GIMA
Geographical Information Management and Applications

Evaluating the applicability of user location inference methodologies to increase the usability of Twitter data in event detection research scenarios

Universiteit Utrecht
TU Delft
University of Twente.
Wageningen UR

GIMA Master Thesis

Author: Joe d'Hont
Supervisor: dr. ir. Arend Ligtenberg
Professor: prof. dr. ir. Arnold Bregt
7th June 2017

Author:
Joe d'Hont

Contact:
j.dhont@students.uu.nl
+31(0)616787084

Supervisor:
dr. ir. Arend Ligtenberg

Professor:
prof. dr. ir. Arnold Bregt

Special thanks to:

Arnold Bregt | *WUR* | *Wageningen, The Netherlands*

…for providing feedback on adding boundaries to my research context.

Sarah Hoffmann | *OpenStreetMap Operations Working Group* | *Zurich, Switzerland*

…for providing information on the rate limitations of the OpenStreetMap Nominatim API.

Derek Karssenberg | *Utrecht University* | *Utrecht, The Netherlands*

…for providing advice on the Python script performance metrics.

Arend Ligtenberg | *WUR* | *Wageningen, The Netherlands*

…for supervising this thesis and motivating me throughout its conduction.

Wei Yang | *Spatial Sciences Institute* | *Los Angeles, United States*

…for providing advice on the health management scenario workflow.

Marco Bonzanini | *Various* | *London, United Kingdom*

Alexander Galea | *Ayima* | *Vancouver, Canada*

Michal Migurski | *Code for America* | *San Francisco, United States*

The users of Stack Overflow | *Various* | *Anywhere*

…for allowing me to use and edit their Python scripts for this thesis research.

“Twitter is not a social network,
it’s an information network.”

Evan Williams, former Chairman and CEO of Twitter

Summary

Social media are one of the defining phenomena of current times. Through social media, people are able to communicate in ways that would have been considered impossible in the past. Social media form an interesting subject to academics from a wide variety of fields as well because of its widespread use and potential to solve societal problems.

Twitter in particular has found interest among scholars due the fact that Twitter data can be gathered and analysed easily. Moreover, tweets can be geotagged and therefore show potential use in GIS research. When geotagging tweets, an accurate geographical reference in the form of a GPS coordinate is attached to it. The problem is that in practice only a small fraction of the tweets posted is geotagged. This results in a vast majority of entries being removed from Twitter data sets used in GIS research due to a lack of an accurate geographical reference available. This is problematic because a lot of semantic information of interest to research is lost this way as well.

To tackle the problem as described above, geolocation inference methods (GIMs) have been developed over the years to increase the usability of Twitter data in GIS research. When using a GIM, the geolocations of either users or tweets are inferred through indirect means. Examples of these means are the content of the tweets or the geolocations of friends and followers of a particular user. While currently plenty of GIMs are available as presented by academics from a wide variety of fields, it is unclear whether these methods are equally applicable in all types of GIS research scenarios. The aim of this thesis research is to provide such clarity by answering the following central question:

To what extent can the usability of Twitter data in event detection research scenarios using this type of data be increased through the application of geolocation inference methodologies?

A content-based method based on text mining ("content-user method") and network-based method based on tie-strength ("network-user method") which are both meant to infer the geolocations of users have been evaluated and compared in a disaster management, health management and topic modelling research scenario to answer the question as mentioned above. These particular GIMs and GIS research scenarios have been selected based on the findings of two systematic literature studies conducted on the subjects respectively as part of this thesis research, in which either the types of GIMs most often presented in academic research and applications of Twitter data most often used in GIS research were determined. A third literature study of a narrative nature on the opportunities and limitations of Twitter data usage in GIS research has been conducted as well, to which the knowledge gained from this respective literature study has been implemented in the research design.

It was found that the output of the content-user method had a relatively good temporal reliability, low overall scale level, high precision and high speed. The output of the network-user method on the other hand was found to have a relatively high spatial reliability, completeness and recall. When comparing the performances of the GIMs among the selected GIS research scenarios some differences could be perceived as well. The content-user method's output was met in particularly with a different spatial reliability, recall and speed while on the other hand the network-user method's output was met with a different spatial reliability, temporal reliability, overall scale and precision. The output of the GIMs have been compared to the unprocessed Twitter API output as well. It was found that when weighting all evaluation metrics equally the unprocessed Twitter API output performed best, but when the completeness of the data was weighted heavier the GIMs' output performed better.

In this thesis research it has been concluded that when using GIMs to increase the amount of geographical references in Twitter data sets the increase of usability is a matter of compromise rather than an overall increase of data usability. While at the same time GIMs have the potential to drastically increase the completeness of Twitter data sets, this goes at the cost of data quality. It depends on the aim of the research performed to what extent the usability of Twitter data can be increased in event detection research scenarios using this type of data. Future research on GIMs and GIS research scenarios not part of this thesis research is advised to enhance the quality of GIS research using Twitter data as a whole.

Table of contents

Appendix

Figures

Tables

1. Introduction

1.1 The rise of social media

Mankind has enabled itself to transfer information between its members more rapidly and freely ever since the beginning of the information age in the early 1970s compared to previous eras (Castells, 1997). The information age can be defined as a period of time in which formally industry-based economies all over the world turned into economies based on the computerization of information. The development of steadily more advanced information and computer technologies have played a key role in the establishment of a network society in which new possibilities for people to communicate with each other for were created. Social media can be considered both a product made possible due these new technologies and a catalysator in the acceleration of the previously described shifts and developments. Social media can be defined as "*internet-based, disentrained, and persistent channels of mass personal communication facilitating perceptions of interaction among users, deriving value primarily from user-generated content*" (Carr & Hayes, 2015, p. 49).

Since their emergence in the early 2000s, social media have enhanced and increased the ways by which people are able to communicate with each other. These interactions take place in social networks that were hardly existent before when considering communication speed, network structure and information content. American TV-host Ellen DeGeneres famously posted a "selfie"[1] with other celebrities made at the 86th Academy Awards on Twitter to response of millions of Twitter users around the world (BBC, 2014). This example shows that social media have enabled users to communicate directly with an audience of millions of people instantly on a mutual level. Social media have also given "regular" citizens the ability to communicate on a horizontal level with prominent figures such as presidents and celebrities directly. A prime example of this is the "IAmA" sub forum on Reddit[2] on which many of these figures have participated in interviews with users of the social medium, who were enabled to ask questions of their own interest (Reddit, 2017a). Examples include former president Barack Obama, business magnate Bill Gates and chef cook Gordon Ramsay (Reddit, 2017b-d). Concluding, social media have in a sense rooted the establishment of new communicational realities that might not have even be considered possible before their inception.

Since the early 2000s many social media have emerged, with a few gaining mass popularity over others. Facebook has just over 1.9 billion users worldwide as of the first quarter of 2017 for example, making it the biggest social medium available today based on its absolute user base (Statista, 2017a). Twitter belongs among the group of most popular social media platforms available today as well. Incepted in March 2006, the social medium has managed to garner a user base of 328 million active users[3] worldwide as of the first quarter of 2017 (Twitter, 2017a; Statista, 2017b). Twitter has become a medium to these users to post whatever there is on their mind, if the content of these posts are within the guidelines set up by the social medium itself (see Twitter, 2017b for an overview of these guidelines). The content of these posts contains information on sentiments of users on topics as broad as politics, sports and music (Twitter, 2016). Users are enabled to support their sentiments with various forms of media such as images, videos and hyperlinks (Twitter, 2017c-e). The posts containing these sentiments are alternatively known as "tweets" (Twitter, 2017f), a term that will be used in the rest of this thesis report. Twitter distinguishes itself from other social media in particular because tweets have a maximum limit of 140 characters and user profiles are public by default, unless specified otherwise by the user (Twitter, 2017f; Twitter, 2017g).

1.2 Twitter in academic research

Academics have garnered an interest in researching this communicational medium as well because social media (and thus Twitter) have taken such a prominent place in daily human life. This is illustrated by the fact that an estimated 2.5 billion people use social media since the year 2017 (Statista, 2017c). There are multiple reasons imaginable that can explain the interest by academics in Twitter specifically (Zimmer & Proferes, 2014, p. 250-251). The first reason is that Twitter data[4] can be gathered easily due to the open and public

[1] A selfie can be defined as a self-portrait photograph (Wikipedia, 2017a).
[2] Reddit is a social medium in which user-submitted news and media is aggregated (Reddit, 2017e).
[3] In this report a Twitter user is defined as someone who has a registered profile on the social medium and uses this medium regularly.
[4] Twitter data typically includes information on user and tweet characteristics, as will be detailed later in chapter 5.

nature of Twitter itself and through the presence of officially developed and well-supported APIs[5] through which this type of data can be gathered (Twitter, 2017h). The second reason is that Twitter data can be processed and analysed easily because tweets are relatively short due their 140-character limit as explained in paragraph 1.1. previously. The interest to research Twitter data comes from a wide variety of academic fields such as information- and computer science, but also geo-related and social sciences as identified by Steiger et al (2015, p. 6-8). Steiger et al have also have identified the three most frequently-used applications of Twitter data in GIS research[6] between the years 2006 and 2013 by conducting an extensive systematic literature study. This is as of writing the most recent and topical one available. Descriptions of these research applications are listed below. They are clarified by a resume of an archetypical research example using this respective type of application of Twitter data as well:

- **Event detection**: This research application revolves around detecting abnormal spatial, temporal and semantic tweet frequencies and patterns in real-time using Twitter as a social sensor for real world events. This research application has been used in disaster-, health- and traffic management in particular[7]. Crooks et al. (2013) have for example conducted research in which they developed an earthquake detection and geolocation system which uses Twitter as its main sensor.

- **Geolocation inference**: When performing geolocation inference, Twitter data is applied to retrieve direct or indirect geolocation information of users or tweets using provided metadata[8] attributes or semantic tweet content. Davis Jr et al. (2011) have used a very straight-forward geolocation inference method in which the city most frequently mentioned by a Twitter user is automatically assumed to be the place of residence of this user for example.

- **Social network analysis (SNA)**: This type of analysis revolves around the investigation of the individual Twitter user's characteristics within a social network and their relationships among each other. Cranshaw et al. (2012) have for example used Twitter data in combination with FourSquare[9] data to identify social cluster groups and compared the social and spatial proximity of the members of these groups.

1.3 Obstacles concerning Twitter data positioning

Over 40% of academic research done using one of research applications as described previously in paragraph 1.2 has some sort of spatial component (Steiger et al, 2015, p. 10). These components are either exclusively spatio-temporal or supported with semantic information. There are multiple ways possible by which the geolocation of tweets and users of the social medium can be derived. The most accurate method would be using the GPS coordinates attached to tweets if the geotagging option is enabled by the user. With the geotagging option enabled GPS coordinates using the WGS84[10] coordinate system are attached as metadata to a tweet automatically (Twitter, 2017i, Twitter, 2015).

This method of geographically positioning tweets and users through the GPS coordinates attached to these tweets sounds more straight-forward than it is in practice. The main reason why this is the case is that only a small portion of tweets is actually geotagged. In data sets used in recently published GIS research typically one to ten percent of the tweets are geotagged, depending on the geographical position of the bounding box[11] from which the Twitter data is derived and the subject matter of the tweets (Pavalanathan & Eisenstein, 2015, p. 2; Widener & Li, 2014, p. 190-191; Sloan & Morgan, 2015, p. 5; Katsuki et al, 2015, p. 4; Lwin et al, 2016, p. 1585). There are two reasons imaginable why this is the case. The first reason is that a user must actively enable the geotagging option to attach GPS coordinates to his or her tweets (Twitter, 2017i). Feeling no need to share their geolocation with anyone, Twitter users might not activate this option as well. The second reason is that many users might not feel comfortable sharing their geolocation with strangers and

[5] An API (Application Programming Interface) is a predefined method of communication between various software components (Wikipedia, 2017b).
[6] GIS research is in this thesis defined as research in which at least one geographical information system is used to come forms its conclusions.
[7] The types of management mentioned here and other types of management will be discussed in more detail in chapter 4 of the thesis report.
[8] Metadata can be defined as "data about data" (Wikipedia, 2017c).
[9] Foursquare is a social medium that recommends users places to visit based on their interests (Foursquare, 2017a).
[10] As defined by the United States National Geospatial-Intelligence Agency (NGA, 2014).
[11] A bounding box is a predefined rectangle-shaped geographical area from which data is derived (Wikipedia, 2017d).

therefore do not enable the geotagging option. Unfortunately, limited research has been done on why users do or do not geotag their tweets to support the arguments as describe on the previous page with.

Because only a limited number of users geotags their tweets this results in a limited share of georeferenced data within Twitter data sets as well. The usability of these data sets to geoscientists is therefore relatively limited without using data processing methods, as will be detailed later in paragraph 1.4. A severe lack of proper geographical references among Twitter data is problematic due the fact that this data may very well include valuable semantic information of interest to their research and is now possibly excluded from the data sets due to a lack of an accurate geographical reference available. Therefore, a need to develop data processing methodologies exists that enable researchers to increase the amount of Twitter data that is accurately geographically positioned one way or another.

1.4 Tackling the obstacles

Academics from different fields of expertise have tried to develop Twitter data processing methodologies to estimate the geolocation of tweets and users in an indirect way to tackle the problem as illustrated paragraph 1.3 previously. These methodologies are defined in this thesis as "geolocation inference methodologies" (GIMs), based on the definition used by Ajao et al (2015). They are methodologies that are used to determine a geolocation of a tweet or user through indirect means (Ajao et al, 2015, p. 2). Characteristics of the tweet or user other than the exact geolocation are used to determine the geolocation of this respective tweet or user. This type of methodologies is known under a lot of different names, however. Examples are "(geo)location prediction" as used by Han et al (2014), Lee et al (2014), Chang et al (2012) and McGee et al (2013), (geo)location estimation as used by Ozdikis et al (2013), Chandra et al (2011) and Ao et al (2014) and (geo)location profiling as used by Li et al (2012a) and Chen et al (2016). The term "geolocation inference" is preferred in this thesis for several reasons. The first reason is that "geolocation prediction" seems to suggest that the future geographical position of users or tweets is estimated while only the current position is of interest to this thesis research. The second reason is that the other variations of the terms as mentioned above are used sporadically than systematically. The final reason is that "location" could possibly refer to other types of locations besides geographical ones, for example the location of toponyms[12] within tweet texts themselves. To prevent any misinterpretation, the term "geolocation" is therefore used in this thesis research.

Academic literature suggests that the most popular GIMs among geoscientists to position Twitter data with are based on SNA[9] and text mining (Ajao et al, 2015; Jurgens et al, 2015; Han et al, 2014). When using SNA, the geolocation of both the followers of the user as the friends (the accounts the user follows) of the user are used to geographically position a tweet or user. It is based on the assumption that relationships in social media are strong indicators of spatial proximity (Jurgens et al, 2015, p.2). McGee et al (2013) have weighted friends within a social network by the amount of mutual interaction, with a higher mutual interaction indicating a closer friendship among each other thus a higher spatial proximity for example. When using text mining the geolocation is determined by analysing the content of the tweets of the users and derive toponyms from these tweets (Han et al, 2014). Davis Jr et al (2011) have developed a GIM in which the place of residence of users is determined by the city name was most often mentioned by those users for example, as previously mentioned in paragraph 1.2.

A limited amount of research has been done on what both the opportunities and limitations of GIMs are and to what extent these methodologies can be applied in different types of GIS research scenarios. A reason for this might be that not much academic literature has been written on the subject currently. Using the term "location inference" and "Twitter" to query for articles in Google scholar results only in approximately 400 entries as of writing for example. Articles on the subject of Twitter data geolocation inference seem to focus on improving individual GIMs without positioning them within the bigger picture (see Jurgens et al, 2015 for examples). Research in which these methodologies are compared and evaluated is therefore necessary to enable geoscientists to apply the appropriate GIM to their research and give them oversight what methodologies are available to enhance the geographical positioning of their Twitter data used in research.

[12] Toponyms can be defined as "synonyms for place names" (Wikipedia, 2017e).

2. Research setup

2.1 Research objectives

The research context as detailed earlier in the first chapter of this thesis report makes clear that there is a need for research on data processing methods to determine the geolocation of non-georeferenced tweets and user locations by other means than GPS coordinates attached to tweets by geotagging. The main objective of the thesis research as presented in this thesis report is therefore to provide a clear overview on the subject of GIMs currently available, their applicability in relevant GIS research scenarios using georeferenced Twitter data and to what extent the output of these GIMs compare to the unprocessed Twitter data. Currently the academic paradigm related to the subject of GIMs lacks such overview or knowledge. The aim of this thesis research is to reach the following sub objectives specifically:

- **Provide clarity on the concept of GIMs in general**: Academic literature on the subject of GIMs currently available does not seem to form one collection of knowledge but rather a scattered whole, as previously explained in paragraph 1.4. An indication that supports this argument are the different terminologies used in academic literature to identify the methods used to position Twitter data missing an accurate geographical reference. Examples are "(geo)location prediction", "(geo)location inference", "(geo)location estimation" and (geo)location profiling as mentioned previously in paragraph 1.4. By providing clarity through providing an overview on GIMs currently available through conducting this thesis research, geoscientists will be enabled to develop a better understanding of the subject in general using the knowledge gained through this thesis research. At the same time, they will be enabled to position their research within an academic framework to which they can add knowledge to.

- **Provide a clear and topical overview on what GIMs are currently available**: The current academic framework on the subject of GIMs is rather scattered, as mentioned above. While academics from multiple fields have developed a limited amount of literature reviews on the subject previously (Jurgens et al, 2015; Han et al, 2014, Ajao et al, 2015), there is no recent overview available on what GIMs are currently available or used in research. Geoscientists will be enabled to develop a better understanding of the GIMs currently available to use in their research by providing them a clear and topical overview on this subject generated through conducting this thesis research.

- **Provide a clear and topical overview on the applicability of relevant GIMs in relevant GIS research scenarios**: Inaudibility on the concept of GIMs and which are currently available will inevitably lead to inaudibility on the applicability of these GIMs in GIS research as well. Knowledge on GIM applicability is therefore vital to ensure that geoscientists use the proper methodology fit to their research. If this is not the results and conclusions from these researches are potentially flawed. Implementation of such flawed research can possibly lead to an accumulation of errors within the academic paradigm on the subject of GIMs. This should be avoided at all costs.

- **Add new knowledge to the subject of GIMs in general**: The concept of GIMs is incredibly relevant in an age in which social media still continue to become a more integral part of human life, even more than a decade after their introduction to the general public. This assumption is made based on the fact that the number of social media users worldwide is estimated to rise to 2.95 billion users in 2020 worldwide (Statista, 2017c). Adding academic knowledge on the subject of GIMs to its paradigm is therefore important to be able to enhance academic research on Twitter and social media in the future. It is important to understand the phenomenon of social media and how it affects the world around us to solve societal problems, as will be detailed later in paragraph 2.3. Therefore, any knowledge related to this subject is helpful to develop such understanding.

2.2 Research questions

To fulfil the research objectives as detailed on the previous page the following central question has been developed and will be answered in this thesis report:

To what extent can the usability of Twitter data in event detection research scenarios using this type of data be increased through the application of geolocation inference methodologies?

It has been pointed out previously in paragraph 1.2 that there were mainly three applications of Twitter data used in GIS research between the years 2006 and 2013 as identified by Steiger et al (2015). The focus in this thesis research will lie specifically on event detection, however. After conducting a literature study on the applications of Twitter data in GIS research it was found that from the years 2013 to 2016 the amount of GIS research focussing on SNA was so small that this research application has been excluded from this thesis research due to the strict time-limit at which the thesis research can be conducted, as will be detailed later in paragraph 2.5. The findings as a result of this specific literature study mentioned above will be detailed later in the fourth chapter this thesis report. Geolocation inference will not be taken into account in any GIS research scenario because the thesis itself is already about this subject. The definitions of event detection, geolocation inference methodologies (GIMs) and Twitter data as used in the central question above have previously been detailed in paragraph 1.2 and 1.4 respectively. In support of answering the central question, the following sub questions will be answered in this thesis report as well:

- ***What are currently the most relevant frequently used application types of Twitter data in GIS research and how is this research structured?***
- ***What are the benefits and drawbacks of using Twitter data in GIS research?***
- ***What geolocation inference methodologies for Twitter data currently exist and how are their workflows structured?***
- ***What are the strengths and weaknesses concerning the applicability of these methodologies in relevant event detection research scenarios using Twitter data?***
- ***How does the geolocation inference methodologies' data output compare to the unprocessed Twitter data's validity?***

GIS research is defined in this thesis as research in which at least one geographical information system is used to form its conclusions, as given in footnote 6. The first, second and third sub question will make up the first part of the thesis research, being a set of literature studies. The second part of the thesis research will be made up of sub question 4 and 5, in which the knowledge gained in the first part of the thesis research will be used to evaluate and compare several GIMs among each other in relevant event detection research scenarios. It has to be pointed out that in the second part of the thesis research the focus will lie exclusively on GIMs that infer the geolocations of users and not ones that infer the geolocation of tweets. This is done because no GIM meant to infer the geolocation of tweets could be developed within the time-limit at which the thesis research could be conducted. This will be argued in more detail in paragraph 2.5 and 7.4 later in this thesis report. The way in which the sub questions part of the first and second part of the thesis research will be answered will be explained later in more detail in the third and fifth chapter of this thesis report respectively. The way in which this thesis report is structured will be explained in more detail later in paragraph 2.6.

2.3 Research relevance

It is important to define the utility of this thesis research to justify its conduction. There are multiple reasons why this thesis research is relevant to the academic field, as previously explained in paragraph 1.4. The first reason is that current knowledge on the subject seems not to be a structured collection of ideas but rather a scattered whole. Academics will be enabled to develop a better understanding on the subject of GIMs and position their research within a scientific framework to which they can add knowledge to after the

conduction of this thesis research. The second reason why this thesis research can be considered relevant is that limited research has been done on the opportunities and limitations of GIMs and their applicability in different event detection research scenarios. Academics will better be enabled to select the appropriate GIM to use in their research after the conduction of this thesis research as well. This will result in more trustful and accurate conclusions from research conducted in the future because the opportunities and limitations of GIMs are taken into consideration and anticipated upon.

This thesis research is relevant from a social point of view in particular because the thesis research benefits the applicability of Twitter data in solving societal problems. Twitter data has so far been used in disaster-, health- and traffic management among many other society-benefiting applications as will be detailed later in chapter 4. When Twitter data can be geographically positioned with more accuracy and certainty, this will benefit these applications and thus improve the liveability of societies as a whole.

2.4 Research scope

A research scope needs to be properly defined to enhance the readability of the thesis report and develop a higher level of understanding of choices made concerning the research design from the perspective of the reader. The research scope has been described below and on the next page. All choices made considering this research scope are argued when needed:

- **Analysis scope:** The purpose of this thesis research is to compare the performance of multiple GIMs among each other and compare their applicability in relevant event detection research scenarios. There is in no way interest in analysing any spatial patterns of users or tweets part of the data sets used in these GIS research scenarios if they are unrelated to the evaluation of the methodologies as described above. Related to this fact is that this thesis research only concerns the use of Twitter data in GIS research. Other types of research unrelated to GIS are not within the scope of this thesis research, even though Twitter data may well be used in other academic fields as well.

- **Data scope**: This thesis research focusses on increasing the use of Twitter data exclusively and not data gathered from any other social media. The geolocation inference of Twitter data is examined exclusively as well, meaning that geolocation estimation of any other type of (spatial) data[13] is not taken into account in this thesis research. Spatial data does play a supportive role in some of the GIS research scenarios, as will be defined in paragraph 7.2 later in this thesis report. Since the role of this type of data is supportive, its use does not conflict the research scope as defined here.

- **Juridical scope:** The thesis research is done within the context of the privacy settings as defined by Twitter at the time of conducting this thesis research (September 2016 – June 2017). Results and conclusions described and detailed in this thesis report might not be applicable when these settings are altered by Twitter in the future.

- **Study area:** The data used in this thesis research will be derived from within bounding boxes surfacing the contiguous United States[14] and the state of California respectively. These specific study areas have been chosen because of multiple reasons. The first reason is that the contiguous United States has the most active Twitter users as to date considering absolute values as of May 2016 (Statista, 2017d), thus a bigger data set can be derived from these bounding boxes compared to the scenario in which a different study area is chosen. More data can lead to a better validity of the results and conclusions found during this thesis research. The second reason is that most academic research, on either GIS using Twitter data as their main data input or research on GIMs, have set the contiguous United States as their (main) study area. This will be described in more detail in paragraph 4.2 and 6.2 respectively later in this thesis report. California is set as a study area of one of the GIS research scenarios used in this research as well because disaster management typically is done on a sub-national level. The latter will be argued in more detail later in chapter 4 of this thesis report.

[13] Spatial data can be defined as data that contains information on the geographical position of (a) certain subject(s) in reality (Wikipedia, 2017f).
[14] The contiguous United States can be defined as the 48 adjoining U.S. states including the federal district of Washington, D.C. (Wikipedia, 2017g).

- **Language scope:** The GIMs will be tested on data that uses the English language exclusively for multiple reasons. The first reason is that the conductor of this thesis research is not proficient with languages other than English or Dutch. The second reason is that most natural language processing packages have a general bias towards the English language, partly due to the ease of tokenization of this particular language (Rodrigues & Teixeira, 2015, p. 15). The exclusive use of the English language in this thesis research is important to consider because the GIMs that will be examined may give different data outputs when used on data using languages other than English. Therefore, results found might not be representative in all GIS research scenarios imaginable. The final reason why English Twitter data is used exclusively is that most research on GIMs had the same language scope specified as well. The language scope of those articles will be detailed later in paragraph 6.2.

- **GIM scope:** As has been pointed out earlier in paragraph 2.2, the focus in the second part of the thesis research will lie exclusively on GIMs that infer the geolocations of users and not ones that infer the geolocation of tweets. The reason for this is the fact that no accurate GIM could be designed to infer the geolocation of tweets within the time-limit at which the thesis research could be conducted, as will be detailed later in paragraph 7.3. Given that the majority of the GIMs described in academic research are meant to infer user locations as well, as will be described later in paragraph 6.3, makes sure that the thesis research as conducted will remain relevant and representative. The reasons behind this specific GIM scope will be described in more detail in paragraph 6.3 and 7.3 later in this thesis report.

The research scope as defined above and on the previous page will have consequences to what extent the conclusions made in this thesis report are representative for other GIS research scenarios or GIMs. Results found through the conduction of the thesis research are representative only for GIMs used to infer geolocations of Twitter users living within the contiguous United States whose native language is English. These results are only representative for disaster management, health management, or topic modelling research scenarios as well. These results have been found within the privacy- and user policies as of spring 2017. When these policies change, the results found by others when using the GIMs in the GIS research scenarios examined in this thesis research might differ from the results found during the conduction of the thesis research.

2.5 Assumptions and constraints

A few assumptions have been defined and clarified below and on the next page to be able to successfully conduct the thesis research as presented in this thesis report. Research assumptions are statements accepted as true (or very plausible) and need to be defined because either these statements are difficult to prove or not even provable at all. An example of such statements are statements about the future. The assumptions as defined for this thesis research have been listed below and on the next page:

- Metadata attributes containing information of the content of tweets or profiles of users can be used to infer the geolocation of these users or tweets.

- GIMs are the most appropriate type of data processing methodologies to use to derive the geolocation of users or tweets apart from using the GPS coordinates as attached to the tweet's metadata attributes by geotagging.

- The (contiguous) United States will stay the country with the most active Twitter users in absolute value over the course of this thesis research, thus remaining the most relevant subject area to derive data from.

- The usability of the Twitter API's data output is sufficient to use in GIMs and this thesis research, even though this software provides only a sample of all tweets posted (Twitter, 2017j).

- During the conduction of the thesis research, the U.S. Twitter's user base will maintain an absolute size that guarantees that a sufficient amount of data can be gathered needed to successfully conduct this thesis research.

- Twitter's policy concerning the gathering of Twitter data will not change during the conduction of the thesis research or at least not hinder the research methodology as presented in this report when these policies are altered.

- The technical expertise and resources of the conductor of this thesis research are sufficient enough to be able to conduct the thesis research properly.

- The selection of articles that is gathered to be implemented in the literature studies conducted to answer sub question 1 to 3 is representative for the total of articles written about the subjects central to these respective literature studies. It is assumed that not every article relevant to the thesis research can be found due to either human error, lack of access possibilities or other miscellaneous reasons.

A couple of research constraints have been defined as well. These constrains need to be defined both to narrow down the scope of the thesis research and prevent the conductor of the thesis research to walk into limitations or obstacles unexpectedly while conducting the thesis research. The constraints as identified for this thesis research have been listed below:

- The Twitter API used to gather data only provides a sample of all tweets available when gathering data by keywords (Twitter, 2017j).

- The Twitter API can be used only to gather data as far back as seven days when gathering data by keywords (Twitter, 2017k).

- There is a limited time period at which the thesis research can be conducted. While this time period can be extended, the aim is to keep this period as short as possible.

- Only readily available or freely available software can be used to conduct this thesis research due to the financial limitations of the conductor of the thesis research.

- Conclusions made are representative only for the (contiguous) United states, tweets in English, users who have set the default language on Twitter to English, certain event detection research scenarios and certain GIMs meant to infer the geolocation of users. Results might be different when a different research scope is defined, as explained previously in paragraph 2.4.

