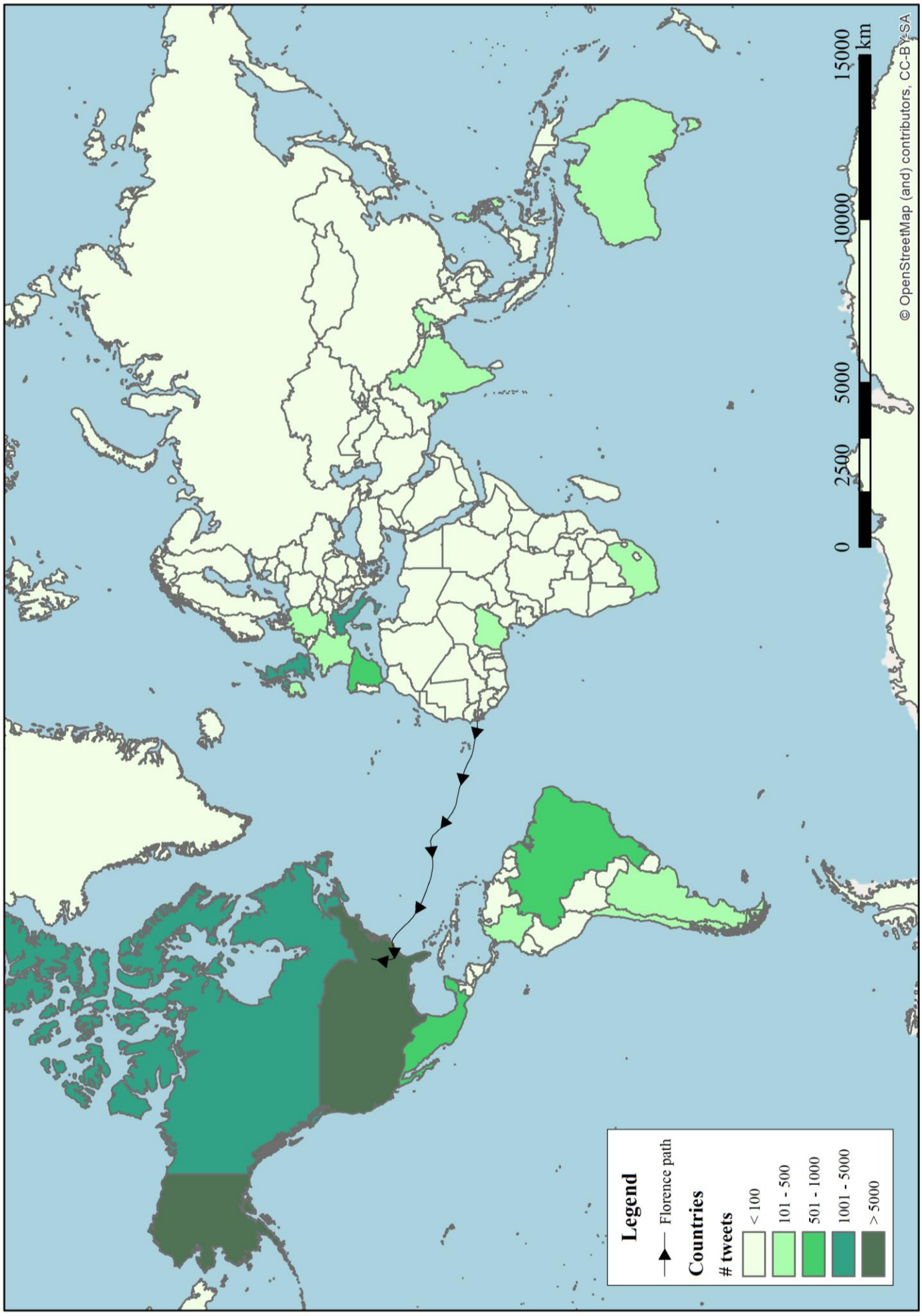

Figure 29 Classification of countries at a global scale by the number of tweets published within their borders

Figure 29 confirms the observations previously made at a continental scale, highlighting the 3 countries characterised by the highest number of Florence-word related tweets for the given time window: United States (69991 records), Italy (2772 posts) and United Kingdom (1925 tweets).The number of social media posts located within the US borders is considerably higher than the ones registered in the other two countries.

Italy's presence in second place does not result surprising because the 'Florence' tweet filtering keyword is associated also to the English toponym of the city of Firenze, located in Tuscany region. Furthermore, it is the eighth most populated Italian city and the fourth one by number of foreign visitors in 2018, as reported by Istituto Nazionale di Statistica (ISTAT). The high level of tourists' activity in posting and sharing photos or videos around Firenze could have indeed influenced the total number of posts in Italy. The level of social media activity recorded in the United Kingdom could instead be associated to the previous consideration about international flights and travels. However, this aspect needs to be further analysed in the next steps of this pre-processing procedure.

Nonetheless, before proceeding with specific analysis, it should be considered the possibility of different temporal daily trend in the tweet publication across the three different countries. Also, different proportion between exactly and approximately geolocated tweets, as much as the one between original and quoted retweet posts, could give additional insights about users' behaviours and could represent proxy indicators of different motivations. Figure 30 shows the different temporal trends, highlighting a significative peak on September 14th, 2018 both on the United States and global timeseries. This last one is strongly influenced by the American tweet activity that represents the 79% of the entire Florence-related Twitter dataset.

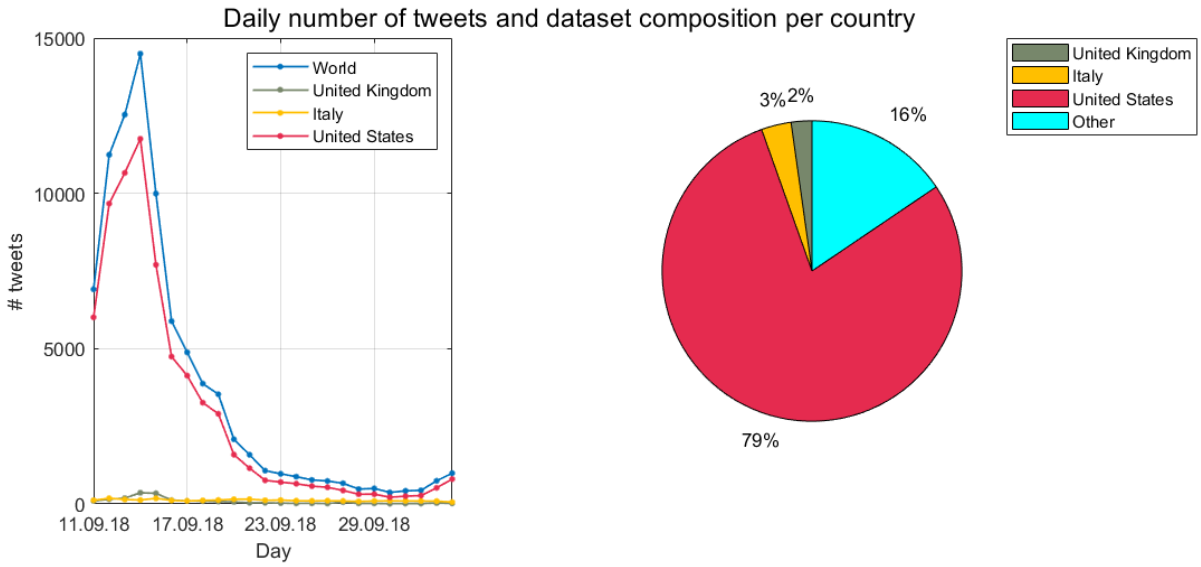

Figure 30 Temporal daily trends and dataset composition for the 3 countries with the highest number of tweets

Normalizing daily tweets by the total number of posts over the observed time window within each country's borders, was possible to have three comparable timeseries. In Figure 31 the different behaviours could be evaluated and compared. After the normalization, the Italian trend have shown significative differences from the other two. It is almost constant for the given time window, attesting that, on average, a portion of 4% of the total number of tweets was published daily. This constant activity could be considered not related to Hurricane because it was not associated to oscillations from a baseline of normal tweet rate. It is then possible to suppose that the tweet publishing in Italy is affected and determined by the city toponym (always constant in time) and not by the hurricane.

In the case of United Kingdom (UK), it is more complex to define a specific trend because the tweet sharing behaviour seemed to partially fit the United States one with a one-day delay. The peak indeed is reached between September 14th and 15th, 2018 and then it is detected a rapid decreasing slope. This small delay could be associated to the post hurricane landfall international media coverage. This could be supported by the fact that three of the main UK newspaper agencies (The Sun, Daily Mail and The Telegraph) published online the majority of Florence-related articles between those two days. These news articles were then shared and commented on Twitter as seen inside the case study dataset. The UK activity then decreases rapidly until September 16th, date time from which the tweeting rate is only slightly reduced with a smaller slope until September 20th. Looking through the posts, this could be due to the meteorological effects on the European Atlantic side provoked by Florence circulation. Indeed, the Met Office – the UK national weather service – alerted citizens through communications and forecasts about hurricane Helene (since September 15th until 18th) and storm Ali, that caused 3 fatalities and heavy rainfall in Great Britain and North Ireland. Additionally, this period preceded the Florence & The Machine singer's gig on September 20th for the Mercury Prize, an annual music prize award broadcasted by BBC national channel that was associated to many UK inside the dataset. Another strange small peak is detected on September 27^{th,} but it is associated to a not relevant strong activity of a single user who published 41 out of the total 58 registered tweets.

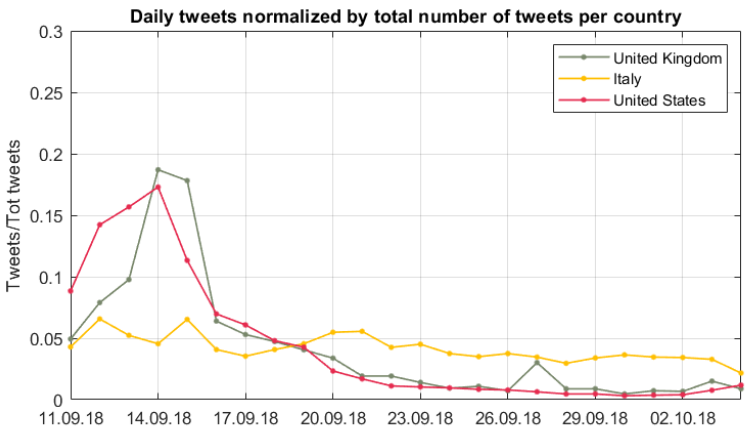

Figure 31 Daily normalized trends for the 3 countries with highest level of activities within their borders

The different behaviours of Twitter user populations of the countries considered can find further evidence with the percentage analysis related to the geolocation mode and the type of tweets published. In Figure 32 the pie charts show the percentages of the type of positioning associated to the first three countries by number of tweets. The percentage of point positioning for the United Kingdom (6%) and the United States (8%) can be considered comparable to that recorded on the complete dataset (10%). In the case of tweets located in Italy, on the other hand, there is a much higher percentage (76%) than the others, another symptom of a different social behaviour in this geographical area maybe due to the fact that tourists in the Firenze area decided to apply specific coordinates to their monuments or historical places photos.

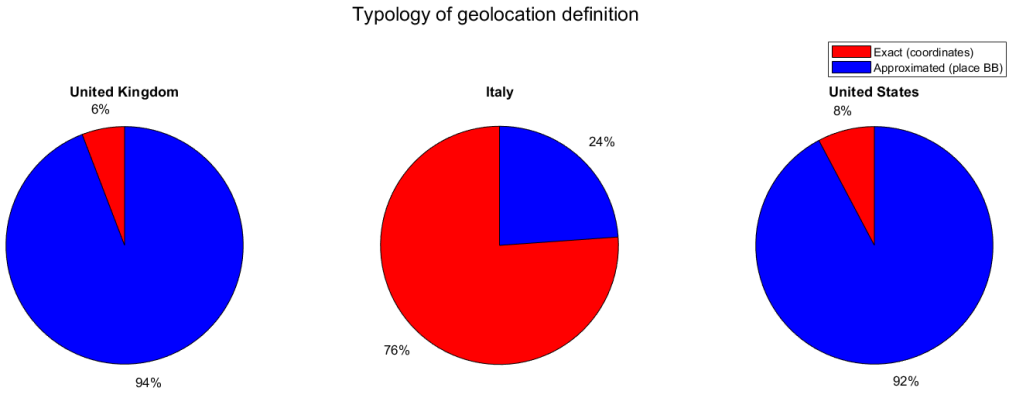

Figure 32 Comparison about the geolocation methods between the 3 countries with the most active users

The evaluation of the percentage composition of the tweets according to their type of publication provides comments on the different factors that may have influenced the active Twitter population within the three countries. In Figure 33 it is possible to observe again values of the subset of Italy very different from those of the global dataset. In fact, a portion of original tweets is recorded equal to 97%, a very high value if compared to the 52% composition of the global Hurricane-related dataset.

On the other hands, the US tweets subset is confirmed to influence and determine the global values while the social contributions localized within the UK borders are strongly influenced by the quoted-retweet portion equal to 58% of the total. This could be associated to a greater interest in sharing third-party contributions (news articles, weather alerting or reporting) and a possible lower personal involvement in the scenario defined by the term "florence" on Twitter, in accordance also with the observation made about the temporal Twitter trend.

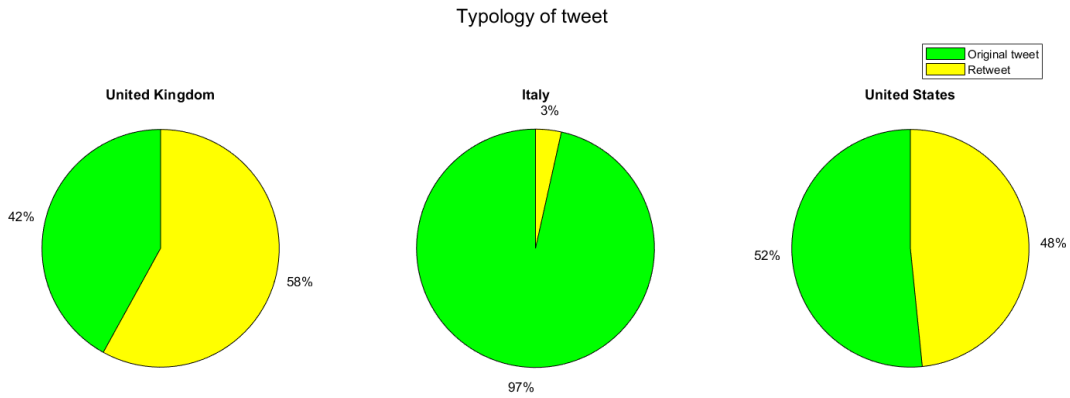

Figure 33 Comparison about the tweet type between the 3 countries with the most active user.

Considering in a more detailed and statistical way the social media's possible reactions (registered 'favorite' endorsing the content of a tweet or 'retweet' sharing an original post to a wider audience) and the user profile types (Table 2), it is possible to find a higher level of interaction in the US subset than in other countries. However, this evidence does not support the hypothesis of a more valuable tweet subset in the United States because in general American user are considered more interactive on social network platform.

Table 2 Twitter statistics about tweet reactions and user profiles.

	Entire dataset		**USA**		**Italia**		**UK**	
	Mean	**Max**	**Mean**	**Max**	**Mean**	**Max**	**Mean**	**Max**
Favorite	6,28	55949	7,06	55949	2,04	1018	2,6	664
Retweet	2,09	29389	2,36	29389	0,28	60	0,75	421
User followers	6220,91	3477249	5669,79	3477249	2840,19	522221	2477,04	177859
User following	2011,03	1229026	1989,66	814318	1046,81	33101	1830,07	111407
User statuses	35140,55	4131691	32809,05	4131691	16377,12	410613	43918,19	1048043

To better understand the nature of the audience of users involved in social activity, the attribute linked to the language automatically recognized by Twitter in the text of the tweets collected was also considered. As predictable, Figure 34, showing the proportion between languages detected in the texts of the entire dataset, highlights the highest use of English also due to the fact that the keywords used for the collecting procedure ('hurricane', 'hurricaneFlorence') were English. The undefined component – Twitter associates 'und' to the language attribute when words from different languages or symbols are detected – is equal to the one that includes all the other languages. In the pie chart two of the most spoken languages at a global scale are absent: Mandarin Chinese and Hindi. However, these absences could be attributed, as previously mentioned, to the lack of Twitter posts geolocated

in the correspondent countries. Hence, the second most detected language inside the case study dataset is Spanish that is very common in the US territory.

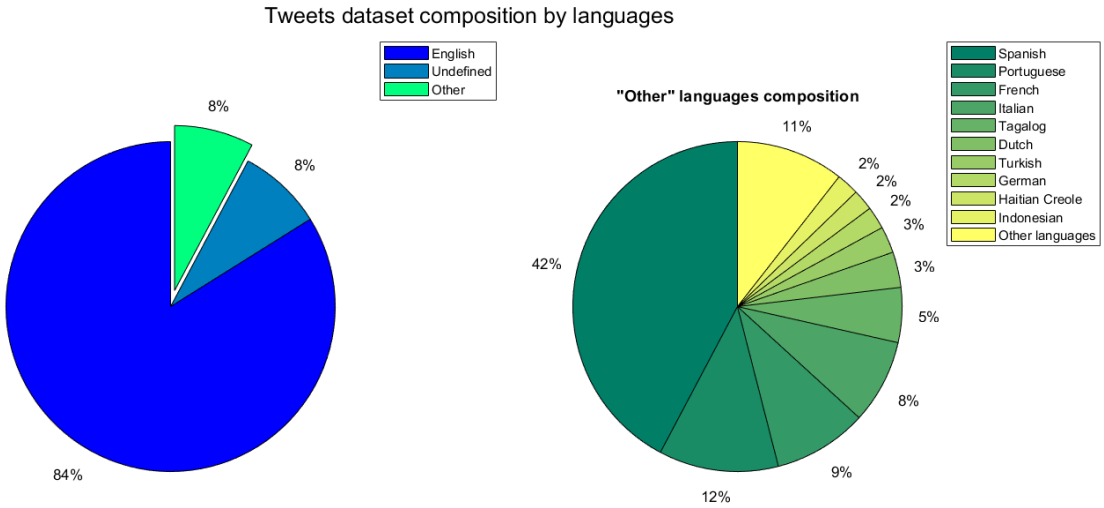

Figure 34 Detected languages for the entire Twitter dataset

Figure 35, 36 and 37 show the language composition of the Twitter post texts geolocated in the United States, in Italy and in the United Kingdom. English is the predominant language for the US and the UK whose users are presumably mainly native speakers. A high percentage of English posts is detected in Italy too, probably because of the significant presence of tourists but also due to the choice of "a more international" type of communication made by some local users. The supposition previously made about Spanish-writing users finds evidences in the US pie chart were the language of the Hispanic community represents almost half of the not-English detected languages. This is confirmed also by the fact that other language-writing users in Italy and UK are present in a more distributed way.

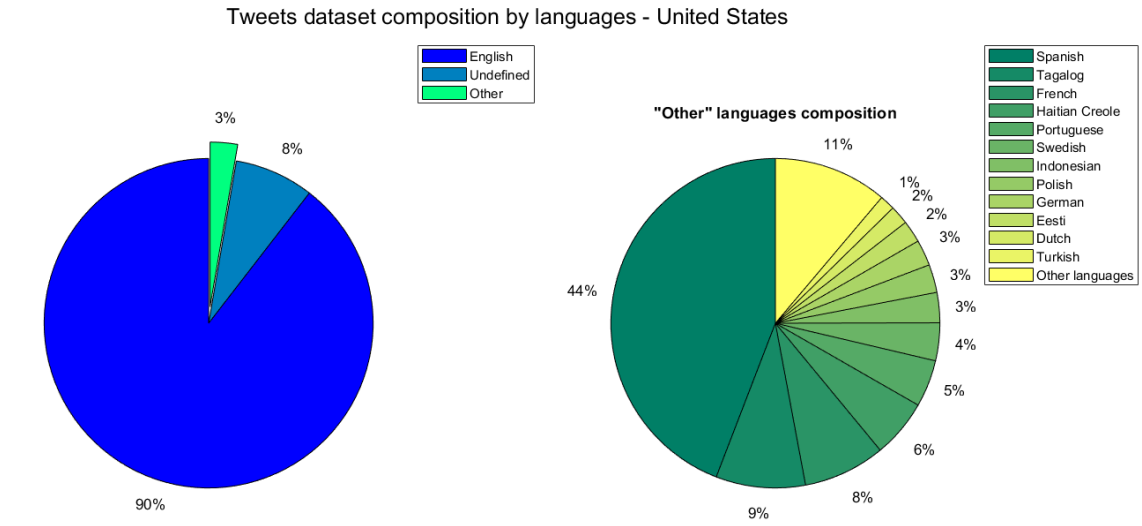

Figure 35 Detected languages for the tweets located in the United States

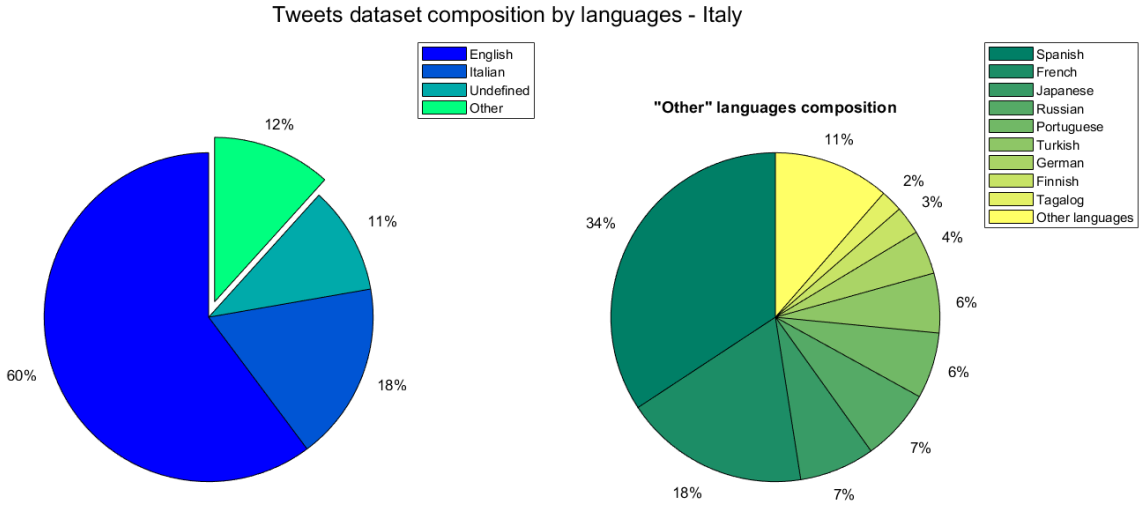

Figure 36 Detected languages for the tweets located in Italy

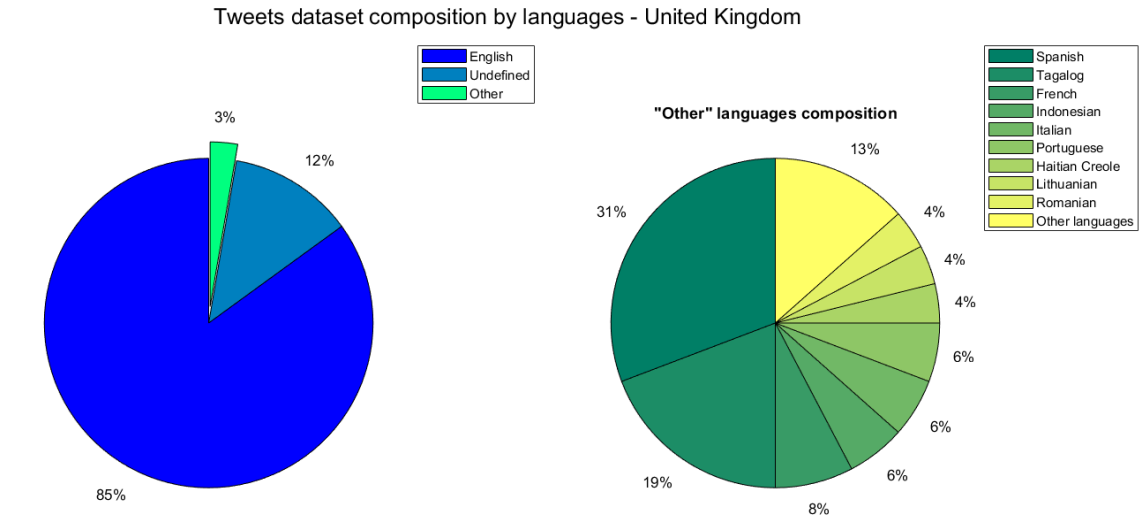

Figure 37 Detected languages for the tweets located in the United Kingdom

Following the previous observations on the composition of the entire dataset and the atypical case of the subset of tweets located in Italy, it was decided to perform a further spatial elaboration able to highlight the influence of the toponym "Florence/Firenze" on the social activity. Figure 39 shows the classification of Italian regions - whose shapefiles are available on the ISTAT website - according to the number of geo-localized tweets in their territory. The influence of the Tuscany region and, consequently, of the tweets located in Florence is clear and is further highlighted by the fact that the 2467 posts located in this area represent the 89% of the total social media activity detected in the case study time window.