2.6 Research structure

This thesis research will consist of two main parts, as briefly explained earlier in paragraph 2.2. In the first part of the thesis research the first three sub questions as defined earlier in paragraph 2.2. will be answered by conducting a set of literature studies. In the second part of the thesis research GIMs will be evaluated and compared in several GIS research scenarios to answer the fourth and fifth sub question as defined earlier in paragraph 2.2 as well. For each of these two parts of the thesis research mentioned above a separate methodology chapter is written. There are multiple reasons why the thesis report is structured this way. The first reason is that typically the literature study serves as an introduction to the research subject and is therefore generally positioned in front of the methodology chapter in academic research reports. Because in the case of this thesis research the literature study is part of the research itself as described earlier in paragraph 2.2, it would simply be illogical to put the methodology chapter of the literature study after the literature study itself. Therefore, the choice has been made to put the methodology chapter on the conduction of the literature study in front of the literature study itself. The second reason why the report is structured in an unorthodox way is that the methodology used to answer the fourth and fifth sub question is based on the findings found through answering sub questions 1 to 3. Therefore, a second methodology chapter is created, taking these findings into account. The thesis research has been divided in two parts to ensure the readability of the report for the reasons mentioned above. The thesis research report will consist of eleven chapters, as listed on the next page:

Research introduction

- **1. Introduction**: The research context is explained, detailed and argued where needed.

- **2. Research setup**: The research objectives, questions, relevance, scope, structure, assumptions and constraints are detailed and argued where needed. Any choices made are argued when needed.

Research part 1: Literature study

- **3. Methodology**: The methodology used to perform the literature studies used to answer the first three sub questions will be defined, detailed and argued. The way in which these literature studies are structured will be argued as well.

- **4. Sub question 1**: The applications of Twitter data in GIS research from 2013 to 2016 will be determined by classifying articles on the subject based on various characteristics and performing descriptive analysis on the results found. This will be done by performing a systematic literature study.

- **5. Sub question 2**: The opportunities and limitations of Twitter data usage in (GIS) research will be determined through academic literature on this subject. This will be done by performing a narrative literature study.

- **6. Sub question 3**: The most popular and relevant GIMs to infer the geolocation of Twitter users and tweets will be determined by classifying articles on GIMs based on various characteristics and performing descriptive analysis on the results found where needed. This will be done by performing a systematic literature study.

Research part 2: Analysis and evaluation

- **7. Methodology**: The methodology used to evaluate and compare the GIMs among each other within the GIS research scenarios as defined will be explained, detailed and argued where needed.
  The technical framework used to successfully conduct these methodologies will be detailed as well.

- **8. Sub question 4**: Several analyses will be performed in which the GIMs will be evaluated and compared among each other based on various evaluation and comparison metrics. The findings found through these analyses will be supported by descriptive analysis, to which the fourth sub question will be answered.

- **9. Sub question 5**: Several analyses will be performed in which data output of the GIMs will be compared to the unprocessed output of the Twitter API. The findings found through these analyses will be supported by descriptive analyses, to which the fifth sub question will be answered.

Research conclusion

- **10. Conclusion**: The sub questions will be answered to which the central question will be answered as well. The answers found will not be interpreted in this chapter.

- **11. Discussion**: The answers found for the sub- and central question(s) will discussed and interpreted. Recommendations for further research will be made as well. Finally, the thesis research as has been conducted will be evaluated and reflected upon.

The findings presented and detailed in this thesis report are supported with additional information such as references, tables, graphs and other materials. Since they do not all fit the lay-out of this thesis report they have been put together in a so-called appendix. To this appendix will be referred to when needed.
The structure of the appendix is as followed:

- **I. References**: In the first part of the appendix all (academic) literature and sources used to support arguments made in the thesis report are put together. This is done to ensure that the reader of the thesis report can read the literature and sources used themselves and validate whether the use of the knowledge presented in these is sufficient and just.

- **II. Article selections**: The article selections used to answer the first and third sub questions put together here. Additional information on the classifications these articles are part of will be detailed here as well.

- **III. Scripts used**: All the programming scripts used during the conduction of the thesis research will be detailed in this part of the appendix. Additional information on the original author(s) and source(s) of the scripts are given when needed as well.

- **IV. Software used**: All software used during the conduction of the thesis research will be detailed in this part of the appendix. Additional information on the original distributor of the software will be given when needed as well.

- **V. Shapefiles used**: Any data used during the conduction of the thesis research will be detailed in the last part of the appendix. Additional information on the original owner(s) and source(s) of the data used will be given when needed as well.

- **VI. USB-content:** All contents of the USB attached to the thesis report will be described.

An USB is attached with all Twitter data and programming scripts used during the conduction of the thesis research is attached to this report to increase the transparency of the thesis research and to enable to reader of the thesis report to validate statements made in this report. The content of the USB can also be accessed by visiting the following hyperlink:

https://www.mediafire.com/?j4kk1hci58ripq1

3. Methodology Part 1: Literature study

3.1 Introduction

This thesis research will consist of mainly two parts, with the first part being the literature study and the second part being the evaluation and comparison of GIMs in several GIS research scenarios respectively. This research structure has previously been detailed and argued earlier in paragraph 2.6. The methodology used to conduct the first part of the thesis research will be presented, detailed and argued were needed in this chapter of the report. The first three sub questions part of the first part of the thesis research as defined earlier in paragraph 2.2 will be answered through a set of literature studies. The following parameters will be detailed and argued in this methodology chapter for each literature study to be conducted respectively:

- The literature study scope
- The nature of the literature study
- The article selection criteria
- The article classification criteria (sub question 1 and 3 only)
- The keyword-concepts used to search for literature to be used in the respective study
- In what way the knowledge gathered will be implemented in the thesis research

The proper conduction of a literature study is important for multiple reasons (Bryman, 2012, p. 98).
The most important reason is that it is a useful tool to develop a general understanding among readers and the conductor of the thesis research on relevant concepts related to the thesis research. The second reason is that a literature study can be used to show the significance of the thesis research to be conducted. The third reason is that it demonstrates the ability of the conductor of the thesis research to engage in scholarly review with others in the same academic field.

3.2 Sub question 1

By answering the first sub question, knowledge will be gained on what the most relevant used types of application of Twitter data in GIS research are and how this research is structured. GIS research is defined within the sub question context as research that uses one or multiple geographical information systems to come to its research conclusions, as previously detailed in paragraph 2.2.

Any research application that uses both GIS and georeferenced Twitter data is hypothetically qualified to be taken into account in this literature study. It is estimated however that most applications will be either focussed on event detection, SNA and geolocation inference based on previously conducted literature studies on the subject as detailed previously in paragraph 1.2. Research applications that do not use georeferenced Twitter data or any geographical information system will not be taken into account in this literature study since they do not fit the research scope as defined earlier in paragraph 2.4. Within the sub question context, the focus lies on recent applications of Twitter data in GIS research. Articles that will be taken into account in this literature study therefore have to be written within a time period of the year 2013 to 2016 since the current literature reviews available on the subject only range until 2013 in detail (Jurgens et al, 2015; Han et al, 2014; Ajao et al, 2015). This literature study to be conducted can serve as a continuation of the previously mentioned reviews.

Articles are gathered based on a cyclic-iterative strategy in which the initial article selection is narrowed down based on multiple characteristics in multiple steps as well. This is the strategy as illustrated in Figure 3.1 on the next page and described there as well:

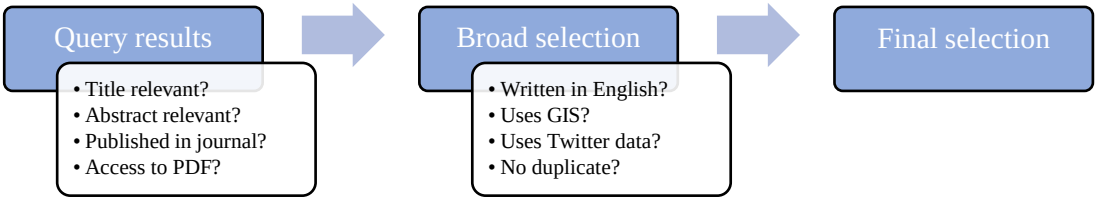

Figure 3.1: Literature selection strategy for sub question 1

- **Broad selection**: Articles will be selected based on title relevance and abstract relevance first. A title in which both Twitter (data) and a GIS-related term is mentioned is perceived to be relevant to be included in the initial article selection. Then it is determined whether a PDF file of the article can be retrieved. Only articles published in journals will be taken into account to ensure the validity of the article's content.

- **Final selection**: The broad selection created through the methodology as detailed above will be evaluated on whether in these articles Twitter data is actually used in a GIS one way or another. Only articles written in English will be taken into account due to the language proficiency of the conductor of the thesis research, previously mentioned in paragraph 2.4 as well. Possibly multiple articles by the same author(s) will be found on a similar subject with a similar research structure. Only the most recent article within this set will be included in the literature study if the articles seem too similar. This is done to prevent a bias to be created in which articles by one (group of) researcher(s) published as a continuation of each other are perceived as two separate instances of research while this is not the case.

The literature study conducted to answer the first sub question will take the form of a systematic review. When performing a systematic literature study an explicit procedure is followed while writing the literature study (Bryman, 2012, p. 102-109). This type of literature study has been chosen to answer sub question 1 specifically because this will lead to easily quantifiable results, which can then be used to design the GIS research scenarios by used to answer the fourth and fifth sub question. The academic articles found according to the methodology as mentioned will be classified based on the characteristics as described below:

- **Year of publishing:** The year in which the article is published.

- **Citations**: The amount of times the article is cited by other researchers as given in Google Scholar.

- **Application methodology**[15]: The method used to convert unprocessed Twitter data to valid results.

- **Application domain**[16]: The field in which the application is meant to be used.

- **Temporal dimension**: Whether the application uses real-time or historic Twitter data.

- **Gather period length:** The period of time over which Twitter data was gathered, if data was not real-time.

- **Study area**: The area that is studied or Twitter data is derived from in the research.

- **Additional sources**: Whether or whether not supplementary data sources besides Twitter were used.

- **Corpus size**: The number of tweets that are part of the data set used in research.

[15/16] The difference between an application methodology and application domain is that the application methodology describes the way in which data is processed while the application domain described for what purpose this data is processed.

All findings will be collected into a CSV file[17] to which these will be analysed using Microsoft Excel as specified in Appendix IV.1. CSV files are used because they create a good oversight of data and are easy to export to a wide arrange of software to create tables for in the thesis report and perform statistical analysis on. The findings found will be supported by descriptive analysis where needed. Other typical characteristics such as the author(s), title of the article and journal in which the article is published will be collected as well but merely to provide clarity to the one conducting the thesis research and are not necessarily of interest to the reader of the report. The main concepts related to the answering of the first sub question including their relevance to the keyword query are the following:

- **Twitter data**: Articles need to use (georeferenced) Twitter data as their main input of data.

- **Geographical information systems (GIS)**: Geographical information systems need to be used to come to conclusions in the articles.

The most frequently used and relevant research applications will be identified in this literature study to which this knowledge will be used to determine what GIS research scenarios will be integrated in methodology used to answer the fourth and fifth sub question. The way in which the data sets will be gathered and how these data sets will be structured will be based on this knowledge as well.

3.3 Sub question 2

The aim of answering the second sub question is to find out what the benefits and drawbacks are of using Twitter data in GIS research. Since Twitter data shows similar characteristics to both social media data[18] in general and Big Data[19] (Tsou, 2015) it is useful to take the benefits and drawbacks of those kinds of data in consideration as well.

Any article that mentions some kind of benefit or drawback of either Twitter data, social media data or Big Data is hypothetically qualified to be taken into account in this literature study. It is however preferred to use articles that focus specifically on the benefits and drawbacks on either of these types of data. Articles do not have to be written within a specific time period though preferably articles published as recent as possible are used to ensure the most recent and relevant findings on the subject are implemented in this literature study. Only articles from respected journals or conferences will be taken into account in this literature study to ensure the validity of the knowledge presented in the articles.

The literature study to be conducted to answer the second sub question will take the form of a narrative review. Such review is intended to gain an initial impression of the topic area that is intended to be understood by the conductor of the thesis research (Bryman, 2012, p. 110-113). It is structured as a narrative in which all concepts relevant to the thesis research are discussed in detail and their relevance positioned within the research context in a descriptive manner. In the case of this thesis research the focus lies on the understanding of what Twitter data is and what the benefits and drawbacks are of using this type of data with a specific focus on its use in GIS research. The main concepts related to the answering of the second sub question including their relevance to the keyword query are the following:

- **Twitter data**: The main focus on this thesis research is Twitter data thus explaining the relevance of this concept within the keyword query.

- **Social media data**: Twitter data is part of the social media data as determined by Tsou (2015). Since both data types share similar characteristics they might share similar benefits and drawbacks as well, pointing out the relevance of the concept to the keyword query.

[17] CSV stands for “Comma Separated Values” (Wikipedia, 2017h).
[18] Social media data is defined in this thesis as data that is derived from any social media website as defined in paragraph 1.1 earlier.
[19] While its definition is not undisputed, Big Data can be defined as data so big or complex traditional data processing techniques are not sufficient to use to process the data (Wikipedia, 2017i).

- **Big Data**: Social media data is part of the Big Data type as determined by Tsou (2015) as well. Since Twitter data is part of the social media data type, Twitter data and Big Data might share similar characteristics and thus benefits and drawbacks as well.

- **Geographical information systems (GIS)**: The focus in this thesis research lies in the benefits and drawbacks within GIS research and not necessarily within other academic fields.

- **Benefits and drawbacks**: This concept is relevant in the keyword query to narrow down articles that do not have any focus on the benefits and drawbacks but rather are about the use of Twitter data, social media data or Big Data in GIS research in general.

A general understanding of the concept of Twitter data within the context of GIS research and the broader scientific spectrum will be developed through this literature study to both the conductor of this thesis research and the reader of the thesis report. This is necessary due to the complex nature of the subject and due the fact that the concepts relevant to this thesis research originate from multiple academic fields due the multidisciplinary nature of the thesis research. Therefore, a slight chance exists that readers with a specific academic background might not be able to grasp the concepts and ideas presented in this thesis report because these are not part of the concepts considered to be common knowledge associated with their academic discipline. Both the benefits and drawbacks identified through this literature study will be taken into account in the methodology used to evaluate and compare of the GIMs in multiple GIS research scenarios, making up the second part of the thesis research.

3.4 Sub question 3

Finally, the third sub question will provide an answer to the question what GIMs used to infer the geolocation of tweets or users currently exists and how their workflows are structured. The definition of GIMs as used in this thesis report has been detailed previously in paragraph 1.4.

Any GIM that uses Twitter data as its main data input will hypothetically be qualified to be part of this literature study. Any GIMs that do not have Twitter data as their main input but instead data gathered from other social media will not be taken into account in this literature study to fit the research scope as previously defined in paragraph 2.4. Through preparatory research it was found that certain academics interested in the subject had written multiple articles in which each consecutive article was an improvement of the GIM presented in an article previously published by the same author(s) (see Li et al, 2012a; Li et al, 2012b for an example). If this is the case the most recent article by these author(s) will be taken into account in this literature study unless the GIMs as presented in both articles are so radically different that they can be considered different GIMs on their own. The focus in this sub question will not lie specifically on recently developed methodologies thus articles from any period can be implemented in this literature study.

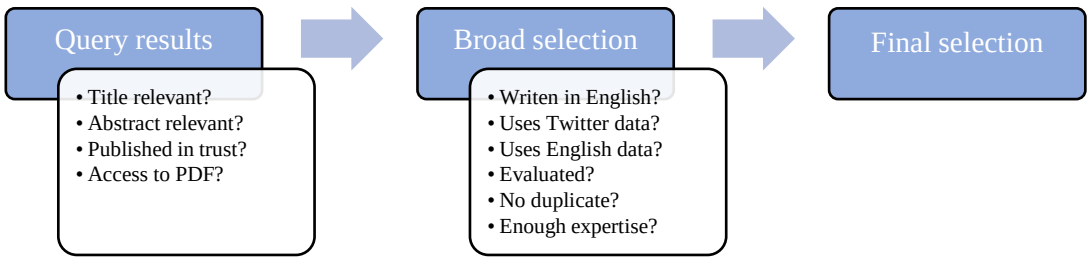

Figure 3.2: Literature selection strategy for sub question 3

Articles are gathered based on a cyclic-iterative strategy in which the article selection was narrowed down based on multiple characteristics in multiple steps as well. This is the strategy as illustrated in Figure 3.2 above and described on the next page. This strategy is similar to the one used in first sub question detailed earlier in paragraph 3.2:

- **Broad selection**: Articles will be selected based on title relevance and abstract relevance first. It is then determined whether a PDF file of the article can be retrieved. Only articles published in journals and conference papers will be taken into account. Conference papers are included as well because it turned out that the article selection would otherwise be too small to base any reliable conclusions on, as will be discussed in more detail later in paragraph 4.2.

- **Final selection**: The broad selection will be evaluated on the fact whether the GIMs as presented in the articles primarily use English Twitter data. Only articles written in English will be taken into account as well due to the language proficiency of the conductor of the thesis research previously explained in paragraph 2.4. Possibly multiple articles by the same author(s) will be found on a similar subject with a similar research structure. If the articles seem too similar, only the most recent one will be included in the literature study. The methodologies used have to match the technical proficiency of the conductor of the thesis research. The articles are excluded if this is not the case.

The literature study conducted to answer the third sub question will take the form of a systematic review which concept has previously been explained in paragraph 3.2. This means that several academic articles will be classified based on multiple characteristics. This type of literature study has been chosen for similar reasons as for sub question 1, mainly because this will lead to easily quantifiable results which can then be used to design the event detection research scenarios used to answer the fourth and fifth sub question.
The classifications used are the following:

- **Year of publishing**: The year in which the article is published.

- **Citations**: The amount of times the article is cited by other researchers

- **GIM-type**: The methodology used to infer geolocations of specified subjects. The definitions as listed below are based on the same literature on GIMs used earlier in paragraph 1.4 (Ajao et al, 2015; Jurgens et al, 2015; Han et al, 2014):

    - *Content-based*: The geolocation is determined by analysing the content of the tweets of the users and derive toponyms from these tweets.

    - *Network-based*: The geolocation of both the followers of the user and the friends of the user are used to geographically position a tweet or user.

    - *Other*: If a GIM-type can neither be classified to the two classifications above.

- **Inference subject**: The subject that is inferred by using the GIM presented in the article. The classification used is based on the findings of Ajao et al (2015, p. 2-3).

    - *User geolocation*: A broad definition of the area or places the user frequently visits.

    - *Tweet geolocation*: The geolocation from which the tweet is tweeted.

    - *Other*: If a subject can neither be classified to the two classifications above.

- **Methodology**: The main method used within the article to infer the geolocation of either tweets or users. Note that multiple methods can be used within this general methodology as will be explained later in paragraph 6.4.

- **Output form**: The type of output generated after using the GIM. This can be a place name in text form or GPS coordinate for example.

- **Scale**: The (maximum) geographical scale at which geolocation of either tweets or users can be derived when using a specific GIM.

- **Amount inferred**: The number of tweets or user locations that can be inferred using a specific GIM.

- **Error distance**: The error distance associated with the output of a specific GIM. A low error distance indicates a high accuracy and vice versa.

All findings will be collected in a CSV file which to which these will be analysed using Microsoft Excel as specified in Appendix IV.1. This is done for the same reasons as described for sub question 1 earlier in paragraph 3.2. The findings will be supported by descriptive analysis where needed. Other typical characteristics such as the author(s), title of the article and journal in which the article is published will be collected as well but merely to provide clarity to the conductor of the thesis research and are not necessarily of interest to the reader of the thesis report. The following concepts as listed below are the main concepts related to the answering of the third sub question, including their relevance to the keyword query:

- **Twitter data**: Articles need to use Twitter data as their main input of data. Data does not have to georeferenced per se because they are about to be georeferenced through using a GIM.

- **Geolocation inference methods (GIMs)**: Articles need to use a type of GIM to geographically reference coordinate-free Twitter data. Therefore, synonyms of GIMs will be used as keywords.

Through this literature study the GIMs that will be evaluated and compared to answer of the fourth and fifth sub question will be identified. The form and structure of these GIMs will be detailed and argued later in the seventh chapter of this thesis report.

4. Sub question 1: Applications of Twitter data in GIS research in 2013-2016

4.1 Introduction

In this chapter, the first sub question will be answered by conducting a systematic literature review as previously explained in paragraph 3.2. Through answering this sub question, it will be determined what the most frequently used and relevant applications of Twitter data in GIS research are. This knowledge will then be used to define the GIS research scenarios used to answer the fourth and fifth sub question. Arguments made in this chapter will be supported with academic literature where needed.

4.2 Data set creation and description

The parameters as defined below have been used to create the first selection of articles to be included in this literature study. The initial querying took place in November 2016:

- **Search engine**: Google Scholar[20] was used to find articles presenting GIS research using Twitter data primarily. Google Scholar was used specifically because other search engines such as Web of Science[21] and Scopus[22] returned few results when using the search query as specified below.

- **Search query**: The search query used in Google Scholar was as followed:

  (GIS OR "geographical* information system*") AND "Twitter data"

  The conductor of the thesis research decided to use the term "Twitter data" instead of "Twitter" as part of the search query because the latter keyword led to articles in which the Twitter account(s) of the author(s) of the article was or were mentioned rather than Twitter data usage. Therefore, the keywords "Twitter data" were used to narrow down the results found in the search engine. It was assumed that Twitter data was georeferenced when used in a GIS. Therefore, variations on the keyword "georeferenced" have not been used in the search query as specified above. The keywords "geographical" and "system" were given an asterisk (*) in the query to ensure than any variation of these words would be included in the search query.

- **Period**: Articles published between 2013 and 2016 have been taken into account exclusively because this literature study is meant to serve as a continuation of the previously written reviews on the subject, as previously explained in paragraph 3.2.

The first article selection consisted of 977 articles using the three parameters as detailed above. Narrowing this selection down to a broad selection using the criteria as specified earlier in paragraph 3.2 resulted in 139 results. The final selection consisted of 81 articles being publishing in (respected) academic journals between the years 2013 and 2016 containing scholarly research incorporating both GIS and Twitter data, after using the criteria as specified earlier in paragraph 3.2 as well. A list of these articles can be found in Appendix II.1, including relevant additional information on these articles.

The number of articles published using Twitter data in a GIS research scenario has increased over time, as can be seen in Figure 4.1 on the next page. A slight decrease can be observed for the year 2016 compared to the previous year. A plausible reason for this is that the article selection had been gathered in November with still one month worth of publishing time. Therefore, any articles published in December 2016 combining Twitter data in GIS have not been incorporated in this literature study. Generally, there seems to be an increasing interest in using Twitter data in GIS research as can be seen in Figure 4.1 on the next page as well. This development will quite possibly continue in the future given the current trend found among the final article selection. Figure 4.1 on the next page also shows that the thesis research as conducted here is relevant and possibly becomes even more relevant in the future, because it is very plausible that the interest

[20] Google Scholar is a web search engine for scholarly literature (Google Scholar, 2017).
[21] Web of Science is a scientific citation indexing service (Clarivate Analytics, 2017).
[22] Scopus is a bibliographic database for academic literature (Elsevier, 2017).

in the use of GIMs to increase the usability in GIS research using Twitter data may increase in the future as well.

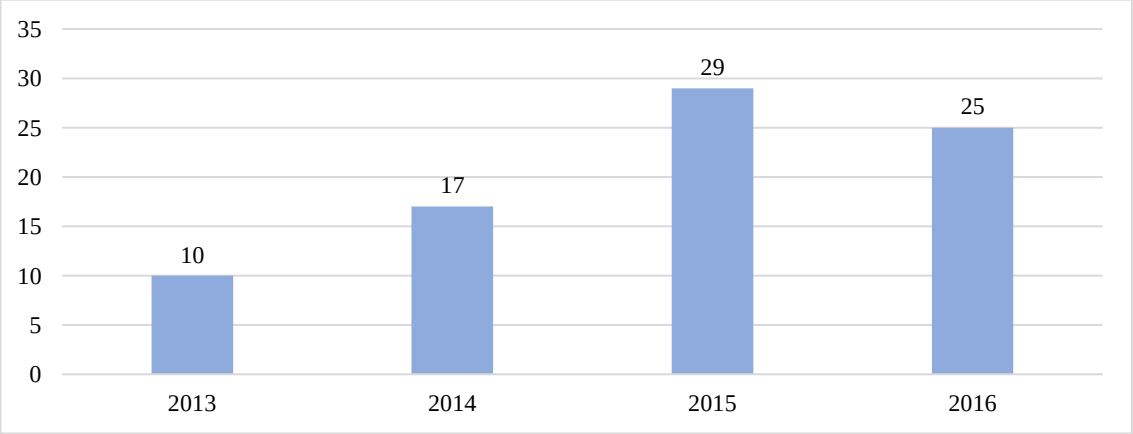

Figure 4.1 Amount of articles part of final sub question 1 article selection by year

In Figure 4.2 below the opposite effect can be observed when looking at the amount of citations per article per year compared to the observations made for Figure 4.1 above. The (average) amount of citations decreases progressively over time. This does not necessarily mean that the interest among the academic field in using Twitter data in GIS research has decreased. There are two plausible reasons that can be given for the differences that are observed when comparing Figure 4.1 and 4.2 above and below respectively. The first reason is that the pool of research combining Twitter data and GIS was relatively small in 2013 compared to the pools of research in the years 2015 and 2016. With a big interest from the academic field but limited scholarly sources to base such research on, the few articles that were available on the subject were therefore cited heavily. As more articles became available through time the few examples already available (especially from 2013) were so heavily embedded in the subject's academic paradigm they were also still being cited often compared to other articles, even though they might not be that topical anymore. The second reason is that it is obvious that through time the amount of citations decreases since articles from 2013 have three years of research to be cited by compared to one year of research for the year 2015 for example.

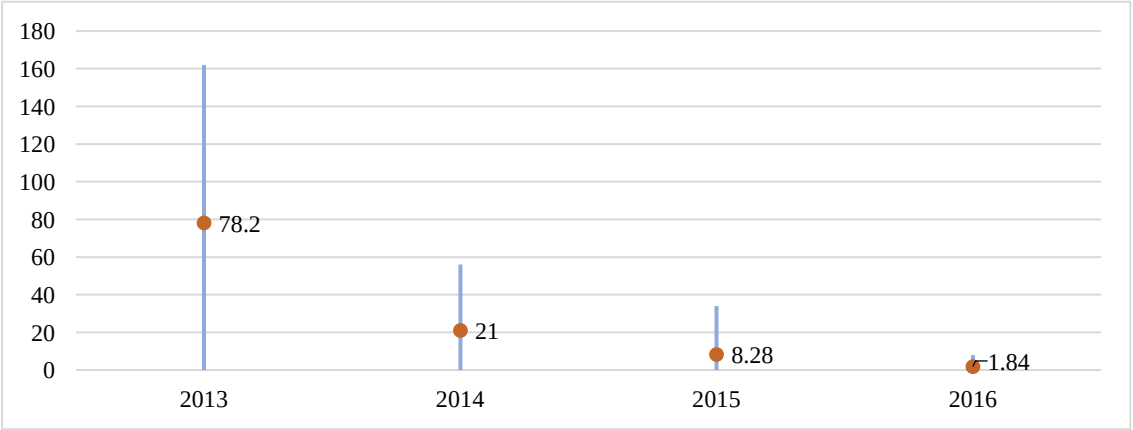

Figure 4.2: Average amount of citations of final sub question 1 article selection by year

When considering study areas, the vast majority of the articles part of the final selection had case studies set in the contiguous United States, as can be seen in Figure 4.3 and 4.4 on the next page. The contiguous United States is the only study area on which research was performed on a sub-national[23] level, such as a state-level[24]. Another relatively large group of articles have set the United Kingdom as their study area.

[23] The sub-national scale is defined in this thesis as the geographical scale on which the focus lies within an extent below the national-level but above lower geographical scales' (administrative) outer boundaries.
[24]The state-level scale is this defined in this thesis as the geographical scale on which the focus lies within an extent of a U.S. state's (administrative) outer boundaries.

When comparing articles written on the contiguous United States and United Kingdom respectively it is interesting to see that the relative number of articles written on a city-level[25] scale is much higher for the United Kingdom than for the contiguous United States. The reason for this difference in distribution could not be determined. If the United Kingdom was set as the (main) study area, the interest of the researchers lied in particularly in the area of Greater London. It has to be noted that for three articles no study area was specified and have therefore not been included in the figure below. In Figure 4.4 below the study areas as defined in the articles part of the selection have been put on a map. As can be seen in that figure there is a concentration of studies that have taken place in either the contiguous United States and Europe. A less concentrated cluster can be observed in East-Asia.

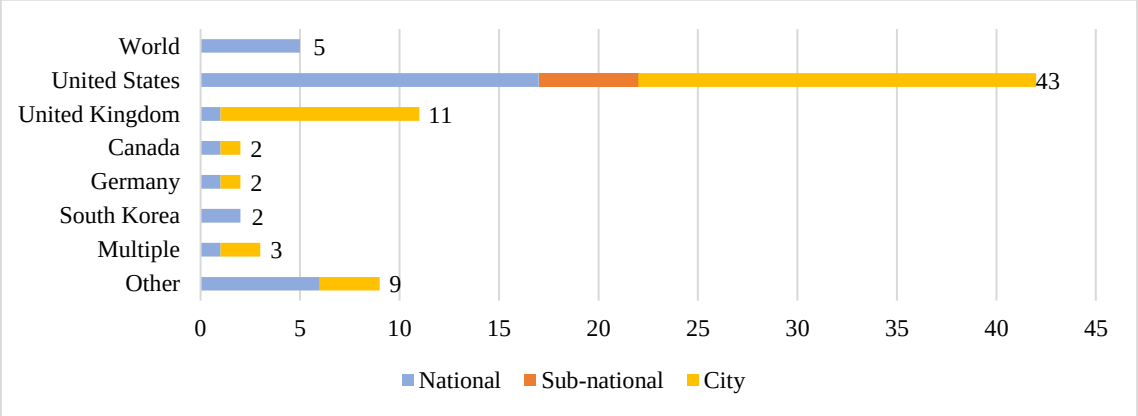

Figure 4.3: Frequency of study areas set in final sub question 1 article selection

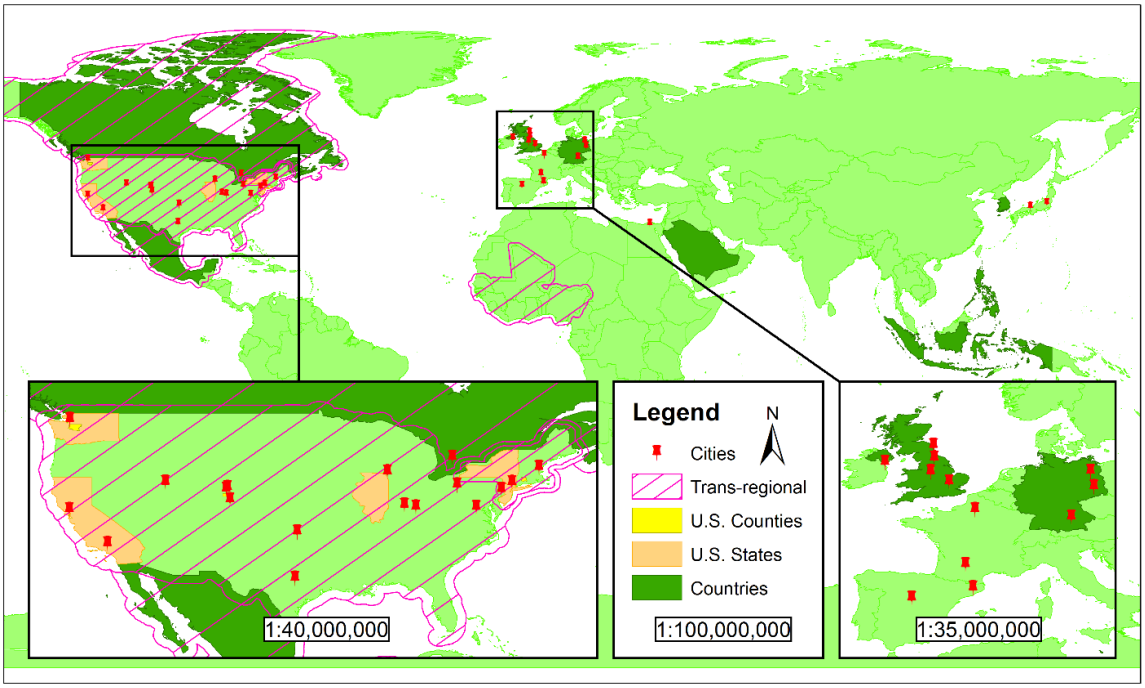

Figure 4.4: Geographical distribution of study areas set in final sub question 1 article selection

4.3 Application methodologies

The articles part of the final article selection have been classified by application methodology first. The class names used have been inspired by the works of Steiger et al (2015), though not their definitions necessarily. The reason for this is that to the opinion of the conductor of the thesis research, SNA in particular was poorly defined. Steiger et al (2015, p. 17) defined SNA as the monitoring of topics, such as political opinions,

[25] A city-level scale is defined in this thesis as the scale on which the focus lies within an extent of a city's (administrative) outer boundaries.

while in the opinion of the conductor of the thesis research these type of research subjects should be classed under the name of event detection. SNA should be about the structure of social networks rather than what opinions are situated within that social network. The latter subject is better situated under the event detection moniker according to the conductor of this research because related topics can be just as dynamic and unique as events typically subject to event detection research. Therefore, topic modelling fits better within the event detection methodology domain as has been done so in this thesis research.

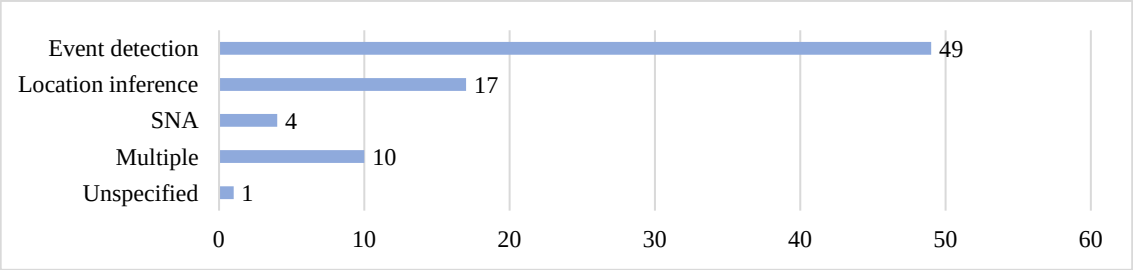

Figure 4.5: Application methodologies frequency in final sub question 1 article selection

The distribution of the application methodologies used in the final article selection has been illustrated above in Figure 4.5. The majority of the articles used event detection as their main application methodology. It is interesting to see that when comparing this literature study to the one conducted by Steiger et al referenced earlier in this paragraph that the number of articles using SNA is much lower than in this literature study. The most plausible reason for this is the different definitions used for SNA during the classification of the articles as described earlier in this paragraph as well. Several articles used multiple methodologies within the same research. Radzikowski et al (2016) used both event detection and SNA techniques to study the cyber and physical characteristics regarding vaccination in the aftermath of the 2015 measles outbreak in the United States for example.

4.4 Application domains

Articles have been classified by their application domain as well. The choice has been made to classify articles that used event detection as their main and only application methodology exclusively. The reason for this is that due the strict time-limit at which the thesis could be conducted there is not much room for other GIS research scenarios by which the GIMs could be evaluated and compared by. Therefore, only the most relevant GIS research scenarios have been chosen to be incorporated in this thesis research. Since the majority of the articles used event detection as their (main) application methodology, these have been identified as being most relevant to the thesis research. The distribution of the application domains among this new article selection can be seen in Figure 4.6 below. Short descriptions of each class can be found in alphabetical order on the next page. Each definition is clarified using an archetypical example of research within each respective application domain as well:

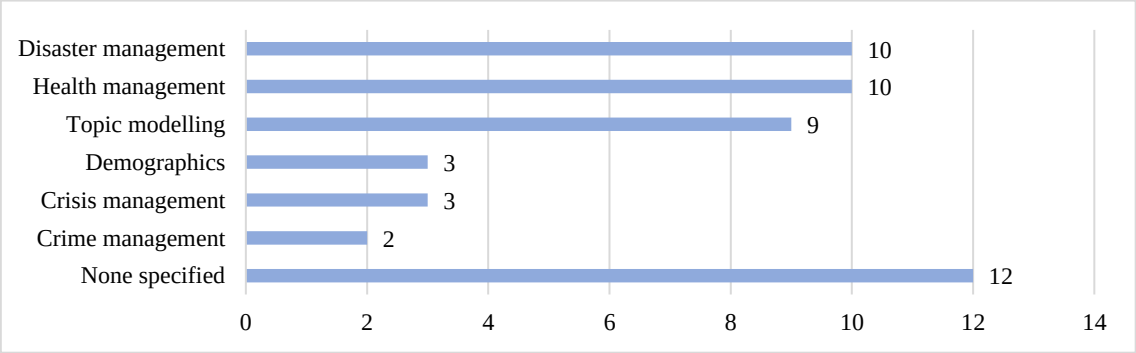

Figure 4.6: Application domain frequency in sub question 1 articles selection using event detection

- **Crime management**: Within this application domain, Twitter data is used to assess crime-related patterns in a certain area with the purpose to decrease the number of crime-related events in that particular area. Malleson and Andresen (2015) used Twitter data to measure the population at risk to violent crime in city of Leeds, England for example.

- **Crisis management**: Within this application domain class, a "crisis" is defined as an abnormal event which results in a state of civil disorder caused by human activity. One can think for example of a traffic accident or a terrorist attack. Within this application domain, Twitter data is used to limit the various negative effects of this civil disorder or even to prevent these events from happening. Gu et al (2016) have created a sensor for traffic incidents in the Pittsburgh and Philadelphia metropolitan areas using Twitter data for example.

- **Demographics**: Within this application domain, Twitter data is used to derive socio-economic, cultural or other demographic patterns in a particular area. Li et al (2013) used Twitter data in combination with Flickr data to explore socio-economic patterns (among others) across social media users in the contiguous United States for example.

- **Disaster management**: Within this application domain class, a "disaster" is defined as an abnormal event which results in a state of civil disorder caused by environmental activity. One can think for example of a hurricane or a flood. Within this application domain, Twitter data is used to limit the various negative effects of this civil disorder or even to prevent these from happening. A prime example is the earthquake sensor created by Crooks et al (2013), as previously mentioned in paragraph 1.2.

- **Health management**: Within this application domain, Twitter data is used to assess health-related patterns in a certain area with the purpose to monitor or decrease the amount of health-related issues in that particular area. Nagel et al (2013) used Twitter data to monitor influenza and pertussis outbreaks in major urban areas of the contiguous United States for example.

- **Topic modelling**: Within this application domain, Twitter data is used to derive patterns of subjective thoughts on various subject such as politics, sports or societal issues. Crampton et al (2013) mapped sentiments on Twitter on the riots following the University of Kentucky's men's basketball team's victory in the 2012 NCAA championship in Lexington, Virginia for example.

The majority of the articles for which an application domain could be determined lied either within the disaster management, health management or topic modelling domain as can be seen in Figure 4.6 on the previous page. Articles on crime management, crisis management and demographics were present as well but form a relatively small group compared to the first three mentioned. The largest groups consist of articles for which no application domain was determined. Most of these present some kind of methodology related to event detection but were not specifically meant to be used within one specific application domain.

4.5 Application data infrastructure

Finally, articles have been classified according to four characteristics related to data input. The first characteristic the articles were classified by was whether the data used in the research as presented was gathered in real-time or not. It was found that among the selection of articles that used event detection as their primary application methodology 33 articles (≈ 67%) did not implement real-time data gathering while 16 articles (≈ 33%) did. The latter group mainly consisted of articles within the disaster management application domain. The second characteristic the articles were classified by was whether additional sources had been used to compliment the Twitter data. Additional sources were used in 13 articles (≈ 27%) that use event detection as their primary application methodology. The distribution of the types of additional data used in these articles has been illustrated in Figure 4.7 on the next page. The majority of the articles use authoritative data[26] as an additional source, followed by social media data and commercial data.

[26] Authoritative data is defined in this thesis as data coming from government (affiliated) organisations.

Commercial data[27] is the only additional source which is not used exclusively as an additional source. Examples of authoritative data used in articles were for example census data from local governments (Nguyen et al, 2016a-b; Lansley & Longley, 2016), hazard-related data from FEMA[28] (Guan & Chen, 2014) and satellite imagery (Cervone et al, 2016). Flickr[29] data was used in all articles that used social media as an additional source. Other social services used were Google Plus[30] and Instagram[31]. These were only used once and within the same research as well, however (see Poorazizi et al, 2015).

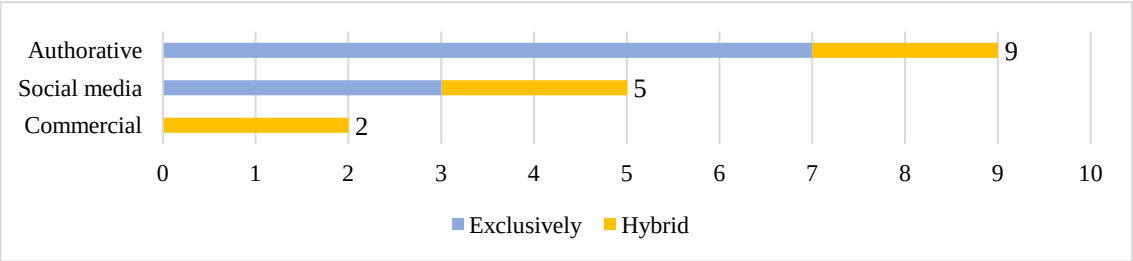

Figure 4.7: Additional sources frequency in sub question 1 articles selection using event detection

The third characteristic related to data input the articles were classified by was the corpus size of the data sets used in research. Some differences can be noted concerning the number of georeferenced tweets used in GIS research among the most frequently used application domains as can be seen in Table 4.1 below, with the highest values found being given in bold. It has to be noted that in the table below only the disaster management, health management and topic modelling application domains have been taken into account because these are the only application domains that will be implemented in the GIS research scenarios used in answering sub question 4 and 5, as will be described later in paragraph 4.6. The maximum corpus size found for the health management application domain is much higher compared to the values for the disaster management and topic modelling application domains. A possible reason for this difference could be that one given research within the health management application domain is an outlier compared to the other absolute corpus sizes found within the same application domain. The fact that the second biggest corpus size within the health management application domain (provided by Nguyen et al, 2016a) only has three percent of the size of the maximum corpus size (provided by Nguyen et al, 2016b) supports this fact. The minimum corpus size for all application domains is relatively similar, being approximately between 400 and 500 tweets.

Application domain	Studies	Maximum	Minimum	Average	Median
Disaster management	7	8000000	440	1198341	141909
Health management	6	**79848992**	402	**13834388**	149118
Topic modelling	**9**	1928937	**516**	1183224	**455981**

Table 4.1: Corpus size statistics for sub question 1 articles selection using event detection

The average corpus size for the health management application domain is significantly higher than the average found for the topic modelling and disaster management application domains. One possible reason for this is the fact that Twitter data gathered for research in the health management application domain is significantly longer than for the other scenarios, as will be described later. Another reason might be that health management research is typically done on a national scale level and thus naturally leads to a higher corpus size because of the use of a bigger bounding box. Perhaps the most important reason is the presence of an outlier as described earlier leading to a relatively high average as described earlier. Another measure that gives a less skewed overview of the corpus size distribution among the application domains is the median, which has been calculated as well for this reason. The table above shows that the median is the highest for the topic modelling application domain while similar for both the health management and disaster management application domains. By comparing the average and the median values of each application domain it can be concluded that especially the distribution of corpus sizes for the health management application domain is heavily skewed.

[27] Commercial data is defined in this thesis as data coming from commercial organisations.
[28] FEMA stands for "Federal Emergency Management Agency", which purpose is coordinate responses to disasters happening in the U.S. (FEMA, 2017).
[29] Flickr is an online photo management and sharing application incepted in 2004 (Flickr, 2017a).
[30] Google Plus is an online social network incepted in 2011 (Google Plus, 2017).
[31] Instagram is an online mobile photo-sharing site incepted in 2011 (Instagram, 2017).

The fourth and final data input characteristic that will be looked into is the total length of the period in which data used in the research presented was gathered. Some differences concerning the length of the time period in which the Twitter data was gathered among the most frequently used application domains can also be noted as seen in Table 4.2 below, with the highest values being given in bold. It has to be noted that in the table below only application domains with more than three articles in it have been taken into account for the same reason as described for Table 4.1 on the previous page previously. The overall pattern seen in the table below is that the health management and topic modelling application domains have relatively similar values while the disaster management application domain has the lowest values for all statistics. The most probable reason why the disaster management application domain has the lowest maximum is that events such as floods and earthquakes researched within this application domain typically happen over a very short period of time as well. The difference between the minimum data gathering period for the health management and topic modelling application domains is relative high because of a low outlier within the topic modelling domain (provided by Kim et al, 2016).

Application domain	Studies	Maximum	Minimum	Average	Median
Disaster management	**9**	36	1	10	8
Health management	8	425	**127**	**261**	**205**
Topic modelling	8	**518**	19	211	138

Table 4.2: Data gathering period statistics for sub question 1 articles selection using event detection

While the patterns found for the corpus size and time period of data gathering are definitely interesting, they might be biased due to these two data input characteristics heavily correlate with each other. Given that a time period of data gathering is relatively long it is obvious that the corpus size is bigger because of that reason as well. Therefore, the number of tweets per day have been calculated as well for each relevant application domain. This value has been calculated by dividing the corpus size by the amount of days in which data was gathered when both values for a specific article where available. Some differences concerning the number of tweets per day among the most frequently used application domains can also be noted as seen in Table 4.3 below. It has to be noted that in the table below only application domains with more than three articles in it have been taken into account for the same reason as described previously for Table 4.1 on the previous page and Table 4.2 above.

Application domain	Studies	Maximum	Minimum	Average	Median
Disaster management	**7**	**222222**	**37**	**60202**	**20174**
Health management	6	187880	2	33768	620
Topic modelling	**8**	101523	5	20567	1860

Table 4.3: Number of tweets per day gathered for sub question 1 articles selection using event detection

Both the maximum and minimum number of tweets per day in datasets is the highest for the disaster management application domain. The same goes for the average and the median. A possible reason for this pattern is that disasters are typically unique events and therefore people are more eager to tweet about this subject than health- or topic-related subjects.

4.6 Summary

After conducting the literature study the following conclusions can be made:

- The contiguous United States had been found to be the most frequently used study area among articles combining GIS and Twitter data. Within the disaster management application domain, the most frequently used geographical scale was on a sub-national level while for the health management and topic modelling application domains a national scale level was primarily used.

- Event detection is the most often used application methodology among articles using Twitter data in combination with GIS.

- Event detection research was meant to be applied mainly in either disaster management, health management and topic modelling if an application domain was specified at all.

- Approximately one third of these articles used real-time data, in particular within the disaster management application domain.

- Approximately one fourth of these articles used additional sources, in particular within the disaster management application domain as well.

- The corpus sizes of the data sets used within the health management and topic modelling application domains were significantly higher compared to other application domains depending on whether the average or median value was during comparison. The time period over which this data was gathered was relatively high for research within health management application domain while being incredibly small within the disaster management application domain. The number of tweets per day was relatively high for the disaster management application domain compared to other application domains.

The GIS research scenarios to be used in the fourth and fifth sub question have been defined following the findings described on the previous page and above and are presented in Table 4.4 below. The research parameters have been detailed as well. The corpus size and period of gathering have been defined by rounding off the medians as found in Table 4.1 and 4.2 previously found in this chapter and subtracting or adding ten percent to these medians. The median has been chosen over the average because median values are less prone to outliers which heavily affected some of the statistics found as previously explained in paragraph 4.5. Corpus sizes have been rounded off to thousands. It has to be noted that these GIS research scenario parameters are merely a guide for the design of the actual GIS research scenarios used in this thesis research. If implementing any parameters leads to problem concerning the conduction of the thesis research this parameter will not be implemented in the final thesis research.

Parameter	GIS research scenario 1	GIS research scenario 2	GIS research scenario 3
Application methodology	Event detection	Event detection	Event detection
Application domain	Disaster management	Health management	Topic modelling
Real-time	Yes	No	No
Additional sources	Yes	No	No
Corpus size	128000 to 156000 tweets	134000 to 164000 tweets	410000 to 502000 tweets
Period of gathering	7 to 9 days	185 to 226 days	124 to 152 days
Study area	Contiguous United States	Contiguous United States	Contiguous United States
Scale	Sub-national/City-level	National-level	National-level

Table 4.4: Preferred GIS research scenarios' parameters for sub question 4 and 5

5. Sub question 2: Opportunities and limitations of Twitter data

5.1 Introduction

A study on academic literature on the benefits and drawbacks of Twitter data and the usability of this type of data in (GIS) research in general will be conducted to answer the second sub question. From this literature study a more detailed research context can be derived both of use to the conductor of the thesis research as the reader of the thesis report to develop a better general understanding of the concepts relevant to this thesis research. The main benefits and drawbacks of the use of Twitter data in (GIS) research will be identified and taken into account in the methodology used to answer the fourth and fifth sub question and the central question.

5.2 Twitter data quantity and quality

Twitter data can easily be gathered in (potentially) vast quantities in a relatively short time depending on the search parameters used. Just over fifty-thousand tweets were found within exactly one hour when looking for tweets containing hashtags[32] within a bounding radius[33] surfacing the contiguous United States and bordering areas when creating a test dataset to evaluate the point made previously. Data quantity is an important indicator of research conclusion validity and Twitter data seems to serve its purpose at least on that part. The quantity of the data is not just big in terms of row entries, but also in the amount of metadata attributes available. This is indicated by the fact that the metadata attributes of one tweet takes forty times the amount of space as just the disk space needed to store 140 characters of one tweet (Russell, 2014, p. 22). Examples of such metadata attributes include user profile information of the person posting the tweet, the communicational purpose of the tweet and miscellaneous information such as the time at which the tweet was posted.

A big quantity of data does not necessarily have to be considered a benefit in research. As the amount of data becomes bigger, the difficulty to process that data increases as well. Storing capacity can become a serious problem in the field of Twitter research, given that the data sets used can easily contain million tweets. Since one tweet's metadata attributes is approximately five kilobytes in size (Russell, 2014, p. 22), data sets can easily become multiple gigabytes in size which still have to be processed somehow. Datasets of this size have become common rather than uncommon in this field of research (Ajao et al, 2015, p. 7; Jurgens et al, 2015, p.3). Someone interested in researching Twitter data should look for alternative storing spaces for his or her data such as (big) databases like Hadoop (Apache, 2017) or external hard disks that enable him or herself to process Twitter data efficiently and effectively. When using these data storing options the conductors of such research are enabled to conduct his or her research to their will and not restricted by the quantity of their data.

The quality of Twitter data has not been undisputed. The main concerns considering Twitter data quality are representativeness issues and the trustworthiness of the data content, as will be explained now. Representative issues exist because in particular the distribution of age of Twitter users is different than that of the real-world population. The differences in the age distributions of the U.S Twitter and real populations in 2015 have been illustrated in Figure 5.1 on the next page for example. People of the age of 18 to 44 years are overrepresented while people over the age of 55 are underrepresented among the Twitter population compared to the real US population in 2015. Sloan and Morgan (2015) found that differences in the distribution of gender, age and ethnicity in particular affect whether Twitter data is geotagged or not. This finding will be gone into more detail in paragraph 5.6 later in this chapter. Other differences in distribution between the Twitter and the real (U.S) population are seen concerning ethnicity and overrepresentation of people from urban areas (Mislove et al, 2011). These differences are problematic because patterns found among the Twitter population cannot be reflected on the real population directly with ease. This means that any research conclusions found by researchers might only be applicable on the Twitter population and not necessarily on the real population.

[32] Hashtags serve as an index enabling users to easily follow topics they are interested in by putting a #-symbol in front of these topics in their posts when mentioned (Twitter, 2017l).
[33] Similar to a bounding box but circlet-shaped instead of rectangular.

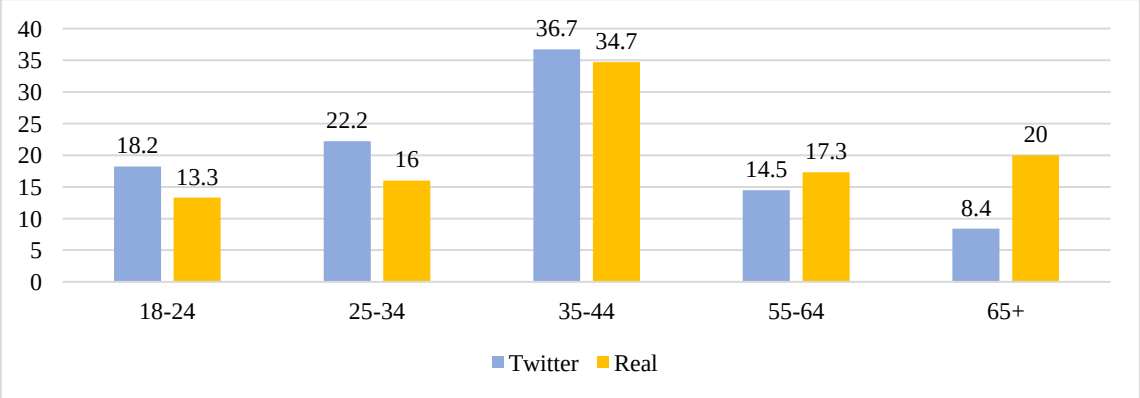

Figure 5.1: Age distribution differences among Twitter and real US population in 2015 in percentages (Statista, 2017e; KFF, 2016)

The second big issue concerning Twitter data quality is that the trustworthiness of the content of the data is sometimes difficult to determine. People sometimes set up fake profiles pretending to be a certain person at a certain geolocation while that person might not even exist, for example. While statistics on fake Twitter profiles could not be found, statistics for other social media were available. Approximately two percent of Facebook accounts are fake for example (Facebook, 2016, p. 4). Instagram initiated a so called "rapture" in which accounts were deleted on characteristics typical for fake accounts in 2014. It led to a heavy decrease of followers among celebrities' accounts especially indicating that the website hosted potentially millions of fake users (Business Insider, 2014). This type of fake profiles is difficult to filter out of data sets. It is essential to filter out fake profiles from data sets because GIS is meant to model the real world and thus needs real observations from real people. Various researchers have tried to develop methodologies to detect these type of Twitter profiles (see Kontaxis et al, 2011; Orita & Hisakazu, 2009; Gurajala et al, 2015 for examples).

A final problem worth to note is that certain uses of language such as humour or sarcasm are hard to interpret without the use of an advanced natural language processing[34] (NLP) package. A similar problem arises when words that can have vastly different meanings within different contexts are used within tweets. This problem is effectively illustrated in a research as conducted by Jung (2014, p. 9) in which he reviewed innovative approaches to study spatially linked social media. He found the following uses of the word "sick" in his database that were not related to "being sick" in a literal way:

> *"I need to call in sick today so I can watch all my shows"* – User A
>
> *"I'm sure his concert was sick as hell"* – User B
>
> *"Sick socks bro"* – User C

There is a significant risk of misinterpretation of data leading to potentially flawed conclusions without advanced NLP algorithms. Various researchers have already tried to develop methodologies to detect the exact meaning of tweets (see Davidov et al, 2010; Maynard & Greenwood, 2014; Ptácek et al, 2014 for examples).

5.3 Twitter data access possibilities and limitations

Twitter data can be gathered relatively easily using the Twitter API in combination with a programming language applicable with this API such as Python[35] or R[36]. People who want to use this API only have to go through a quick and relatively easy authentication process to gain the legal rights to gather this type of data (Russell, 2015, p. 12-15). This can be considered ironic given that this type of data could easily be

[34] Natural language processing (NLP) can be defined as the field of science concerned with programming computers to process big amounts of data sources containing natural language.
[35] Python is a programming language with a wide range of possibilities, often called the equivalent of a "swiss knife" (Python Software Foundation, 2017).
[36] R is a programming language meant for statistical computing and developing statistical graphics (The R Foundation, 2017).

considered privacy-sensitive. The data output of this API provides information on more than a few dozen characteristics per tweet. These include ones that are not even visible to regular Twitter users such as the default language specified by the user and identifiers of several metadata attribute objects as they are featured in the database of the Twitter corporation. This wide variety of metadata attributes that can be accessed enables researchers to do a wide variety of research using this type of data as well. There are typically two main types of data gathering freely available through the Twitter API (Twitter, 2017m). These are the following:

- **Historic sampling**: Twitter activity that has happened in the past will be gathered using a "fish-net" method. Using this method, a certain amount of tweets is being "caught" at random from a large pool of past tweets. This way of gathering data is technically easy to set up but has some temporal limitations that might not make this method useful in certain research scenarios.

- **Real-time streaming**: Twitter activity happening at the very moment of streaming is being processed. This type of data gathering can be used as a data input of a real-time application or be used in phenomena from which the time span at which they happen is known on beforehand.

Despite the fact that there is a wide variety and degree to which Twitter data can be gathered certain limitations have been put up by Twitter in data access for technical reasons and to ensure the privacy their user base and not break any (international) laws. These limitations are the following:

- **Data sampling**: Researchers are not able to gather all tweets posted but rather a one percent random sample of all tweets posted (Twitter, 2017j). This is done through an iterative process in which a set of one hundred tweets is gathered at random from the big pool of tweets until a certain quantity limit is reached to which the sampling is forbidden for 15 minutes. Other limits such as a maximum time at which tweets are gathered can be set by the researcher if preferred.

- **Rate limiting**: Only data from the time of gathering until seven days prior can be gathered (Twitter, 2017k). This can be considered problematic for researchers who are interested to research phenomena that happen over a longer period of time in the past.

- **Default settings**: The geotagging option is disabled by default to ensure that when users use this option they do it intentionally (Twitter, 2017i). Therefore, not all tweets are automatically tweeted with a GPS coordinate attached to it. This directly influences the usability of Twitter data in GIS research as will be described later in paragraph 5.6.

There are multiple ways to overcome these limitations. One option is using the official Gnip service as set up by Twitter itself (Gnip, 2017). Gnip provides the almost the same possibilities as the regular Twitter APIs. The main difference is that there are no limitations when it comes to gathering data. The problem is that this service costs a certain amount of money depending on the amount and type of data needing to be gathered. The case whether someone is able to use this service is completely dependent on financial resources. These might be scarce within certain academic contexts (like the one in which this thesis is conducted).
A second option is working around these rate limitations through smart programming and data processing. A reasonable amount of expertise on these methodologies is necessary to succeed though.