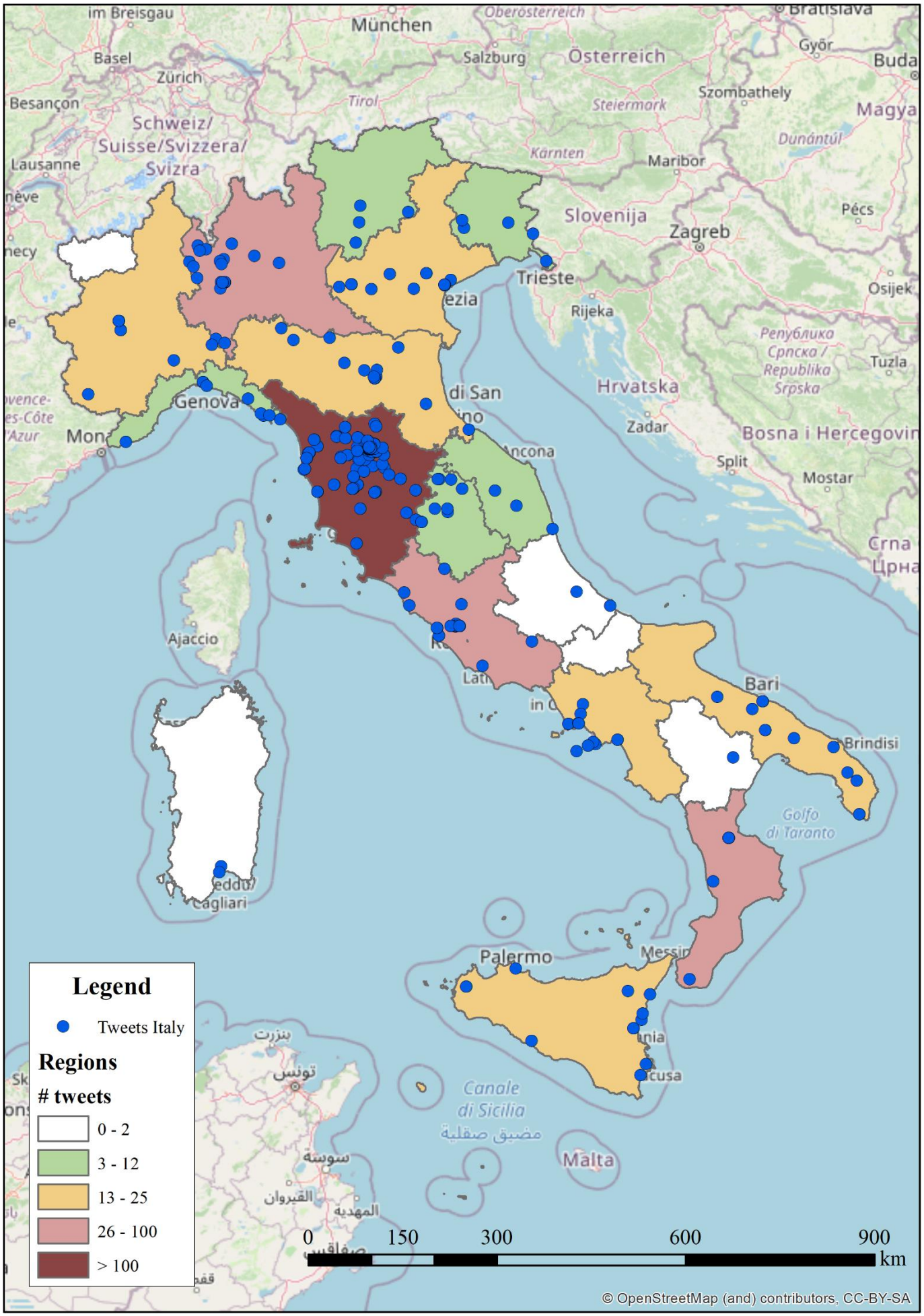

Figure 38 Distribution of downloaded geolocated tweets across the Italian territory

After these first geo-spatial elaborations, it was then possible to affirm that, based on the dataset composition, spatial distribution and on the comparison between the three most active countries, at a global scale the social media activities in the United States are significatively higher than the other countries, also in comparison with the second and the third country with the highest number of tweets located within their borders. In addition to what previously remarked, it is important to highlight the fact that the temporal peak associated to the US tweets (September 14th, 2020) is perfectly consistent with the landfall date of hurricane Florence.

The temporal trend analysis and the dataset composition highlighted different users' behaviours for Italy and the United Kingdom too. For the first case, a constant activity in time located almost exclusively in the region of Firenze, the total number of Twitter geolocated post is directly linked to a toponym influence and to a tourists' activity pattern. On the other hand, the social media trend detected in the UK is a complex context that requires more considerations referring to international travel connections, media coverage and ambiguous meaning of the chosen keywords (e.g. 'florence' can be mismatched with not-hurricane related posts about music exhibitions of the indie rock band Florence + The Machine).

In conclusion, the next steps of the spatio-temporal elaborations and analyses need to focus only within the United States borders, where a proposal of hurricane detection workflow is defined in order to identify the most affected area and to perform specific geo-statistical analyses.

Chapter 4

Hurricane Florence case study: geo-statistical and temporal analysis results

The preliminary analyses about the global tweet distribution helped detecting and contextualising the areas characterised by a high number of geolocated tweets within their borders. After having identified the United States as the country most affected by the Florence-related social media activity, it was necessary to define a procedure at a regional scale level in order to further reduce the study area and evaluate the reliability of Twitter data at a local scale. This chapter, after a first evaluation of the tweets' distribution within the US borders, presents the definition of a possible process to identify the US states with the most significant social media activity based on specific criteria. Then the Twitter subset associated to the resulting areas is explored through temporal and geo-statistical indexes and tools for pattern detection procedures and hot-spot analyses.

4.1 Tweets distribution across the United States

The Twitter Florence-related dataset contains 69991 posts geolocated within the official United States borders. Figure 39 illustrates the distribution of unprocessed tweets in Central and North America suggesting in particular some areas of interest. It is clearly visible that the majority of tweets was qualitatively located on the East Coast of the United States between Maine and Florida while in the rest of the country the Twitter posts resulted more dispersed. However, on the southern part of West Coast it was possible to see an evident concentration of SMGI in California. Considering that only the southern states of the US East Coast were directly affected by the hurricane , further quantitative elaborations were required to understand the possible reasons for the presence of tweets in un-affected areas and the key factors of Twitter activities. Additionally, it is important to remark the position of 13 tweets off the North Carolina coast that corresponded to the hurricane Florence path before the landfall. These geolocated tweets – original posts with exact coordinates - have been published between September 11th 02:51:57 UTC and September 13th 20:40:24 UTC by two accounts in real-time storm evolution. The spatial accuracy of these tweets could be motivated by the fact that both accounts are linked to weather-tracking application and science dissemination projects focused on natural disasters that are equipped with high resolution instruments.

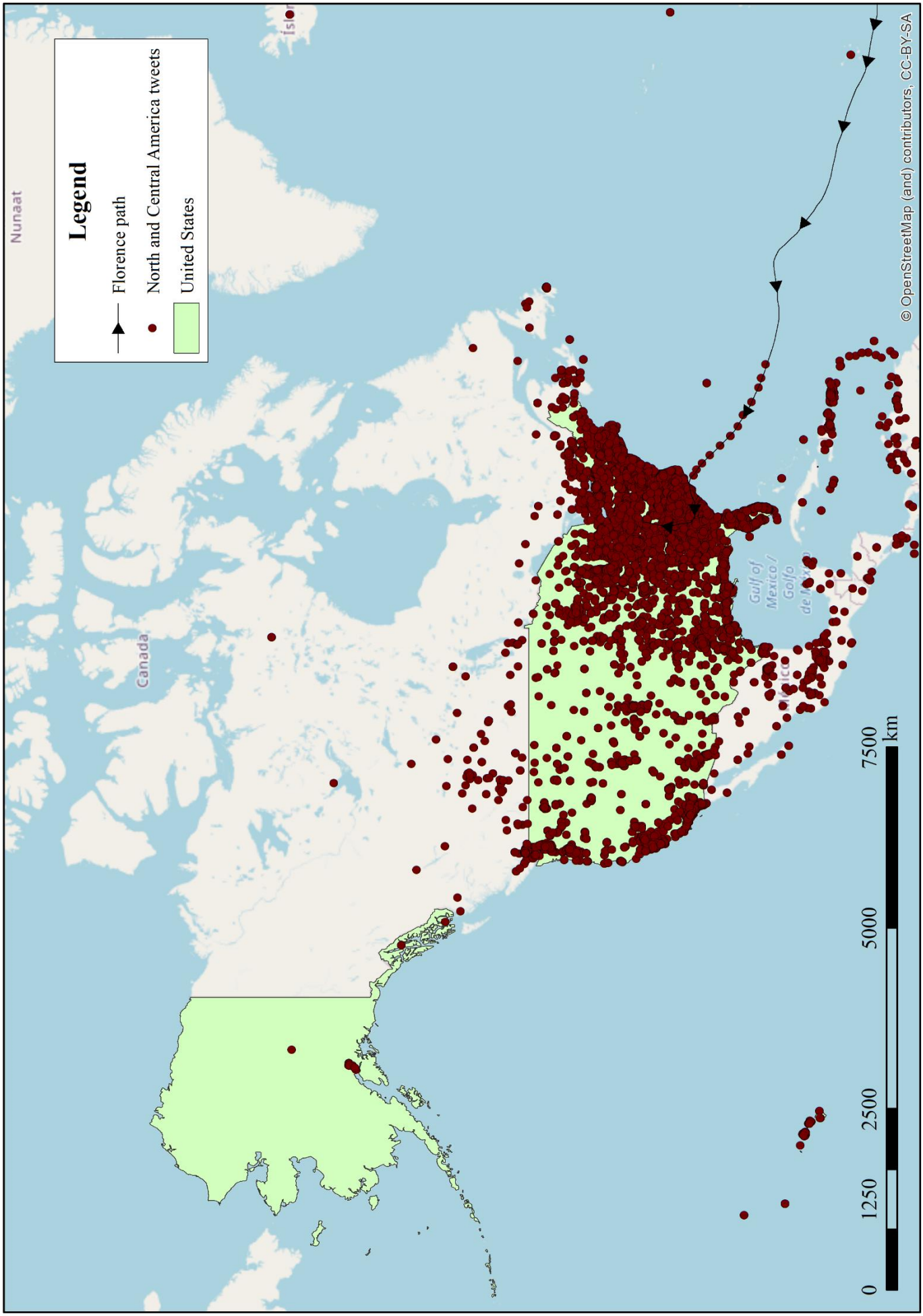

Figure 39 Distribution of geolocated tweets across the United States

A more detailed and quantified evaluation of US tweets distribution was given by the classification of US states and territories by the number of tweets geolocated within their 2018 borders made available as vector files by the United States Census Bureau. In this way it could be possible to effectively highlights the most influenced areas. Consequently, the classification illustrated in Figure 40 is the first possible step of a scaling procedure needed to obtain a preliminary quantification of the hurricane Florence influence on Twitter users. However, after this simple procedure, states like California and Texas resulted as areas characterised by a significant tweet activity. This result demonstrated that more factors, besides proximity, influence the distribution of social media posts, detected also across states that were not directly affected by the hurricane Florence event.

Based on the not proven hypothesis that the active user population of a specific geographic area is linearly dependent on the real population, an alternative classification criterion was defined normalizing the total number of tweets for the population. Figure 41 shows the results of this operation. The comparison with Figure 40 highlights important differences regarding the identification of the most active states, that after the normalization were mainly located on the East Coast. The normalized criterion, indeed, gave less relevance to states like California and Texas that, according to the United States Census Bureau were between the 5 most populated US States in 2018. However, both the North and South Carolina are identified as the states with the most hurricane relevant tweets during the given time window.

Even if this simple procedure gave significant results and remarked the importance of the relationship between Twitter users and the real population, it was not valid enough to support an SMGI scaling workflow because it relied on ambiguous assumptions and poor documentation. The linear dependence of the Twitter user population, indeed, has still to be investigated and, at the current state of art, can not be considered true due to the crucial influence of the digital division that affects internet availability under social and demographical perspective and could vary geographically (Blank, 2017 and Mellon & Prosser, 2017). For this reason, the definition of a workflow based only on the intrinsic peculiarities of the Twitter dataset was required. Its results could then be integrated with authoritative geographic information (e.g. population data) in order to perform a reliability evaluation of given SMGI.

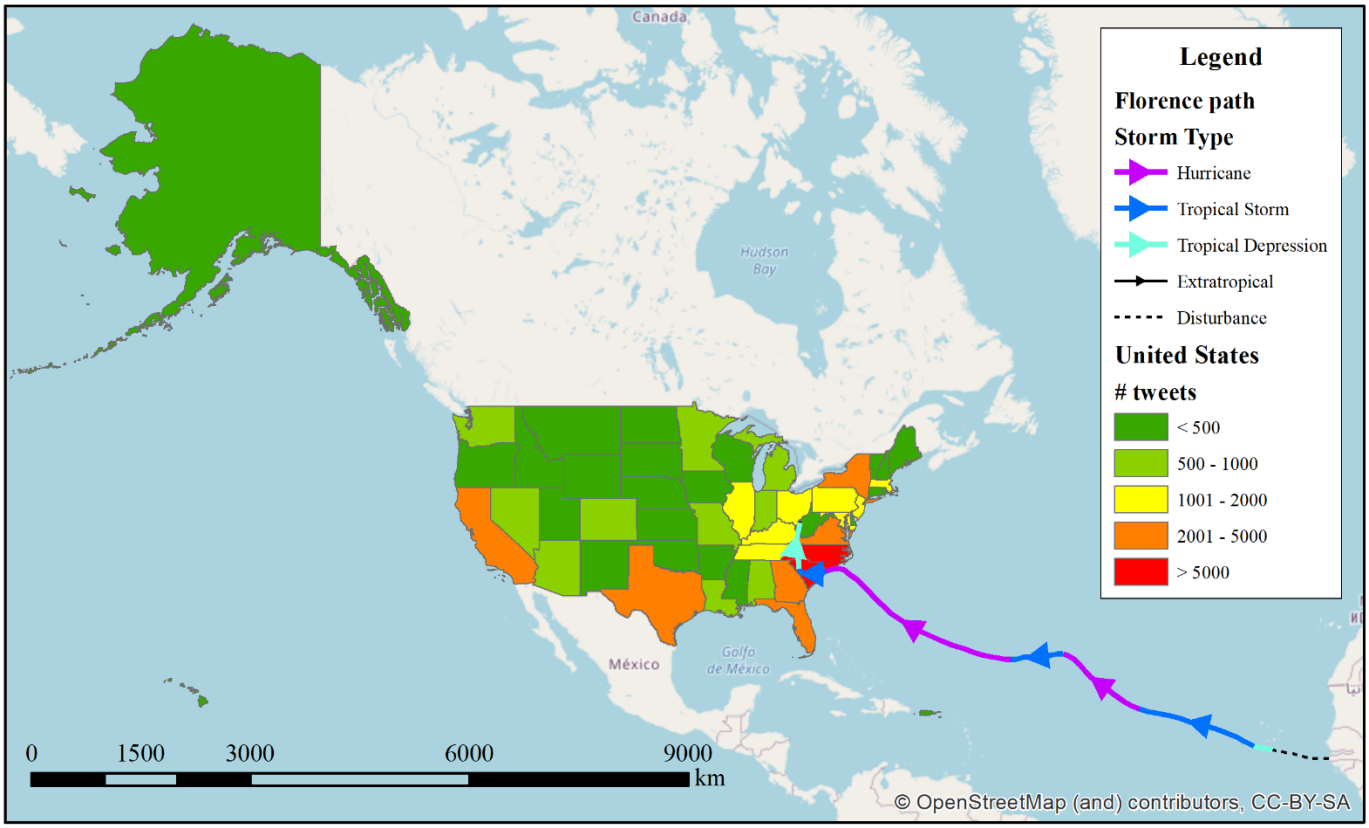

Figure 41 US states classified by the number of tweets geolocated within their borders

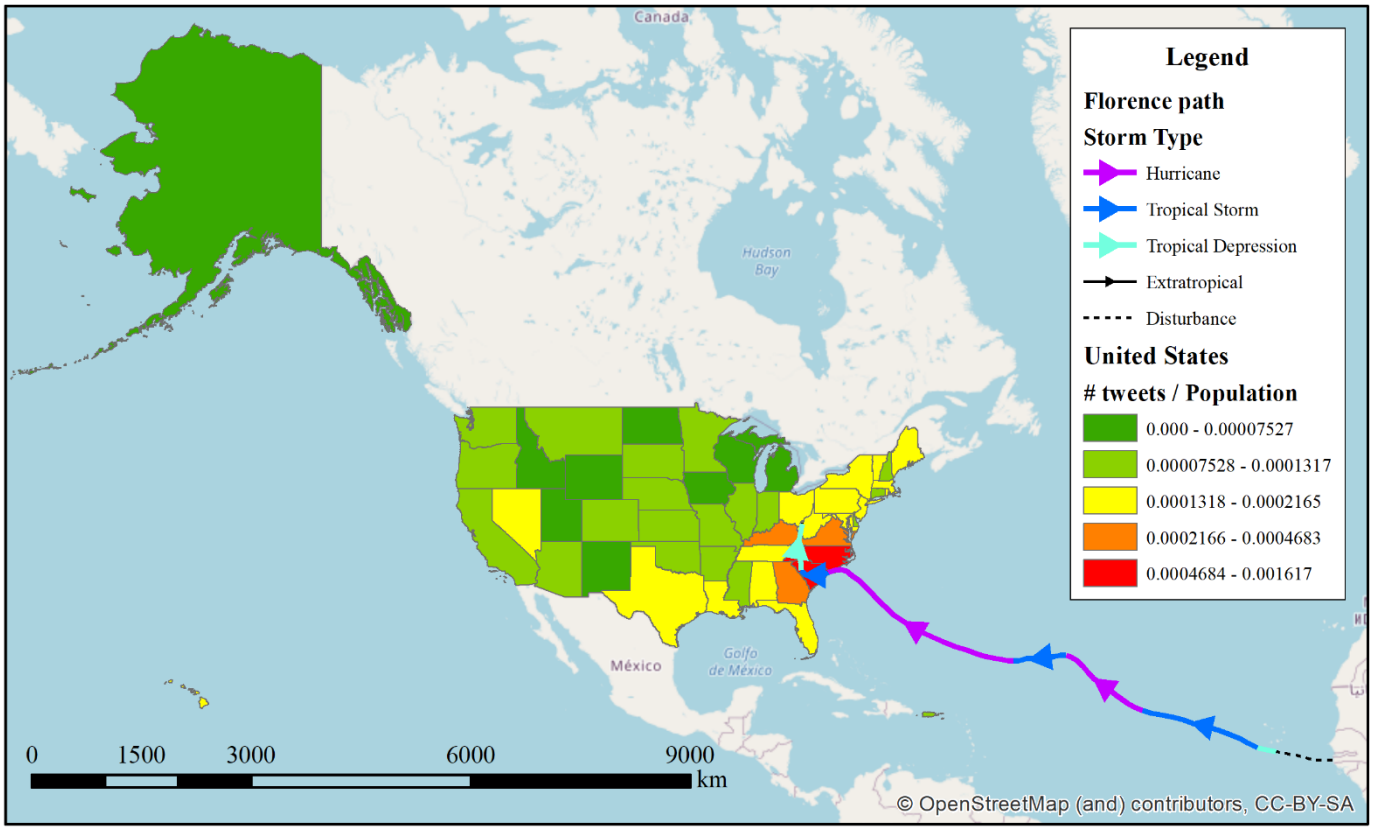

Figure 40 US states classified by the number of tweets geolocated within their borders normalised for their 2018 population

4.2 Identification of the states with the most significant Twitter activity

Following the observations made of the tweets' distribution across the entire US territory, it has been defined a process to identify the states with the most significant Twitter activity. This procedure is based on three filtering criteria linked to the tweet object itself, as depicted in Figure 42.

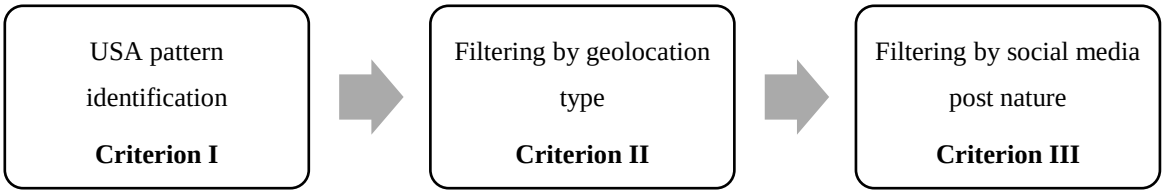

Figure 42 Procedure for the identification of tweets over the US states characterized by significant rates of social media activity

The 3 criteria have been defined as follows:

- For the USA pattern identification phase (Criterion I), the total number of tweets located within state borders has been calculated for each state in the US. Subsequently, the national average of this value was calculated (1271 tweets). Only states with a total number of tweets above the average reference value were considered as relevant.
- A filter by geolocation type has been applied, defying as a Criterion II threshold the percentage of tweets with exact coordinates for the entire US Twitter dataset (8%, as previously reported in Figure 32). This step concerned exclusively the states which satisfied the first criterion. Only states with exact position of the posts whose percentage was higher than the national one were considered for further analyses.
- Finally, the remaining US territories have been filtered considering the nature of their post composition (original contents or quoted retweets). The national percentage of original social media posts (52%) has been used as threshold for Criterion III that considered significant US states with a percentage higher than the reference one.

It is important to remark that this procedure is based on the assumption that high concentrations of geolocated posts with a not null *coordinates* attribute are more relevant than others. For example, a social media post containing user's comment and multimedia with a point positioning method has been considered more important because it has been assumed that the directly affected user needed to share a first-hand experience or observation about the hurricane. On the other hand, a quoted retweet could include additional valuable information about the event through discussions or comments on the retweeted content. However, this consideration would require a more complex

reliability evaluation on the Twitter users network in relation with a sentiment analysis on the textual content that could need external reference data not directly inspectable from the tweet object.

4.2.1 USA pattern identification

The US states that satisfied the Criterion I are illustrated in Figure 43. Only 13 out of the total 50 states (26% of the United States political entities) registered a total number of tweets higher than the national average. North Carolina resulted as the state with the highest number of social media posts geolocated within its borders, representing the 24% of the entire US Twitter Florence-related dataset. The bar plot also shows the difference with the other considered states that were characterised by a lower rate of tweet publication. The second most active state was, indeed, South Carolina with the 9,43% of the collected tweet in the US.