5.4 Platform openness and privacy issues

Twitter is content-wise a relatively open platform, on which users often reveal personal information to the public. Examples of these are opinions, personal health and the whereabouts of these users. A lot of research using Twitter data to research these kinds of subjects have therefore been able to be conducted, as has been illustrated with the various examples in the thesis report so far. Moreover, this information is relatively easy to access even with limitations incorporated similar to the point made earlier in paragraph 5.3.
Ironically enough this provides new research possibilities especially in which the research subject is considered sensitive. Examples incorporating GIS include detection and remote monitoring of HIV outcomes (Young et al, 2014), detecting depressed users (Yang & Mu, 2015) and exploring the political discourse of

users (Nelson et al, 2015). Twitter data can serve as a good alternative to for example surveying because the latter is in particular less anonymous.

Twitter data is often personal data and therefore privacy-sensitive. The results of certain research such as in the field of geolocation inference can be considered unethical or illegal when privacy-ensuring measures have not been made by researchers, such as anonymizing the data. An example would be analysing and visualising lifestyle patterns of individuals such as research conducted by Huang and Wong (2016) in which they have linked home- and work locations of Twitter users to the socioeconomic status of users on an individual level. It is therefore necessary for researchers to set ethical and legal boundaries on beforehand to prevent themselves from to breaking these, quite possibly preventing them from conducting their research as a whole. Several academics have tried to develop such framework with the aim to enable research to set up the proper boundaries for their research (see Conway, 2014; Crawford & Finn, 2015; Henderson et al, 2013 for examples).

5.5 Academic interest and expertise

Social media are interesting phenomena affecting the world on a global scale and have therefore found interest among scholars from multiple academic fields as well. Social media has also been used in a wide variety of applications, as discussed previously in the fourth chapter of this thesis report. The fact that Twitter data is applicable on such a wide variety of subjects possibly plays an important factor in this, as illustrated previously in paragraph 4.4.

Some level of technical expertise concerning the web and social media in general is required to be able to understand the subject, however. More importantly adept expertise on gathering, storing and analysing Twitter data is required in (several) programming languages such as Python and R. The problem is that in (especially) alpha sciences such as social sciences researchers often do not have this level of technical expertise. The reason for this is that (big) data processing and database management have never been an important part of the research paradigms associated with these specific academic fields until recently and in the past overly qualitative methodologies were used such as surveys and interviews. A strong indicator that proves this statement is the fact that neither programming, data processing or database management are discussed in often-referenced methodology handbooks for social sciences (see Bryman, 2015; Bernard, 2012; Tashakkori & Teddlie, 2010; Punch, 2014; May, 2015 for examples). Researchers from these fields are often not able to conduct research to their will or unaware of certain research possibilities without self-learning the essential skills or getting help from scholars from other disciplines.

5.6 Twitter data usability in GIS research

In this paragraph, a focus will be laid on the spatiality of Twitter data since this thesis is about the use of Twitter data in GIS research scenarios specifically. Twitter data provides multiple attributes of metadata on the whereabouts of the user or tweet on different scale levels. The most relevant ones for use in GIS research are listed below and on the next page:

- **Tweets** (Twitter, 2017n): Direct geolocation information can be derived when users have enabled the geotagging option meaning that to all tweets posted a geolocation on GPS level will be attached. Geolocation information can indirectly be gained using information on what language is used in the tweet.

- **Users** (Twitter, 2017o): Direct geolocation information can be derived from the user-defined profile location and in which time zone the user has specified to be living in. Geolocation information can indirectly be gained using information from the user description or self-declared user interface language.

- **Places** (Twitter, 2017p): Users can attach their tweets to certain places using a gazetteer. Tweets do not necessarily be issued from that geolocation but could also potentially be about that geolocation. Metadata attributes such as the street address can be derived from these places as well as the

bounding box associated with the specific place. Other information such as the city and country in which the place is settled can be derived from the metadata attributes.

This wide range of spatial data available on Twitter users and their tweets should enable geoscientists to rapidly gather such information for their own research in theory. Accurate geolocation information on users and tweets is sparse in reality though. A possible reason causing this is that most users do not like to share their geolocation to ensure their privacy. When filtering Twitter data sets on the availability of geolocation information researchers might potentially have to throw away the majority of their found data. Possibly the amount of data after post-processing is not enough to satisfy the data quantity needs in certain GIS research scenarios. The distribution of most relevant geolocation-related metadata attributes of tweets among the test dataset as described earlier in paragraph 5.2 has been detailed in Table 5.1 below.

Attribute	Description	Absolute (n)	Relative (%)
location	User-defined geolocation for account's profile.	42253	83,9
time_zone	The time zone the user declares him or herself to be in.	30080	59,7
country	The country related to a tweet.	1853	3,7
full_name	Full human-readable representation of place's name.	1853	3,7
name	Short human-readable representation of the place's name.	1853	3,7
coordinates	GPS coordinates of tweet	156	0,3

Table 5.1: Distribution of geo-related metadata attributes of tweets in test dataset

Some geo-related metadata attributes are relatively well represented while others are hardly represented in the test dataset at all as can be seen in the table above. Metadata attributes on user-defined geolocations and time zones are well represented. The data quality turned out to be poor, however. A significant portion of users entered a geolocation that was not even real (such as Middle-Earth) while some time zones were not even part of the United States or surrounding areas. The latter can be explained due the fact that these specific users might be tourists visiting the United States. Metadata attributes on country, place names and coordinates are poorly presented among the test dataset but are more trustworthy. When one uses Twitter data as their main or additional source in GIS research it is important to find a healthy balance between data quantity on one side and data quality on the other side.

It has to be noted that differences in the availability of geo-related metadata attributes exists among Twitter users depending on their characteristics. Sloan and Morgan (2015) found that in particular the native language of the Twitter users played an important role in the number of users enabling the geotagging option. Especially users who spoke Turkish, Portuguese and Indonesian as their native language enabled this option while for example Russian-tongue users hardly did. This is an important factor to consider even though the exact reason behind this pattern could not be determined by the researchers.

5.7 Summary

The following opportunities and limitations have been identified through studying academic literature and will be taken into account in the answering of the sub question 4 and 5:

- While Twitter data can be gather relatively quickly and in potentially big quantities, the data quality cannot be guaranteed.
- While Twitter data is relatively easy to access certain, access limitations need to be taken into account by researchers when developing a research design.
- Twitter has an open nature but may potentially threat privacy of users when using this type of data in certain research contexts.
- While there is a big interest in Twitter from different academic fields, researchers from some fields do not have the necessary skills to process Twitter data within their research context into valid results and conclusions.

- While Twitter data metadata attributes may contain multiple indications of geolocation on different scale levels, these metadata attributes are in reality only sparsely featured in such data sets.
  This proves that the incorporation of GIMs is relevant and necessary, showing the relevance of this thesis research.

This knowledge will be taken into account in the answering of sub question 4 and 5 in the following way:

- Data quantity and quality metrics will be used to measure both the data quantity and quality validity of the data output of the GIMs evaluated in this research. This way it can be determined whether the data output of a specific GIM meets the data quantity and quality needs of a specific GIS research scenario.

- Methodologies to work around rate limitations set up for the Twitter API will be implemented in the technical framework of this thesis if necessary to be sure that a sufficient amount of data can be worked with in this thesis research. Maximizing the amount of data used is especially beneficial for the validity of the conclusions made in this thesis report.

- The privacy of users part of the Twitter data sets used in this research will be ensured by incorporating appropriate measures such as anonymizing data where possible.

- The GIMs that will be evaluated and compared in this research will be designed in such a way that they can be used by an averaged skilled GI-scientist without nullifying these methodologies due to oversimplification.

6. Sub question 3: Popular methods to spatially infer Twitter users and tweets

6.1 Introduction

The third sub question will be answered in this chapter by conducting a systematic literature review in which the most often used and relevant methods to infer the geolocation of tweets or users will be determined. This information will then be used to define what GIMs will be evaluated and compared in the fourth and fifth sub question according to the GIS research scenarios as defined previously in the fourth chapter of this thesis report. Arguments made will be supported with academic literature where needed.

6.2 Data set creation and description

For the conduction of this literature study the methodology as explained earlier in paragraph 3.4 has been applied. The parameters as detailed below have been used to create the first selection of articles to be included in this literature study. The initial querying took place in December 2016:

- **Search engine**: Google Scholar was used to look up articles. The reason for this choice is that alternative search engines such as Web of Science and Scopus only gave very few results when using the search query, as previously detailed in paragraph 4.2 as well.

- **Search query**: The search query used in Google Scholar to look up articles was:

  ```
  ("location prediction*" OR "location estimation*" OR "location profiling*" OR "location inference*" OR geoinference*  OR "geolocation prediction*" OR "geolocation estimation*" OR "geolocation profiling*" OR "geolocation inference*") AND Twitter
  ```

  This query was created based on a cyclic-iterative process in which over time synonyms for geolocation inference from article abstracts were added to expand the search query. In contrast with the search query used to look up articles for sub question as specified earlier in paragraph 4.2, Twitter was used as a keyword instead of "Twitter data". The main reason was that when using the latter set of keywords a lot of articles turned out to be excluded from the selection, including some often-cited ones. Therefore "Twitter" was chosen as a keyword to be sure those type of articles were included as well. The keywords "prediction", "estimation", "profiling", "inference" and "geoinference" were given an asterisk (*) in the query to ensure than any variation of these words would be included in the search query.

- **Period**: Articles from any year have been taken account in this first selection.

The query parameters as detailed above resulted in a first selection of 2890 articles. The broad selection resulted in 85 results using the criteria as previously detailed in paragraph 3.4. The final selection consisted of 60 articles being publishing in (respected) academic journals (≈ 22%) or conferences (≈ 78%), written in English and using Twitter data exclusively to infer a subject. The geolocation of Twitter data was inferred using an experimentally verified methodology in these articles as well. A list of these articles can be found in Appendix II.2, including relevant metadata on these articles as well.

The number of articles and papers published on the subject of GIMs has increased over time as can be seen in Figure 6.1 on the next page, especially in 2014. The reason for sudden peak found for 2014 and the following years is unknown.

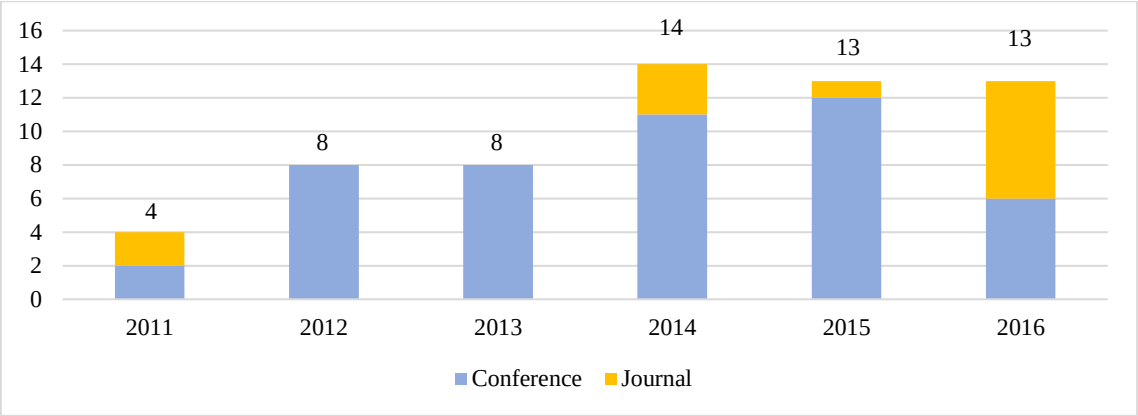

Figure 6.1: Amount of articles part of final sub question 3 article selection by year

For the amount of citations of articles used, the opposite effect can be observed as illustrated in Figure 6.2 below compared to the pattern observed in Figure 6.1 above. The amount of citations decreases progressively through time. This does not necessarily mean that the interest among the academic field in using GIMs research has decreased. Two possible reasons can be determined, similar to the ones found for the amount of citations over time for the selection for the first sub question as detailed earlier in paragraph 4.2. The first reason is that the pool of research in the period of 2011 to 2013 was small compared to the ones in the years 2014 to 2016. The second reason is that it is obvious that through time the amount of citations decreases since articles from 2013 have three years of research to be cited by compared to one year of research for the year 2015 for example. Another interesting effect over time is that all the articles selected in the period of 2011 to 2013 have at least one citation. The reason for this might be the same reason given for the relatively high amount of citations meaning that the pool of methods available was relatively small but the interest-level was high.

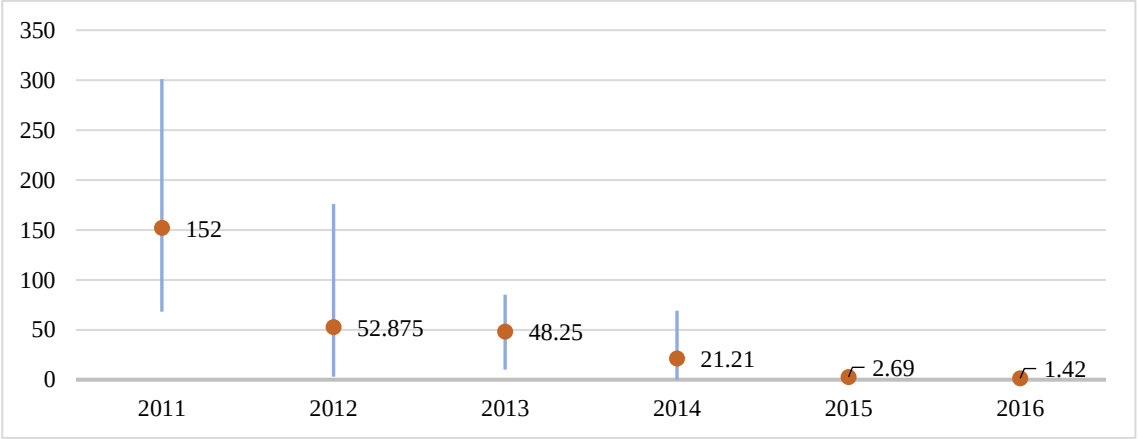

Figure 6.2: Average amount of citations of final sub question 3 article selection by year

The vast majority of the articles part of the final selection had case studies set in the contiguous United, as can be seen in Figure 6.3 and 6.4 on the next page. It has to be noted that the results as presented in the figures are a bit skewed since plenty of articles used the same data sets. An example of this is the data set as gathered by Eistenstein et al (2010) which is also used for the evaluation of proposed GIMs by Han et al (2014), Duong-Trung et al (2016) and Liu and Inkpen (2015) among others. The majority of the case studies have been done on a national scale [37] while articles that used a city-level scale often used multiple cities as their study area. Rodrigues et al (2016) used multiple cities in Brazil, Paraskevopoulos and Palpanas (2016) used multiple cities in Italy while Kinsella et al (2016) used the top ten of cities with most Twitter users in a specific year for example. While Figure 6.4 shows clusters in the contiguous United States, Europe and Brazil this visualisation is a bit misleading because the cities in Brazil are all part of the same case study. It has to be noted that no study area was or could be specified for 13 articles.

[37] The "national scale" is defined in this thesis as the geographical scale on which the focus lies within an extent of a nation's (administrative) outer boundaries.

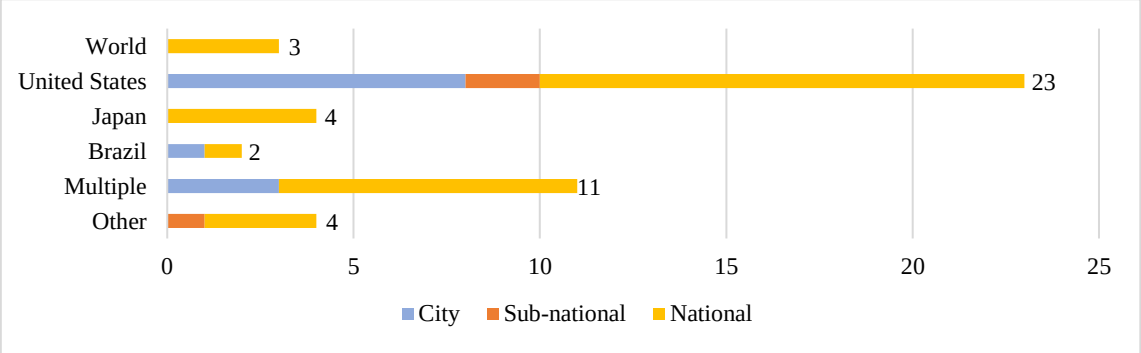

Figure 6.3: Frequency of study areas set in final sub question 3 article selection

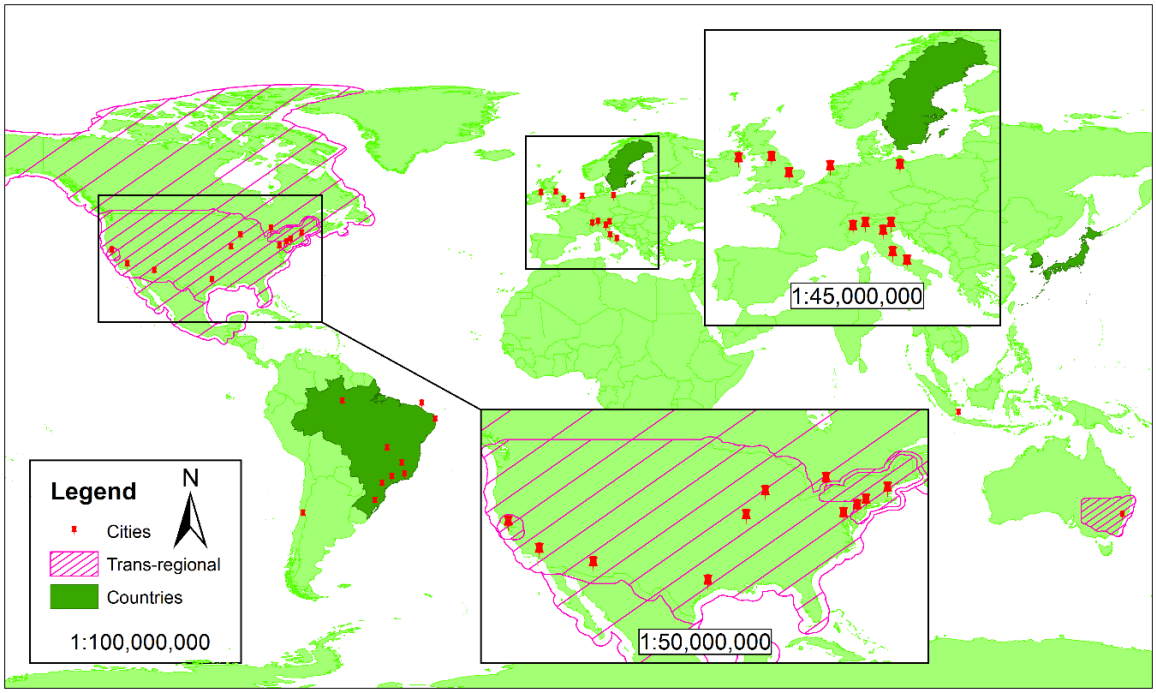

Figure 6.4: Geographical distribution of study areas set in final sub question 3 article selection

It has been defined earlier in paragraph 2.4 that Twitter data in English will be used in this thesis research exclusively. This is done because most natural language processing packages have a bias towards this language and the fact that the conductor of this thesis research is proficient with this language but not with other languages generally used in this type of research. Therefore, the choice has been made to discard content-based or hybrid[38] methodologies that used different languages. For network-based methodologies language is not necessarily important because geolocations are determined rather by toponyms who are not heavily influenced by language used either by the user or in the tweets themselves. The distribution of languages in Twitter data across the article selection has been visualised in Figure 6.5 on the next page. The main language featured in the Twitter data could generally poorly be determined since for 40 articles (≈ 66%) no language was specified. If no language was defined the native language of that particular country or area has been assumed to be the main language of the Twitter data as well, if a study area for these articles was known. Articles of which case study was set in the (contiguous) United States are considered using English Twitter data while articles with a case study set in Japan are considered using Japanese Twitter data for example. In a few articles multi-lingual Twitter data was used. If tweets used in research were overly in English, these articles were considered to meet the research scope as defined earlier in paragraph 2.4 and thus remain included in the article selection as well. 17 articles have been excluded because they did not use English Twitter data exclusively or primarily.

[38] Hybrid-based LIMs are methodologies that use a combination of both content-based and network-based techniques.

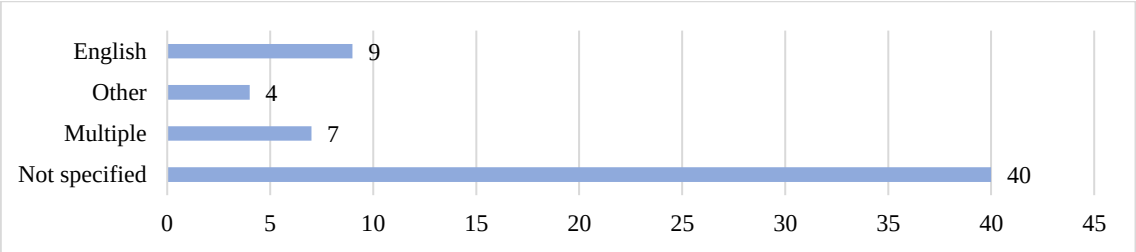

Figure 6.5: Twitter data language frequency in final sub question 3 article selection

6.3 GIM-types

The articles that use English Twitter data have been classified by GIM-type first. The distribution of the different GIM-types is illustrated below in Figure 6.6. The majority of the methodologies are content-based, meaning that they primarily use the texts of tweets, user descriptions or other content to determine the geolocation of either users or tweets. Hybrid methodologies combined both content and network metadata to determine the geolocation of either user or tweets. Zhang et al (2015) used for example a method in which they first determined the city of residence through a network-based method and secondly derived a more detailed sub-city geolocation by relating a specific user to another user from which such detailed information was available from. Another example comes from Gu et al (2012) who estimated a geolocation through a text-based and graph-based methodology and combined these findings to determine the final geolocation. What can be seen in Figure 6.6 below as well is that for network-based and hybrid GIMs the main purpose is to locate users while for content-based GIMs there are plenty of methods available for both the inference of the geolocation of tweets and users. The reason for this is unknown. Perhaps network and hybrid approaches are not perceived to be suitable for inferring the geolocation of tweets among the scientific discourse of GIMs.

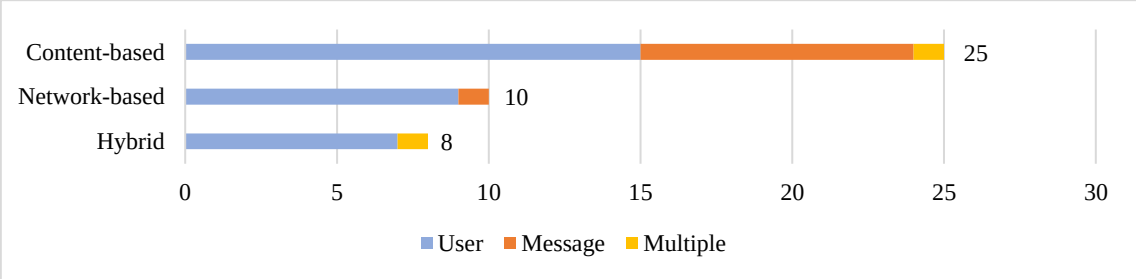

Figure 6.6: GIM-type frequency in sub question 3 article selection using English Twitter data

A more detailed inference subject was defined in articles that presented GIMs specifically aimed at inferring the geolocation of users (≈ 23% of the complete final selection). The first group (among articles of all GIM-types) were methodologies specifically meant to infer the home location of users. Li et al (2012a) used a hybrid approach combining both tweet content and the user's social network to determine the home location of that specific user for example. A second group of mainly content-based GIMs were methodologies specifically meant to infer mobility patterns of users. Huang and Wong (2016) used a content-based approach in which they linked activity patterns to the socio-economic status of users using Twitter data for example. Articles that had a mobility-centred subject have been excluded in the next sections of the report because only "static" subjects are generally used within event detection research (Steiger et al, 2015). This led to six articles being excluded (≈ 10%) leaving a new total of 34 articles to be taken into consideration for the determination of GIMs to be evaluated and compared in the second part of this thesis research.

6.4 General workflow

Classifying articles based on the methodology used within those articles to infer the geolocation of specified subjects was difficult. The main reason for this was that often multiple methodologies were used to serve the article's aim which made it difficult to filter out the main methodology used. Cha et al (2015) used for example sparse coding, dictionary learning and pattern recognition in their hybrid user-centred GIM.

Another difficulty arose by the fact that different techniques could be used to serve an article's aim. The tie-strength of two Twitter users can be determined through various methodologies for example. Examples are low density graphs as used by Apreleva and Cantarero (2015), a total variation minimization technique as used by Compton et al (2014) and an iterative model as used by Chen et al (2016). The aim of these methodologies was the same while the workflow structure of these methodologies was very different. Different synonyms were used for methodologies that practically did the same thing as well. A typical example are the terms "text mining" as used by Ren et al (2012) and Cheng et al (2013) and "text analysis" as used by Lingad et al (2013). A similar example comes from the term "tie-strength" as used by Chen et al (2016) and Zhang et al (2015) which is also known under the name "social closeness" as used by Liu & Huang (2016). Therefore, the choice has been made to classify articles on general workflow rather than on the exact methodology used to work around the difficulties as specified above on and the previous page. To derive the main workflow the following procedure has been followed:

- Information on the general workflow from the article keywords were derived first. Lingad et al (2013) used "text analysis", "named entity recognition" and "social media mining" as keywords to index their article for example.

- Information on the general workflow from the article categorization or descriptors were derived next. Beside the keywords mentioned above, Lingad et al (2013) also use both "natural language processing" and "text analysis" as subject descriptors for example.

- Information from the article abstract was derived when the previous two possibilities did not lead to any results. Zhang et al (2015) do not use keywords or subject descriptors for their article but do mention a methodology incorporating user tie-strength in the abstract of their article.

- If all three methodologies mentioned above did not reveal any workflow information the article was read in detail to derive workflow information that way.

The general workflows used within the article selection concerning inferring static subjects has been illustrated in Figure 6.7 below. It is important to notice that in this figure GIM-types that were unique rather than part of a bigger GIM-type group have been excluded in this figure and the rest of this thesis report. The GIM as developed by Davis et al (2014) has not been included for example due the fact it was the only article that used a network-based approach to infer tweets. For both content-based GIMs text mining is most often used to derive the geolocation of either users or tweets. This pattern is logical due the fact that content on Twitter is mainly in text-form. Among articles different techniques were used under the umbrella of text mining. Lingad et al (2014) and Zhang and Gelernter (2014) used similar techniques ("named entity recognition" and "named entity disambiguation") in which toponyms were automatically derived from tweet content using a gazetteer[39]. Laylavi et al (2016) used an iterative-model in which the tweet's geolocation was not just based on tweet content but also profile information or the place as given by the user itself when possible.

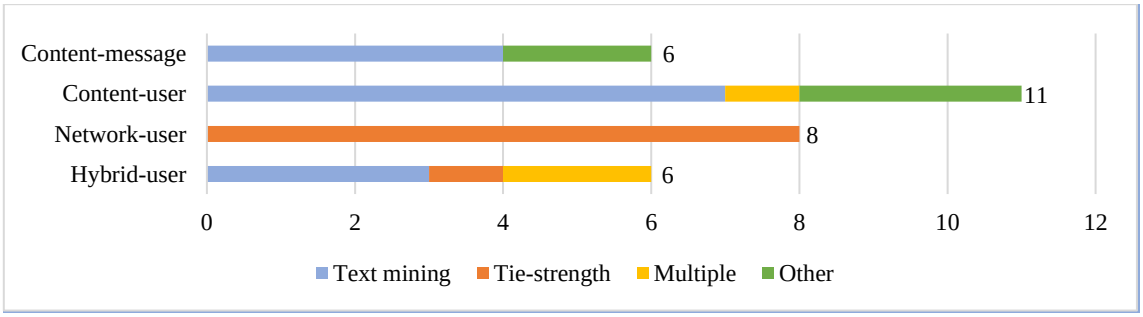

Figure 6.7: Workflow frequency in sub question 3 article selection using English Twitter data

[39] A gazetteer can be defined as a directory containing toponyms (Wikipedia, 2017j).

Network-based GIMs exclusively used to infer the geolocation of users most often calculated one way or another the level of friendship between two individuals to determine the tie-strength between these individuals. Geolocations of friends can be weighted through these tie-strength for example.
Different methodologies were used to determine the tie-strength between two users. Yamaguchi et al (2013) used the concept of "landmark users", which were users that are central nodes within social network from which the geolocation was known from. Depending on the tie strength of a particular user to these landmark users the geolocation can be derived. Chen et al (2016) and Kong et al (2014) both created a metric by which friendships are weighted. Two users' tie-strength is weak when they just follow each other back for example but strong if they follow each other, retweet each other and mention each other in tweets. Aprevela and Cantarero (2015) used a combination of graph theory, Gaussian distributions and diffusion processes to determine the tie-strength

Among hybrid user-centred GIMs many different workflow structures are used to determine the geolocation of a user. There is therefore not necessarily one workflow that is typical for this type of GIMs. Cha et al (2015) use a combination of sparse coding, dictionary learning and pattern recognition to determine the geolocation of users. Li et al (2012a) use a unified discriminative influence model which iteratively determines what users within a social network can be used to determine one's geolocation. Finally, Kotzias et al (2016) used a social graph of a social network which was based both on text mining and tie-strength.

Other methodologies used are worth mentioning as well while not taken account into the fourth and fifth sub question necessarily. Huang et al (2014) used a GIS-driven approach in which they clustered points of activity and intersected that cluster with activity zones in the city of St. Louis, MO. Duong-Trung et al (2016) used a GIS-driven approach as well combining matrix factorization, regression and models of learning to infer the geolocation of users based on tweet content. Roller et al (2012) used k-tree clustering to derive geolocations of messages from tweets. While these methodologies will not be taken into account in this thesis research, they are definitely worth researching in future work.

6.5 Data output quality

On beforehand a methodology was thought of in which articles on GIMs would be classified by scale, the amount of data from which the geolocation could be inferred and the distance error of the GIM output as detailed earlier in paragraph 3.4. These results would then be presented in a style similar to the way in which the results were presented in chapter 4 earlier in the thesis report. It turned out that for the majority of the articles these characteristics were not specified at all which makes it impossible to draw generalizing conclusions on each of the GIM types as described earlier. Therefore, the choice has been made to present all articles that have specified the majority of the parameters as specified earlier in paragraph 3.4 in one table as has been done in Table 6.1 below. The values as given in the articles themselves have been presented and not been rounded off. When error distances (E.D.) have been given in miles these values have been converted to kilometres while maintaining the same amount of accuracy.

Presented by	Based on	Subject	Method	Output	Scale	Inf.[1]	Avg. E.D.[2]	Med. E.D.[2]
Cheng et al (2013)	Content	User	Text mining	City name/GPS	City	54.26	760.03	
Yamaguchi et al (2013)	Network	User	Tie-strength			85.0	297.739	3.804
Compton et al (2014)	Network	User	Tie-strength	City name/GPS		81.9	289.00	6.38
Han et al (2014)	Content	User	Text mining	City name/GPS	City	49.0		9
Cha et al (2015)	Hybrid	User	Text mining		State	41	581	425
Krishnamurthy et al (2015)	Content	User	Text mining	City name/GPS	City	54.48	690.41	
Laylavi et al (2016)	Content	Tweet	Text mining	City name/GPS	Suburb	87	12.2	4.5

[1] in %
[2] in km

Table 6.1: Data quality parameters in selected studies part of final sub question 3 article selection

Within the selection as presented in the table above the majority of the articles output form consisted of a city name with a GPS coordinate attached to it. Typically, the centroid of the city's boundaries was chosen as the source of this coordinate. Considering geographical scale, the majority of the selection was on a city scale of a scale was specified at all. Cha et al (2015)'s study differs from the offers being the only study having a state-level as the lowest scale level scale.

The GIM as developed by Laylavi et al (2016) seems to outperform all other GIMs presented in the Table 6.1 on the previous page when comparing studies based on the amount of data from which the geolocation could be inferred and the average and median value found for the distance errors. This is especially the case for the average D.E. value found. A possible reason for this difference is the fact that outliers play a huge role in the composition of the average E.D. Whether this was the case for Laylavi et al's study could not be determined. Cha et al (2015)'s GIM performed poor compared to the other GIMs as presented in the table on the previous page. It scored the lowest on scale-level, amount of data inferred and median D.E. Especially the latter value is radically higher than the other values found within the table selection. A reason for this could not be determined other that the GIM as presented in the article was poorly designed compared to the others in the table selection. It is interesting to see that the network-based GIMs by Yamaguchi et al (2013) and Compton et al (2014) perform well considering the amount of data inferred and average E.D. compared to the other GIMs from the table selection. This possibly indicates that network-based GIMs perform better compared to other GIM-types.

6.6 Summary

The following GIM-types to be taken into consideration for this thesis research given the results found earlier in paragraph 6.2 to 6.5 as detailed in Table 6.2 below. Primarily the frequency of occurrence of each attribute within the article selection has been used to validate the choices made. Given that the output form for most articles within the selection as presented in Table 6.1 on the previous page have been on a city-level this is the preferred scale-level as well. The amount of data that has to be inferred and the preferred accuracy has not been specified due the fact that not enough articles were available to determine these values as previously explained in paragraph 6.5. These GIMs will be evaluated and compared in the fourth and fifth sub question according to the methodology as will be specified in the next chapter of this thesis report.

Parameter	GIM 1	GIM 2	GIM 3	GIM 4
Alias	Content-message method	Content-user method	Network-user method	Hybrid-user method
GIM-type	Content	Content	Network	Hybrid
Inference subject	Message	User	User	User
Methodology	Text mining	Text mining	Tie-strength	Text mining + tie-strength
Output form	City name and GPS	City name and GPS	City name and GPS	City name and GPS
Scale	City-level	City-level	City-level	City-level
Amount inferred	Unspecified	Unspecified	Unspecified	Unspecified
Accuracy	Unspecified	Unspecified	Unspecified	Unspecified

Table 6.2: Preferred thesis GIMs' parameters

7. Methodology Part 2: Evaluation and Comparison

7.1 Introduction

This thesis research consists of two parts, as previously detailed and argued in paragraph 2.6. The first part consists of a set of literature studies while the second part consists of the evaluation and comparison of GIMs in various GIS research scenarios. Both the methodology used to conduct the first part of the thesis research and the results that followed through conducting this methodology have been discussed in chapter 3 to 6 previously in this thesis report. In the current chapter the methodology used to conduct the second part of the thesis research will be presented, detailed and argued where needed. The second part of this thesis research consists of answering sub question 4 and 5, which have been previously detailed in paragraph 2.2. Through answering these sub questions it becomes clear what the strengths and weaknesses are of the relevant GIMs as selected for this thesis research through answering sub question 3, to what extent they are applicable within the relevant GIS research scenarios as selected through answering sub question 1 and to what extent they compare to the unprocessed Twitter API data output. The knowledge gained by answering sub question 2 on the opportunities and limitations of Twitter data in (GIS) research will be taken into account in the design of the methodology used in the second part of the thesis research as well.

First, the GIS research scenarios used to evaluate and compare the GIMs by will be detailed and argued. Secondly, the design of the GIMs will be presented and argued. Then the metrics used to evaluate and compare the GIMs respectively will be detailed and argued. Finally, the methodology used to answer sub question 4 and 5 will be detailed and argued and needed as well. For each of these parts the technical framework used to perform that part of the methodology will be detailed and argued as well when needed. The scripts used in this thesis research have been detailed in Appendix III and will be referred to when needed. The main and sub software packages used within the technical framework have been detailed in Appendix IV and will be referred to as well. It has to be noted that the technical framework as presented in this chapter is not necessarily as efficient or effective as it theoretically could be. The main reason for this is that while the programming proficiency in languages such as Python or R of the conductor of the research is sufficient it is not necessarily on an expert level. Therefore, some data processing steps might seem inefficient or devious to those proficient in these particular programming languages. This does not mean that the technical framework as will be presented later in paragraph 7.2.4 and 7.4.4 respectively leads to flawed results or errors in the data, however. It is merely a case of working around limitations rooted by a certain lack of programming proficiency by the conductor the thesis research.

7.2 GIS research scenarios

The GIMs as determined through answering sub question 3 in chapter 6 will be evaluated and compared in the GIS research scenarios as determined through answering sub question 1 in chapter 4 to answer the fourth sub question. These GIMs will be compared to the regular Twitter API data output using the same GIS research scenarios to answer sub question 5. GIS research scenarios are used to determine whether the content of the data as derived within the context of these scenarios is different and if therefore the applicability of the GIMs within these scenarios differs as well. It has to be pointed out that by answering sub question 1 only the preferred parameters of these scenarios have been defined and not the final parameters that will be used in the thesis research. The reason for this difference is that the preferred parameters might conflict with the research scope or constraints as defined earlier in paragraph 2.4 and 2.5 respectively. The three GIS research scenarios determined will be defined in more detail later in paragraph 7.2.1 to 7.2.3 respectively. The following components of the research design will be detailed for each GIS research scenario in these paragraphs:

- **Research scenario parameters**: Both the preferred and final parameters used in the thesis research will be detailed. Any differences between both of these parameter groups as a result of choices made will be argued where needed.

- **Research scenario objective**: A short description will be given on what the central aim of the GIS research scenario is. This central aim will be argued on the basis of academic literature with a similar research design within the same application domain. This is done to make sure that the GIS research

scenarios as defined here are representative for their academic counterparts within the same application domain.

- **Workflow:** The workflow used in the GIS research scenarios will be detailed step by step and argued where needed. Any relevant technical details not detailed in paragraph 7.2.4 will be detailed here as well.

The GIS research scenarios as presented in the current chapter have been kept as simple as possible while at the same time it has been made sure these are representative for their (more complex) application domain counterparts as well. Creating GIS research scenarios that are too complex would lead to spending unnecessary extra time on conducting these scenarios instead evaluating and comparing the GIMs, being the aim of this thesis research. It has to be noted that the pre-processed data will not be analysed further than needed since this does not serve the aim of this thesis as defined earlier in paragraph 2.1. There is an exclusive interest in analysing the data sets to measure the increase in usability of the Twitter data within the GIS research scenarios and the differences within the data sets in for example data quantity and quality. These differences will be measured according to certain evaluation and comparison metrics which will be defined later in paragraph 7.4.2. Spatial patterns of the phenomena the Twitter data is about (either being disasters, health or popular topics in the case of this thesis research) might be interesting to determine in future academic work but are currently outside the scope of this thesis research as defined earlier in paragraph 2.4.

7.2.1 Disaster management research scenario

The first GIS research scenario the selected GIMs will be evaluated and compared by is a disaster management research scenario. Within this application domain, a "disaster" is defined as an abnormal event which results in a state of civil disorder caused by environmental activity. The purpose of this application domain is to monitor or decrease the amount of disaster-related issues in a particular area, as previously defined in paragraph 4.4. In Table 7.1 below the preferred GIS research scenario parameters as defined earlier in paragraph 4.6 have been listed next to the parameters that will be used in the final thesis research:

Parameter	Preferred	Final
Application methodology	Event detection	Event detection
Application domain	Disaster management	Disaster management
Real-time	Yes	No
Additional sources	Yes	Yes
Corpus size	128000 to 156000 tweets	128000 to 155000 tweets
Period of gathering	7 to 9 days	7 days
Study area	Contiguous United States	Contiguous United States
Scale	Sub-national/City-level	Sub-national level

Table 7.1: Preferred and final parameters for disaster management research scenario

One difference is noticeable among the preferred and final GIS research scenario parameters for the disaster management scenario as presented in the table above. This difference lies in the use of real-time data particularly. The reason why real-time data is not used in the final disaster management research scenario is because of the lack of expertise by the conductor of this thesis research to create an application that handles such type of data. Knowledge on programming languages such as HTML[40], CSS[41], jQuery[42] and Python as used by Bonzanini (2015), comSysto (2012) and Aghabozorgi (2016) is required to be able to create these types of applications. Within the current time period at which the thesis research can be conducted there is simply no time to learn how to code the programming languages as mentioned above. Not using real-time data does not affect the validity of the results found through this GIS research scenario, even though it might not be as representative for the general academic discourse among disaster management research as preferred. The only difference between real-time and historic data is the time at which the data is gathered and at which rate the data access is limited. In terms of structure or available metadata both types of data are exactly the same. Therefore, this GIS research scenario can still be considered representative and this way

[40] HTML (Hyperlink Markup Language) is a markup language used to create web pages and applications.
[41] CSS (Cascading Style Sheets) is a style sheet language used to describe the presentation of a markup language.
[42] jQuery is a library that enables the user to add additional functions to their web apps such as animations and plug-ins.

the use of historic data be argued. The size of the final corpus size as presented in the table on the previous page has been based on the median number of tweets per day for the health management domain as given in Table 4.3 in paragraph 4.5 earlier in this thesis report. The median amount was chosen over the average amount because the first value is less skewed and prone to outliers as previously explained in paragraph 4.5 as well. The minimum and maximum value have been estimated by taking the median value as found in Table 4.3 and respectively subtract or add ten percent to that value and round it off to thousands, similar to the method used earlier in paragraph 4.6.

Within the disaster management discourse, additional sources are typically used to define the catchment areas used gather Twitter data used in research. Examples are Albuquerque et al (2015) and Guan and Chen (2014), who have used authoritative data to define flood catchment areas to validate the geographical distribution found among the gathered Twitter data. This is done primarily to remove noisy data from data sets. It is for example possible that users tweet about the disaster of interest while they are not in the vicinity of this disaster. The way in which the catchment areas used in this GIS research scenario are defined will be detailed later in this paragraph.

A sub-national scale level has been chosen over a city-level scale level for this GIS research scenario. While this does not necessarily depict a difference in the preferred and final parameters as described on the previous page it is important to argue this choice to validate the results found through this GIS research scenario. The main reason why a sub-national level has been chosen is because the data output of the GIMs is a city name, as will be detailed and argued previously in paragraph 6.6. More detailed user location information is necessary to be able to conduct a GIS research scenario on a city-level such as neighbourhood or street names. Since this is not the case for the GIMs as evaluated and compared in this thesis research a sub-national scale level has been chosen.

Given that disasters are difficult to predict on beforehand it has been hard to define a disaster management research scenario for the same reason. Major news outlets were checked upon on a daily basis to discover any occurrence of a disaster on the contiguous United States soil. From the 17th to 22th February 2017 a set of storms hit the state of California, leading to all kinds of disruptions in the area. The disaster led to millions of dollars in damage because of extreme weather leading to floods, but also led to many wounded and fatal casualties (CNN, 2017a-h). While the disaster is of a serious nature it can serve as an excellent case study for the disaster management research scenario as used in this thesis research. The disaster management research scenario consists of a simple design in which Twitter data is gathered on the disaster and tweets are placed on a map. It has been based on a much-referenced article by Crooks et al (2013), who have used a similar technique to create a real-time earthquake detection system. The workflow of this GIS research scenario will be the following, as illustrated in Figure 7.1 below:

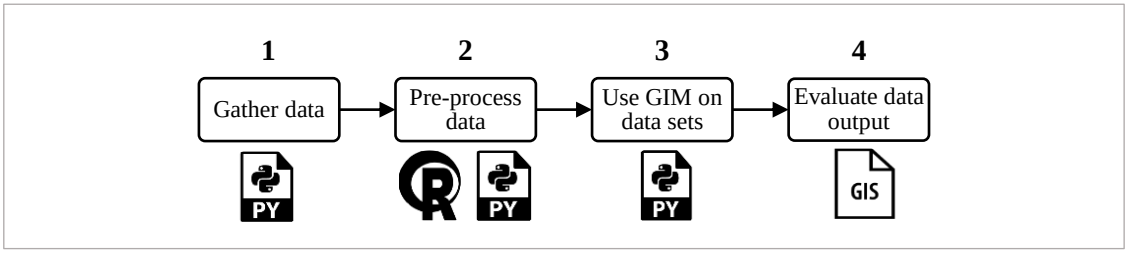

Figure 7.1: GIS research scenarios' workflow

Step 1: Gathering data: Tweets will be gathered in JSON[43]-format using a Python script as presented in Appendix III.1 and which will later be detailed in paragraph 7.2.4. Only tweets containing specific keywords related to the 2017 Californian flood are gathered. The keywords that will be used in the first step are difficult to define given that disasters themselves are difficult to predict, as previously explained in this paragraph. Therefore, the methodology used to select the keywords used in the disaster management research scenario will be detailed first based on academic literature within the disaster management domain.
The keywords used in the disaster management scenario will be based on the following three subjects as listed on the next page:

[43] JSON (JavaScript Object Notation) is file format typically used to contain data used in JavaScript applications.

- **Synonyms of the disaster phenomenon**: The first set of keywords will be based on synonyms of the disaster of interest. This approach has been used by Albuquerque et al (2015), Chae et al (2014), Crooks et al (2013), Shook and Turner (2016) and Shelton et al (2014) as well. Crooks et al (2013) used both the hashtags "#earthquake" and "#quake" to find tweets referencing an earthquake that occurred on 23rd August 2011 near Mineral, VA, United States.

- **Nickname(s) given to the disaster phenomenon**: The second set of keywords will be based on nicknames given to a particular disaster, when present. This approach has been used by Chae et al (2014), Poorazizi et al (2015), Shook and Turner (2016) and Shelton et al (2014) as well. A typical example are the nicknames given to the hurricanes in the Atlantic hurricane seasons as done by World Meteorological Organization (WMO, 2017).

- **Names of preventive measures**: The third and final set of keywords will be based on names of measures that are typically taken to prevent the negative effects of the subject disaster. Albuquerque et al (2015) used this approach to gather tweets on the River Elbe Flood of June 2013 in Germany. They used for example the German equivalent of "dike" and "sandbag".

Using the three subjects as described above a set of keywords have been defined to gather Twitter data by. Since no nickname(s) have been given to the disaster of interest, synonyms and names of preventive measures related to the 2017 Californian flood have been used exclusively.
Specifically, articles published by CNN[44] on the subject have been used to define keywords to use in this disaster management research scenario (CNN, 2017a-h). The keywords as used are listed below:

"storm", "downpour", "flood", "rain", "sinkholes", "mud", "landslide", "snow", "evacuation", "dam", "spillway"

A bounding radius from within the Twitter data used in this GIS research scenario will be gathered has been defined as well. This bounding radius has been illustrated in Figure 7.2 on the next page. Within this bounding radius the state of California and surrounding areas are contained. The central point from which the radius is defined using the WGS84 coordinate system standards and has a latitude of 37.2, longitude of approximately -119.6 and a diameter of 700 kilometres. The central point and diameter of the radius have been defined by calculating the median point for the contiguous United States and the minimal distance needed to contain this area within one buffer using the ArcMap software package, as specified in Appendix IV.1. Using this bounding radius, Twitter data outside the disaster management research scenario study area will be gathered as well. Twitter data that originates from outside the study area will be excluded from the data set using GIS, as will be described later in paragraph 7.2.4. The data itself has been gathered on the 23rd and 24th February 2017.

Step 2: Pre-processing data: The data will be pre-processed by the steps as will be detailed later in paragraph 7.2.4. The data set will be clipped by catchment area using GIS techniques. This is done to make sure that only tweets originating from the actual place of disaster are part of the final data set, as explained previously in this paragraph. Clipping data by area is especially difficult for this GIS research scenario given that there is no data (as of writing) available on the exact areas in California that have been affected by the floods of February 2017. For that reason, self-made and -defined boundaries have been used in this particular GIS research scenario instead as illustrated in Figure 7.3 on the next page. Specifically, the administrative boundaries of counties in which toponyms named in CNN articles (CNN, 2017a-h) on the disaster reside in are used in this GIS research scenario to clip the data by. The data used is a shapefile distributed by the U.S. Census Bureau representing all administrative boundaries on a county-level. The relevant metadata on this particular shapefile has been detailed in Appendix V. Places or counties not mentioned in the CNN articles are excluded from this shapefile by hand using the ArcMap software package, as specified in Appendix IV.1. This is done under the assumption that these areas have been unaffected by the disaster subject to this GIS research scenario.

[44] CNN (Cable News Network) is an American based news channel broadcasting both nationally and globally (CNN, 2017g).

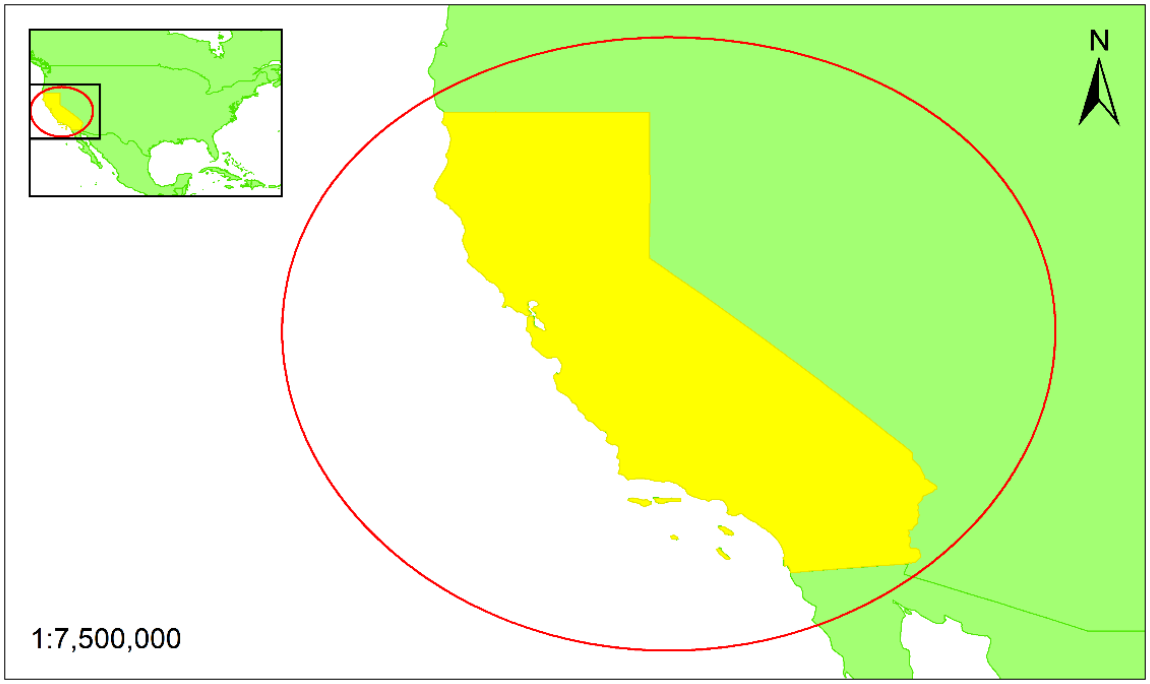

Figure 7.2: Bounding radius used in gathering Twitter data for disaster management research scenario

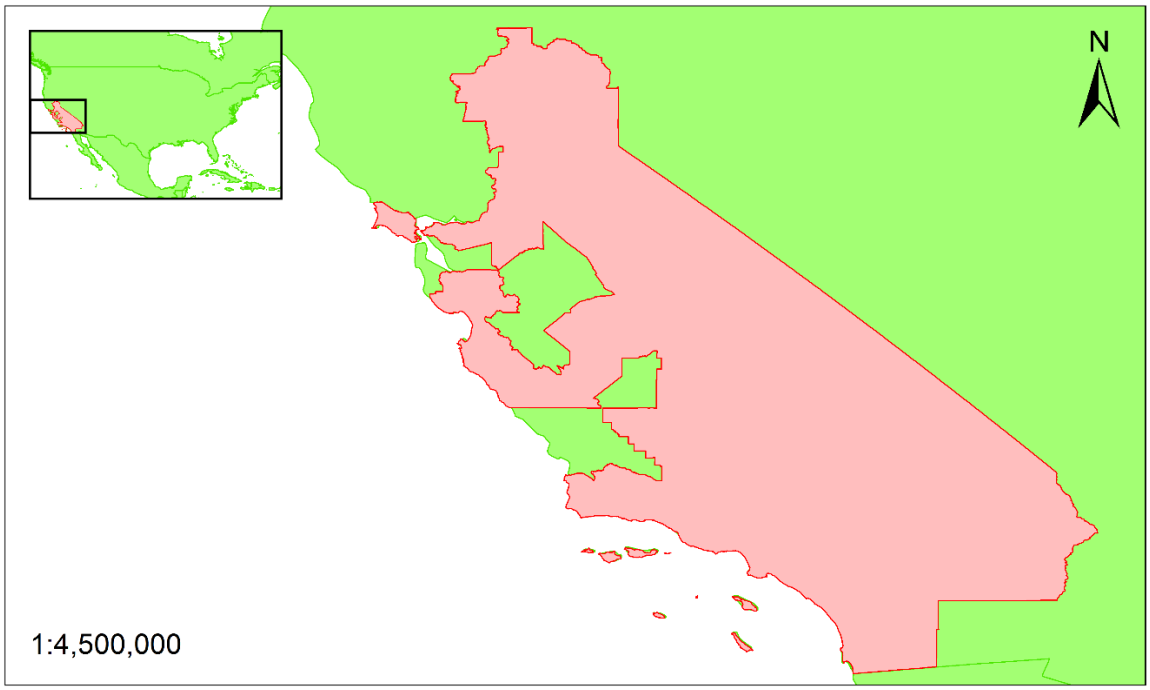

Figure 7.3: Catchment area used in disaster management research scenario

Step 3: Using the GIMs on the data sets: The GIMs will be used to redetermine the geolocation of the users part of the data set as gathered and pre-processed in step 1 and 2 respectively. The workflows of these GIMs will be explained in more detail later in paragraph 7.3. By comparing the original given user location and inferred user location, the increase in usability of the Twitter data by the GIMs can be determined according to the evaluation metrics that will be detailed later in paragraph 7.4.2.

Step 4: Evaluate data output: The resulting data set will be used to evaluate the GIMs and compare them among each other using the framework as will be presented later in paragraph 7.4. The output of the GIMs will be compared to unprocessed output of the Twitter API as well.

7.2.2 Health management research scenario

The second application domain the selected GIMs will be evaluated and compared by is a health management research scenario. In this application domain Twitter data is used in combination with GIS to assess health-related patterns in a certain area with the purpose to monitor or decrease the amount of health-related issues in that particular area, as previously defined in paragraph 4.4. In Table 7.2 below the preferred GIS research scenario parameters as defined earlier in paragraph 4.6 have been listed next to the parameters that will be used in the final thesis research.

Some differences are noticeable among the preferred and final GIS research scenario parameters for the health management modelling scenario as presented in the table below. These differences lie in the corpus size and period of gathering specifically. The reason for this difference is the strict time-limit at which the thesis research can be conducted, as previously mentioned earlier in paragraph 2.5. Originally 185 to 226 days (or approximately 6 to 7.5 months) were preferred at which Twitter data would be gathered based on the findings found through answering the sub question 4. This period is longer than the preferred amount of time available at which the thesis research can be conducted (approximately June 2017 as of writing). Therefore, a different period of gathering of 7 days (or one week) is set for multiple reasons. The first reason is that typically the Twitter API lets users only gather data up to 7 days old (Twitter, 2017k). If tweets are older extra measures have to be taken to work around these limitations, as explained previously in paragraph 5.3. When setting the period of gathering of 7 days there is no need to develop such measures. The second reason is that there is no indication that Twitter data gathered over a shorter period of time is less sufficient than data gathered over a longer period of time, as long as the preferred median number of tweets per day as previously detailed in Table 4.3 in paragraph 4.5 is reached. The third reason is that there is no interest in researching spatio-temporal health-related patterns within the GIS research scenario. For that reason, the fact that the period of gathering is shorter does not matter when making sure the results found through analysis are meaningful within this research scope.

Parameter	Preferred	Final
Application methodology	Event detection	Event detection
Application domain	Health management	Health management
Real-time	No	No
Additional sources	No	No
Corpus size	134000 to 164000 tweets	4000 to 5000 tweets
Period of gathering	185 to 226 days	7 days
Study area	Contiguous United States	Contiguous United States
Scale	National-level	National-level

Table 7.2: Preferred and final parameters for health management research scenario

As stated previously, the preferred corpus size is different from the final corpus size that will be used in the second part of the thesis research. The reason for this is that the corpus size correlates with the period of gathering, meaning that a shorter period of gathering automatically leads to a smaller corpus size. The size of the final corpus size as presented in the table above have been based on the amount of median number of tweets per day for the topic modelling domain as presented earlier in Table 4.3 in paragraph 4.5 for the same reason as previously described in paragraph 7.2.1 and calculated the same way as well.

Following the preferred GIS research scenario parameters as determined earlier in this paragraph a health management research scenario has been designed in which Twitter data and GIS will be used to define the spatial distribution of breast cancer occurrences among Twitter users in the contiguous United States for a certain week. The GIS research scenario was initially inspired by the works of Paul and Dredze (2011) who have determined the most-often mentioned health-related topics on Twitter among English tweets posted from May 2009 to October 2010 in the United States. They found that the terms “cancer” and especially “breast cancer” were among the top-3 most popular health-related topics Twitter users talked about in the (contiguous) United States in that period. Other researchers investigated the subject of breast cancer within the context of Twitter as well. Examples include Lee et al (2013), Sugawara et al (2012) and Himelboim and Han (2014). It has to be noted that the spatial component of the data used in these articles was not necessarily prominent. However, breast cancer occurrences and the spatial distribution of these occurrences might very well be a plausible GIS research scenario within the health management domain.

Therefore, the health management research scenario as presented here can be considered representative for the health management application domain overall. Research on the subject of breast cancer as specified above will be taken into account in the design of the health management research scenario. The workflow of this GIS research scenario will be the same one as used for the disaster management research scenario and illustrated in Figure 7.1 previously in paragraph 7.2.1. While the steps set for the health management scenario are more or less the same as the ones used for the disaster management research scenario, some differences exist among the parameters used in these steps. These differences will be detailed below and on the next page:

Step 1: Gathering data: Only tweets containing the key words "breast cancer" are gathered. A different bounding radius has been defined from within the Twitter data used in this GIS research scenario will be gathered compared to the disaster management research scenario detailed earlier in paragraph 7.2.1 as well. This bounding radius has been illustrated in Figure 7.4 below. Within this bounding radius the contiguous United States and surrounding areas are contained. The central point from which the radius is defined using the WGS84 coordinate system standards and has a latitude of 39.8 and longitude of approximately -97.4 with a diameter of 2600 kilometres. The bounding radius has been defined in the same way as described earlier for the bounding radius used in the disaster management scenario in paragraph 7.2.1. Using this bounding radius Twitter data outside the health management research scenario study area will be gathered as well. Twitter data that originates from outside the study area will be excluded from the data set using GIS as will be described later in this paragraph. The data itself has been gathered on April 5th, 2017.