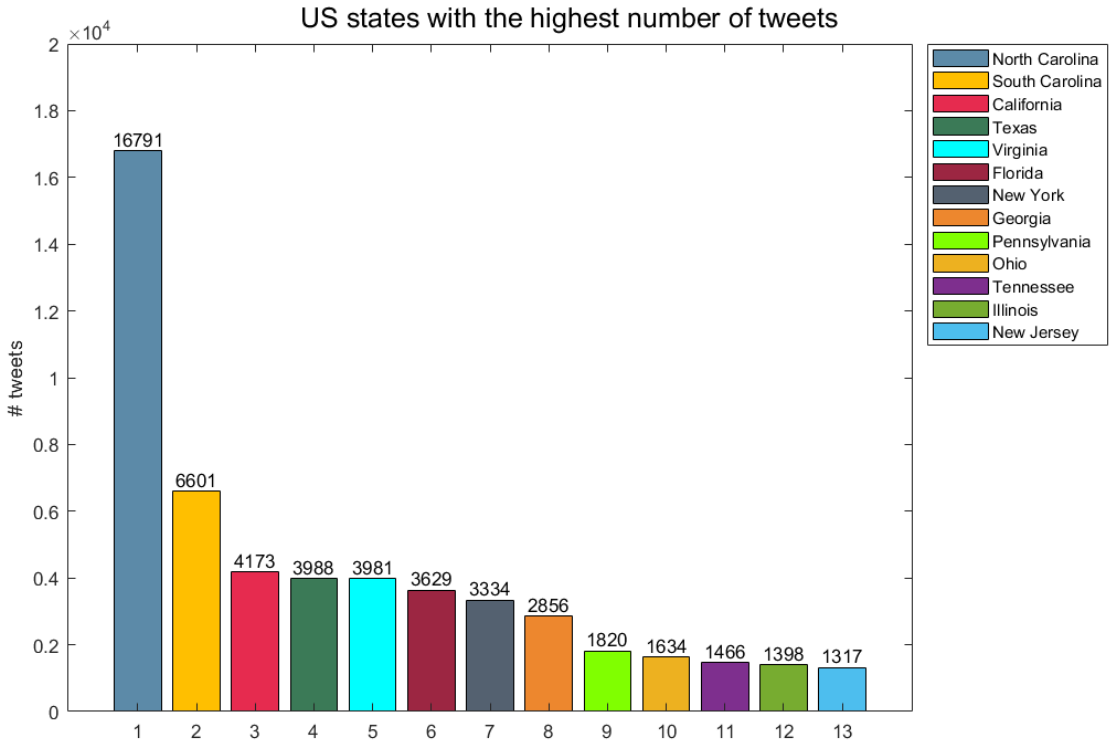

Figure 43 US States that satisfied the first criterion associated to the total number of tweets geolocated within their borders

Most of the resulting US territories are located along the Eastern part of the United States with the solely exceptions of California and Texas. The presence of these last two could be effectively associated to the proportion of their active Twitter user groups, involved in the indirect reporting of news and information about the hurricane. However, as previously mentioned, the direct dependence of Twitter users on real population is not proven. Although, a possible proxy of the type of California and Texas social media activities could be further explored through the next criteria of the procedure.

The most directly affected countries according to the NOAA report (North Carolina, South Carolina and Virginia) have been found within the first five position. The high rate of North Carolina could then be explained by the relevant amount of reported damages and fatalities, more than the double of the ones reported in the other two neighbouring states, as shown in Figure 19.

Considering the evolution in time of the hurricane Florence whose strength gradually decrease after the landfall on September 14th, it is important to evaluate the temporal trends that drove the Twitter activity. Figure 44 shows the daily trend of the Twitter activities highlighting a strong peak in correspondence of September 14th for North Carolina where hurricane Florence made its landfall. Before the peak day, the Twitter publication activity increased constantly. A similar behaviour could be observed for South Carolina whose peak was recorded on September 13th. However, in this case it has been detected a smoother peak with significant number of tweets distributed mainly over four days (September 12th-15th). Another important difference between the Carolinas curves is represented by the decreasing slope after the peak. In South Carolina, the Twitter activity rapidly decreased and attested on a low number of tweets per day after September 17th, when Florence was declassified as tropical depression in North Carolina the activity remained relevant but lower than the one during the first alert days. This could be motivated by the damage reporting activities that influenced more the North Carolina territory. Nonetheless, after October 2nd, the South Carolina series reported a small increase whose evolution can not be completely explored because the tweet collecting period ended on October 4th. This behaviour should be further investigated.

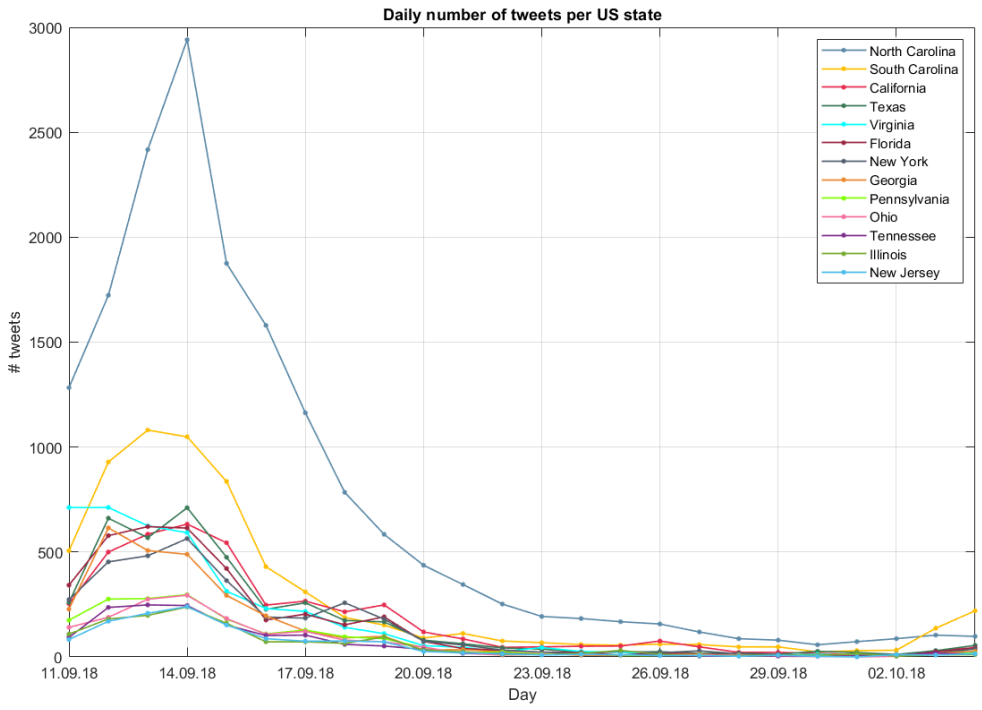

Figure 44 Temporal daily trend for tweets published in the US states that satisfied the Criterion I

The other US states recorded less sharp peaks with smaller variation from the Twitter activity rate associated to the days after the hurricane alert dismission. The maximum values and peak days for each US state are shown in Table 3. Many of the states registered their maximum number of geolocated tweets on the landfall day. Virginia, Florida and Georgia had their Twitter posts peaks on the days before, the reason for that could be due to the effect of the storm surges occurred along their Atlantic coasts when hurricane Florence was still offshore but already caused relevant wind events, as mentioned by the NOAA report. Eventually, it is important to highlight that Florence was already declassified to a tropical depression when it crossed Virginia territory. Tropical depression is a meteorological event whose effects are considered less dangerous than the ones of a hurricane or of a tropical storm. This could be the reason of the different moderate Twitter trend of this state in comparison with the most relevant ones detected in North and South Carolina.

Table 3 Peak values and days for the US states that satisfied the Criterion I

US state	**Maximum number of tweets**	**Peak day**
North Carolina	2941	September 14th
South Carolina	1081	September 13th
California	633	September 14th
Texas	711	September 14th
Virginia	712	September 12th
Florida	621	September 13th
New York	564	September 14th
Georgia	615	September 12th
Pennsylvania	297	September 14th
Ohio	294	September 14th
Tennessee	248	September 13th
Illinois	238	September 14th
New Jersey	240	September 14th

4.2.2 Filtering by geolocation type

A directly affected user is probably more prone to share an exact location through the *coordinates* attribute of a tweet that gives the information about the occurrence of a severe weather event. According to this assumption, the Criterion II gives more importance to the US countries whose percentage of tweets geolocated through the exact method is greater or equal than the national one (8%). Only 3 of the 13 states satisfied the requirement: North Carolina, South Carolina and Virginia. Figure 45, 46 and 47 illustrate the composition of the Twitter subset associated to the US states resulting from the second step of the procedure.

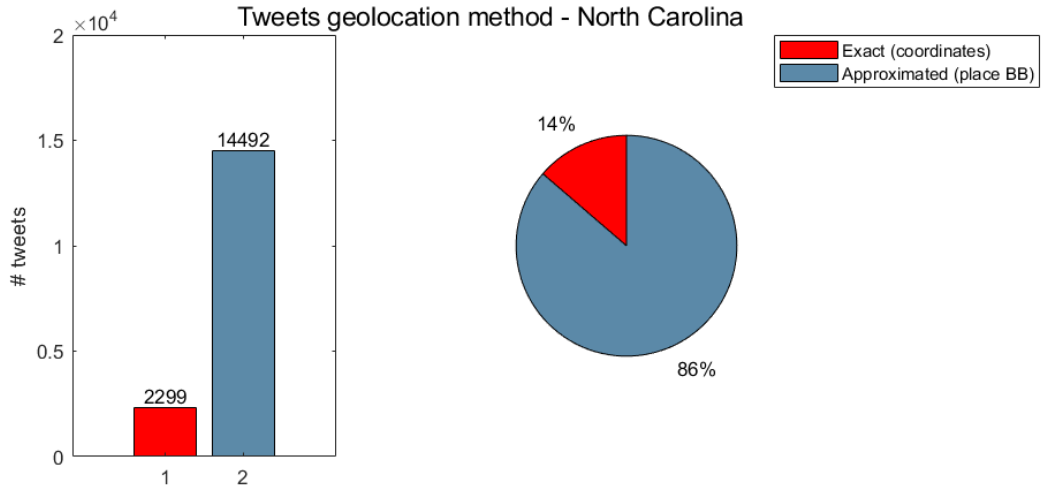

Figure 45 Composition of geolocated tweets in North Carolina

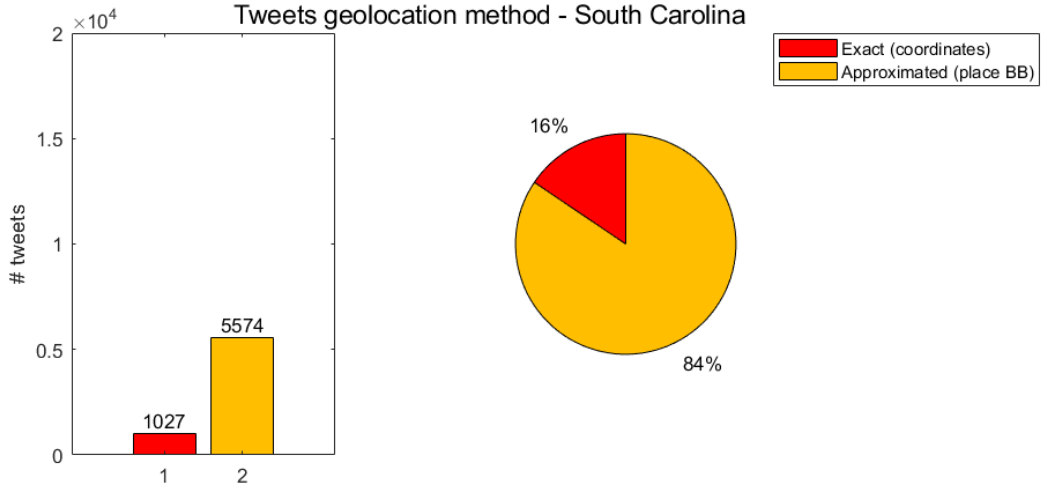

Figure 46 Composition of geolocated tweets in South Carolina

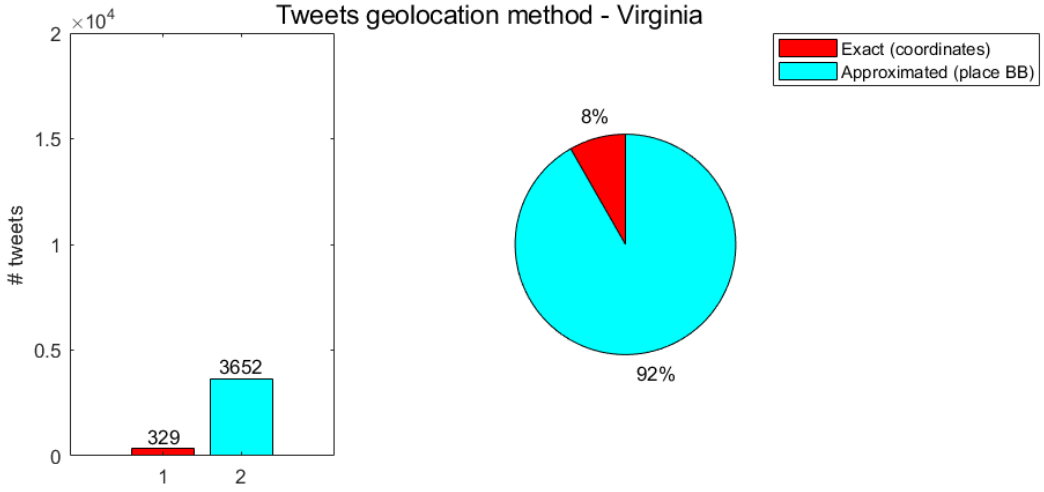

Figure 47 Composition of geolocated tweets in Virginia

Both North Carolina and South Carolina significantly exceeded the threshold, with the second one that doubled the percentage value of the US national dataset. On the other hand, Virginia recorded the same percentage, satisfying the criterion without a wide margin. At this stage, it is relevant to compare these three US states with the other two among the original 5 most active US states per number of tweets: California and Texas. The graphs in Figure 48 and 49 show that both countries were characterised by percentages of social media posts that were significantly lower than the reference one. This could be explained by the fact that California and Texas users, who were not directly affected by hurricane Florence, were posting generic comment about the event and the emergency. Consequently, they logically did not care about sharing their exact position because it would have resulted useless if associated to the content of their tweets.

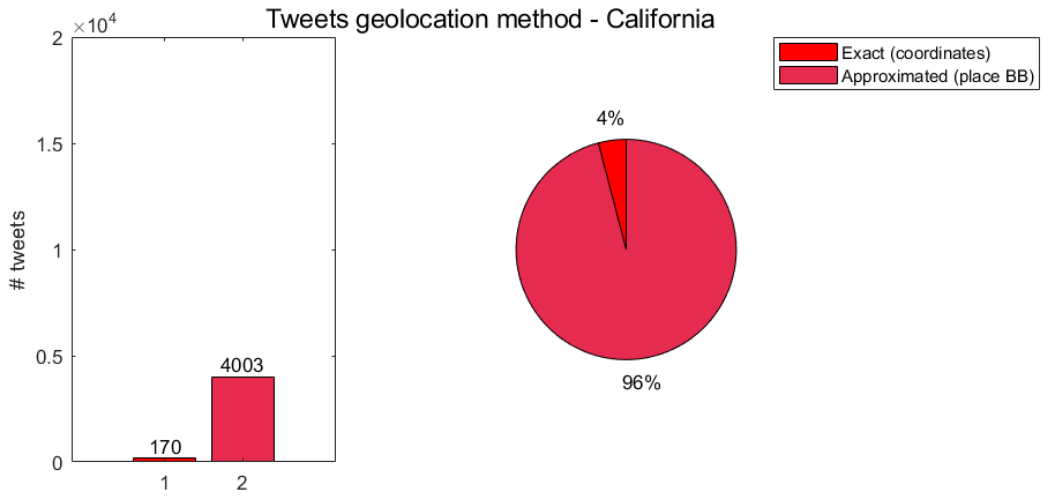

Figure 48 Composition of geolocated tweets in California

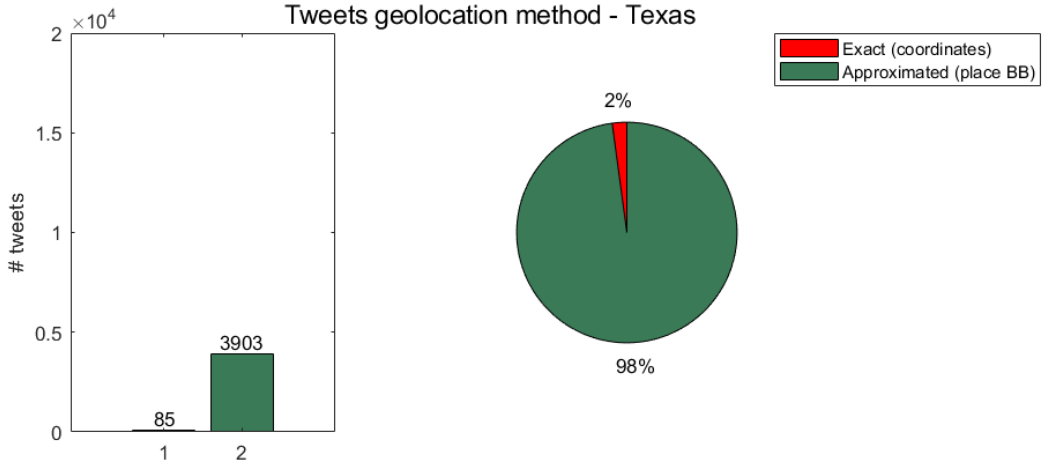

Figure 49 Composition of geolocated tweets in Texas

4.2.3 Filtering by social media post nature

In a similar way to Criterion II, the percentage threshold for original tweets posted during the collecting time window was satisfied again by North Carolina, South Carolina and Virginia as depicted in Figure 50, 51 and 52. For all the three states, original tweets represented the majority of the dataset, suggesting that Twitter users were more interested in contributing to the social media conversation with their original and personal contributions about hurricane Florence. As already noticed in the previous steps, Virginia satisfied the filtering procedure but without the wide margin that characterised both North and South Carolina.

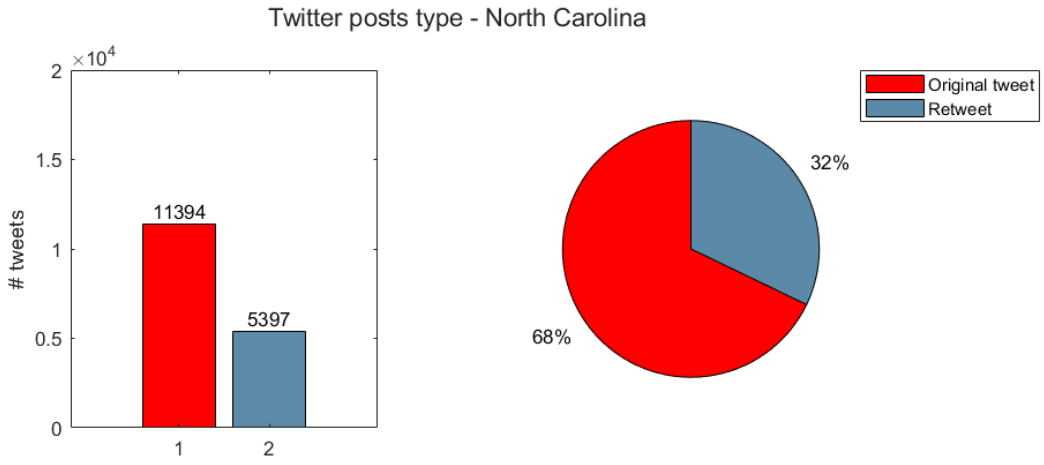

Figure 50 Typology of Twitter posts geolocated in North Carolina

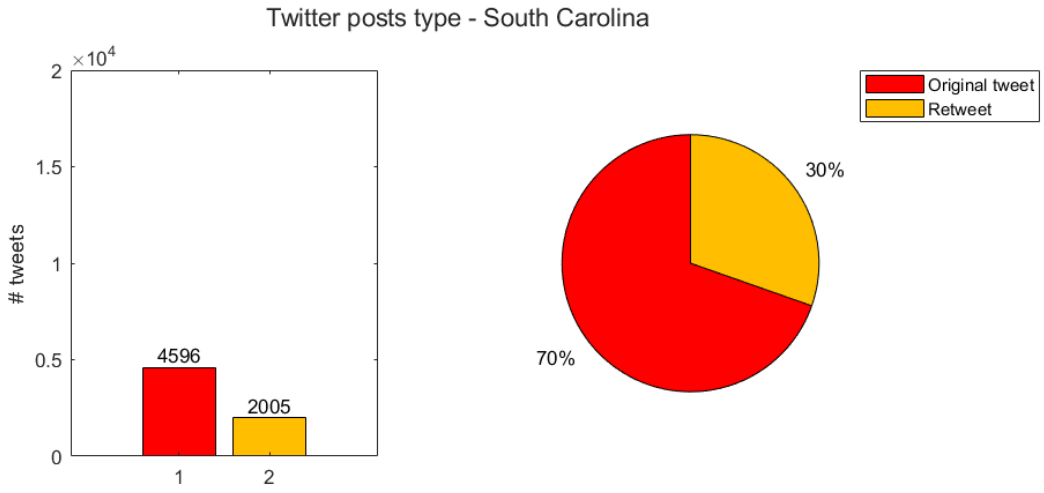

Figure 51 Typology of Twitter posts geolocated in South Carolina

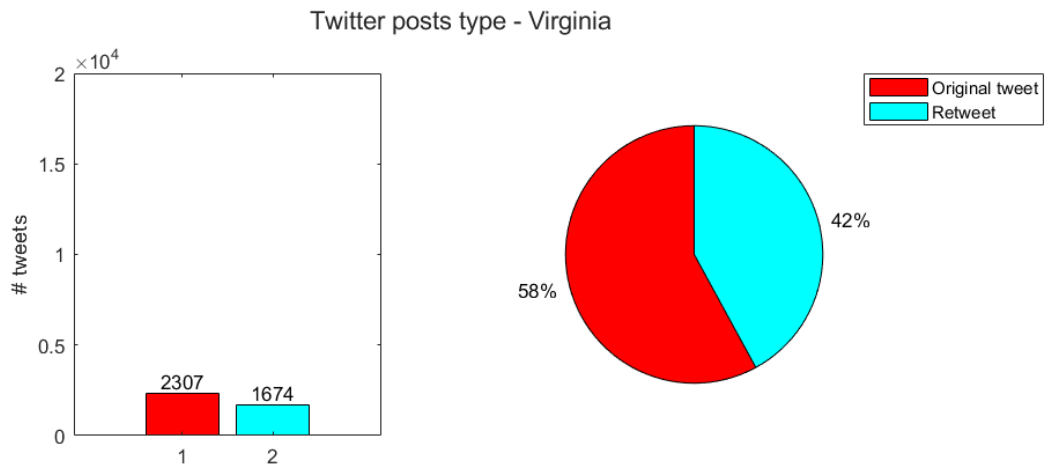

Figure 52 Typology of Twitter posts geolocated in Virginia

An additional comparison with California and Texas - that were filtered out in the previous step - revealed another different behaviour of the users who posted tweets located within the borders of this US states. Figures 53 and 54 illustrate the composition of the Twitter subsets by nature of the post and highlight the fact that in the case of these two countries the quoted retweet component is way more relevant than the original one. These observations supports the assumption that not affected users are more likely to share or re-post contents created by other users, powering an SMGI dissemination process associated to the recirculation of the most impressive hurricane-related tweets or of the official emergency alerts and reports.

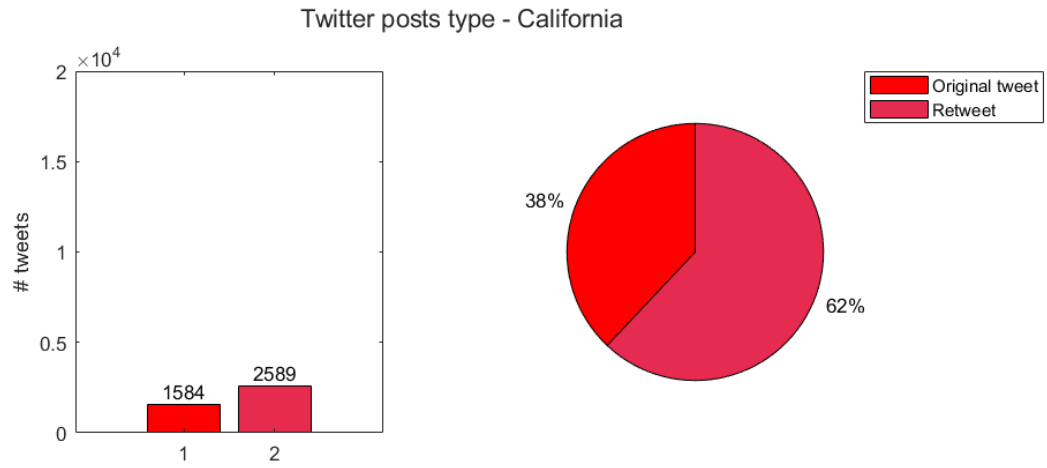

Figure 53 Typology of Twitter posts geolocated in California

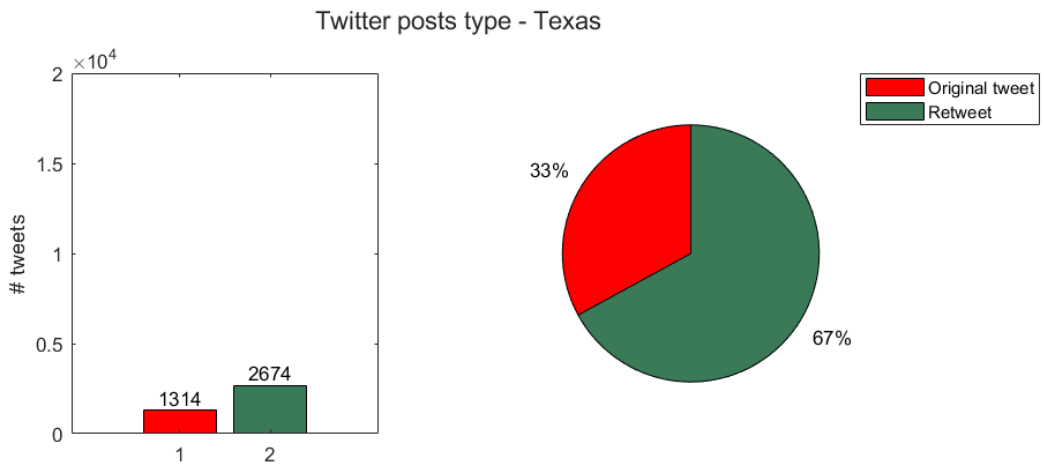

Figure 54 Typology of Twitter posts geolocated in Texas

In conclusion, only three neighbouring US states satisfied all the criteria of this procedure: North Carolina, South Carolina and Virginia. Having identified this area on the East coast, it has been possible to have a first evaluation of the spatial distribution of the Twitter dataset, computing the total number of tweets geolocated in each county during the collecting time window. These cartographic elaborations used the same method of Figure 40 but at a regional scale. Exploring these three Twitter subsets, it has been found that all these US countries contained some tweets geolocated only through the *place* geotag associated to a boundary box referring to the entire countries ('North Carolina, USA', 'South Carolina, USA' and 'Virginia, USA'). Consequently, in the cartographic representation, the approximation of the tweets coordinates – corresponding to the centroid of the place bounding box - results in a complete and meaningless overlap of all the concerned social media posts in a single point that could negatively affect the next spatial analyses. For this reason, all the tweets associated to the three geotags previously specified have been removed from the subsets of North Carolina, South Carolina and Virginia. As it could be seen in Table 4, the removed tweets represented relevant percentages of the Twitter subset of each US state, highlighting again the complexity of the spatial accuracy of social media posts.

Table 4 Characteristics of the national generic place geotags for North Carolina, South Carolina and Virginia

***Place* geotag**	**# tweets with only approximated coordinates**	**% of the entire state subset**
North Carolina, USA	3938	23,45
South Carolina, USA	2121	32,13
Virginia, USA	819	20,57

However, the removal of these tweets does not affect the previously explained procedure because these geotags, whose *place.type* is of administration level, are valuable at a regional scale, identifying a specific area across the entire United States. This observation highlights additionally the need of analyses at multiple spatial scales. Moreover, within the actual Twitter geotag hierarchy, a *place* attribute could significantly vary its meaning and values at different scales, implying specific approaches and considerations.

Finally, after the outliers filtering procedure, Figure 55, 56 and 57 illustrate the classification of North Carolina, South Carolina and Virginia counties by the resulting number of tweets geolocated within each counties' borders. The cartographic representations also include the Florence path with its classification and the hurricane wind swath, whose dimensions – calculated by NOAA with meteorological models based on observations and statistical processes – represent the area directly affected by Florence hurricane .

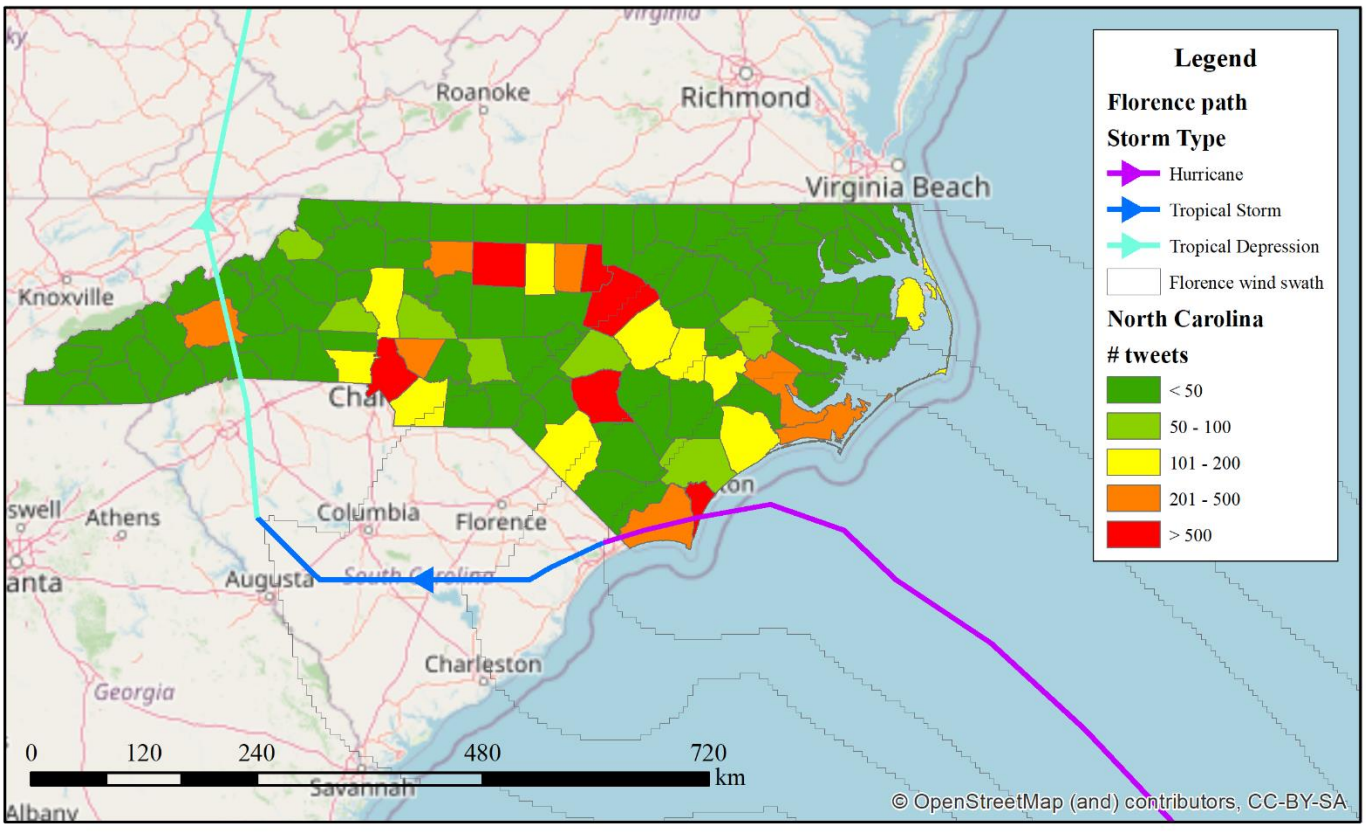

Figure 55 North Carolina counties classified by the number of tweets geolocated within their borders

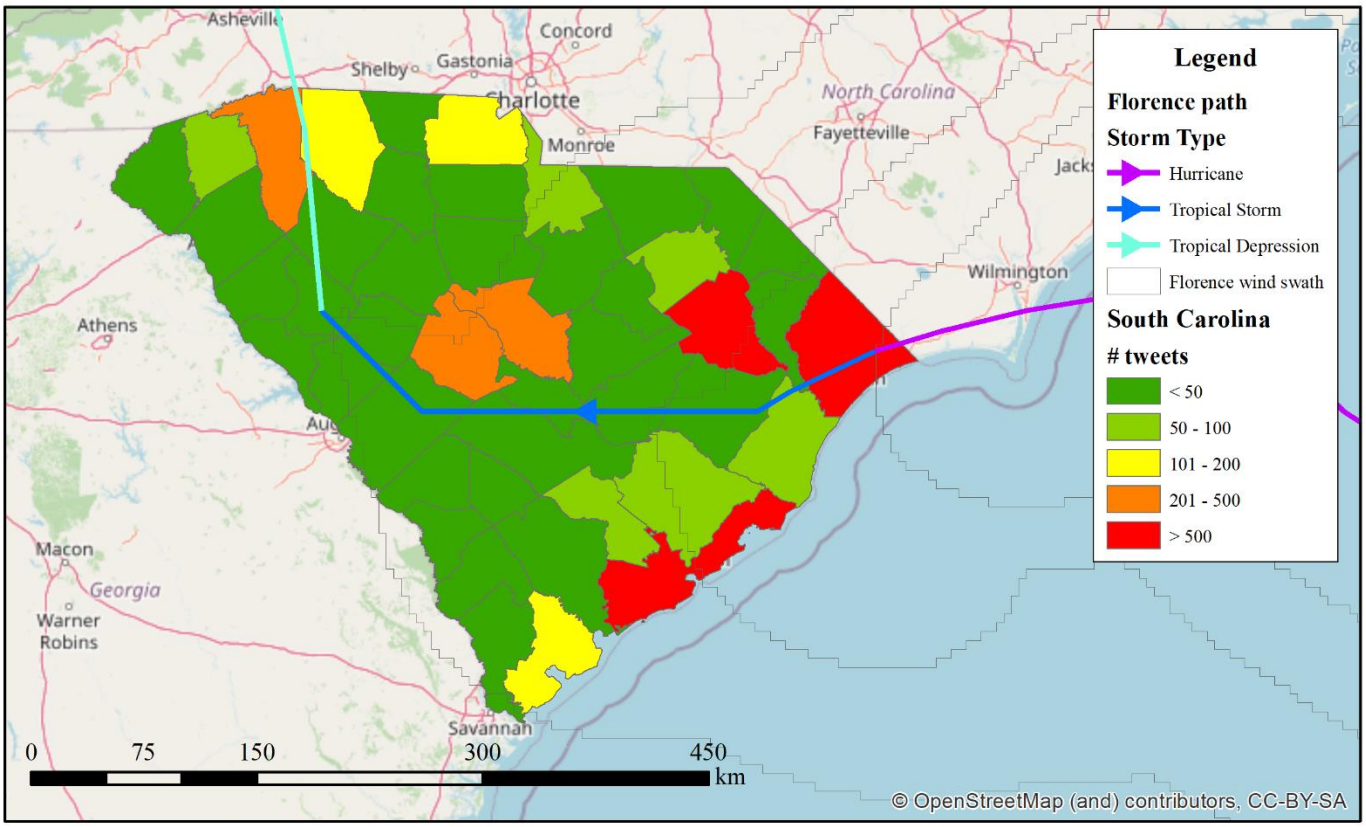

Figure 56 South Carolina counties classified by the number of tweets geolocated within their borders

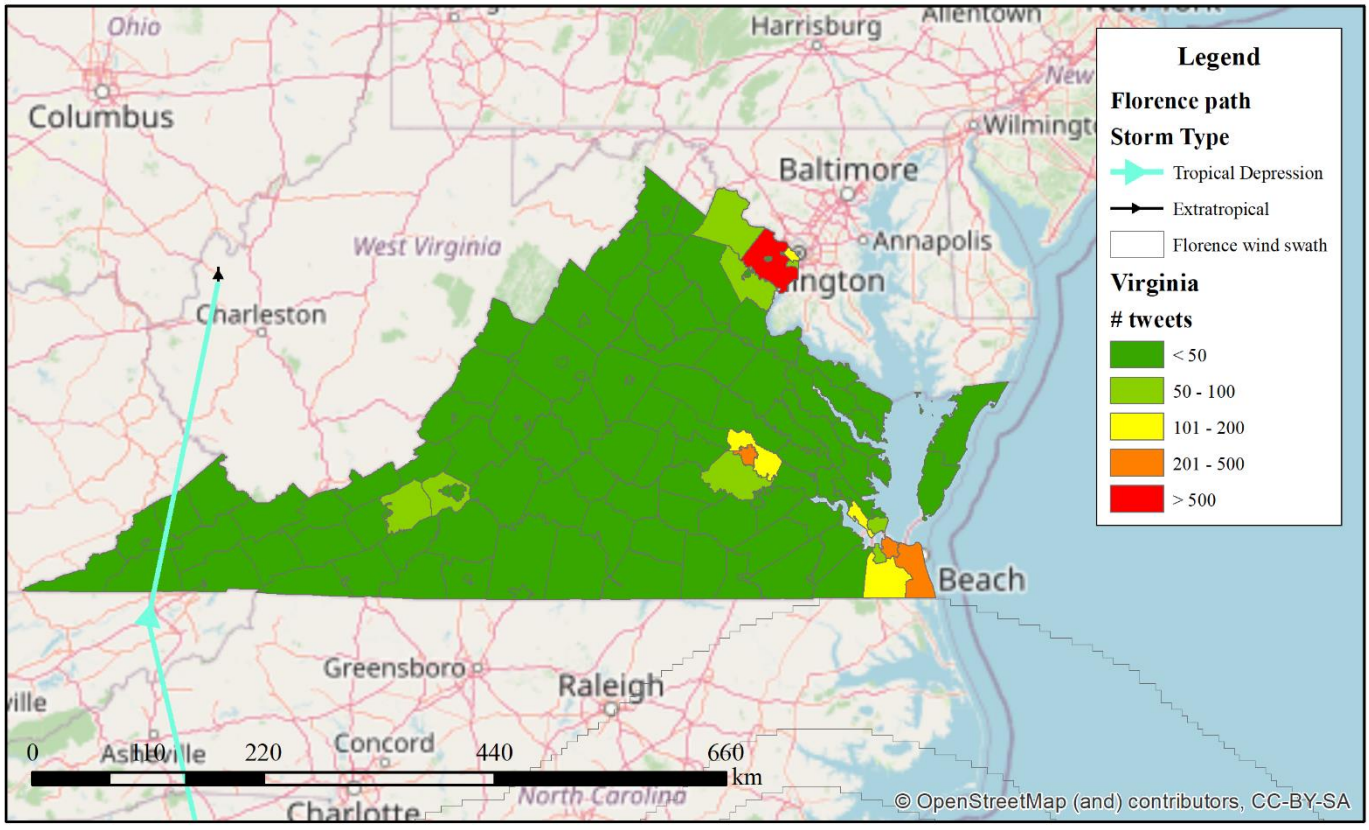

Figure 57 Virginia counties classified by the number of tweets geolocated within their borders

Table 5 lists the counties that registered a relevant Twitter activity – depicted in red colour in Figure 55, 56 and 57. In this final consideration about the identification process it is finally needed to include the value of the 2018 population for each counties in order to understand the influence of the demographic factor at a regional scale.

Table 5 Counties with more than 500 tweets in North Carolina (NC), South Carolina (SC) and Virginia (VA)

County	State	#tweets	2018 population
Wake	NC	2880	1091273
Mecklenburg	NC	2506	1093750
Horry	SC	982	344105
New Hanover	NC	826	232256
Durham	NC	753	316979
Charleston	SC	656	406222
Cumberland	NC	642	333430
Guilford	NC	622	532607
Fairfax	VA	566	1148463
Florence	SC	534	138277

In the state of North Carolina, 6 out of 100 counties registered more than 500 tweets with the special cases of Wake and Mecklenburg that both had more than 2000 tweets within their borders. It is important to notice that Mecklenburg was not included in the Florence wind swath. In 2018 this county was also the most populous one and its county seat, Charlotte, with its surrounding area was the largest metropolitan area in the Carolinas, as reported by US Census Bureau. Its high level of Twitter activity could then be linked to the population. However, the total number of tweets could be linked to effect of the hurricane like power outages and traffic interruptions. The second most populous county, Wake, was partially covered by the hurricane wind swath, so the total number of 2880 geolocated posts could be motivated by direct effects on its main city Raleigh, North Carolina capital. New Hanover, instead, is the county were Florence made its landfall as a category 4 hurricane. According to the NOAA report, electricity was cut down for more than 90% of the county and its county seat Wilmington reported 2 direct deaths. Also, Cumberland was directly affected by Florence. Nonetheless, both Durham and Guilford, whose 2018 population were higher than the average of North Carolina county population, 103816 inhabitants, were out of the hurricane radius but their activity could be motivated by the fact that their user population were involved in commenting and sharing information related to what was happening in the neighbouring areas.

South Carolina Twitter subset identifies 3 counties out of 46. All of them are within the wind swath and Horry was directly crossed by hurricane Florence. On the other hand, Charleston could have seen its Atlantic coasts impacted by storm surges and sea swells. Florence county, according to NOAA report, was affected by serious flooding along the Lynches river causing the evacuation of some cities. However, it is important to consider that the Twitter activity in this area could be affected by the toponym effect similarly detected in Italy for the city of Firenze, as previously detected in Chapter 3.

Fairfax was the only county of Virginia who was characterised by more than 500 geolocated tweets while the other counties had an average of 20 tweets. The high rate of activity in this area, 2018 most populous Virginia jurisdiction, could be explained by its proximity to Washington D.C., the US capital were many of the national emergency agencies involved in the rescue and safety operations had their head offices.

In conclusion, as previously mentioned, it is important to remember that when Florence was classified as a severe hurricane and then as a dangerous tropical storm, Virginia was not directly hit. Also, the Florence wind swath, that represents the footprint of this severe weather event, did not cover Virginia territory. For this reason, in the next analyses, only North and South Carolina have been considered. However, the population factor needs to be included in the further more detailed spatio-temporal analyses in order to detect with greater efficiency the areas most affected by the hurricane.

4.3 Spatio-temporal evolution of Twitter activity in North and South Carolina

The complexity of the Florence-related SMGI required additional tools able to combine the spatial and the statistical factors for understanding the dataset distribution, considering the proximity and not only the position. In order to achieve this goal, only geolocated tweets posted in North and South Carolina between September 13th and 17th, 2018 have been considered. This choice relies on the fact that between those days Florence was classified as a hurricane and then as a tropical storm crossing the Carolinas territory. After September 17th, Florence has dissipated its energy as an extratropical event.

During the data preparation, it has been detected the presence of 56 possible outliers published by 47 separate users in the area between Charlotte and Raleigh in North Carolina. All these tweets associated to the *place* geotag “North Carolina, USA” but with additional and exact *coordinates* attribute resulted overlapped on a single point with coordinates 35,5; -80. From the comparison with Bing and Maxar satellite imagery, these latitude and longitude values are associated to a rural area without traces of visible buildings or residential man-made objects. This observation suggested that

these tweets were suspicious elements that could affect the analyses results. The reason associated to this overlapping could be explained in the fact that these users, who published only these tweets for the entire time window, retrieved the post position from their profile location or tweeted using the same cell-site of the mobile whose centroid corresponded to the coordinates couple (Earl et al., 2012). Moreover, it has been excluded the use of a same IP address considering the absence of man-made features in this area. However, considering the ambiguity on the geolocation method of these tweets, the concerned 56 Twitter posts have been filtered out and not used in the next elaborations.

Finally, Figure 58 shows the total number of the remaining tweets for each day and illustrates the temporal trend for the social media posts geolocated in the Carolinas. The peak day for the entire area is in correspondence of September 14th, 2018 for both tweets with exact and approximated geolocation. The portion of tweets geolocated through the coordinates attribute was constant during these five days, representing approximately the 21% of the daily total number of geolocated social media posts in North and South Carolina. This percentage is less than a quarter of the SMGI subset, but it is greater than the one of the North and South Carolina for the entire period from September 11th to October 4th (Figure 45 and 46). Then, this difference could be again motivated by the fact that Twitter users during the hurricane were more engaged in sharing their exact location.

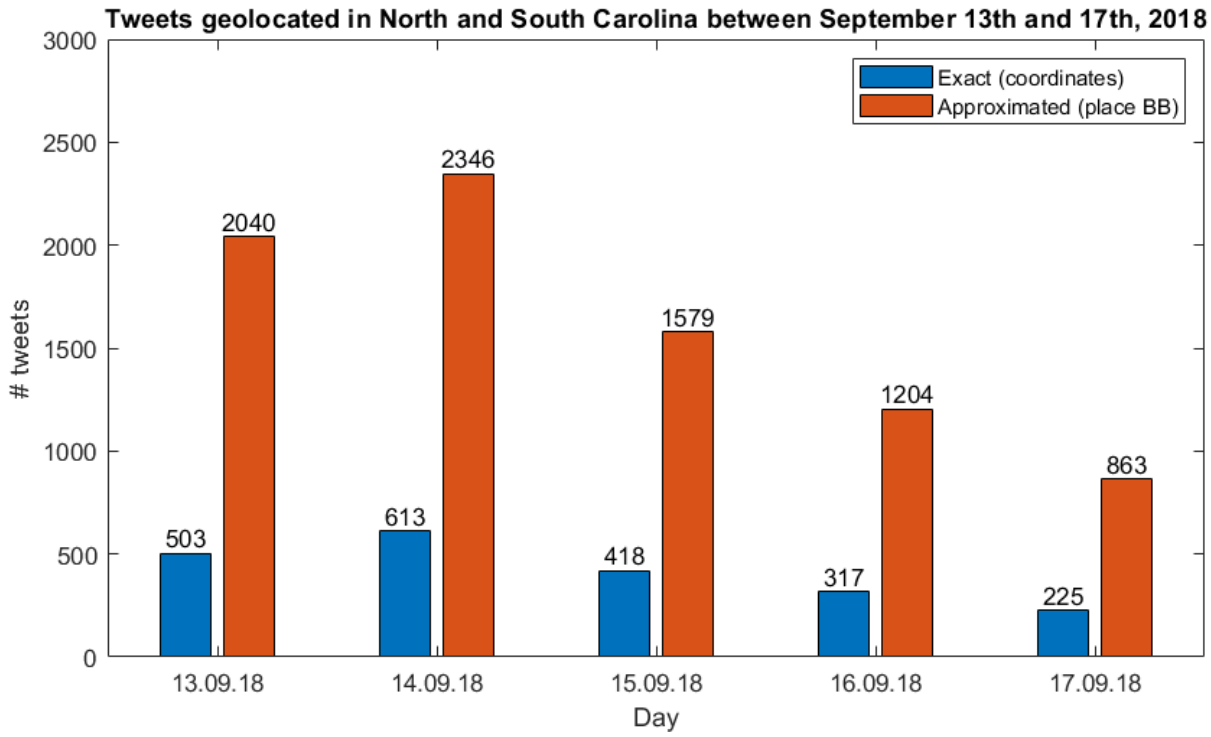

Figure 58 Temporal trend of Carolinas tweets classified by geolocation typology

4.3.1 Point Pattern Analysis

In order to have a first concise evaluation of the spatial characteristics of the dataset, the Nearest Neighbour Index (NNI) have been calculated for each day performing a Point Pattern Analysis (PPA). NNI is indeed an effective indicator for evaluating the dispersion of a dataset (Clarks & Evans, 1954). This index is calculated using the following formula:

$$NNI = \frac{D_o}{D_E}$$

with:

$$D_O = \frac{\sum_{i=1}^{n} d_i}{n}; \quad D_E = 0{,}5\sqrt{\frac{A}{n}}$$

where D_O is the average observed distance between each point and the nearest point, while D_E is the expected distance in the case of random distribution (random, hypothesis H_0). The result of this calculation is interpreted in relation to the H_0 hypothesis:

- NNI > 1 implies a *dispersed* distribution for which specific trends are not identified in the input subset.
- NNI < 1 classifies the subset as *clustered*, characterised by the significant presence of grouped tweets.
- NNI=1 identifies a *random* intermediate dispersion.

The total area of North and South Carolina, equal to 430316,36 km², has been used as the input surface value A. In this way the values obtained are comparable with each other. NNI has been calculated for each day and the calculated indexes are reported in Table 6.

Table 6 Nearest Neighbour Indexes for geolocated tweets posted between September 13th and 17th in North and South Carolina.

Day	NNI
13/09	0,178426
14/09	0,152383
15/09	0,183239
16/09	0,204280
17/09	0,218055

NNI detected a clustered distribution for all the five days considered. September 14th, when the hurricane Florence landfall occurred in North Carolina, was identified as the day with the most clustered distribution. In the days after the value of NNI gradually increased but resulting always smaller than 1. These results could be motivated by major concerns about the hurricane severity in the preparation phase the day before its arrival and by the weather updates during the day of the

landfall. However, these low values could be simply motivated by the fact that, as previously mentioned, the Twitter population is mainly composed by users in urban areas. Actually, a great concentration of active users in bigger city may results in big SMGI clusters that overrepresent that context in spite of rural areas that, even if possibly more affected by the event, could register a more dispersed level of Twitter activity do to their morphology or service accessibility. This representation bias is a crucial representativeness issue also for the next geo-statistical analyses that, in addition to the proximity of each SMGI element, considers also a thematic input parameters that is as a key element for understanding the geolocated tweets distribution in the Carolinas.