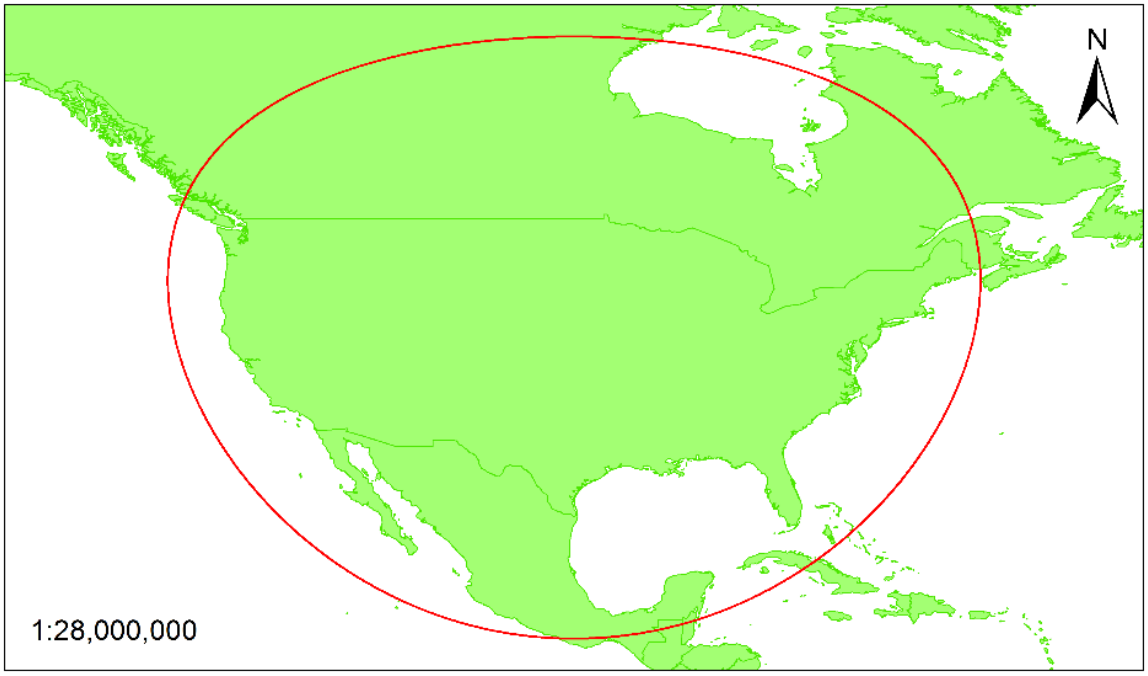

Figure 7.4: Bounding radius used in gathering Twitter data for health management and topic modelling research scenarios

Step 2: Pre-processing data: The administrative boundaries of the contiguous United States are used to clip the data by instead of disaster catchment areas. This catchment area has been visualised on the next page in Figure 7.5. This is done because this is the specified case study area for the health management research scenario as well. The data clipping is performed to a shapefile distributed by the U.S. Census Bureau representing all administrative boundaries on a state-level. The relevant metadata for this particular shapefile has been detailed in Appendix V. The state of Alaska, Hawaii and off-shore territories such as Puerto Rico are excluded from this shapefile by hand using the ArcMap software package, as specified in Appendix IV.1. This is done because these areas are not part of the contiguous United States and therefore not of interest to the health management research scenario.

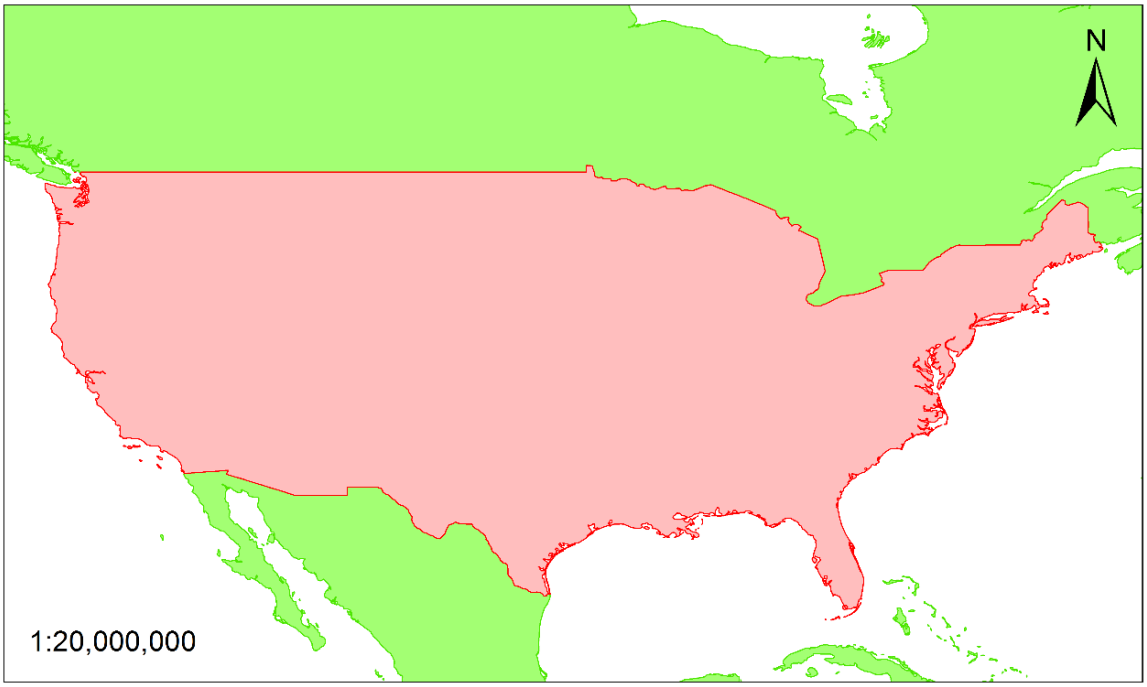

Figure 7.5: Catchment area used in health management and topic modelling research scenario

7.2.3 Topic modelling research scenario

The third application domain the selected GIMs will be evaluated and compared by is a topic modelling research scenario. In this application domain Twitter data is used to derive patterns of subjective thoughts on various subject such as politics, sports or societal issues as previously defined in paragraph 4.4. In Table 7.3 below the preferred GIS research scenario parameters as defined earlier in paragraph 4.6 have been listed next to the parameters that will be used in the final thesis research.

Parameter	Preferred	Final
Application methodology	Event detection	Event detection
Application domain	Topic modelling	Topic modelling
Real-time	No	No
Additional sources	No	No
Corpus size	410000 to 502000 tweets	12000 to 14000 tweets
Period of gathering	124 to 152 days	7 days
Study area	Contiguous United States	Contiguous United States
Scale	National-level	National-level

Table 7.3: Preferred and final parameters for topic modelling research scenario

Some differences are noticeable among the preferred and final GIS research scenario parameters for the topic modelling scenario as presented in the table above. These differences and the reasons behind these differences are similar to the ones described for the health management scenario earlier in paragraph 7.2.2. These differences lie in the corpus size and period of gathering specifically, for the same reason as described for the health management scenario as well. Originally 124 to 152 days (or approximately 4 to 5 months) were preferred at which Twitter data would be gathered for the topic modelling research scenario.
This period is longer than the amount of time available at which the thesis research can be conducted. Therefore, a different period of gathering of 7 days (or one week) has been set for the same reasons as given for the health management scenario. The preferred corpus size is different from the final corpus size due the fact that it correlates with the period of gathering. The size of the final corpus size as presented in the table above have been based on the amount of median number of tweets per day for the topic modelling domain as presented earlier in Table 4.3 in paragraph 4.5 for the same reason as previously described in paragraph 7.2.1 and calculated the same way as well.

From the final parameters as set in the table on the previous page a topic modelling research scenario has been developed in which Twitter data and GIS will be used to determine the spatial distribution of popular hashtags used by Twitter users in the contiguous United States within a certain week. This GIS research scenario has been inspired by research on the spatial distribution of certain topics on Twitter. Kamath et al (2012, 2013a-b) have conducted plenty of research on the subject in which they researched the spatio-temporal dynamics of memes (online inside jokes) through Twitter data for example. Lansley and Longley (2016) have done similar research in which they derived the most popular hashtags used in the Greater London area within a certain time period and derived spatial patterns from them. The workflow of this GIS research scenario has been illustrated in Figure 7.1 previously in paragraph 7.2.1. The same steps of the health management research scenario are more or less followed for the topic modelling management research scenario. The only difference is that different keywords are used within the keyword query in step 1, as will be described now.

Only tweets containing popular hashtags are gathered since tweets using these hashtags are of interest to this GIS research scenario exclusively. The same bounding radius as defined for the health management scenario is used for the topic modelling scenario as well, as previously detailed in paragraph 7.2.2. The hashtags used in this GIS research scenario have been determined by using the search query "hashtag AND twitter" in Google and narrow down the results to make sure they were posted in the last 7 days. Hashtags had to be mentioned by (multiple) news outlets to be valid for this GIS research scenario. The initial query has been used on 29th March 2017 and resulted in the following hashtags:

- **#GOPDnD**: The hashtag, as introduced by One Shot Podcast co-host James D'Amato, is used in tweets depicting the way in which the Republic Party tried to get the TrumpCare health plan through the House of Representatives as a Dungeons and Dragons board game (Daily Kos, 2017; Slate, 2017; The Mary Sue, 2017; Washington Post, 2017; Dorkly, 2017; Gizmodo, 2017).

- **#BigThighTwitter**: The hashtag was used to initiate body positivity over girls and women with big thighs (Refinery29, 2017; Popsugar, 2017; Seventeen, 2017; TeenVogue, 2017; Yahoo! Beauty, 2017; Huffington Post, 2017; Elite Daily, 2017).

It has to be noted that some hashtags initiated as a result of the 2017 Westminster terrorist attack [45]have been considered as well. Examples are "#prayforlondon", "#wearenotafraid" and "#prayformuslimban" (Metro, 2017; USA Today, 2017; ITV, 2017; news.com.au, 2017). The reason why these have not been included into the topic modelling research scenario is because these hashtags refer to events outside the (contiguous) United States and their validity for implementation in this thesis research can therefore be questioned.
The data itself has been gathered on 29th March 2017.

7.2.4 Technical framework: data gathering and pre-processing

With the GIS research scenario designs being detailed and argued in paragraph 7.2.1 to 7.2.3 earlier in this chapter, the technical framework used within these GIS research scenarios can be detailed now.
Attention will be given specifically to the way in which the data used in this thesis research will be gathered and pre-processed. This will be done according to the steps as illustrated in Figure 7.6 below and described below and on the next page as well:

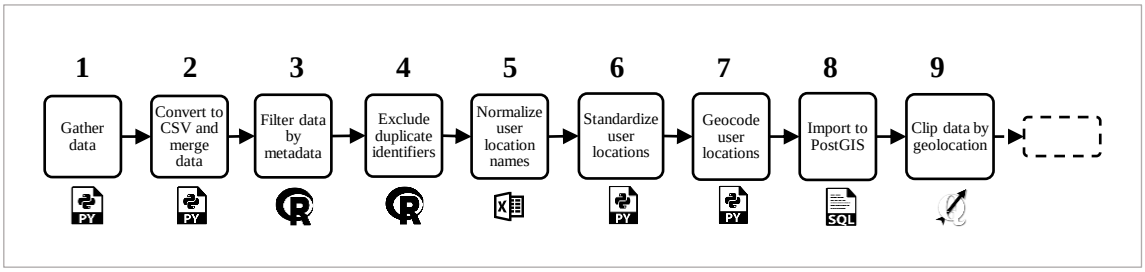

Figure 7.6: Data gathering and pre-processing workflow

[45] A terrorist attack that took place on 22nd March 2017 in which an Islamic extremist drove into pedestrians on the Westminster Bridge in London, leaving 6 fatal casualties and dozens of people wounded.

Step 1: Gathering Twitter data: The Twitter data used in this thesis research will be gathered using a Python script that enables the conductor of the thesis research to connect with the Twitter API and gather historic data in a JSON file-format. This script is run within the Canopy software package as detailed in Appendix IV.1 and has been detailed in Appendix III.1. Within the script certain parameters can be set when gathering data, which are the following:

- *Search phrase*: By setting this parameter, the user of the script can set specific keywords that need to be part of the tweets texts for these tweets to be gathered. For each keyword, a separate JSON file will be created, containing metadata on tweets using these specific keywords posted within the other parameters defined by the user of the script. The exact keywords used for each GIS research scenario have been detailed earlier in paragraph 7.2.1 to 7.2.3 respectively. It has to be noted that only tweets containing the exact keyword(s) as specified will be gathered using this script. When a tweet contains the hashtag "#ilikedogs" but the keyword has been set to "dogs" this specific tweet will not be part of the final data sets for example.

- *Time limit*: The maximum number of hours the script will gather Twitter data. This parameter is especially useful when gathering tweets with popular keywords, a process that can potentially take days or even weeks to complete. In this thesis research the maximum number of hours has been set to 24. This is done to make sure that all tweets part of the data sets within a period of 7 days to fulfil the requirements for the GIS research scenarios as previously detailed in paragraph 7.2.1 to 7.2.3 respectively.

- *Date range*: The minimum and maximum age of the tweets in days that will be looked for relative to the time at which the script is run. The maximum value for this parameter is 7 days as a result of the rate limitations set up by the Twitter corporation, previously detailed in paragraph 5.3. For all GIS research scenarios, the minimum amount of days has been set to 0 while the maximum amount of days has been set to 7 to fulfil the requirements for the GIS research scenarios as previously detailed in paragraph 7.2.1 to 7.2.3 respectively.

- *Bounding radius*: A bounding radius can be set to make sure only tweets within a certain area are gathered. For this bounding radius the longitude, latitude and diameter in kilometres can be set. The radius is defined using the WGS84 coordinate system standards, the same system used for the GPS coordinates attached as metadata to tweets by the geotagging option if it is enabled by the user (Twitter, 2015). The bounding radiuses used in each GIS research scenario has been detailed previously in paragraph 7.2.1 to 7.2.3 respectively.

Step 2: Converting to CSV and merging files: The pre-processed data will be converted from JSON-format to CSV-format using two scripts as detailed in Appendix III.2 and III.3 respectively. The scripts have been detailed in Appendix III.2 and Appendix III.3 respectively. These scripts are run within the Canopy software package as detailed in Appendix IV.1. The data is converted to this specific file format because CSV-formatted data is found to be more easily to manage and query than JSON-formatted data by the conductor of the thesis research. The tweet metadata attributes are filtered automatically in the process, meaning that after this step exclusively metadata attributes needed for this thesis research are part of the data sets. The conversion of the JSON files and merging of the CSV files will be done according to the two steps as detailed below and on the next page:

- *Step 1*: First, each separate JSON file with tweets containing a specific keyword will be converted to a CSV file format using the script as detailed in Appendix III.2. During this process the id_str(Users), lang(Users), lang(Tweets) and location(Users) metadata attributes will be the only attributes that will be included in the CSV files (Twitter, 2017n-o). The id_str(Users) metadata attribute will be used later within the GIMs themselves to derive user-specific metadata and gather tweets posted by these users. The identifier metadata attribute is chosen over the screen_name(Users) metadata attribute of the users part of the data sets for multiple reasons. The first reason is to ensure the anonymity of the Twitter users part of these data sets. The second reason is that identifiers in string format specifically are less prone to give errors than identifiers in an integer format (Twitter, 2017o). The location(Users) metadata attribute will mainly serve as the ground-truth of this thesis research described later in paragraph 7.4.1. Any users that have not specified a geolocation on their

profile will be excluded from the final data sets. This is done because both the original user location and inferred user location are needed to evaluate and compare the GIMs among each other.
The remaining metadata attributes will be used to determine the validity of each user to be part of the data sets and make sure they fit the research scope as defined earlier in paragraph 2.4.

- *Step 2*: Secondly, the resulting CSV files will be merged into one CSV file using a script as detailed in Appendix III.3.

Step 3: Filtering data based on metadata attributes: The data will be filtered based on certain metadata attributes using self-written R code[46]. The code used to do this has been detailed in Appendix III.4.
The code was written and run within the RStudio software package as detailed in Appendix IV.1.
The data sets are filtered to fit the research scope as defined earlier in paragraph 2.4. The data sets will be filtered according to the metadata attributes as listed below:

- *Language*: Any tweets not written in the English language will be excluded from the data sets. Any users who have not set English as their default language will be excluded as well. This will be done according to the lang(Users) and lang(Tweets) metadata attributes Twitter, 2017n-o).

- *Location*: Users that have not specified a user location will be excluded from the data sets. This will be done according to the location(Users) metadata field in the CSV files (Twitter, 2017o).
The availability of user location information is necessary to be able to verify and evaluate the output of the GIMs as will be explained in more detail later in paragraph 7.4.

Step 4: Excluding duplicates: Following the three steps as detailed above and on the previous pages, data sets will be created containing the identifiers of Twitter user part of the data sets in a string format accompanied with the user-specified user location in the same format. The problem is that these data sets will contain duplicates of identifiers because these identifiers have been gathered through tweets part of the original data sets. This means that if certain tweets in the original data sets have been posted by the same user, this user is also part of the data sets pre-processed so far multiple times. Therefore, these duplicates have to be deleted to make sure that the final data sets contain unique identifiers exclusively. This is done to reduce the processing time of the rest of the analysis to be conducted. With fewer entries to be processed, the time needed will automatically be reduced as well. The duplicate users will be excluded from the data sets using a self-written R code as specified in Appendix III.4. The code was written and run within the RStudio software package as detailed in Appendix IV.1. After using this code, data sets are created containing all unique identifiers part of the original data sets and the user-specified user location specified for each identifier.

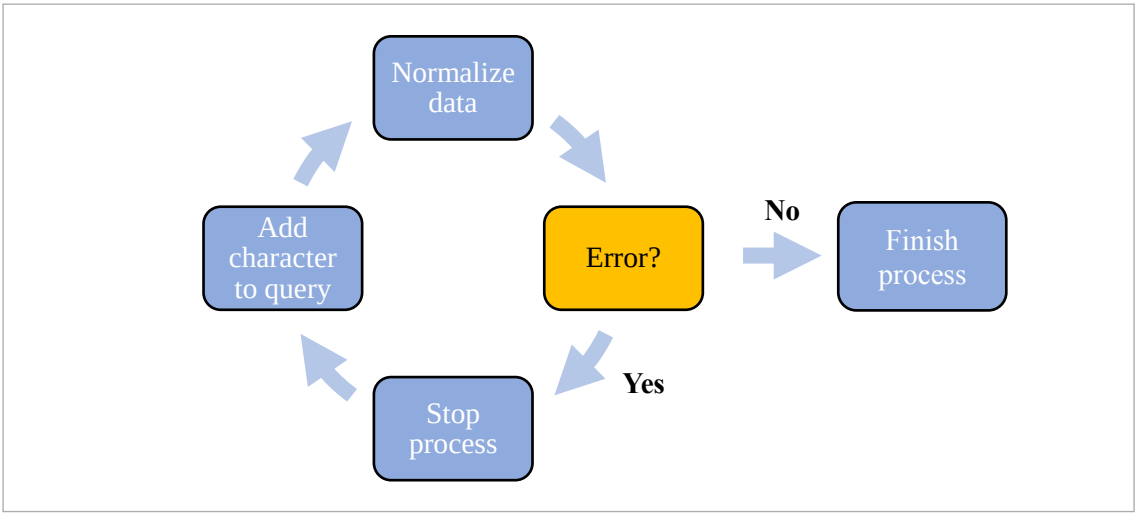

Figure 7.7: Normalization process for Twitter data sets used in thesis research

[46] It has to be noted that this programming language works with just coding rather than scripts.

Step 5: Normalizing[47] user-specified user locations names: To prevent the script used to perform step 6 and 7 to run into any errors when performing these respective steps the user-specified user locations names will be normalized using the KuTools for Excel software package as specified in Appendix IV.2 within the Microsoft Excel software package as specified in Appendix IV.1. Any characters that cannot be interpreted by Python will either replaced by a similar alternative (é → e) or completely deleted from the row entry when no similar alternative is available (º, †, ›› et cetera). This is done by converting the user locations names from UTF-8 Unicode[48] to characters similar to the ASCII-format. The KuTools Excel software package provides a wide variety of alternatives to characters that might cause errors but does not include all characters that are used within Twitter user profiles. Therefore, additional unwanted characters are added to the package using a cyclic-iterative process in which characters causing errors when performing step 6 and 7 are manually added to the KuTools Excel package by the conductor of the thesis research. This process has been illustrated in Figure 7.7 on the previous page.

Step 6: Standardizing user locations: The user-specified user locations are standardized[49] using the GeoPy Python package as detailed in Appendix IV.2 and put into a separate column in the CSV files. The GeoPy Python package provides several gazetteers to geocode possible toponyms by services such as OpenStreetMap[50], ESRI ArcGIS[51], Google Maps[52] and many others (GeoPy, 2017). In the case of this thesis research, the OpenStreetMap gazetteer has been used because access to this API is met with relatively few rate limitations and access is very straight-forward compared to the alternatives provided by ESRI and Google for example. This is important to consider due the fact that thousands of geolocations need to be geocoded in this thesis research. Whenever a user-specified user location cannot be geocoded, this entry is excluded from the final data sets using the R code as detailed in Appendix III.6. This specific code was written and run within the RStudio software package as detailed in Appendix IV.1. The standardization will be done using a self-written Python script as specified in Appendix III.5. This script is run within the Canopy software package as detailed in Appendix IV.1. The standardization is done to ease the comparison of the user-specified user locations with the inferred user locations specifically. When a user has for example listed "L.A." as its geolocation and "Los Angeles, CA" has been found as the inferred user location, there is a slight chance that while these two place names refer to the same geolocation, they are interpreted by the Python script used as two different geolocations. Therefore, standardization takes place to prevent this from happening. User locations are standardized per 1000 entries in the data sets as a safety measure to prevent rerunning Python scripts over and over again due to hard to avoid errors occurring and automatically stopping the script from completing its run. One user location is standardized per 1.1 second to meet the OpenStreetMap geocoding API requirements (OSM Foundation, 2017), meaning that it takes approximately just over 18 minutes to standardize 1000 entries in the data sets. The data sets are divided in smaller parts containing 1000 data entries through the script as detailed in Appendix III.7. This script is run within the Canopy software package as detailed in Appendix IV.1.

Step 7: Adding GPS coordinates to data entries: The user-specified user locations will be geocoded and attached a GPS coordinate to representative to the centroid of the given geolocation. The GPS coordinates follow the WGS84 coordinate system standards because it is the same one as used in the Twitter metadata (Twitter, 2015). The user locations are geocoded using the self-written Python script as detailed in Appendix III.5 and is the same one used for step 6. The script is run within the Canopy software package as detailed in Appendix IV.1. Step 6 and 7 are performed using one script to decrease the amount of time needed to pre-process the data sets. The GPS coordinates are added to be able to filter the data sets in such way that only users who have specified a user location within the contiguous United States are part of the data sets. This is done to fit the research scope as defined earlier in paragraph 2.4. The way in which these GPS coordinates will be used to filter the data further will be described later in this paragraph. Since step 7 is executed within the same script as used for step 6, GPS coordinates are added per 1000 data set entries at the time and taking 1.1 second per entry for the reasons explained above as well. When all parts of the data sets have been standardized and added GPS coordinates to these will be merged again using the script as detailed in Appendix III.3 and mentioned earlier for step 2.

[47] When normalizing, a certain value is made more "normal" according to a certain standard.
[48] This unicode is the main unicode used on the web currently.
[49] When standardizing data, (nominal) values using different forms of notation are converted to new collectively used standard.
[50] OpenStreetMap is an open-source collaborative project which main aim is to create a freely available editable map of the world (OpenStreetMap, 2017).
[51] ESRI ArcGIS is a popular proprietary GIS software package (ESRI, 2017a).
[52] Google Maps is a web mapping service (Google Maps, 2017).

Step 8: Importing and managing in SQL-database: The data sets will be imported into a PostGIS database using the pgAdmin III software package as detailed in Appendix IV.1 and given geometry to using the GUI[53] provided by the software itself and self-written SQL code. The SQL-code used to perform the actions as described above has been detailed in Appendix III.8. This type of database has been chosen because it is considered very suitable to store CSV-formatted geographical data in by the conductor of the thesis research. By giving the data geometry the data becomes spatial and can be mapped accurately or implemented in a GIS for example. The data is specifically imported into a database to enable faster and easier processing of the data when mapping and performing analysis given that the data sets contain thousands of entries.

Step 9: Clipping[54] data by area: The data is clipped according to certain catchment areas set, either being administratively-based or determined by the conductor of the thesis research himself. This is done using the QGIS software package as detailed in Appendix IV.1. This software package is used specifically because this software program enables users to map their data from PostGIS databases in a relatively straight-forward way compared to alternatives such as ArcMap. The geographical boundaries by which the data sets will be clipped by differs per GIS research scenario and have been detailed earlier in paragraph 7.2.1 to 7.2.3 respectively. It has been noted earlier in this paragraph that the user-specified user locations will be clipped exclusively. The reason for this is to ensure that only users that actually live in the contiguous United States are part of the data sets. The inferred user locations are not clipped because these geolocations might represent errors as a result of using the GIMs. These errors are of interest when evaluating and comparing the GIMs among each other and should therefore be kept within the data sets.

7.3 GIMs' workflows

The central aim of this thesis research is to evaluate a set of relevant GIMs within multiple GIS research scenarios considered relevant as well. With the GIS research scenarios being defined and argued previously in paragraph 7.2 the same will be done for the GIMs in paragraph 7.3.1 to 7.3.3. For these GIMs the following characteristics will be defined, detailed and argued where needed:

- **Metadata attributes used**: The metadata attributes used in the GIMs (and the reason why) will be listed and argued where needed. The way in which these metadata attributes are used will not be explained in this part.

- **Workflow**: The steps taken to infer the geolocations of Twitter users will be described and detailed. Choices made will be argued with (academic) literature where needed. Figures will be used as well to clarify the workflow's structure where needed. The way in which the metadata attributes are used will be explained in this part as well.

- **Parameters**: Within the GIMs certain parameters have been defined which can be altered when inferring the geolocations of Twitter users by the will of the user of the GIM. An example is the minimal number of tweets, followers and friends needed to infer a valid and accurate user location. Defining these parameters is important especially when defining the to-be-performed sensitivity analysis that will be detailed later in paragraph 7.4.3.

Originally four GIMs had been defined earlier in paragraph 6.6 to be researched in this thesis research. The choice has been made to include GIMs used to infer user locations exclusively and exclude the one used to infer the geolocation of a tweet through the user's content. The main reason why this choice has been made is that while it was possible to infer the geolocation of a tweet through this method, the output of this method was considered too inaccurate to be worth evaluating in this thesis research. While it is definitely possible to create such GIM, a lot of time and effort is needed to be able to create an accurate iterative. Within the current time-limit defined for this thesis research as defined earlier in paragraph 2.5 this was not considered possible by the conductor of this thesis research. Given that the majority of the GIMs as presented in academic writing is focussed on inferring users as well, as previously detailed in Figure 6.6 in paragraph 6.6, this choice does not affect the thesis research's validity to an extent that it becomes

[53] A GUI (Graphical User Interface) is a user interface that lets users interact with the software through visuals primarily instead of text.
[54] When clipping data, data entries that overlay with a specified geographical area are extracted.

unrepresentative for the majority of the academic framework concerning GIMs. From now on the focus in this thesis research will therefore lie on GIMs that infer the geolocation of Twitter users exclusively.

7.3.1 Content-user method

When using the content-user method, tweet content and certain profile information is used to infer the geolocation of a user when not being given by the user themselves. Similar GIMs have been proposed by Ahmed et al (2013), Xie et al (2014) and Mahmud et al (2014) for example. The inference of the user location will be done according to the set of metadata attributes as detailed in Table 7.4 below.
These metadata attributes will be implemented according to the workflow illustrated in Figure 7.8 below as well. All metadata attributes will be gathered using the Tweepy Python package as detailed in Appendix IV.2 in combination with the official Twitter API using the id_str(Users) metadata attribute (Twitter, 2017o). This is done to be sure the latest description, amount of statuses posted by the user and tweets posted are taken into consideration when inferring the user's geolocation. The GIM has been written using the Python programming language with the script being run within the Canopy software package as detailed in Appendix IV.1. The script containing the content-user method itself has been detailed in Appendix III.9.

Object	Field name	Type	Description
Users	id_str	String	String representation of the user's identifier.
	description	String	The user-defined description for their account.
	statuses_count	Integer	The number of tweets issued by the user.
Tweets	text	String	The actual text of a status update.
	place	Object	Various info on the place associated with the tweet.