4.3.2 Spatial autocorrelation, hot-spot analysis and Kernel density

In order to further explore the Carolinas Twitter subset through their social media peculiarities, some spatial autocorrelation tools that combine the tweets spatial proximity with the influence of a specific input weight attribute assumed relevant for a geolocated Twitter activity.

The Getis Ord Gi* local index helps detecting clusters characterized by high value of a given weight attribute (hot-spots) and those with low values (cold-spots). Formulated in the early 1990s (Getis & Ord, 1992), it is measured with the following formula:

$$G_i * = \frac{\sum_{j=1}^{n} w_{i,j} x_j - \bar{X} \sum_{j=1}^{n} w_{i,j}}{S \sqrt{\frac{n \sum_{j=1}^{n} w_{i,j}^2 - \left(\sum_{j=1}^{n} w_{i,j}\right)^2}{n-1}}}$$

With:

$$S = \sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - \overline{X^2}}$$

Where x_j is the attribute value for feature j while $\bar{X}$ is its average value. The $w_{i,j}$ element is the *spatial weight* between feature i and j and n is equal to the total number of features.

$G_i *$ positive values that are gradually higher correspond to hot-spots of increasing intensity, while similarly negative values define cold-spots of different intensity. The index result is associated to a dotted graphic representation with colours varying from red to blue depending on whether the value is positive or negative. Elements that are not classified in either category are shown in white.

Additionally, useful insights about the tweets distribution and its interpretation are given by the Kernel density map. This tool provides a spatial evaluation of the Kernel Density Estimation (KDE), a key statistic for Exploratory Spatial Data Analysed (ESDA) that corresponds to the probability estimation for a random variable whose observations' positions are known (Silverman, 1986). This statistic tool needs as input a point dataset (Carolinas tweets subset) and gives as result a raster map representing the KDE variability for the area within the input features are distributed.

KDE is computed applying to each tweet point feature a so-called *kernel* function that represents a uniform curve surface whose peak value equal to 1 is calculated in correspondence of the point position. Increasing the distance from a given point, the kernel function value decreases and becomes null when the input search radius distance is reached. Finally, the density is calculated for each raster grid cell, summing the resulting kernel values of the overlapping surfaces linked to the input point features.

The key elements for the KDE map computation are:

- The *output cell size*, that defines the dimensions of the output regular grid cells. Consequently, a bigger cell size would include more point features and the resulting map uniform and less detailed. On the other hand, smaller cell sizes imply few points and a density map whose level of detail is so high that a spatial trend could be hardly detected.
- The *search radius* which defines the spatial range and smoothness of the kernel function.
- The *population field*, a weight attribute whose influence is considered relevant for the observed spatial phenomena.

Considered the level of detail required for the analyses in North and South Carolina, for the Kernel density map computation a cell size of 2 km and a search radius of 50 km have been chosen.

In order to evaluate only the tweets spatial distribution in time, a first weight attribute has been defined using as reference the social media post sharing time. Increasing *Day* attribute values was defined as reported in Table 7. The cartographic product resulting from the computation of the hot-spot analysis using this weight is presented in Figure 59.

Table 7 "Day" population field values assigned for tweets posted during the given days

Tweet sharing datetime	***Day***
13/09	1000
14/09	2000
15/09	3000
16/09	4000
17/09	5000

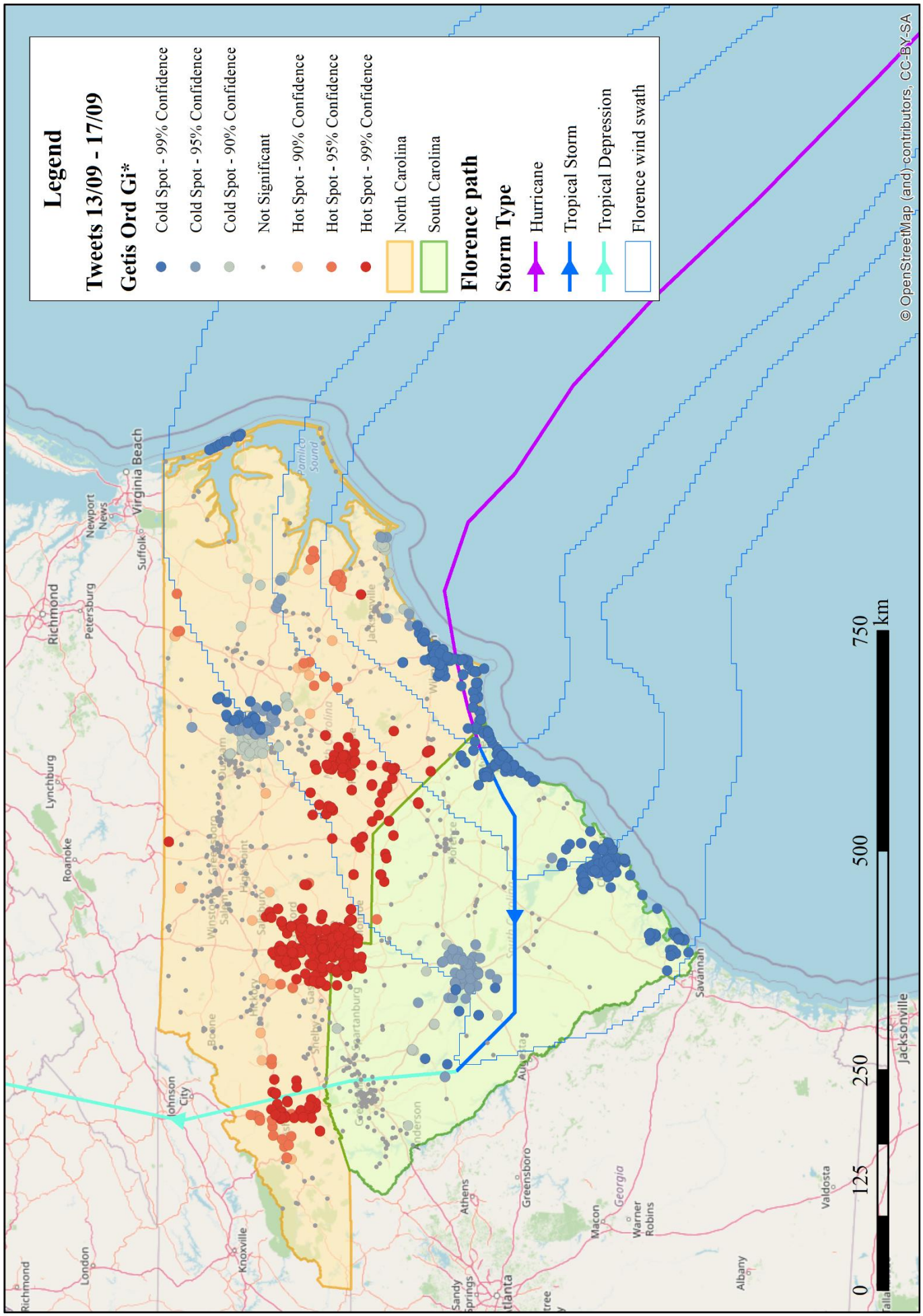

Figure 59 Twitter hot-spots and cold-spots distribution defined using the "Day" attribute for North and South Carolina

The resulting cold-spots (lowest values of Getis Gi*) have been identified in correspondence of tweets located along the Atlantic coast of North and South Carolina and in particular in the areas of Wilmington (NC), Myrtle Beach (SC) and Charleston (SC). This means that many of the Florence-related geolocated tweets posted during the first days of the considered temporal window – the ones with lowest value of the *Day* field – were geolocated in the Eastern area of the Carolinas. On the other side, the hot-spots resulted in the Western region of North Carolina and in the Charlotte (NC) area denotes that many tweets were posted between September 16th and 17th, reflecting the Florence circulation from East to West.

The previous analyses have found additional insights with considerations associated to a temporal analysis for the 4 most active counties (Table 5) and for some relevant counties crossed by the hurricane centre in different days. Figure 60 shows the comparison for the daily Twitter activity – calculated as the daily number of tweets with coordinates normalized by the total number of geolocated social media posts posted in the considered time window - in the most active counties in the Carolinas in order to possibly identify a different trend for territories directly crossed by Florence, covered by or out of the hurricane wind swath.

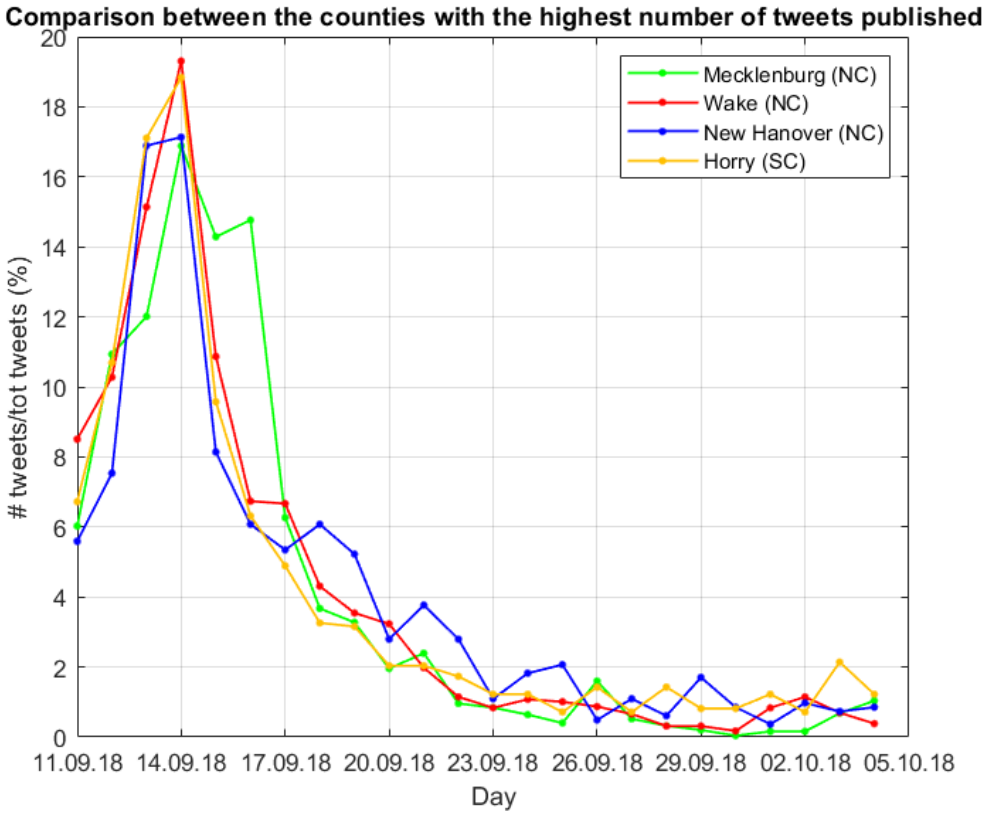

Figure 60 Daily trend for the 4 counties with the highest number of published tweets in the Carolinas

All the 4 counties registered their peaks on September 14th with a higher rate for Wake (NC) and Horry (SC). However, the shapes of the timeseries are different. In particular, the Wake and Horry series are characterized by a narrow peak band and rapid decrease of the normalized number of tweets after it. In a similar way, New Hanover (NC), the county where Florence made its landfall, registered a main peak but also other small peaks after September 17th, suggesting that the interest on the past hurricane event was still high also on the response and recovery phase especially in this area that was within the most damaged one. On the opposite side, Mecklenburg (NC), the most active county based on its total number of tweets, is characterised by two main peaks on September 14th and 16th with a larger peak band. This could be due to the relevant media coverage and attention on the event at the moment of the landfall but also to the previously mentioned fact that between September 15th and 17th Charlotte, major city of Mecklenburg, NC, was indirectly affected by the effects of Florence.

The daily trend has been specifically analysed from 8 counties grouped as eastern, inland, and western counties as listed in Table 9 and illustrated in Figure 61.

Table 8 Carolinas counties chosen for the daily trend analyses

Trend groups	Counties
Eastern	New Hanover (NC), Brunswick (NC), Horry (SC)
Inland	Florence (SC), Richland (SC), Lexington (SC)
Western	Greenville (SC), Buncombe (SC)

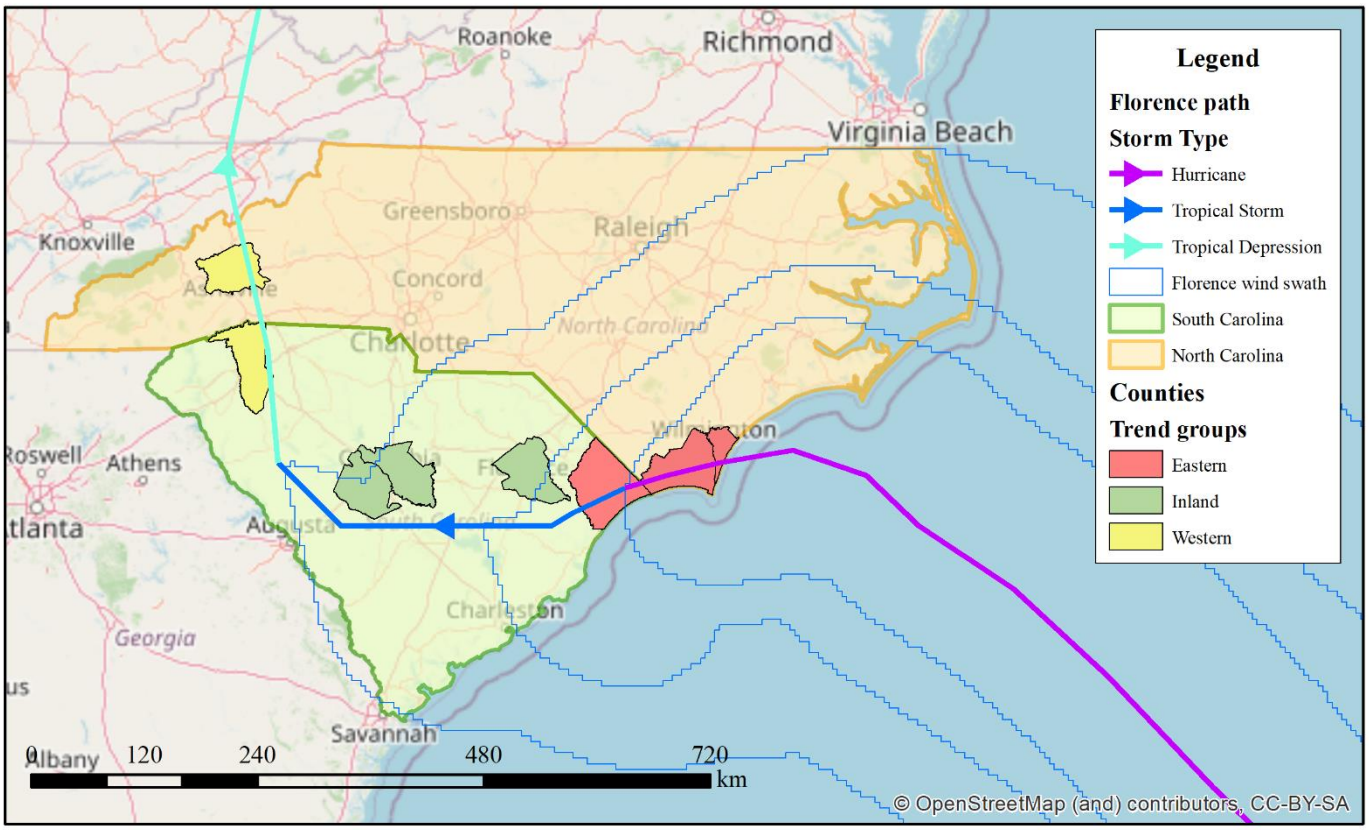

Figure 61 Distribution of the Carolinas counties chosen for the daily trend analyses

Eastern counties on the Atlantic Coast are characterised by a significant peak associated to September 14th (Figure 62). As previously highlighted for New Hanover, in the case of Brunswick, located at the border between North and South Carolina, the decreasing slope shows an oscillating behaviour with smaller peaks in the considered time window, probably due to Florence long-term consequences reported in the area.

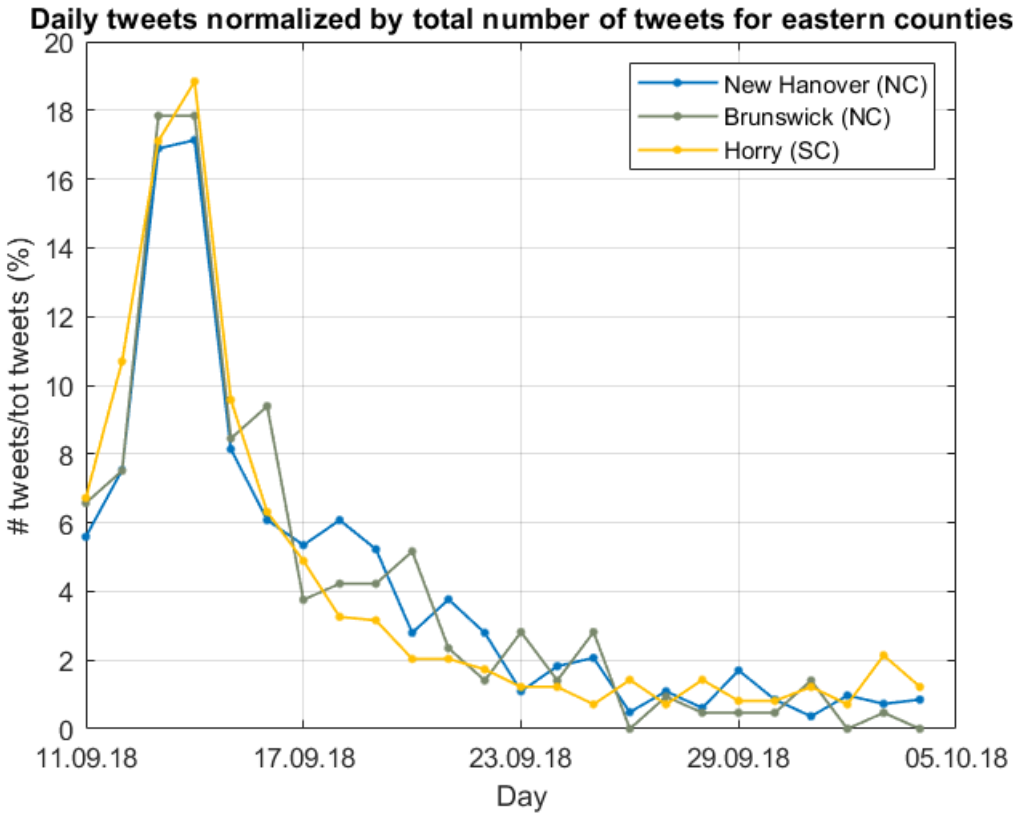

Figure 62 Daily trend for the chosen eastern counties in the Carolinas

The chosen inland counties were not directly crossed by Florence path but were all covered by its wind swath. The reasons of this choice were mainly associated to the fact that Florence county was within the 10 most active counties in this area and the problem with the toponymy brought many questions associated to its possible influence on the tweet collection while Richland and Lexington are the two counties within the most active in South Carolina. Also, within their territory it is located urban area of Columbia, the capitol of South Carolina, which was directly affect by flash flooding events (NOAA report). Figure 63 shows the resulting daily trends for the chosen inland counties. The graph shows peak bands around September 14th that are larger than the eastern ones. The social media rate remained indeed significantly higher than 10% also on September 15th. Additionally, it highlights a particular temporal behaviour for Florence that is different from the previous ones. It shows a similar peak around September 4th, but it identifies the highest percentage in correspondence to October 4th,

the last day considered for the tweet collection. Moreover, this unexpected peak could not be motivated by the toponym influence because, as shown before in chapter 3 for Firenze in Italy, it is a factor that is almost constant in time. It is indeed observable in the period between September 17th and October 2nd, with the time series oscillating between 2 and 4% instead of gradually approaching a zero value. The textual content of tweets published in Florence between October 3rd and 4th indicated that the cause of this abrupt increase is a shooting happened in the city of Florence where two law enforcement officers were killed and other 10 persons were injured, as reported by the CNN. The tweet activity in Florence on the entire period is then strongly affected by this event happened weeks after the hurricane circulation.

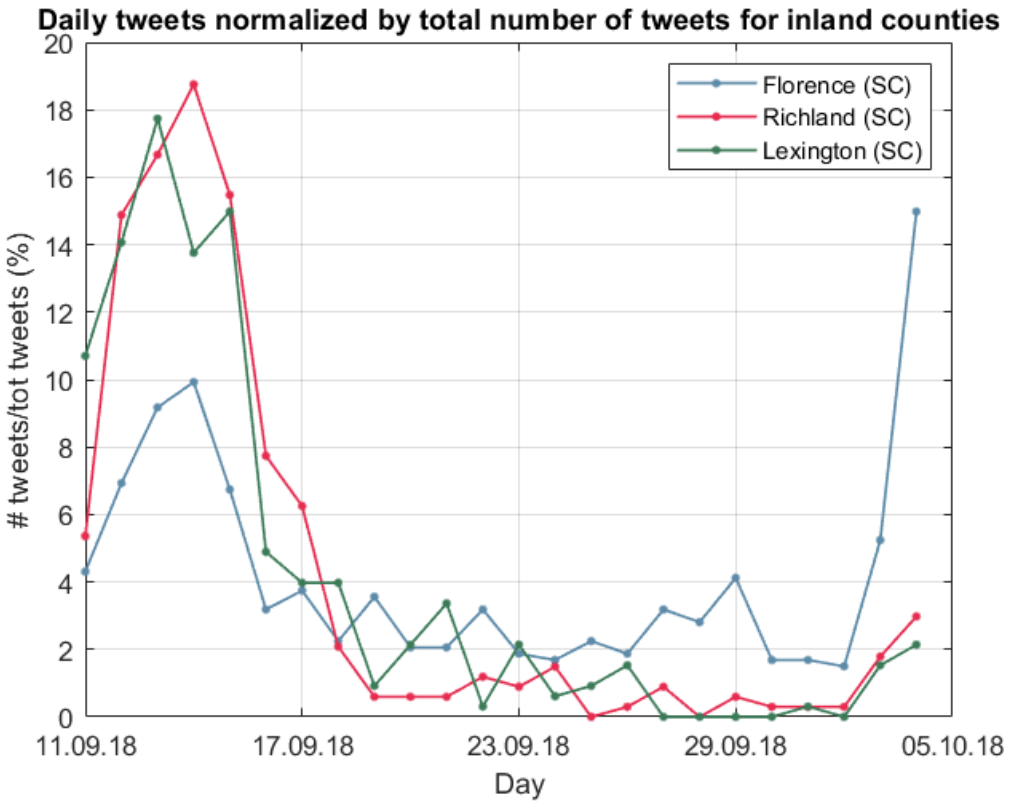

Figure 63 Daily trend for the chosen inland counties in the Carolinas

Eventually, the western counties of Greenville (SC) and Buncombe (NC) illustrates again a shape that is similar to the ones illustrated before but is characterized by a slightly larger peak band for which the Twitter activity rate remained greater than 10% of the total in correspondence of September 16th (Figure 64). It is important also to highlight that in this case both the counties, after the decreasing phase, reached the 0% social activity rate, suggesting that after Florence they were not involved in hurricane-related emergency and recovery operations. Also, when it crossed Greenville and

Buncombe, Florence was already declassified as a tropical depression, having lost most of its devastating energy.

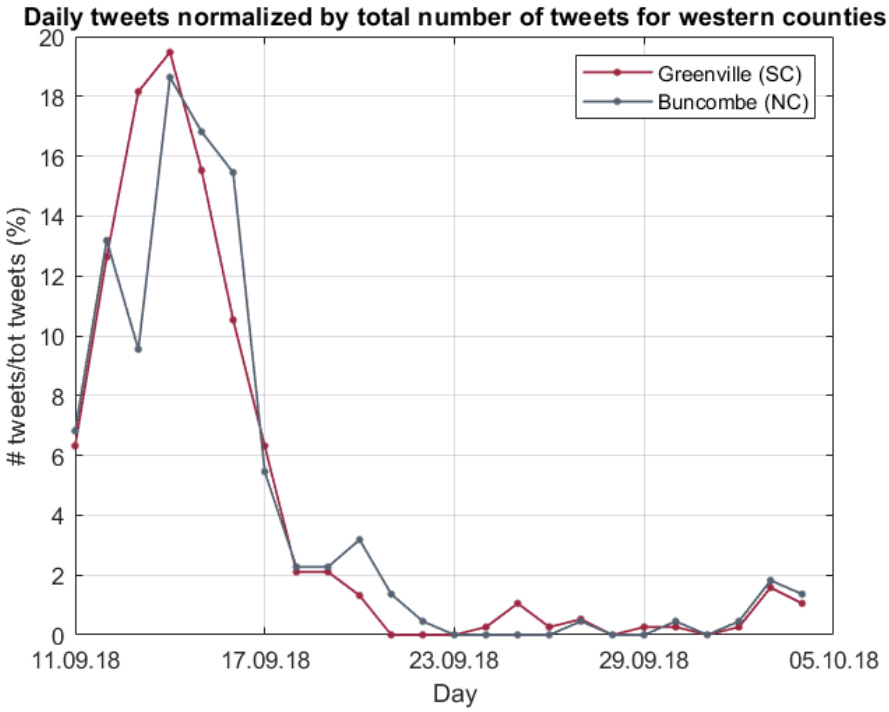

Figure 64 Daily trend for the chosen western counties in the Carolinas

The results on the temporal component of the considered North and South Carolina subset recommended a specific approach for the next steps, supporting the need of defining a smaller time window focused only on the days during which Florence crossed the Carolinas territory. For this reason, the previously defined 5-days window from September 13th and 17th has been considered also for the next analyses. These days were chosen considering the necessity of analyse both the influence of the alerts and warnings immediately before the landfall and the effect of the hurricane on the Twitter activity, assumed as a proxy of the Florence damages.

Before proceeding with more detailed Kernel Density computations, the main Twitter interaction attributes, *favorites_count* (total number of favorite reaction registered by each tweet) and *retweet_count* (total number of retweet action registered by each tweet), were analysed in order to understand if they would had been representative of the hurricane impact but especially of the Twitter population. The descriptive statistics associated to these two attributes are presented in Table 9.

Table 9 Descriptive statistics for the Twitter interaction attributes

Tweet attribute	Mean	Maximum	Standard Deviation
favorites_count	1,81	829	21,16
retweet_count	0,33	52	1,74

The two fields were then not used in the analyses after observing that most of the tweets that registered *favorites_count* and *retweet_count* values that were higher than the average – 63,3% of them - were published by journalists (CNN, FOX and weather channel) with verified Twitter accounts (premium user profiles with additional options for sharing contents and increment popularity on the social media platform) located mainly in Charlotte, NC and Raleigh, NC. It was indeed assumed that they were not representative of the majority of the Twitter population.

A more detailed hot-spot analysis and KDE evaluation has then been given by the $W_{RT,pop}$ defined through the nature of each tweet geolocated with exact *coordinates* field and the population of the county within each tweet is located. Social media posts with approximated *place* coordinates have been removed from the subset due to their ambiguous positioning method.

Then, the $W_{RT,pop}$ has been calculated for each tweet with the following formula:

$$W_{RT,pop} = \frac{W_{RT}}{P_{cat}}$$

Where W_{RT} is a weight based on the nature of the tweets that is equal to 100 if the social media is an original contribution, otherwise it is equal to 50 for quoted retweet. Instead, the P_{cat} is a parameter that considers the total population of the county inside which each tweet is positioned. Considering the high variability of 2018 population values in North and South Carolina, 5 population classes and their corresponding values have been defined according to the number of county inhabitants reported by US Census Bureau (Table 10).

Table 10 Population classification and corresponding P_{cat} values

Population (inhabitants)	P_{cat} value
≤ 120000	1
> 120000 & ≤ 250000	2
> 250000 & ≤ 400000	3
> 400000 & ≤ 750000	4
> 750000	5

The graphical results of the Getis Ord Gi* and Kernel Density computations are depicted in Figure 65, 66, 67, 68, 69, 70, 71, 72, 73 and 74.

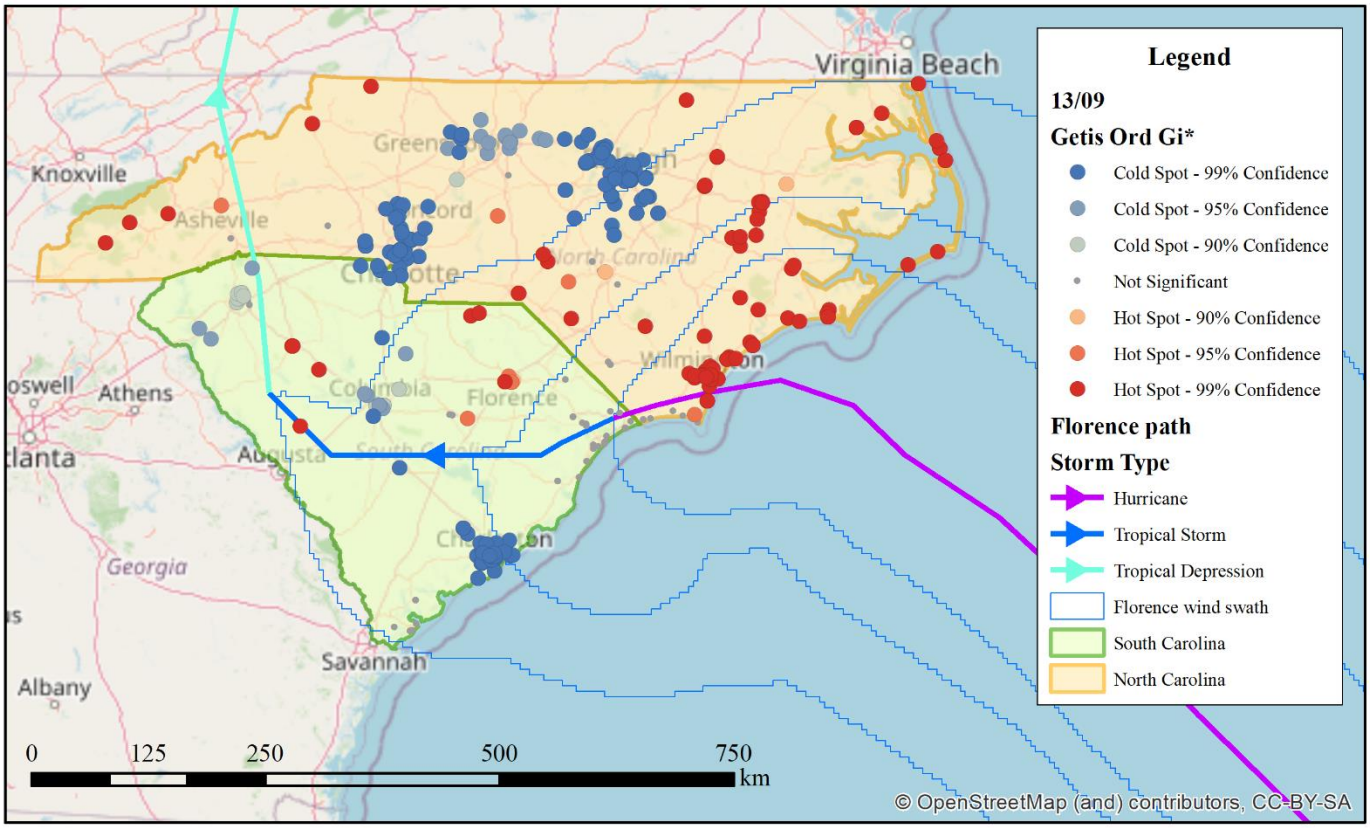

Figure 65 Hot-spots and cold-spots detected with the $W_{RT,pop}$ attribute for September 13th

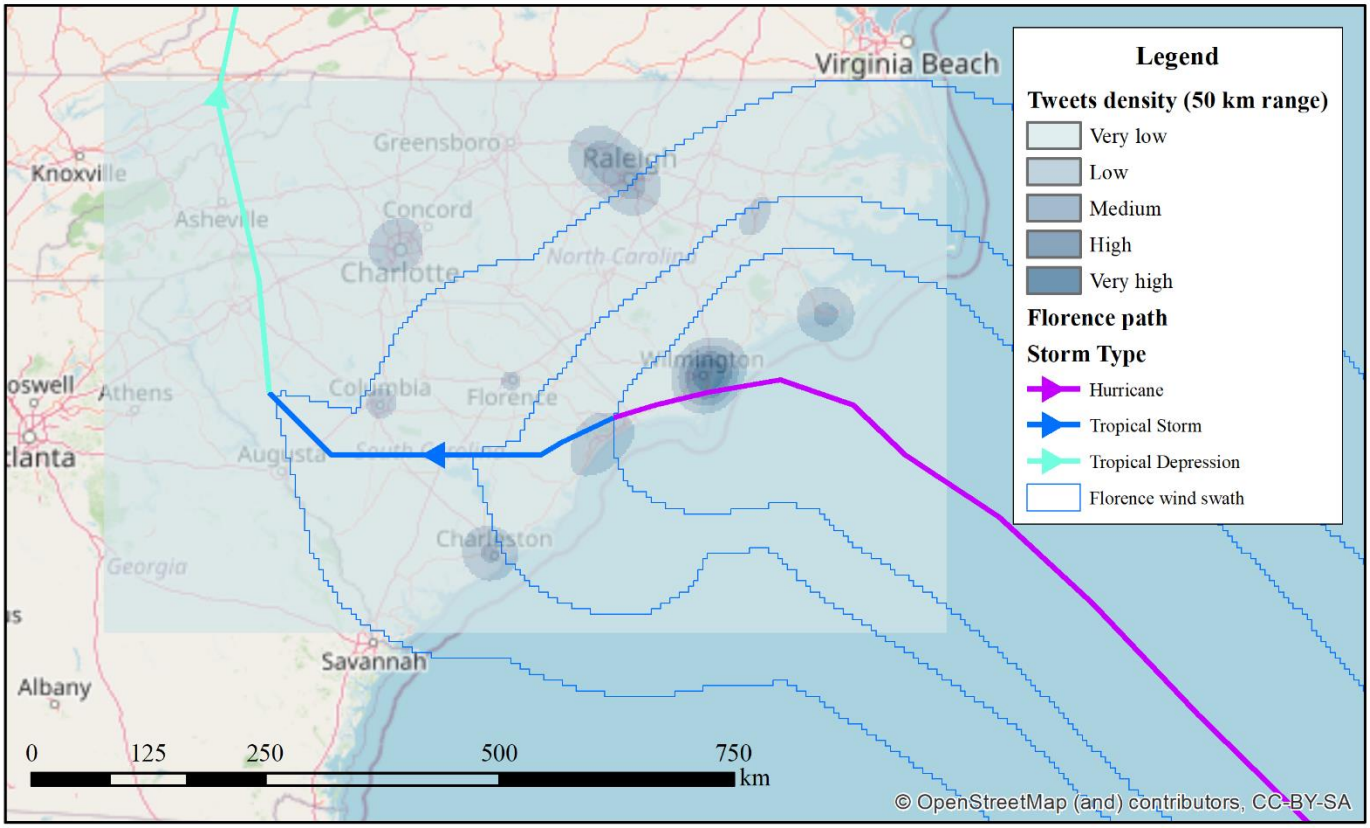

Figure 66 Kernel Density Map computed with the $W_{RT,pop}$ population field for September 13th

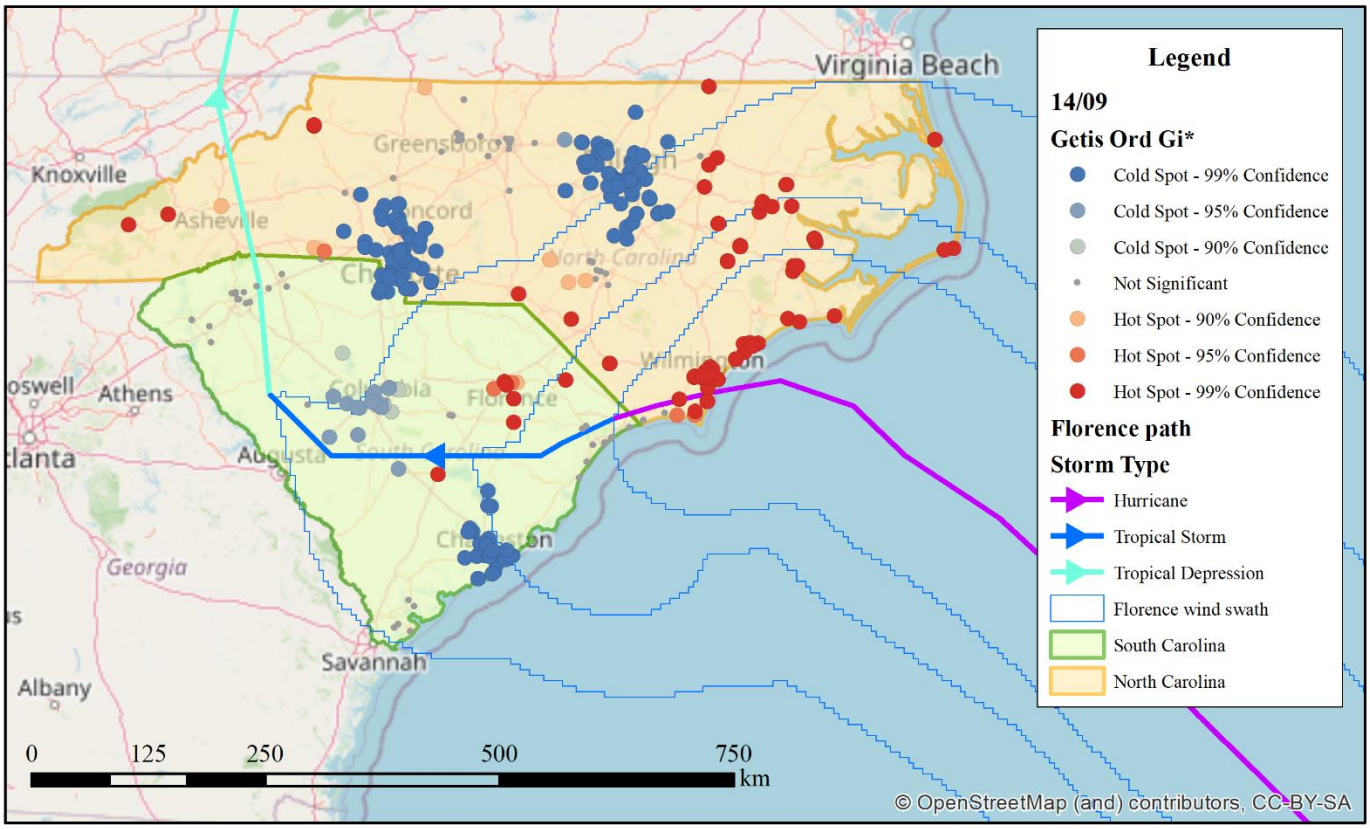

Figure 67 Hot-spots and cold-spots detected with the $W_{RT,pop}$ attribute for September 14th

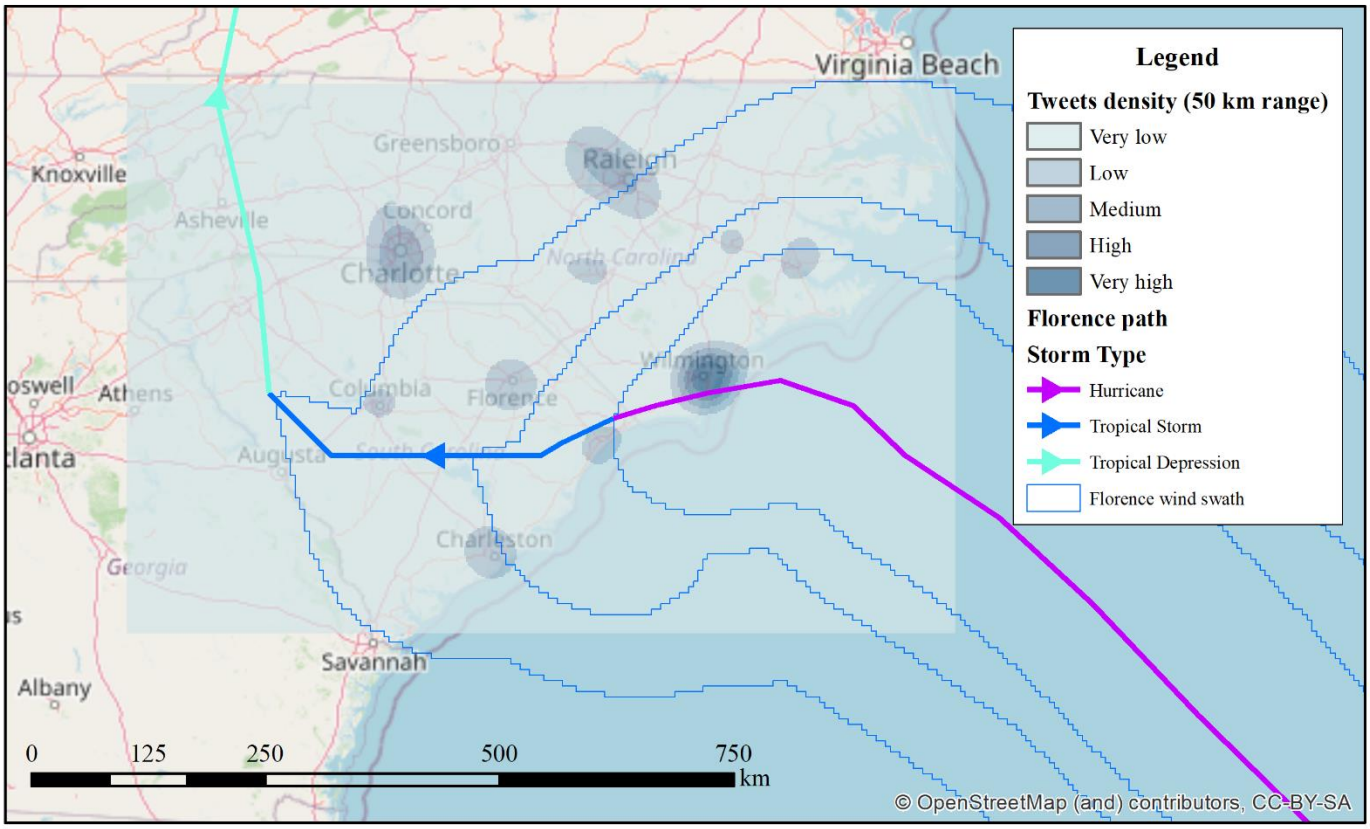

Figure 68 Kernel Density Map computed with the $W_{RT,pop}$ population field for September 14th

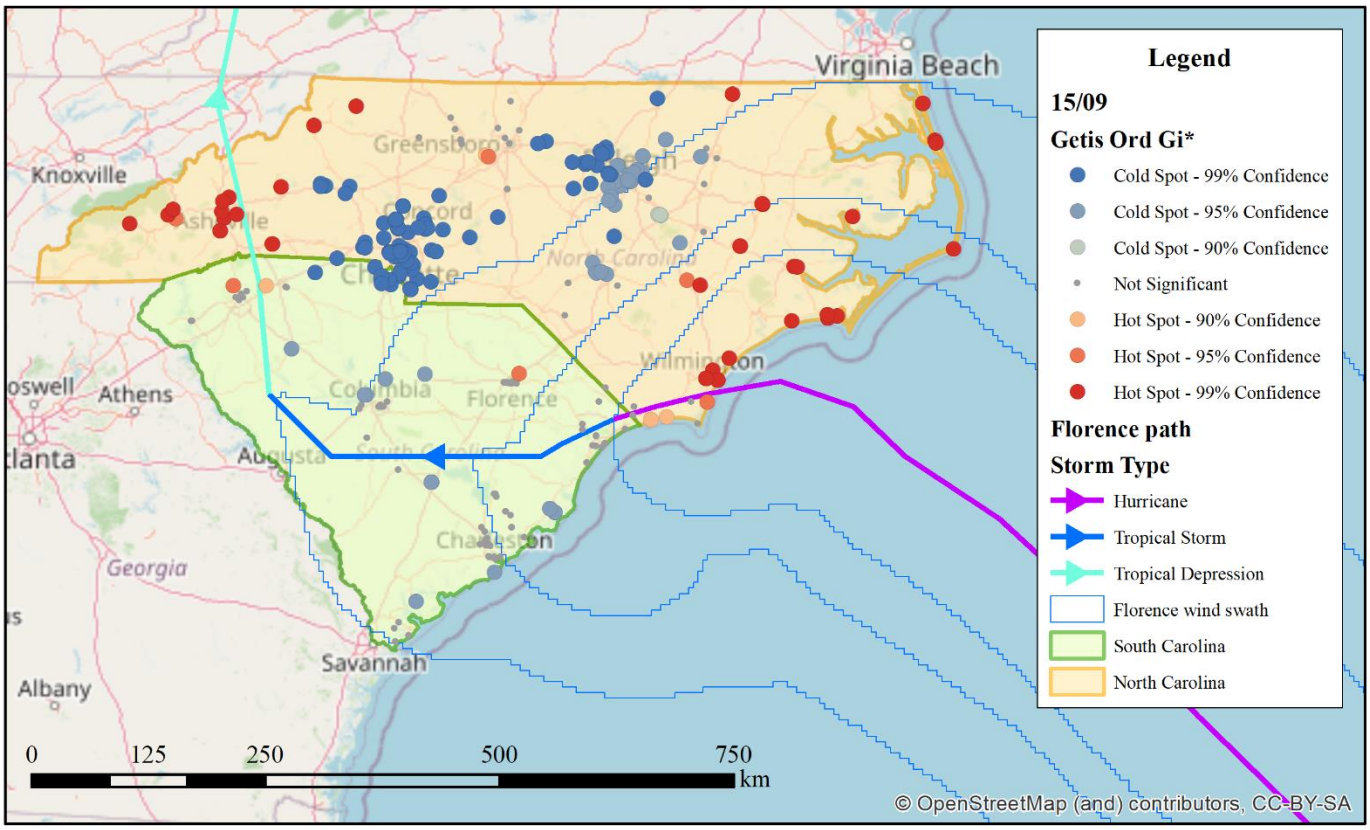

Figure 69 Hot-spots and cold-spots detected with the $W_{RT,pop}$ attribute for September 15th

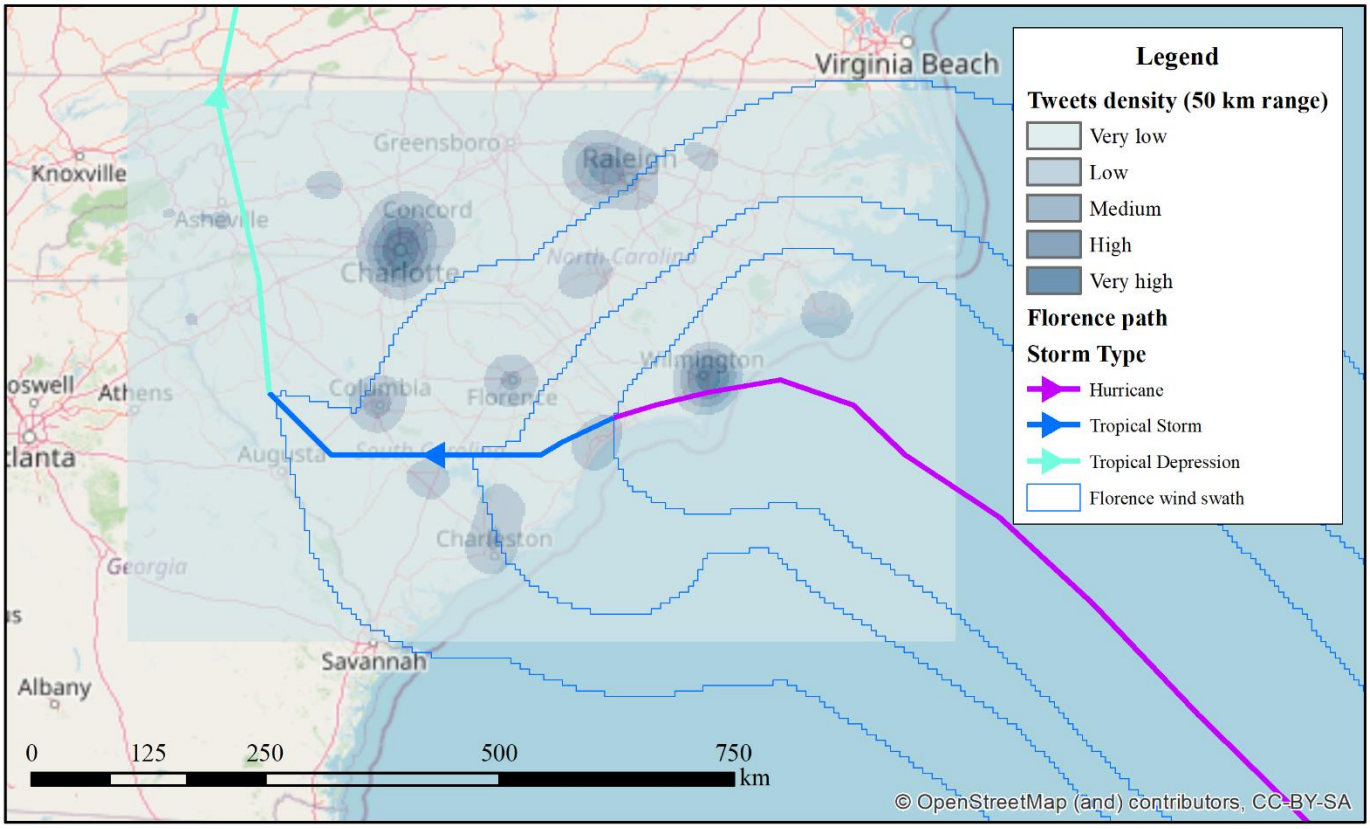

Figure 70 Kernel Density Map computed with the $W_{RT,pop}$ population field for September 15th