Table 7.4: Metadata attributes used in content-user method (Twitter, 2017n-o)

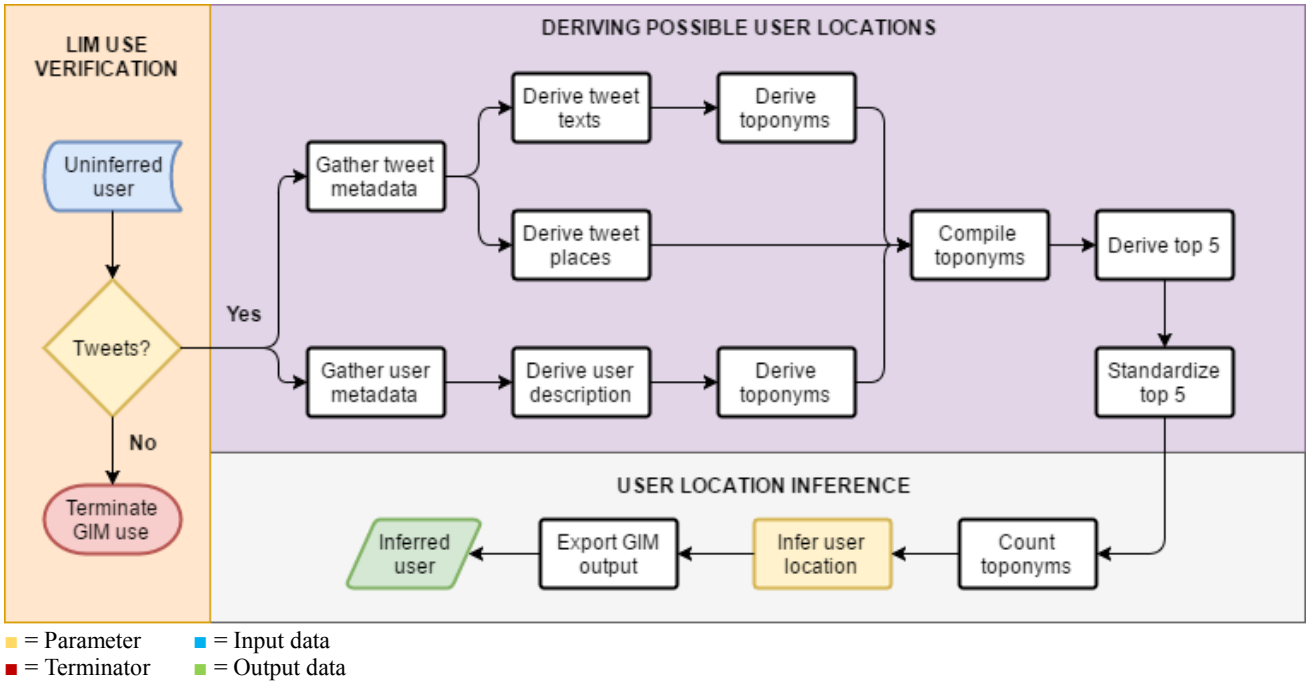

Figure 7.8: Content-user method workflow

Step 1: GIM use verification: The validity of the GIM's usage will be examined first based on the amount of tweets the user of which the geolocation will be inferred has posted. If the user has posted a certain minimal number of tweets, this user is considered to be sufficient to use this GIM on. If this is not the case the GIM will not be used on that specific user because it is then assumed that not enough content can be derived to infer the user's geolocation. The minimum number of tweets needed can be set according to the will of the one using the GIM and is one of the parameters that will be described later in this paragraph.
The number of tweets that have been posted by the user is derived from the user's metadata using the status_count(Users) metadata attribute (Twitter, 2017o). Whenever the metadata of a user cannot be gathered through the id_str(Users) metadata attribute, this user is skipped and the geolocation will not be inferred.

Possible causes might be that the user has deleted its account or has protected its tweets in the time between gathering the tweets and actually analysing the tweets and its metadata can therefore not be accessed.

Step 2: Deriving possible user locations: Three sources of content will be used to derive the user's geolocation when the GIM use has been considered to be sufficient in the previous step. These are the description(Users), text(Tweets) and place(Tweets) metadata attributes and have been described earlier in Table 7.4 on the previous page. These specific metadata attributes are used because all of these attributes possibly contain toponyms that are representative for the user's geolocation or its whereabouts. The three metadata attributes as described above will be gathered using the id_str(Users) metadata attribute. The string-format identifier is preferred used over the integer-formatted alternative because the latter is prone to giving errors as previously explained in paragraph 7.2.4. Due to rate limitations set up by Twitter it is only possible to gather metadata attributes for the latest 200 tweets by the respective user (Twitter, 2017q). It is assumed that this number of tweets will generate enough content to infer the geolocation of the user by.
Toponyms will be derived from the description(Users) and text(Tweets) metadata attributes using the GeoText Python package as detailed in Appendix IV.2. With this package city names (among other type of toponyms) can be derived from any text source using natural language processing through regular expression, based on the official GeoText library (GeoText, 2017). Other Python packages with similar functionalities exist as well, such as GeograPy (GeograPy, 2017). The reason why GeoText has been chosen over the other alternatives is that is based on regular expression rather than tokenization and is therefore faster. When testing this specific GIM and testing differences in performance between these packages it turned out that the performance of both packages is very similar. Toponyms for places(Tweets) do not have to be derived through this package because this metadata attribute already has a sufficient geolocation indicator based on FourSquare (Twitter, 2017i).

The toponyms derived from the three sources described above will be standardized using the GeoPy Python package as detailed in Appendix IV.2. This is done in a similar way as described earlier in paragraph 7.2.4 for step 6. The standardization will take place for two reasons. The first reason is that the output from the GeoText Python package is rather simplistic and not very detailed. By standardizing through using the GeoPy Python package the toponyms found will be added detail to by adding the state and country the found toponym is in for example. The second reason is that by standardizing toponyms it becomes easier to "add up" the amount of times a toponym is mentioned in the user' content. The toponym formatting used in the GeoText Python package through which toponyms within user descriptions and tweets are derived is different from the formatting used for the FourSquare toponyms. Therefore, all toponyms are standardized to make sure that two toponyms representing the same place are not accidentally seen as being two different geolocations because of the fact that they are differently formatted. When a user has for example listed "L.A." as its geolocation and "Los Angeles, CA" has been found as the inferred user location, there is a slight change that while these two place names refer to the same geolocation they are interpreted by the Python script used as two different geolocations as previously explained in paragraph 7.2.4. Only the five toponyms that are most frequent among the user's content will be standardized. Given that it takes just over 1 second to standardize the toponyms because of the OpenStreetMap's API rate limitations as mentioned earlier in paragraph 7.2.4., it would otherwise take too much time to standardize toponyms that occur only a few times in the user's content and will not be considered a possible user location anyway. The toponyms in the steps found will finally be compiled in a data frame.

Step 3: User location inference: The amount of times each toponym occurs within the compiled data frame will be determined first. The user location will then be inferred based on the occurrence of these particular toponyms. This can be done using different methods. The most straight-forward method is to simply assume the most frequently mentioned toponym is the user's geolocation. Other factors such as a minimal number of occurrences can be set as well to increase the inferred geolocation's validity. The multiple options that can be used to infer a user's geolocation in this step will be detailed later in this paragraph given that it is a parameter. The two most-often occurring toponyms will finally be exported to the data set the user is part of to which the process will be repeated for the next user within the data set. The reason for this is that the most often found toponym might be the result of errorous found toponyms using the NLP package. An example is "Clinton county", which is found if someone tweets about Hillary Clinton. The error distances between the standardized user-specified user location and top 2 inferred user locations, the age of the inferred user location observations, the number of commas in the inferred user location's names and the time needed to

infer the user locations are exported into separate columns as well. When all user locations are inferred the script used will automatically stop.

When using the content-user method there are two types of parameters that can be set:

- **Minimal number of tweets**: The validity of the user location inferred correlates with the amount of content that is used to infer this geolocation. With more content to be analysed more toponyms can be found by which the user location can be determined. Given that this is a content-based approach this means that more tweets gathered will automatically lead to more content which results in more toponyms found which means more validity of found results and conclusions, but not necessarily completeness of data.

- **Minimal toponym count needed for inference**: The validity of the user location inferred correlates with the total amount toponym found in the user's content. Also, the most straight-forward method is to simply assume the most-often mentioned toponym is the user location as previously explained. There are situations where the validity of this method can be questioned however. These are the following:

    - Very few occurrences of the same toponym are found. While in theory the inferred user location may be right this way the validity of this outcome is difficult to determine due the fact that only one or two sightings are the basis of this conclusion.

    - Two or multiple toponyms occur approximately the same amount of times in the content. When this happens, multiple geolocations can be considered as a valid user location. Given that the scope of this thesis research is to find one geolocation additional steps have to be taken to select the correct user location from this set of geolocations.

    - Two or multiple toponyms occur exactly the same amount of times in the content. This is problematic due the same reasons as explained in the previous situation.

To determine the values for the parameters as mentioned above used in this particular GIM, academic literature as collected to answer sub question 3 presenting content-based GIMs inferring user locations have been referred to. Cheng et al (2013) have made sure that the tweets part of their data set are posted by users with more than 1000 status updates. They performed a sensitivity analysis as well in which they inferred user locations based on 10, 100 and 1000 tweets containing geolocation indicators. Krishnamurthy et al (2015) used a similar approach given that they used the data set as compiled by Cheng et al though they not perform a sensitivity analysis. Han et al (2014) only inferred the geolocation of users if these users have at least posted 10 tweets containing geolocation indicators. Hecht et al (2011) used a similar approach. Mahmud et al (2014) only collected the latest 200 tweets from the users in their data set. Xie et al (2014) did not specify a minimal number of either tweets or geolocation mentions within those tweets necessary to infer the geolocation of the user.

According to the academic literature as detailed above it has been determined that for the geolocation to be inferred users should have posted at least 10 tweets. This has been based on the works of Han et al (2014), who determined that within the user's content at least a pool of 10 toponyms should be available for the user's geolocation to be inferred. This way it is made sure that there is enough content available to infer users using the method as presented in this paragraph. When two toponyms occur exactly the same amount of times within the user's content it is assumed that the user's geolocation cannot be inferred. In other cases, the toponym mentioned most often in the user's content is assumed to be the user's actual geolocation.

7.3.2 Network-user method

When using the network-user method, a user's geolocation is determined by its social network on Twitter. Similar GIMs have been proposed by Apreleva & Cantarero (2015), Chen et al (2016) and McGee et al (2014) for example. The user locations will be inferred according to a certain set of metadata attributes, which have been detailed in Table 7.5 on the next page. This metadata will then be used according to the

workflow illustrated in Figure 7.9 below as well. These parameters will be detailed later in this paragraph. The script containing the network-user method itself has been detailed in Appendix III.10.

Object	Field name	Type	Description
Users	id_str	String	String representation of the user's identifier.
	followers_count	Integer	Number of followers of the user.
	friends_count	Integer	Amount of accounts the user follows.
	location	String	The user-specified user location as defined by the user.

Table 7.5: Metadata used in network-user method (Twitter, 2017o)

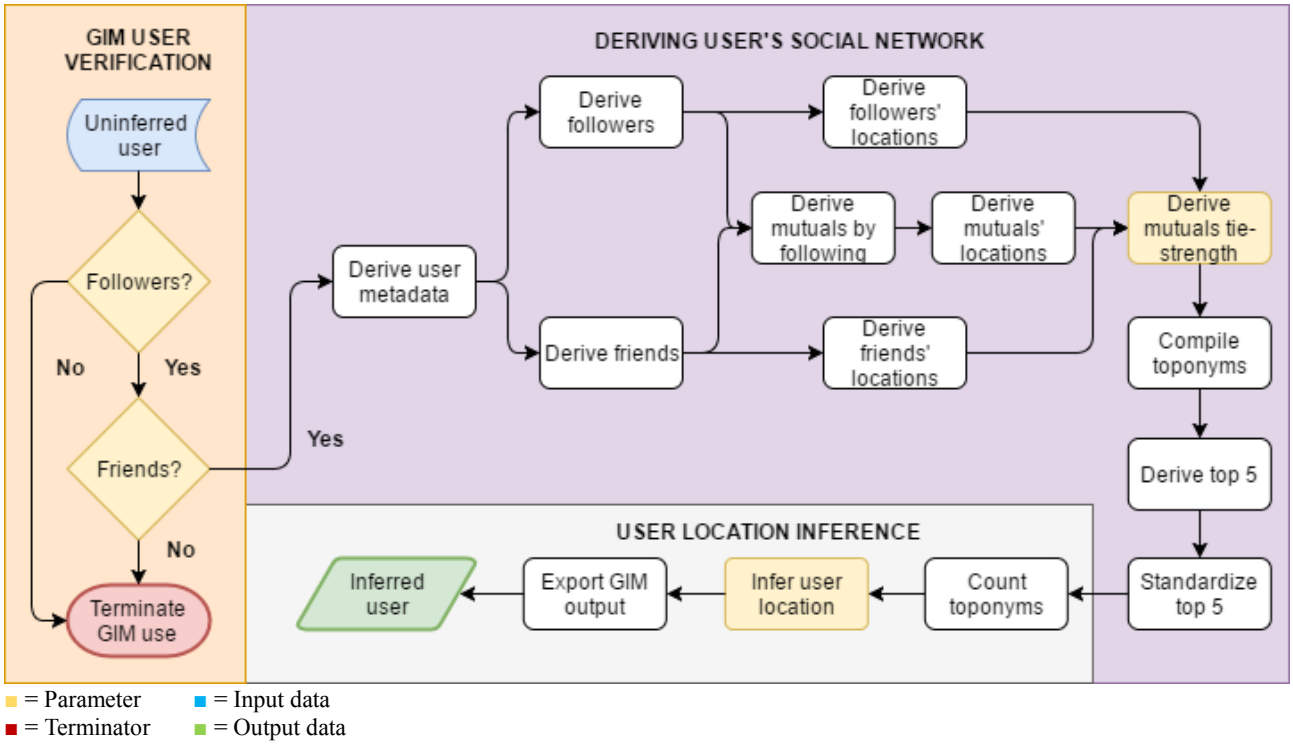

Figure 7.9: Network-user method

Step 1: GIM use verification: Uninferred users will be evaluated first based on the number of followers and friends they have before the GIM is actually used to infer their geolocations. If they have a certain minimal number of friends or followers they are considered to be sufficient to use this GIM on. If this is not the case the GIM will not be used on that specific user because it is then assumed that not enough mutual connections can be derived to infer the user's location. The minimum number of followers and friends needed can be set according to the will of the one using the GIM and is one of the parameters that will be described later in this paragraph.

Step 2: Deriving possible user locations: Two sources of content will then be used to derive the user location. Through these sources of content three types of mutual connections will be derived. These are the following:

- **Connection by mutual following**: This type of connection is considered mutual if the followers or friends of the user which geolocation will be inferred is following back these people. Whether this is the case will be determined using the Followers(Users) and Friends(Users) metadata attributes.
  If users are among both lists the user from which the geolocation will be inferred and the follower or friends are connected mutually by following.

- **Connection by singular following**: This type of connection is considered singular if the user to be inferred follows a certain other Twitter user but is not followed back by this user. Whether this is the case will be determined using the Followers(Users) and Friends(Users) metadata attributes. If users

are not among both lists the user from which the geolocation will be inferred and the follower are considered to be connected singular instead of mutual.

- **Connection by singular friendship**: This type of connection is considered singular if the user to be inferred is followed by a certain other Twitter user but does not follow back this user. Whether this is the case will be determined using the Followers(Users) and Friends(Users) metadata attributes. If users are not among both lists the user from which the geolocation will be inferred and the user the to-be-inferred user is followed are considered to be connected singular instead of mutual.

The two metadata sources as described on the previous page will be gathered using the id_str (Users) metadata attribute. From each of these two sources users that have had some kind of mutual connection with the subject user will be gathered and put into a data frame. For each unique user in this data frame the geolocation will be derived. These geolocations will be weighted according to the tie-strength of each connection. These weights can be set to the will of the user of the GIM and, given that it is a parameter, will be detailed later in this paragraph.

Step 3: User location inference: The user location will be inferred according to the same method as described earlier in paragraph 7.3.1 for the content-user method. Additionally, other statistics described for the content-user method will be exported as well. When all user locations are inferred the script used will automatically stop.

When using this GIM there are four types of parameters that can be set:

- **Minimal number of followers/friends***:* To be able to infer a user's geolocation through the network-user method it has to be made sure that a user actually has a social network on Twitter. The number of followers and friends is a good indicator to measure this. With more mutual connections that can be derived a more accurate estimation can be made of the user's geolocation. Depending on the preferred validity one can set this parameter either high or low.

- **Weights of connections***:* Connections are weighted because they differ in strength. Users that follow each other back are assumed to have a stronger connection than users that do not follow each other back for example. Therefore, the geolocations of each connection found are weighted to take the differences in tie-strength into account when inferring the user location. Different weights can be used based on the preferences of the one using the GIM.

- **Minimal toponym count needed for inference**: This parameter has previously been explained in paragraph 7.3.1.

A minimal number of followers or friends needed to infer the geolocation of Twitter users is not explicitly mentioned in any of the articles within the selection for sub question 3 presenting network-user GIMs. McGee et al (2013) and Yamaguchi et al (2013) mention that both a lack and surplus of mutual connections can lead to inaccurate geolocation inference, however. A lack of mutual connections will lead to a lack of toponyms that can be derived from the user's mutual connections geolocations. This lack of content will then negatively influence the validity of the inferred user locations. The reasoning behind why a surplus of mutual connections leads to inaccurate geolocation inference is that it is impossible for a Twitter user to know all his thousands of followers and friends in real life. In that situation, the assumption that mutual connections are people that know each other in real life as well (and live near each other) would be irresponsible to make. Concluding, while there is not a minimum and maximum number of followers or friends needed to infer the geolocation through social network mentioned in the scientific literature on the subject a certain lack or surplus of mutual connections will directly influence the validity of the inferred user locations.

The minimal of toponyms needed to infer a geolocation is set the same as for the content-user based method, being 10, for the same reasons as mentioned earlier in paragraph 7.3.1 as well. While no minimal number of followers or friends needed to infer the geolocation of Twitter users has not been defined in any of the scientific literature on the subject this does not mean that no minimum amount for both metadata attributes has been set as well. Twitter user part of the data sets need at least 10 mutual connections to be inferred (because the toponyms are derived from these connection), thus need at least 10 friends and followers.

A weight of 1 is given to either followers or friends that do not follow the respective user back, or are not follow by this respective user. When the connection of the respective user with its followers or friends is mutual this connection is given a value of 2. The idea behind is that this type of connection is stronger, thus a higher value is given to this stronger connection as well.

7.3.3 Hybrid-user method

When using the hybrid-user method a user's geolocation is determined both by the content from a user's profile and his or hers tweets and the social network of that particular user. Similar GIMs have been presented by Gu et al (2012) and Ren et al (2012). This will be done according to a certain set of metadata attributes, which are detailed in Table 7.6 below. This metadata will then be used according to the workflow illustrated in Figure 7.10 below. These parameters will be detailed later in this paragraph. The script containing the hybrid-user method itself has been detailed in Appendix III.11.

Object	Field name	Type	Comment
Users	id_str	String	Used to derive user and tweet metadata.
	description	String	The user-defined description for their account.
	statuses_count	Integer	The number of tweets issued by the user.
	followers_count	Integer	Number of followers of the user.
	friends_count	Integer	Amount of accounts the user follows.
	location	String	The user-specified user location as defined by the user.
Tweets	text	String	The actual text of a status update.
	place	Object	Various info on the place associated with the tweet.

Table 7.6: Metadata used in hybrid-user method

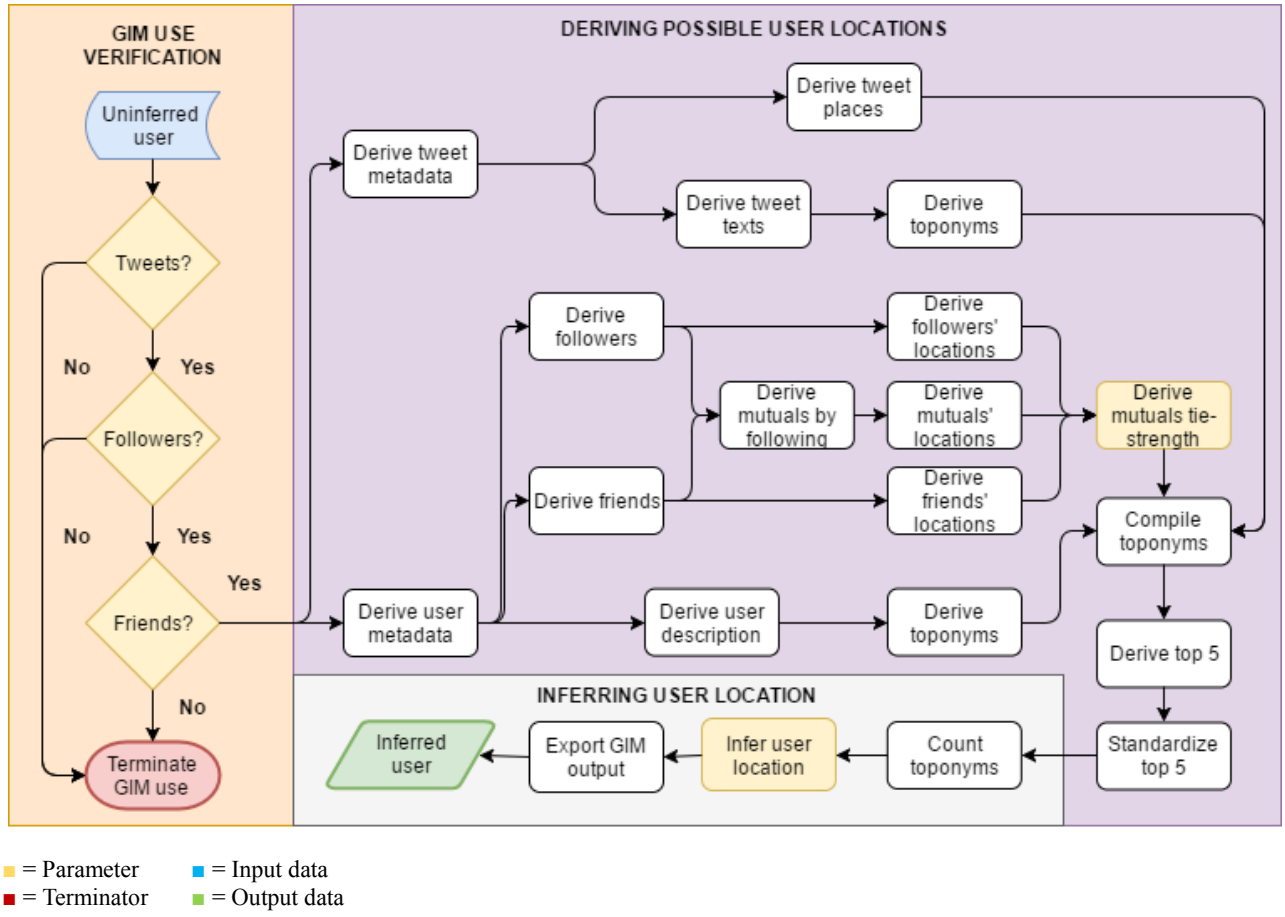

Figure 7.10: Hybrid-user method

Step 1: GIM use verification: Uninferred users will be validated first before the GIM is actually used based on the number of followers and friends they have and the amount of tweets they have posted. If they have posted a certain minimal number of friends or followers and posted a certain number of tweets they are

considered to be sufficient to use this GIM on. If this is not the case the GIM will not be used on that specific user because it is then assumed that either not enough mutual connections can be derived to infer the user's geolocation or that not enough content can be derived to infer the user's geolocation. The minimum number of followers, friends and tweets needed can be set according to the will of the one using the GIM and is one of the parameters that will be described later in this paragraph.

Step 2: Deriving possible user locations: Five sources of content will then be used to derive the user location. This will be the ones used for the content-user method and the network-user method. These sources, and how this information has been derived, has been detailed earlier in paragraph 7.3.1 and 7.3.2.

Step 3: User location inference: The user location will be inferred according to the same method as described earlier in paragraph 7.3.1 for the content-user method. Additionally, other statistics described for the content-user method will be exported as well. When all user locations are inferred the script used will automatically stop. When using this LIM there are four types of parameters that can be set:

- **Minimal number of tweets***:* This parameter has previously been described in paragraph 7.3.1.
- **Minimal number of followers/friends***:* This parameter has previously been described in paragraph 7.3.2.
- **Weights of connections***:* This parameter has previously been described in paragraph 7.3.2.
- **Minimal toponym count needed for inference**: This parameter has previously been explained in paragraph 7.3.1.

Given that the hybrid-user method is a hybrid of the content-user and network-user method as specified earlier in paragraph 7.3.1 and 7.3.2 the same values set for the parameters for those GIMs are set here as well. This means that users need at least have posted 10 tweets, 10 followers and friends and from the user's content and social network at least 10 toponyms need to be able to be derived before the geolocation of the user can be inferred.

7.4 Evaluation and comparison

Given that the central aim of this thesis is to evaluate and compare several GIMs among each other in various GIS research scenarios it is essential that the way in which this evaluation and comparison takes place is properly defined and argued. This means that certain metrics need to be defined by which the GIMs' design and output can be evaluated while at the same time a system needs to be developed by which the GIMs can be compared among each other. The methodology used to do this will be defined in this paragraph. Any choices or statements made will be supported by (academic) literature where needed.

7.4.1 Ground-truth definition

Before defining any evaluation or comparison metrics it is important to define what the "real" geolocations of Twitter users are. These geolocations will form the "ground-truth". The ground-truth as used in GIS research can be defined as the actual geolocation on earth of a certain subject (Pickles, 1995, p. 179). Defining the ground-truth is necessary to be able to validate whether the inferred user location found through the use of a GIM corresponds with the actual geolocation of this user. The ground-truth is also used in the calculation of for example the error distances which will be detailed later in paragraph 7.4.2. The problem with Twitter user locations is that these geolocations are incredibly hard to verify. A user can for example pretend that he or she lives in New York, NY while the user may actually live in another city or even another country. Within the thesis context it is assumed that the standardized user-specified user locations as part of the datasets are true as long as they correspond to names of existing places in the contiguous United States. These geolocations are then considered as the ground-truth in this thesis research. The ground-truth is derived using the pre-processing methods as described earlier in paragraph 7.2.4.

7.4.2 Definition of metrics

The evaluation and comparison metrics have been defined through literature from three academic fields. The first academic field is GIS research, in which spatial data is used as the main input of geographical information systems to answer certain research questions. Within this academic field it is important that the data quality of this data is sufficient since insufficient quality of data automatically leads to flawed results from the said system. To prevent this, spatial data quality standards and evaluation metrics have been set up over the years by various researchers. A detailed historic overview on this subject has been written by Devillers et al (2010). The second academic field is research on GIMs in which (spatial) data is used to infer the geolocation of a certain subject. The aim of GIMs is to increase the data quality of spatial data by inferring the geolocations of subjects and thus creating more valid data entries to use in research. To estimate the validity of the data output from these GIMs, evaluation metrics have been set up for these methods as well to ensure that the inferred geolocations are correct or at least sufficient. A general overview on the subject has been written by Ajao et al (2015, p. 861). The third academic field metrics are derived from is research on performance of coding scripts. When used for certain applications within certain contexts it is vital for coding scripts to run within a certain time and within a certain memory usage for example. Therefore, metrics to measure the script's performance are needed to be able to make sure these scripts run within the maximum allowed values specified. While academic literature on these types of evaluation metrics is scarce, plenty of blog posts from data scientists are available (see Anant, 2014; Nguyen, 2013; Rossant, 2012 for examples).

The evaluation and comparison metrics as will be used in this thesis research will be based on the three sources as mentioned above. Metrics related to GIS research have been based on the metrics as defined by Veregin (1999). This specific set of metrics has been chosen because they are easy to quantify and normalize, relatively easy to implement and have been referenced often indicating a widespread use by other geoscientists as well. Other academic literature on the subject of spatial data quality metrics have been considered as well but not implemented in this thesis research for various reasons. Devillers et al (2002) have developed an interesting hierarchical system based on detail to evaluate spatial data quality for example. The problem is that this system is of a qualitative nature and therefore difficult to normalize, making it hard to compare GIMs among each other in the GIS research scenarios. Metrics related to research on GIMs have been based on the overview on evaluation metrics as given by Ajao et al (2015, p. 861) previously mentioned. This specific overview is based on a great amount of academic literature on GIMs and topical enough to be implemented in the thesis research as well. The metrics related to script performance were initially based on a blog post by Nguyen (2013a). The reason why no academic literature was used to base these metrics on is because there was a severe lack of scientific literature on the subject of performance analysis of Python scripts. The reason why this specific source was chosen is because while other blog posts only measure script performance through temporal means while Nguyen also integrates another metrics, namely memory usage. The source can considered trustworthy given that Nguyen is a senior research engineer at Yahoo and Flickr and thus is assumed to have a fair knowledge on the subject of performance analysis of scripts (Nguyen, 2013b). In the final analysis memory usage has not been taken into account as an evaluation metric however as will be detailed later in this paragraph. The metrics that will be used have been presented in Table 7.7 below and will be described and argued now.