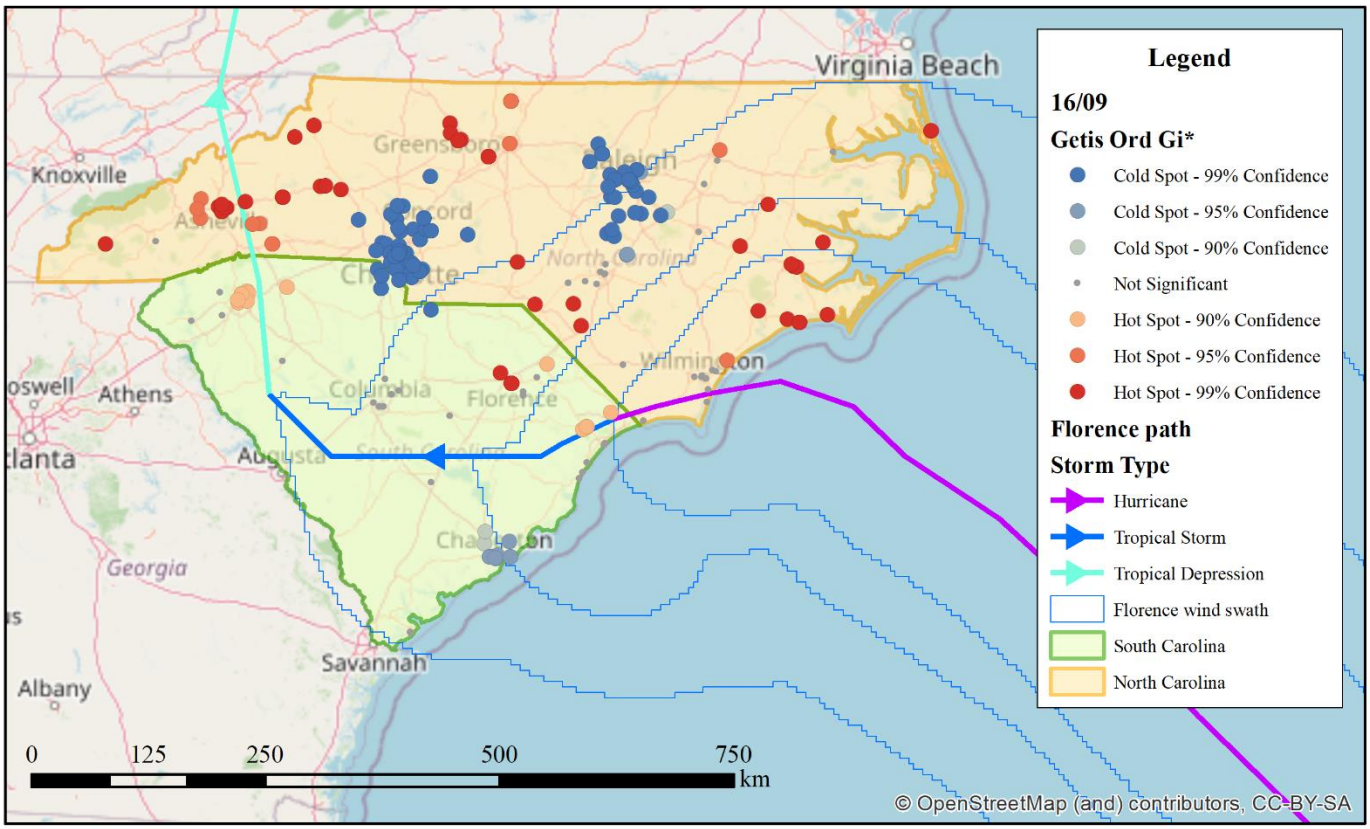

Figure 71 Hot-spots and cold-spots detected with the $W_{RT,pop}$ attribute for September 16th

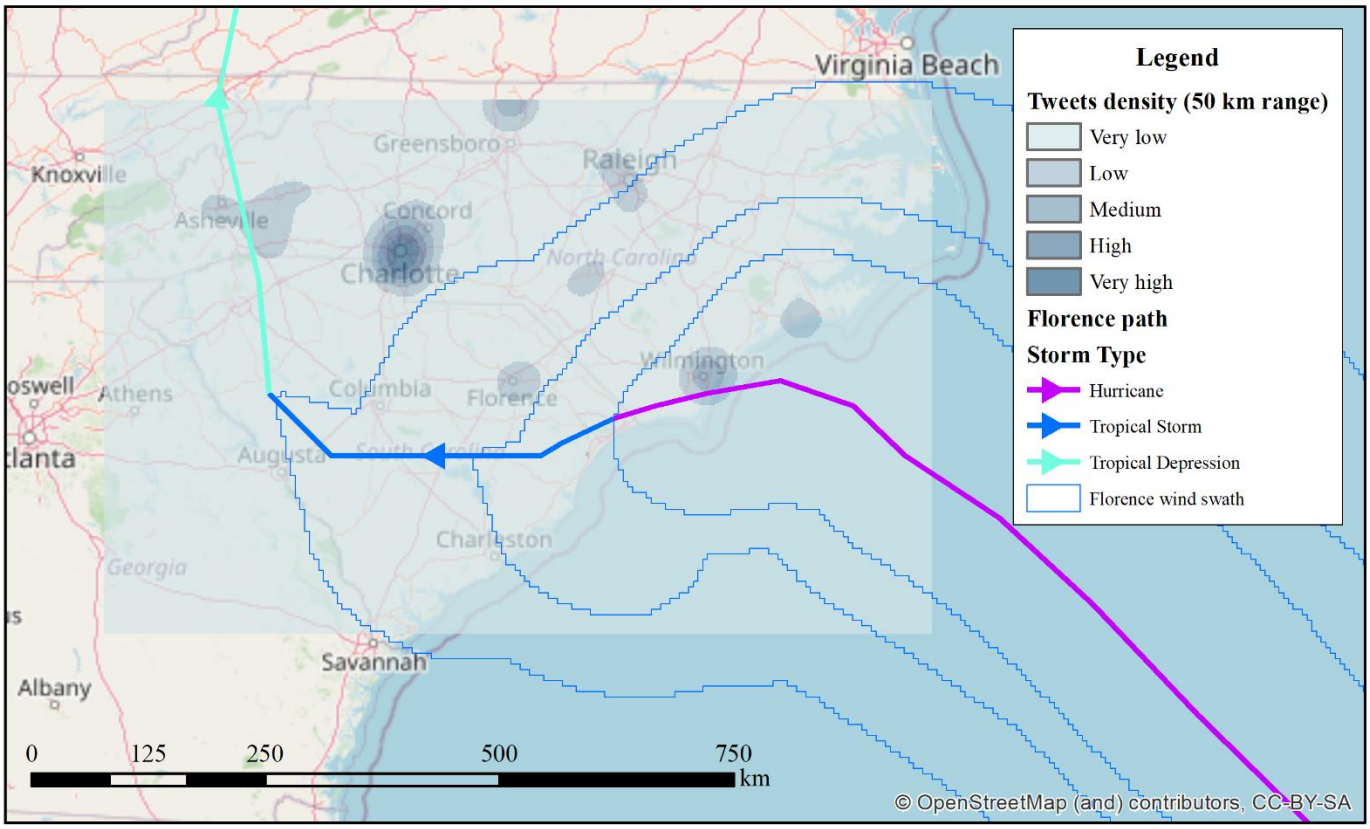

Figure 72 Kernel Density Map computed with the $W_{RT,pop}$ population field for September 16th

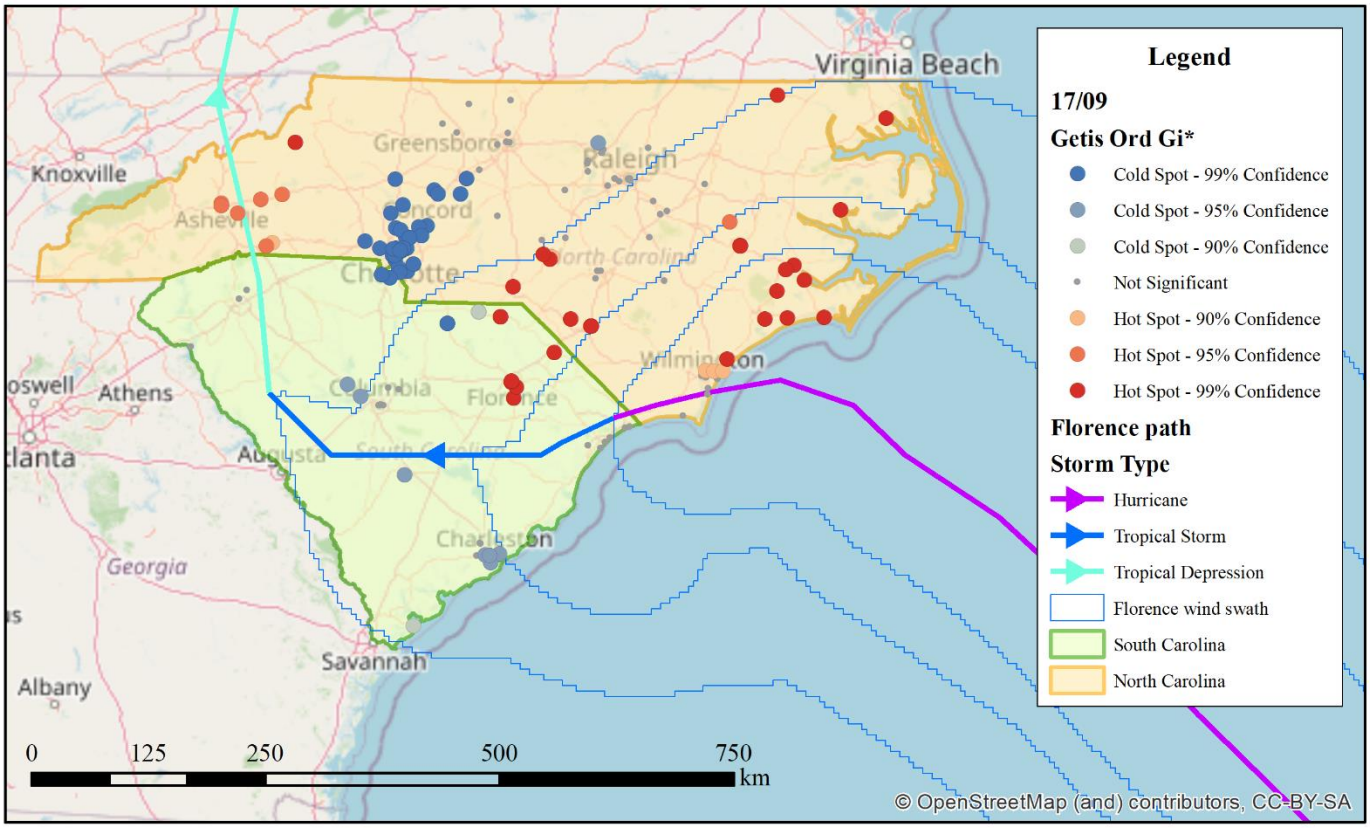

Figure 73 Hot-spots and cold-spots detected with the $W_{RT,pop}$ population field for September 17th

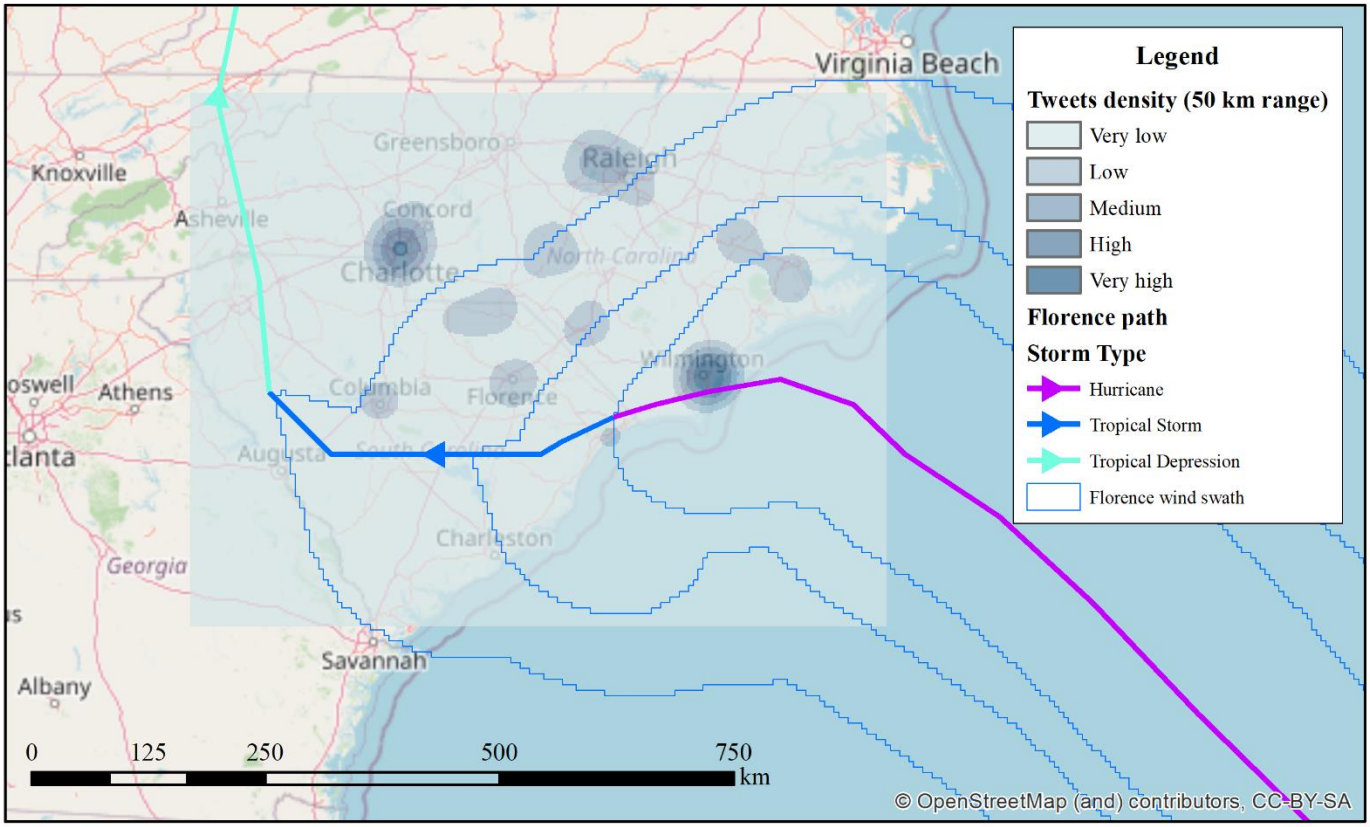

Figure 74 Kernel Density Map computed with the $W_{RT,pop}$ population field for September 17th

During the first days considered (September 13th, 14th and 15th), the Getis Ord Gi* index and Kernel Density maps have detected a relevant high concentration of geolocated tweets and 99%-confidence hot-spot primarily around Wilmington (New Hanover county, NC) near the site where Florence made its landfall. The tweets density in this area resulted high until September 17th but in the last two days only few low-confidence hot-spots were detected. This change in time could be motivated by the fact that the Twitter active population is more interested in posting original social media updates during the hours immediately before the hurricane arrival or on the day when the hurricane hit in that area. The presence of few hot-spot in Wilmington on September 16th and 17th after the circulation of Florence in New Hanover county could then be supported by the fact that Twitter users were involved more in a quoted-retweeting activity, re-posting the most relevant content or multimedia that documented the hurricane impact, then in an original contribution about their personal experience after the severe weather occurrence.

Other areas of low and medium density have been identified along the Atlantic coast on both North and South Carolina by the Kernel Density tool. However, the tweet density gradually decreased after the first day, supporting the hypothesis that these area were merely affected during the hours immediately before the Florence landfall thorough strong winds and major storm surges and waves caused by the hurricane approaching the coast. Most of the hot-spots has been detected along the North Carolina coast while in South Carolina 99%-confidence cold-spot have been found in the area of Charleston on September 14th. However, through the comparison with the Kernel Density maps, it has been possible to understand that the majority of the NC hot-spots are not linked to relevant tweet density values. Consequently, some isolated hot-spots in North Carolina could be motivated by the fact that they were published in counties with low population values (P_{cat} equals to 1 or 2) that resulted in high values of the weight parameter $W_{RT,pop}$.

Besides Wilmington, the other counties with medium and high density during the five considered days were Wake and Mecklenburg, both located in North Carolina. In correspondence of Raleigh, county seat of Wake, it has been detected a medium-high tweet density that had its peak intensity on September 15th (Figure 70) the day after the hurricane landfall. Wake county was partially covered by the hurricane wind swath and in the city of Railegh many heavy raining events were reported during the Florence circulation, as mentioned in the NOAA report. The cold-spot detected in this area were mainly linked to the relevant presence of quoted-retweet posts and to the high population value of this county.

Mecklenburg county is instead characterized by a different behaviour. As previously seen in Figure 59, the area was affected by a relevant social media activity mainly concentrated after September 14th,

2018. Consequently, the high-density area around Charlotte has been detected from September 15th. This could be associated to the fact that in those days, Florence was gradually dissipating its energy as a tropical storm and then as a tropical depression heading North and causing indirectly rainfall and consequently flooding events in Charlotte. Additionally, the NOAA report states that in Mecklenburg area power outages caused significant problems to inhabitants that had a limited access to energy for some days. For these reasons, despite the presence of cold-spots related to the presence of quoted-retweet posts in a densely populated county, the Twitter activity registered in Charlotte (Mecklenburg, NC) could be considered relevant even if its area was not included in the Florence wind swath.

In conclusion, a final comparison has been made between the NOAA total rainfall report for September 13th-17th and the tweet density computed for the entire 5-days Carolina subset (Figure 75 and 76). The hot-spots associated to the weight $W_{RT,pop}$ are mainly concentrated in the high-tweet density area surrounding Wilmington and New Hanover county along the North Carolina Atlantic coast. This result is in accordance with the NOAA total rainfall map that describes that area as one with the highest total rainfall values (between 0,63 and 0,89 meters). Additionally, FEMA considered this county as one of the most affected by the hurricane and provided both individual and public assistance services for debris removal, emergency protective measures and permanent work (FEMA-4393-DR).

The other two medium and high tweet density areas are instead associated to Charlotte and Raleigh in North Carolina. However, tweets located in these two cities resulted as 99% confidence cold-spots, suggesting the major presence of quoted-retweet posts in very populous counties. They are indeed associated to areas with low-medium total rainfall values in the range of 0,13-0,25 meters. Considering that FEMA did not assign any recovery and assistance program, it could be assumed that the high tweet density was mainly associated to a Twitter activity driven by users not directly affect by the most devastating consequences of Florence circulation.

In conclusion, it is necessary to highlight the absence of Horry county (SC) – the third county in the Carolinas for the number of geolocated tweets – in the results of both the hot-spot analysis and the Kernel Density Map. Indeed, FEMA provided the same assistance services of New Hanover to this area, whose main Atlantic coastal city is Myrtle Beach (FEMA-4394-DR). The presence of few 90% confidence cold-spots suggests that this area was interested by a mixed behaviour of both original and quoted retweet posts considering that Horry was classified as a county with a medium population value (P_{cat}=3).

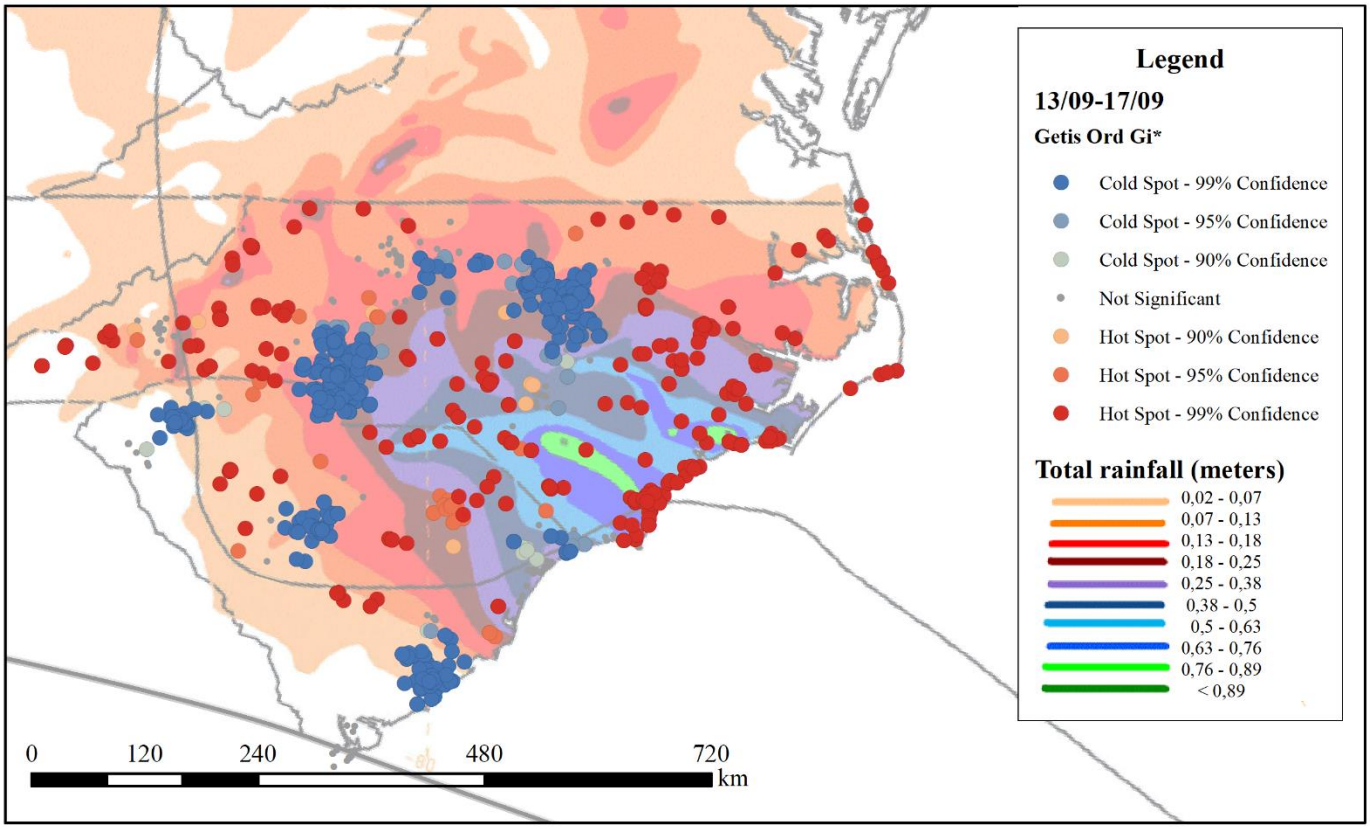

Figure 75 Comparison of NOAA total rainfall map with hot-spot computed with the $W_{RT,pop}$ weight for September 13th-17th

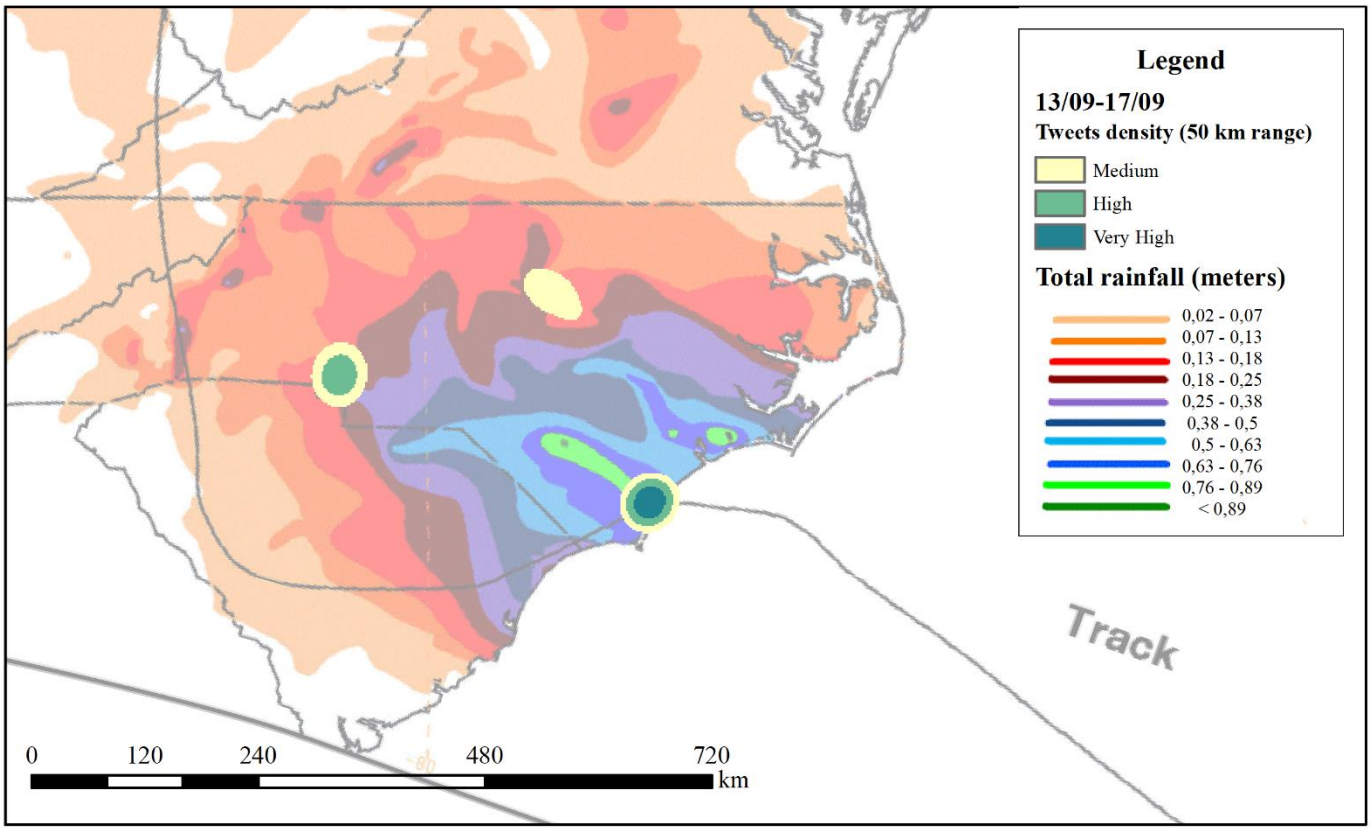

Figure 76 Comparison of NOAA total rainfall map with Kernel Density Map computed with the $W_{RT,pop}$ population field for September 13th-17th

Conclusions and outlook

The objective of this thesis was to explore the potentials of geolocated Twitter posts in the context of a natural disaster in particular hurricane Florence. In the previous chapters, after a preliminary data cleaning and pre-processing phase, a workflow has been presented for the relevant SGMI posted between September 11th and October 4th, 2018. The aim of the proposed workflow was to identify the areas with the significant Twitter activity and compare them with reference data from NOOA, NWS and FEMA.

The geolocated Twitter dataset (Wrubel, 2019) has been analysed at different scales. At a global scale, through spatial observations and temporal trend analyses, the most active area of the United States of America, directly affected by the hurricane, has been detected. Additionally, it has been possible to observe the different behaviour of a timeseries influenced by a toponym, like in the case of Firenze in Italy. A more detailed workflow, based on the posts geolocation methods and on the tweets typology (original or quoted retweet), has been defined and adopted in order to identify the most influenced states. The resulting US states, North and South Carolina, corresponded to the area crossed by Florence when it was classified as hurricane and tropical storm and covered by its wind swath. In this phase, it has been pointed out the possible influence of the population on the rate of Twitter content creation, assuming that a greater US states population implies a greater number of active Twitter users.

At a local scale, the spatio-temporal trend of geolocated tweets across North and South Carolina have been analysed through geo-statistical tools such as the Nearest Neighbour Index, the Getis Ord-Gi* and the Kernel Density Map. The results of the performed hot-spot and tweets density analyses has highlighted the main strengths and weakness of SMGI. The Carolinas spatio-temporal trend for geolocated tweets was coherent with the Florence circulation from East to West and identified Wilmington, in the county of New Hanover (NC), as the area most affected by this severe weather event, in accordance with the NOAA total rainfall map and FEMA report about most damaged areas. Moreover, the Florence-related SMGI gave a detailed picture of the hurricane landfall and of the day that preceded it, identifying spikes and peaks in correspondence of them and suggesting that Twitter users are more into posting the geolocated tweets in the hurricane alert phase. However, the geolocated tweets at county level revealed to be strongly affected by the social media activities registered in the major metropolitans area like Charlotte and Raleigh in North Carolina, where most

of the media operators and weather channel journalist – considered as premium users and opinion leaders - are located and involved in dissemination information.

The current and the future researches still have to face major issues linked to the use of SMGI as proxies of natural disaster damages. They represent a valuable source of information in the preparation phase and in the detection of a hurricane landfall but the lack of availability of geolocated social media posts (less than the 5% of the complete dataset) remains a relevant that presents important challenges for new methodologies to retrieve a tweet position from its semantic content. Moreover, the geolocated component of Twitter object is significantly affected by the *place geotag* use (approximately the 90% of all the posts with a spatial reference), that, as remarked in the previous analyses, represents a biased element with its approximated definition. Additionally, the relation between the real population and the total number of Twitter active users in a specific area still require to be investigates. This aspect would help to understand in a more complete way the Twitter demographic, giving possible new tools and methodologies to compensate the consequences of the Digital Divide and assure the representativeness of Twitter data. However, a social media census approach may be strongly affected by many web privacies issues such the location privacy.

In conclusion, additional works could improve the results obtained for the hurricane Florence case study. For example, a sentiment analysis on the tweet semantic content would help to obtain additional insights about the users' thoughts and opinions about natural disasters during the preparation phase but also after the hurricane landfall. Alternatively, the social interaction Twitter attributes (*favorites_count*, *retweet_count* and *followers_count*) could be integrated in a weight parameter, paying attention to the influence of media operators that are usually associated to verified accounts and have a total number of followers that is higher than the average.

Spatio-temporal analyses could then be applied to single city like Wilmington, Charlotte and Raleigh, aiming to detect specific patterns at urban level. Also, elaborations could take advantages of the multimedia content of the tweet object (Figure 75). Indeed, pictures, videos, audios, and URLs attached to social media posts represent valuable sources of information that could integrate a possible sentiment analyses giving a more precise feedback on the emergency. They also could help defining a weight based on the presence or absence of multimedia. Moreover, it is important to notice that many geolocated tweets include link to content published on Instagram. So, on the Twitter objects only the URL is saved whilst the image or video content associated to the posts is visible only in the Instagram environment. Consequently, the integration of these two different social media sources could provide more contents for geolocated multimedia analyses, although this would enhance the complexity of the SMGI and the issue associated to the difficulties in understanding the

characteristics of the social media population whose demographic might vary from one platform to another.

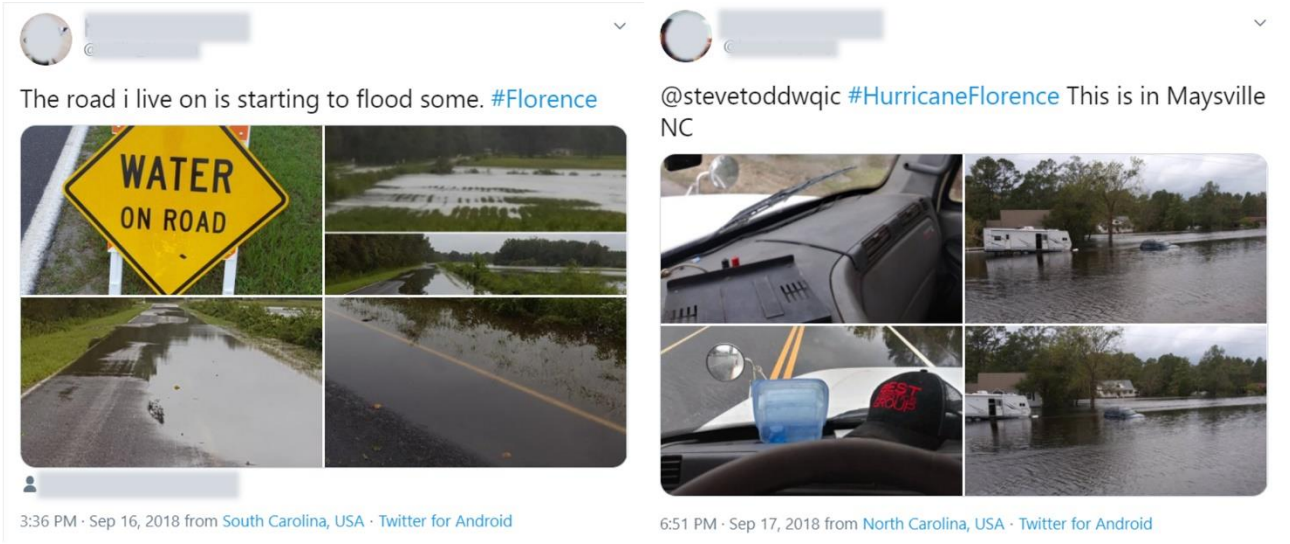

Figure 77 Example of Carolinas geolocated tweets with pictures

In conclusion, the potential and the complexity of using SMGI in crisis scenarios need to be completely investigated through the evolving relationship between their three main components (spatial, temporal and user) that have to be critically examined and interpreted. In particular, it is important to remark that future studies necessarily need to address the user component influence on tweets dissemination: most vulnerable social classes could have limited access to internet connection and consequently geo-statistical results may not reflect the on-the-ground realities following disasters due to consolidated and long-standing patterns of socio-spatial inequality (Shelton et al., 2014). Additionally, spatio-temporal analyses will need to understand the composition and the variety of active Twitter users, considering the influence of verified accounts, media operators and agencies during emergencies. Hence, identifying social media behaviours for different users' classes could be useful for implementing methodologies able to take into account the heterogeneity of the social media population itself.

Bibliography

Acar, A., & Muraki, Y. 2011. Twitter for crisis communication: lessons learned from Japan's tsunami disaster. *International Journal of Web Based Communities*, *7*(3), 392-402.

De Albuquerque, J. P., Herfort, B., Brenning, A., & Zipf, A. (2015). A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. *International journal of geographical information science*, *29*(4), 667-689.

Avvenuti, M., Del Vigna, F., Cresci, S., Marchetti, A., & Tesconi, M. 2015. Pulling information from social media in the aftermath of unpredictable disasters. In *2015 2nd International Conference on Information and Communication Technologies for Disaster Management (ICT-DM)* (pp. 258-264). IEEE.

Bamman, D., O'Connor, B., & Smith, N. 2012. Censorship and deletion practices in Chinese social media. *First Monday*.

Blank, G. 2017. The digital divide among Twitter users and its implications for social research. *Social Science Computer Review*, *35*(6), 679-697.

Budhathoki, N. & Haythornthwaite, C. 2012. "Motivation for Open Collaboration: Crowd and Community Models and the Case of OpenStreetMap" *American Behavioral Scientist*.

Campagna, M. 2016. Social Media Geographic Information: Why social is special when it goes spatial?. In: Capineri, C, Haklay, M, Huang, H, Antoniou, V, Kettunen, J, Ostermann, F and Purves, R. (eds.) *European Handbook of Crowdsourced Geographic Information*, Pp. 45–54. London: Ubiquity Press. DOI: http://dx.doi.org/10.5334/bax.d. License: CC-BY 4.0.

Capineri, C. 2016. The Nature of Volunteered Geographic Information. In: Capineri, C, Haklay, M, Huang, H, Antoniou, V, Kettunen, J, Ostermann, F and Purves, R. (eds.) *European Handbook of Crowdsourced Geographic Information*, Pp. 15–33. London: Ubiquity Press. DOI: http://dx.doi.org/10.5334/bax.b. License: CC-BY 4.0.

Caragea, C., Squicciarini, A. C., Stehle, S., Neppalli, K., & Tapia, A. H. (2014, May). Mapping moods: Geo-mapped sentiment analysis during hurricane sandy. In *ISCRAM*.

Chen, M., Mao, S., & Liu, Y. 2014. Big data: A survey. *Mobile networks and applications*, *19*(2), 171-209.

Clark, P.J., Evans, F.C. *Distance to Nearest Neighbor as a Measure of Spatial Relationships in Populations.* Ecology, Vol. 35, N° 4, 1954

Cohn, J. P. 2008. Citizen science: Can volunteers do real research?. *BioScience*, *58*(3), 192-197.

Coleman, D. J., Georgiadou, Y., & Labonte, J. 2009. Volunteered Geographic Information: the nature and motivation of produsers. *International Journal of Spatial Data Infrastructures Research*, *4*(1): 332–358.

Crawford, K., & Finn, M. 2015. The limits of crisis data: analytical and ethical challenges of using social and mobile data to understand disasters. *GeoJournal*, *80*(4), 491-502.

Earle, P., Guy, M., Buckmaster, R., Ostrum, C., Horvath, S., & Vaughan, A. 2010. OMG earthquake! Can Twitter improve earthquake response?. *Seismological Research Letters*, *81*(2), 246-251.

Elwood S., Goodchild M. and Sui D. 2012. Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice, *Annals of the Association of American Geographers*, 102:3, 571-590, DOI: 10.1080/00045608.2011.595657

FEMA, B. Public Assistance Applicant Handbook.

Howe, J. 2006. The rise of crowdsourcing. *Wired magazine – June 2006,* 1-4.

Getis, A., Ord, K. 1992. *The Analysis of Spatial Association by Use of Distance Statistics: Geographical Analysis 24*, pp. 189-206.

Goodchild, M. 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69: 211–221.

Grinberg, N., Naaman, M., Shaw, B., & Lotan, G. 2013. Extracting Diurnal Patterns of Real World Activity from Social Media. In *ICWSM*.

Haklay, M., Singleton, A., & Parker, C. 2008. Web Mapping 2.0: The Neogeography of the GeoWeb. *Geography Compass, 2*(6), 2011-2039. doi: https://doi.org/10.1111/j.1749-8198.2008.00167.x

Harrison, S. E., & Johnson, P. A. 2016. Crowdsourcing the disaster management cycle. *International Journal of Information Systems for Crisis Response and Management (IJISCRAM)*, *8*(4), 17-40.

Herfort, B., de Albuquerque, J. P., Schelhorn, S. J., & Zipf, A. (2014). Exploring the geographical relations between social media and flood phenomena to improve situational awareness. In *Connecting a digital Europe through location and place* (pp. 55-71). Springer, Cham.

Hossmann, T., Legendre, F., Carta, P., Gunningberg, P., & Rohner, C. 2011. Twitter in disaster mode: Opportunistic communication and distribution of sensor data in emergencies. In *Proceedings of the 3rd Extreme Conference on Communication: The Amazon Expedition* (pp. 1-6).

Kankanamge, N., Yigitcanlar, T., Goonetilleke, A., & Kamruzzaman, M. 2019. Can volunteer crowdsourcing reduce disaster risk? A systematic review of the literature. *International journal of disaster risk reduction*, *35*, 101097.

Klonner, C., Marx, S., Usón, T., Porto de Albuquerque, J., & Höfle, B. 2016. Volunteered geographic information in natural hazard analysis: a systematic literature review of current approaches with a focus on preparedness and mitigation. *ISPRS International Journal of Geo-Information*, *5*(7), 103.

Lampos, V., De Bie, T., & Cristianini, N. 2010. Flu detector-tracking epidemics on Twitter. In *Joint European conference on machine learning and knowledge discovery in databases* (pp. 599-602). Springer, Berlin, Heidelberg.

Leetaru, K., Wang, S., Cao, G., Padmanabhan, A., & Shook, E. 2013. Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday*, *18*(5). https://doi.org/10.5210/fm.v18i5.4366

Levy, P. 1994. L'Intelligence collective. Pour une anthropologie du cyberespace. Paris: La Découverte.

Li, J., & Rao, H. R. 2010. Twitter as a rapid response news service: An exploration in the context of the 2008 China earthquake. *The Electronic Journal of Information Systems in Developing Countries*, *42*(1), 1-22.

Mellon, J., & Prosser, C. 2017. Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users. *Research & Politics*, *4*(3), 2053168017720008.

Migliaccio, F., Pagliari, D., Carrion, D., & Gaspari, F. 2018. Geostatistical and temporal analysis of Instagram data. In *ICTIC-The 7 th International Virtual Scientific Conference on Informatics and Management Sciences* (pp. 101-105)

Migliaccio, F., Carrion, D., & Ferrario, F. 2019. Semantic validation of social media geographic information: A case study on instagram data for expo Milano 2015. In *4th ISPRS Geospatial Week 2019* (Vol. 42, No. 2, pp. 1321-1326). International Society for Photogrammetry and Remote Sensing

O'Reilly, T. 2007. "What is Web 2.0: Design patterns and business models for the next generation of software. Communications & Strategies" 1, 17.

Plewe, B. 2007. "Web cartography in the United States." *Cartography and Geographic Information Science* 34.2: 133-136.

Roick, O., Heuser, S. 2013, *Location Based Social Networks – Definition, Current State of the Art and Research Agenda*, Transactions in GIS 17.5, pp. 763-784

Saffir, H. S., 1973: Hurricane wind and storm surge. *Military Engineering*, 423, 4–5.

Shelton, T., Poorthuis, A., Graham, M., & Zook, M. 2014. Mapping the data shadows of Hurricane Sandy: Uncovering the sociospatial dimensions of 'big data'. *Geoforum*, *52*, 167-179.

Silverman, B. W. 1986. *Density estimation for statistics and data analysis* (Vol. 26). CRC press.

Soden, R., & Palen, L. 2014. From crowdsourced mapping to community mapping: The post-earthquake work of OpenStreetMap Haiti. In *COOP 2014-Proceedings of the 11th International Conference on the Design of Cooperative Systems, 27-30 May 2014, Nice (France)* (pp. 311-326). Springer, Cham.

Spasenovic, K., Carrion, D., Migliaccio, F., & Pernici, B. 2019. FAST INSIGHT ABOUT THE SEVERITY OF HURRICANE IMPACT WITH SPATIAL ANALYSIS OF TWITTER POSTS. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*.

Steiger, E., Westerholt, R. and Zipf, A. 2016. Research on social media feeds – A GIScience perspective. In: Capineri, C, Haklay, M, Huang, H, Antoniou, V, Kettunen, J, Ostermann, F and Purves, R. (eds.) *European Handbook of Crowdsourced Geographic Information*, Pp. 237-254. London: Ubiquity Press. DOI: http://dx.doi.org/10.5334/bax.r. License: CC-BY 4.0., 2016

Stewart, S.R. and Berg, R. 2019. Hurricane Florence (AL062018), *National Hurricane Center Tropical Cyclone Report*

Sui, D., Goodchild, M., & Elwood, S. 2013. Volunteered Geographic Information, the Exaflood, and the Growing Digital Divide. In D. Sui, M. Goodchild, & S. Elwood (Eds.), *Crowdsourcing Geographic Knowledge* (pp. 1-12). Springer, Dordrecht. doi: https://doi.org/10.1007/978-94-007-4587-2_1

Tobler, W. R. 1970. A computer movie simulating urban growth in the Detroit region. *Economic geography*, *46*(sup1), 234-240.

Wang, Z., Ye, X., & Tsou, M. H. 2016. Spatial, temporal, and content analysis of Twitter for wildfire hazards. *Natural Hazards*, *83*(1), 523-540.

Wang, S., Schraagen, M., Sang, E. T. K., & Dastani, M. 2020. Dutch General Public Reaction on Governmental COVID-19 Measures and Announcements in Twitter Data. *arXiv preprint arXiv:2006.07283*.

Wrubel, L., 2019, "Hurricane Florence", https://doi.org/10.7910/DVN/GSIUXQ, Harvard Dataverse, V1

Web resources

British Civil Aviation Authority: https://www.caa.co.uk/Data-and-analysis/UK-aviation-market/Airports/Datasets/UK-Airport-data/Airport-data-2018-09/
Last check: 20/08/2020

CNN | Florence shooting October 3rd: https://edition.cnn.com/2018/10/04/us/florence-south-carolina-shooting/index.html
Last check: 30/08/2020

Daily Mail: https://www.dailymail.co.uk
Last check: 20/08/2020

Facebook: https://www.facebook.com
Last check: 05/08/2020

FEMA | North Carolina Hurricane Florence (FEMA-4393-DR): https://www.fema.gov/disaster/4393
Last check: 31/08/2020

FEMA | South Carolina Hurricane Florence (FEMA-4394-DR): https://www.fema.gov/disaster/4394
Last check: 1/09/2020

Flickr: https://www.flickr.com
Last check: 05/08/2020

Geofabrik: https://www.geofabrik.de/
Last check: 20/08/2020

George Washington archive | Hurricane Florence tweets: https://tweetsets.library.gwu.edu/dataset/c940c3d2?reload=true#datasetExports
Last check : 19/08/2020

Hydrator | GitHub repository : https://github.com/DocNow/hydrator
Last check : 19/08/2020

Instagram: https://www.instagram.com
Last check: 05/08/2020

Istituto Nazionale di Statistica (ISTAT): https://www.istat.it/it/
Last check: 20/08/2020

Mapillary: https://www.mapillary.com
Last check: 05/08/2020

Mercury Prize: https://www.mercuryprize.com/
Last check: 20/08/2020

Met Office: https://www.metoffice.gov.uk/
Last check: 20/08/2020

National Bureau of Statistics of China: http://data.stats.gov.cn/english/
Last check: 20/08/2020

National Oceanic and Atmospheric Administration (NOAA): https://oceanservice.noaa.gov
Last check: 19/08/2020

NOAA | Hurricane Florence data: https://www.nhc.noaa.gov/data/tcr/index.php?season=2018&basin=atl
Last check: 24/08/2020

Open Government Data Platform India: https://data.gov.in/
Last check: 20/08/2020

OpenStreetMap: https://www.openstreetmap.org
Last check: 05/08/2020

Reuters | Hurricane Sandy: https://in.reuters.com/article/storm-sandy-twitter/over-20-million-tweets-sent-as-sandy-struck-idINDEE8A10AX20121102
Last check: 16/08/2020

Runtastic: https://www.runtastic.com
Last check: 05/08/2020

Statista | Twitter: Monthly active users worldwide (2019):
https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/
Last check: 10/08/2020

Statista | Social media usage in India (2020): https://www.statista.com/study/59959/social-media-usage-in-india/
Last check: 20/08/2020

Telegram: https://telegram.org
Last check:05/08/2020

Telegraph: https://www.telegraph.co.uk
Last check: 20/08/2020

The Sun: https://www.thesun.co.uk
Last check: 20/08/2020

Tweepy GitHub repository: https://github.com/tweepy/tweepy
Last check: 12/08/2020

Twitter: https://twitter.com
Last check: 05/08/2020

Twitter Developer | Geo Object: https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/geo-objects#coordinates-dictionary
Last check: 13/08/2020

Twitter Developer | Objects: https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/intro-to-tweet-json
Last check: 13/08/2020

Twitter Developer | Twitter geospatial metadata: https://developer.twitter.com/en/docs/tutorials/tweet-geo-metadata
Last check: 13/08/2020

United States Census Bureau | Mapping files: https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html
Last check: 24/08/2020

United States Census Bureau | Population data: https://www.census.gov/data/datasets/time-series/demo/popest/2010s-counties-total.html
Last check: 24/08/2020

Waze: https://www.waze.com
Last check: 05/08/2020

WhatsApp: https://www.whatsapp.com
Last check: 05/08/2020

WikiMapia: https://wikimapia.org
Last check: 05/08/2020

Acknowledgements

At the conclusion of this work, I would like to express my sincere gratitude to my supervisor Prof. Federica Migliaccio for giving me the opportunity to deepen my knowledge of Social Media Geographic Information and guiding me throughout this work.

I wish to acknowledge my co-supervisor Katarina Spasenovic, Ph.D Candidate, for inspiring with enthusiasm my interest in the project and for providing precious advices and support during the whole thesis.

I am extremely grateful to my mother and grandparents who unconditionally supported me during my studies, encouraging me through all the years spent at Politecnico di Milano.

Finally, I would like to thank all the friends and classmates with whom I have shared the best memories of these years.