Metric	Sub-metric	Description	Range	Unit
Reliability	Spatial reliability	Correctness of geographically positioning.	0 - 100	Percentage
	Temporal reliability	Up-to-datedness of data.	0 – ∞	Days
Scale	Average scale	Sum of values divided by the number of values	0 – 4	None
	Median scale	Value separating upper and lower halves of data sample.	0 – 4	None
Completeness		Level of omission of real-world information.	0 – 100	Percentage
Effectiveness	Precision	Index of perceived amount of inferred user locations.	0 – 1	None
	Recall	Index of actual amount of inferred user locations.	0 – 1	None
	F-measure	Harmonic mean of precision and recall.	0 – 1	None
Speed		Time needed to infer a user location.		
	Average speed	Sum of values divided by the number of values.	0 – ∞	Seconds
	Median speed	Value separating upper and lower halves of data sample.	0 – ∞	Seconds

Table 7.7: Descriptions for evaluation and comparison metrics

Not all metrics described in the three sources as mentioned on the previous page are integrated in the evaluation and comparison metrics used in this thesis research. The reason why varies for each metric.
The thematic accuracy as defined by Veregin (1999, p. 181) has not been included because whether the inferred user locations are true or not is already evaluated through the "effectiveness" metric. The choice has been made to use the term "reliability" instead of "accuracy" because the ground-truth as defined earlier in paragraph 7.4.1 is hard to evaluate on the fact whether it is a true real-world representation of the user locations. The three types of resolution (spatial, temporal and thematic) as defined by Veregin (1999, p. 181-182) are not are not integrated in this thesis research because the terms "spatial resolution" and "thematic resolution" of user locations represented as point data do not fit very well for this type of data structure.
The spatial resolution is defined by Veregin (1999, p. 180) as "*the minimum size of objects on the ground that can be discerned*". Given that within this thesis research point data is used exclusively it is difficult to determine a minimal size of objects they represent. The main reason is that these points are spatial objects that do not exists in the real world. Another reason is that it is incredibly difficult to determine the "size" of vector data (which these point data represent) compared to raster data as also discussed by Veregin (1999, 180-181). Thematic resolution is defined by Veregin (1999, p. 182) as the resolution "*in terms of fineness of category definitions*". Within the context of this thesis research it is meant as the amount of detail the inferred user locations given on the whereabouts of the user. Both these metrics have been converted to a new metric called "scale", which depicts the scale of the geographical object the point represents. The reason for this is that both spatial resolution and thematic resolution as applied on to user locations represented as point data basically mean the same thing. Temporal resolution has not been implemented in this thesis research because the resolution is based on the data from the Twitter API, which is used for all GIMs as has been detailed earlier in paragraph 7.3. Therefore, all data outputs have the same temporal resolution and it is illogical to compare them among each other since they are the same. Consistency as defined by Veregin (1999, p. 182-183) has not been implemented in this thesis research as well. The reason for this is that consistencies in topology, temporal and thematic aspects of the data are not probable. The inferred user locations are based on OpenStreetMap data, which is countlessly verified and evaluated constantly. Inconsistencies in the final data sets used in this thesis research are therefore assumed to be so low they are not worth to be evaluated. As previously mentioned, memory usage was considered as an evaluation metric but not implemented in the final analysis as performed. The main reason was that the memory usage was incredibly low for all GIMs part of this thesis research and therefore considered not worth evaluating, similar to the previous argument made for the exclusion of the consistency metric above.

Spatial reliability: According to Veregin (1999, p. 179) spatial accuracy "[…] *refers to the accuracy of the spatial components of a database*". By measuring the spatial accuracy, one can find out how well the objects part of the data sets they are working with are geographically positioned compared to their actual position in the real world. Within the context of this thesis research it determines how far the inferred user location is away from the standardized user-specified user location. Veregin describes how various metrics have been developed to determine the spatial accuracy of spatial data and not necessarily every metric is applicable in every GIS research scenario. Therefore, the metrics related to the spatial accuracy of GIMs were initially implemented as described by Ajao et al (2015, p. 861) are used to measure the spatial accuracy.
These metrics are the result of a (relatively) topical and extensive literature study on the subject of GIMs. They were therefore considered representative for the metrics used in GIMs research. Two metrics were initially used to measure the spatial reliability of the data. The first was the average error distance, which is the average distance between the perceived geolocation of a subject and the real geolocation of a subject as defined through the ground truth in meters. The median error distance was calculated as well. The reason why the median error distance was measured as well is because the average error distance can be prone to skewed value distribution, giving a skewed average error distance as well. During analysis, it turned out that using this methodology it was difficult to normalize the values found however. Therefore the choice was made to instead of calculating the average or median error distance, the percentage of geolocation inferred below 100 kilometres is considered to be representative for the spatial reliability of the GIMs.
This error distance is considered as sufficient in the majority of GIS research done using Twitter data (being on a national-level primarily). This specific value has also been used in the evaluation of plenty of GIMs in academic literature as a "benckmark" (see Zhang et al; 2015, Chandra et al, 2011; Cheng et al, 2013; Mahmud et al, 2014; Chang et al, 2012; Liu & Inkpen, 2015; Ren et al, 2012 for examples). The metric is calculated using the GeoPy Python package as detailed in Appendix IV.2. In this calculation, the Vincenty distance is calculated using the WGS84 ellipsoidal model of the earth. The Vincenty distance is based on the assumption that the figure of the Earth is an oblate spheroid (Vincenty, 1975). This particular distance is

chosen because this is the most accurate one available to calculate through the GeoPy Python Package. WGS84 is chosen because this is the same coordinate system used in the Twitter metadata (Twitter, 2015).

Temporal reliability: This type of reliability is defined by Veregin as the "currentness" of the spatial data (1999, p. 180). When spatial data is not temporally accurate, the objects part of these data sets are geographical positioned according to a relatively old observation of these specific objects. The geolocation of the objects perceived in this observation might not be current anymore. It has to be noted, like earlier in this paragraph, that the term "reliability" is used over "accuracy" for reasons previously explained in this paragraph as well. Within the context of this thesis research this means that the inferred geolocation might not be the current user location anymore and is actually a different geolocation. It is therefore important that spatial data is temporally accurate. The currentness of inferred user locations part of the data sets are determined by taking the latest time at which the inferred user location is mentioned in the tweets of this respective user.

Scale: The "scale" metric has been based on both the spatial resolution and thematic resolution metrics as defined by Veregin (1999, p. 180-182). The lowest scale of the geographical object the point represents (city, county, state et cetera) is used to determine the spatial resolution in this thesis research. The scale is determined by counting the number of commas used in the name of the inferred user location. The output of the GeoPy Python package typically consists of place names in which each scale level or level of geographical detail is separated by a comma (for example: "*LA, Los Angeles County, California, United States of America"*). The scale each value represents is listed below in Table 7.8. Values of 4 or above are put in the same scale class ("sub-city and lower"). This is done because the scale is determined through the number of commas used in the standardized user-specified user location. While country names, state names, county names and city names can easily be linked to a scale name, it becomes difficult for sub-city and lower toponyms.

Scale	Value
National	0
State	1
County	2
City	3
Sub-city and lower	4+

Table 7.8: Value representation for scale metric

Completeness: Veregin (1999, p. 183) defines multiple definitions for "completeness" in his work. In this thesis specifically "data completeness" is measured, being the amount of omission of data in the data sets. Within the context of this thesis research this means the amount of user locations that could be inferred. This metric is measured by the percentage of user location that could be inferred through a specific GIM.

Precision: The precision is the perceived number of inferred geolocations by the GIM used. Perceived inferred geolocations are either True Positives (correctly inferred geolocations stated as correctly inferred) and False Positives (falsely inferred geolocations stated as correctly inferred). The precision is typically calculated using an index by the formula as defined below by Ajao et al (2015, p. 861). The way in which the True and False Positives found will be determined will be described later in paragraph 7.4.4.

$$Precision = \frac{True\ Positives}{(True\ Positives + False\ Positives)}$$

Recall: The actual amount of inferred user locations by the GIM used. The actual inferred geolocations are the True Positives (previously described) and False Negatives (correct inferred geolocations stated as falsely inferred). The recall is typically calculated using an index by the formula as defined below by Ajao et al (2015, p.861). The way in which the True Positives and False Negatives found will be determined will be described later in paragraph 7.4.4.

$$Recall = \frac{True\ Positives}{(True\ Positives + False\ Negatives)}$$

F-measure: The harmonic mean of both the precision and the recall index. The F-measure is typically calculated using an index by the formula below as defined by Ajao et al (2015, p. 861). The metric's purpose is to decrease the influence of errors in determining the spatial reliability of a GIM. The way in which the Precision and Recall have been calculated has been described in the previous page:

$$F - measure = \frac{2 * Precision * Recall}{(Precision + Recall)}$$

Speed: The processing speed is the amount of time needed to infer one user location in seconds.
The processing speed is measured in Python using the time module, which is provided by default by Python 2.7 as used in this thesis research and detailed in Appendix IV.2. Two metrics are used to measure the processing speed of the scripts. The first is the average processing speed, which is the average time needed to infer the geolocation of a user part of the data sets. The median processing speed is calculated as well.
The reason why the median processing speed is measured as well is because the average processing speed can be prone to skewed value distribution, giving a skewed average processing speed as well.

While the metrics currently defined may be useful to evaluate a single GIM they are not insufficient to use to compare the GIMs among each other in different GIS research scenarios. There are a couple of reasons why this is the case, listed below:

- **Lack of oversight**: Currently 5 metrics are determined by which a single GIM will be evaluated. It is difficult to compare multiple GIMs using these 5 metrics due a lack of oversight, however. It would be preferable to be able to see what GIM performs best using a single value instead of 5 metrics.

- **Values are misleading**: The second reason is that for some metrics it is difficult to determine whether performance of the GIM is good or bad. When the lowest error distances found are 10 kilometres and the highest are 40 kilometres this does not necessarily mean that the first value given is good or the latter is bad because there is no minimum or maximum value to which these values can be compared to.

- **Difficulties creating index**: The third reason is that the metrics as specified use different units and therefore cannot easily be included within one GIM-performance index. The metrics related to effectiveness use a value between 0 and 1, temporal reliability uses an integer value in days and speed uses an integer value in seconds. These values cannot simply be thrown as input in a formula to calculate the performance of the GIMs.

To be able to successfully compare the GIMs among each other in the different GIS research scenarios the values found for the metrics need to be normalized first to be able to create a performance index.
When normalizing values statistically, values with different outer values are converted to an index using the same outer values (ESRI, 2017b). Metrics using different outer values originally can then be compared for example. These values are calculated using the following formula in which P_n represents the normalized value, P_o the original value and P_{min} and P_{max} representing the minimum and maximum value respectively. The minimum and maximum value determined for each evaluation metric has been presented in Table 7.9 on the next page as well. Any choice made while defining the minimum and the maximum values will be argued on the next page.

$$P_n = \frac{P_o - P_{min}}{P_{max} - P_{min}}$$

Metric	Sub-metric	Sub question 4		Sub question 5	
		Minimum	Maximum	Minimum	Maximum
Reliability	Spatial reliability	0.000	1.000	0.000	1.000
	Temporal reliability	0.000	2546	0.000	2546
Scale		0.000	4.000	0.000	4.000
Completeness		0.000	1.000	0.000	1.000
Effectiveness	Precision	0.000	1.000	0.000	1.000
	Recall	0.000	1.000	0.000	1.000
	F-measure	0.000	1.000	0.000	1.000
Speed		3.775	17.720	0.000	17.720

Table 7.9: Maximum and minimum values defined for normalization process of evaluation metrics

The minimum and maximum values determined for the spatial reliability, scale, completeness and all three effectiveness-related metrics are natural given that these values are already determined by a 0 to 1 scale. The maximum value of the temporal reliability has been defined by calculating the average amount of days a U.S. citizen lives at the same geolocation. As long as the topicality of the inferred user location is below this maximum value it is assumed that the user has not moved and thus rendered the inferred user location to be false. The maximum value of 2546 days has been determined by dividing the average life expectancy of U.S. citizens in 2014 as determined by the U.S. Center for Disease Control and Prevention with the average amount of times U.S citizens move in their lifetime as determined by statistics blog FiveThirtyEight using statistics from 2010 and 2013 (CDC, 2014; FiveThirtyEight, 2015). The minimal and maximum speed has been determined by taking the average of the top 5 lowest and highest speed times found over all data sets respectively. The top-5 has been chosen to make sure the effect of any possible outliers is decreased.

The normalized values found for the main metrics as described on the previous page will be put together to form a GIM-performance index represented as P_{GIM} in the formula below. n_c indicates the amount of main metrics used to evaluate the GIMs. w indicates the weight attached to each metric. The way in which these weights will be distributed will be described later in paragraph 7.4.3. As a result of this formula an index will be calculated by which the GIMs can be compared to.

$$P_{GIM} = \frac{w_{rel} * P_{rel} + w_{sc} * P_{sc} + w_{com} * P_{com} + w_{eff} * P_{eff} + w_{sp} * P_{sp}}{n_c}$$

7.4.3 Sensitivity analysis

Determining the influence of each value set for the GIM-parameters on the output of these GIMs and each evaluation and comparison metrics is important to make sure the overall performance value is not misinterpreted. Therefore, two sensitivity analyses will be conducted to find out to what extent each parameter influenced the GIMs and each main metrics influences the overall performance value respectively.

In the first sensitivity analysis, the parameters set for the GIMs as defined earlier in paragraph 7.3.1 to 7.3.3. will be set differently to see to what extent user locations are inferred differently. This sensitivity analysis has been based on the works of Cheng et al (2013), who are one of the few researchers who have performed sensitivity analysis on the GIM they present in their article. They increased and decreases the number of tweets used to infer a user's geolocation to see to what extent setting these parameters influence the GIM's output. With this in mind the following values will be set for the parameters as defined in 7.3.1 to 7.3.3 as presented in Table 7.10 below:

Parameter	Low	Mid	High
Min. no. tweets	1	10	100
Min. no. followers/friends	1	10	100
Min. no. toponyms	1	10	25
Weights	1.5	2	3

Table 7.10 Sensitivity values set for first sensitivity analysis

The second sensitivity analysis will be done according to six weight scenarios in which the metrics will be differently weighted. These have been presented in Table 7.11 below. In weight scenario 2 to 6 the weight of one of the main metrics has been double from 0.20 to 0.40. Different weights have been included to the main metrics exclusively and not the sub-metrics because adding different weights to the sub-metrics would only hardly influence the GIM performance index since there are 11 metrics to weight in that case.
Weight scenario 1 is used as a default sample where all main metrics are equally weighted. If the overall performance value found for each metric in weight scenario 2 to 6 are (radically) different from the weight scenario 1 this indicates that the sensitivity for one or more metrics is relative big. If this is the case this finding will be taken into account in writing the results chapter and forming conclusions in this thesis research.

Metric	W.S. 1	W.S. 2	W.S. 3	W.S. 4	W.S. 5	W.S. 6
Reliability	0.20	0.40	0.15	0.15	0.15	0.15
Scale	0.20	0.15	0.40	0.15	0.15	0.15
Completeness	0.20	0.15	0.15	0.40	0.15	0.15
Effectiveness	0.20	0.15	0.15	0.15	0.40	0.15
Programming	0.20	0.15	0.15	0.15	0.15	0.40
Total	1.00	1.00	1.00	1.00	1.00	1.00

Table 7.11: Weights scenarios used in second sensitivity analysis

7.4.4 Technical framework: Evaluation and comparison of GIMs

After conducting the GIS research scenarios as described earlier in paragraph 7.2, various values are found related to the GIMs. These are for example the inferred geolocations and the time needed to infer the geolocation. The problem is that these values are meaningless without analysing them first. The performance of the GIMs will be determined according to various evaluation and comparison metrics as defined earlier in paragraph 7.4.2 and sensitivity analysis as defined earlier in 7.4.3. The technical framework to calculate these values will be described in the current paragraph. The performance of the GIMs will be analysed according to the workflow as illustrated in Figure 7.11 below and detailed there as well:

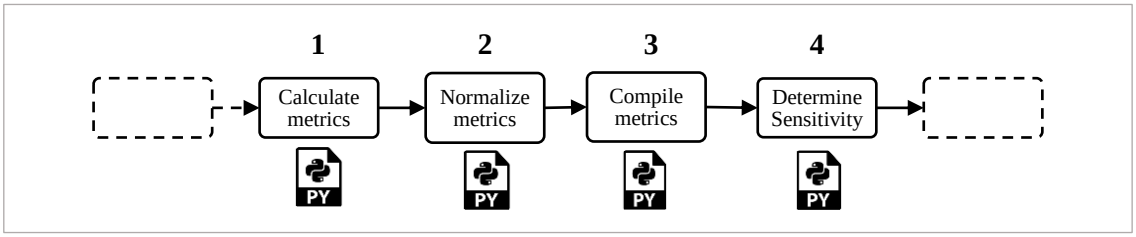

Figure 7.11: Analysis workflow

Step 1: Calculating metrics: For each evaluation metric, the value is calculated using Python. The code to do this has been detailed in Appendix III.12. The code was written and run within the Canopy software package as detailed in Appendix IV.1. The way in which each evaluation metric is calculated is described below and the next page and argued where needed:

- **Spatial reliability**: The spatial reliability is calculated by determining the amount of error distances between the inferred user location and the standardized user-specified user locations found below 100 kilometres. This error distance is a default output of the GIMs as detailed earlier in paragraph 7.3. The Vincenty distance is calculated using the WGS84 coordinate system standards specifically, as has been argued earlier in paragraph 7.4.2. As detailed earlier in paragraph 7.3 the top-2 inferred user locations have been determined. The inferred user location with the lowest error distance is considered to be the final inferred user location and will be taken into account in the determination of the spatial reliability.

- **Temporal reliability**: The temporal reliability is calculated by determining the age of the latest observations of a toponym referring to the same geolocation as the inferred user location for a specific user in the metadata. Therefore, the latest time at which a tweet was posted by a specific

user mentioning the toponym which refers to the same geolocation as the final inferred user location for that specific user is considered to be latest observation. For this particular method, the created_at(Tweets) metadata attribute is used to indicate this specific time (Twitter, 2017i).
As detailed earlier above the top-2 inferred user locations have been determined. The inferred user location with the lowest error distance is considered to be the final inferred user location and will be taken into account in the determination of the spatial reliability.

- **Scale**: The scale is calculated by counting the number of commas in the inferred user location's name. This is done because typically the scale levels of the inferred user location are separated by a comma. This methodology has been argued in more detail earlier in paragraph 7.4.2. Both the median and average scale for all user locations in each respective data set is calculated to prevent misinterpretation of the found values for this particular metrics due to skewed distribution.

- **Completeness**: How "complete" the data set is will be determined by counting the amount of user locations that could not be inferred and subtract that value from the total amount of user locations part of the respective data set.

- **Effectiveness**: The effectiveness is calculated through the Precision, Recall and F-measure as has been detailed earlier in paragraph 7.4.2. To be able to calculate these measures the fact whether the inferred user locations are true positives, false positives or false negatives needs to be determined first. This will be done according to the standards as listed below:

    - *True Positive*:

        - If the error distance is below 0.01 kilometres or the inferred user location is the same as the standardized user-specified user location or the compact inferred user location is contained in the standardized user-specified user location.
        - The compact inferred user location is observed at least once in the latest 200 tweets of the user.
        - The scale of the inferred user location is at least 3 (city-level).

    - *False Positive*:

        - If the error distance is below 0.01 kilometres or the inferred user location is the same as the standardized user-specified user location or the compact inferred user location is contained in the standardized user-specified user location.
        - The compact inferred user location is not observed in the latest 200 tweets of the user.
        - The scale of the inferred user location is at least 3 (city level).

    - *False Negative:*

        - If the error distance is above 0.01 kilometres or the inferred user location is not the same as the standardized user-specified user location or the compact inferred user location is not contained in the standardized user-specified user location.
        - The compact inferred user location is observed at least once in the latest 200 tweets of the user.
        - The compact inferred user location occurs more than 10 times in the latest 200 tweets of the users and more often than the compact standardized user-specified user location.

    - *True Negative:*

        - All inferred user locations that are not considered a true positive, false positive or false negative using the criteria as detailed above.

- ***Speed***: Both the median and average processing speed for all user locations in each respective data set is calculated to prevent misinterpretation of the found values for this particular metrics due to skewed distribution.

Step 2: Normalizing metrics: The values found for each evaluation metric are normalized according to the formula as detailed earlier in paragraph 7.4.2. This is done to create oversight, make sure the different metrics can be added up and compiled as metrics and prevent the values found for the metrics to be misleading. These reasons have been argued in more detailed earlier in paragraph 7.4.2 as well. This is done according with the code as detailed in Appendix III.13 and run within the Canopy software package as detailed in Appendix IV.1.

Step 3: Compiling metrics: The metrics are then compiled to one final metric. This will make the evaluation and comparison of the GIMs among each other easier and better. This is done according with the code as detailed in Appendix III.13 and run within the Canopy software package as detailed in Appendix IV.1.

Step 4: Determine sensitivity: Finally, two types of sensitivity analysis will be performed to determine whether the results found are influenced heavily by the parameters set within the GIMs. This will be done according to the methodology as previously described in paragraph 7.4.3. For the first type of analysis separate data sets will be created to which they will be analysed with GIMs using different sets of parameters. How this is done is explained in more detail in paragraph 8.3.2 later. For the second sensitivity analysis the script as detailed in Appendix III.13 is used.

7.5 Sub question 4

The aim of answering the fourth sub question is to find out what the strengths and weaknesses are of the GIMs in event detection research scenarios. The GIMs to be evaluated have been defined and detailed in paragraph 7.3 previously while the same has been done for the GIS research scenarios in paragraph 7.2 earlier in the thesis report as well. The same has been done for the evaluation metrics in paragraph 7.4.2. First, the GIS research scenarios as have been described will be conducted step by step. Secondly, the values for the evaluation and comparison metrics will be calculated according to the technical framework as defined earlier in paragraph 7.2.4. Finally, the performance of the GIMs in each GIS research scenario will be compared according to the normalized overall performance values. The findings will be presented in figures and tables mainly. Where needed descriptive analysis will be done to support these figures and tables. Where needed arguments made will be support by (academic) literature where needed.

7.6 Sub question 5

The aim of answering the fifth and final sub question is to find out how the data output of the GIMs compares to the geotagged Twitter data found through using the Twitter API by default. This is done using the same methodology as described earlier for sub question 4 in paragraph 7.5. The main difference is that geotagged Twitter data is incorporated as well. The geotagged Twitter data will be derived from the unprocessed data sets gathered for the respective GIS research scenarios as detailed earlier in paragraph 7.2.4. The focus will lie on the differences between the values found and to what extent the GIMs have increased the data quality of the geotagged Twitter data. The findings will be presented in figures and tables mainly. Where needed descriptive analysis will be done to support these figures and tables. Where needed arguments made will be support by (scientific) literature where needed. The metrics as defined earlier in paragraph 7.4.2 are slightly differently interpreted for the geotagged data than for the inferred Twitter data. The differences are the following:

- **Reliability**: Even GPS coordinates attached to the Twitter metadata contain errors. The degree of the error is influenced by various factors, such as the way the smartphone is positioned. An example is given by a much-referenced article by Zandbergen (2009, p, 5-11) on the subject in which it was pointed out that positioning smartphones through GPS, WiFi and cellular position given different accuracies. Different types of metropolitan areas seem to give different accuracies as well,

as researched by Zandbergen (2012) as well. The way in which the coordinates attached to the Twitter metadata is positioned cannot be determined from this metadata. Therefore, the assumed error has been determined through previously conducted scientific research on the subject.
The academic literature considered has been listed in Table 7.12 on the next page. The choice was made to base the spatial reliability on a study by Garnett and Stewart (2015) because this study was set within an urban context, relatively topical and being published in a respected journal. Because of the latter two characteristics the value used can be considered accurate. Garnett and Stewart (2015, p. 5) found an average distance error of 6.5 meters when using GPS for positioning smartphones and this value will be the assumed spatial reliability for the geotagged Twitter data as well.

- **Scale**: GPS coordinates are assumed to be equal to the maximum scale of the GIMs, being 4. Therefore, the value given for the scale metric is 4 as well

- **Completeness**: The amount of completeness is determined in a similar way as done for sub question 4. In this case however, the number of tweets not georeferenced are subtracted from the total amount of tweets part of the data sets.

- **Effectiveness**: Given that the GPS coordinates are assumed to be true positives the Precision, Recall and F-measure are all assumed to be the highest possible, being a value of 1.

- **Programming speed**: The processing speed and memory usage are calculated the same way as for sub question 4.

Source	Device(s)	Method(s)	Average (m)	Median (m)	Context(s)
Zandbergen (2009)	Apple 3G iPhone	A-GPS WiFi Cellular		1.4 to 1.7 (outdoors) 74 (indoors) 599 (indoors)	All methods were tested in the Albuquerque, NM metropolitan area.
Zandbergen & Barbeau (2011)	Motorola i580 Sanyo SCP-7050	Assisted Autonomous	≈ 5.9 to 10.1 (Assisted, static outdoors) ≈ 3.0 (Assisted, dynamic outdoors) ≈ 15.1 (Assisted, static indoors) ≈ 9.8 (Autonomous, static outdoors) ≈ 5.3 to 7.0 (Assisted, static outdoors) ≈ 1.8 (Assisted, dynamic outdoors) ≈ 8.8 (Assisted, static indoors)		A cloverleaf intersection between Interstate 4 and US Highway 301.
Zandbergen (2012)	Apple 3GS iPhone	WiFi		42.6 to 46.4 (San Diego, CA) 38.9 to 41.6 (Miami, FL) 79.7 to 92.4 (Las Vegas, NV)	Starbucks locations in multiple US cities.
Kos, Brčić, Musulin (2013)	Samsung GT-S5570	GPS	≈ 5.1 to 5.4	≈ 3.8 to 4.2	Split, Republic of Croatia; Stable conditions
Musulin, Kos & Brčić (2014)	Samsung Galaxy Mini 1 Samsung Galaxy Mini 2 Sony Xperia 8 iPhone 5 Samsung Galaxy Note	A-GPS GLONASS	2.25 (A-GPS) 2.79 (A-GPS) 2.18 (GPS) 2.72 (A-GPS/GLONASS) 4.18 (A-GPS/GLONASS)		Bay of Zaton, Republic of Croatia.
Park et al (2014)	Samsung Galaxy Note 1	GPS WiFi	53.5 22.0	48.4 21.8	Central business district area of Seoul, South Korea.
Garnett & Stewart (2015)	iPhone 4S	GPS	6.5		Lakehead University, Thunder Bay, Ontario
Zhuang et al (2016)	Samsung Galaxy SIII	WiFi	5.3 to 5.7		Unknown.

Table 7.12: Considered scientific literature to base spatial reliability on

8. Sub question 4: Evaluating and comparing GIMs among each other

8.1 Introduction

The fourth sub question will be answered in this chapter by performing an analysis in which the performances of the GIMs will be evaluated and compared in the GIS research scenarios using the evaluation and comparison metrics. These specific components have been detailed earlier in paragraph 7.2, 7.3 and 7.4.2 respectively. A description of the data sets used in this analysis will be given first to which the results of the analysis will be presented in tables and figures. Descriptive analysis will be performed and arguments made will be supported by academic literature where needed. A sensitivity analysis will be conducted as defined earlier in paragraph 7.4.3 as well. Finally, all findings will be summarized at the end of this chapter.

8.2 Data set descriptions

As described earlier in paragraph 7.2, three data sets based on an equal amount of GIS research scenarios have been created to evaluate and compare the performances of the GIMs presented earlier in paragraph 7.3 by. Before looking into the performances of the GIMs within these GIS research scenarios, attention will be given on the way in which these data sets are structured and how they have been created. In Table 8.1 below the corpus sizes as a result of each pre-processing step as defined earlier in paragraph 7.2.4 have been given for each application domain part of this thesis research respectively. It has to be noted that only the pre-processing steps that affected the corpus sizes have been included in the table below. In this table "Tot. (n)" represents the absolute total number of rows in the data sets, "Dec. (%)" the relative amount of decrease of rows after conducting each pre-processing step and "Tot. (%)" the relative number of rows in the data sets after conducting each pre-processing step compared to the original data set corpus sizes. The highest value found for each pre-processing step in each column has been made bold. The geographical distribution of the data sets has been visualised in Figure 8.1 on the next page.

Processing step	Disaster management			Health management			Topic modelling		
	Tot. (n)	Dec. %	Tot. (%)	Tot. (n)	Dec. (%)	Tot. (%)	Tot. (n)	Dec. (%)	Tot. (%)
Gathering data	**262226**		100	15536		100	30085		100
Filter by metadata	**203988**	-22.2	**77.8**	11665	**-24.9**	75.1	22880	-23.9	76.1
Derive users	**102572**	**-49.7**	39.1	8961	-23.2	57.7	17576	-23.2	**58.4**
Standardizing	**84750**	-17.3	32.3	7796	-13.0	**50.2**	10789	**-38.6**	35.9
Clip by area	**44238**	-47.8	16.9	6649	-14.7	**42.8**	8517	**-21.0**	28.3

Table 8.1: Corpus sizes during pre-processing steps of Twitter data sets used in thesis research

When looking at the table above, a few observations can be made concerning the similarities and differences in which the data sets to be used in analysis are structured during each pre-processing stage of this thesis research. These observations will be described below and on the next pages:

- A similar hierarchy of normalized corpus sizes was found through academic literature as detailed earlier in Table 4.3 in paragraph 4.5 compared to the hierarchy as seen in Table 8.1 above. The absolute corpus sizes found for the disaster management data set are (much) higher than the ones found for the health management and topic modelling data sets.

- When filtering the data sets by metadata (by either the language used by the user/within tweets and whether the user has a geolocation specified), the Dec. (%)-values are almost equal with a difference of 2.7% at max. This indicates that no matter the content of the tweets the relative number of users using the English language and specifying a user location is similar.

- When excluding duplicates (and thus deriving the individual users part of the data sets), the Dec. (%)-value found for the disaster management data set is higher than the same kind of values found for the health management and topic modelling data sets. This indicates that users part of the disaster management data set are represented with more tweets in the unprocessed data set than in the other two data sets.

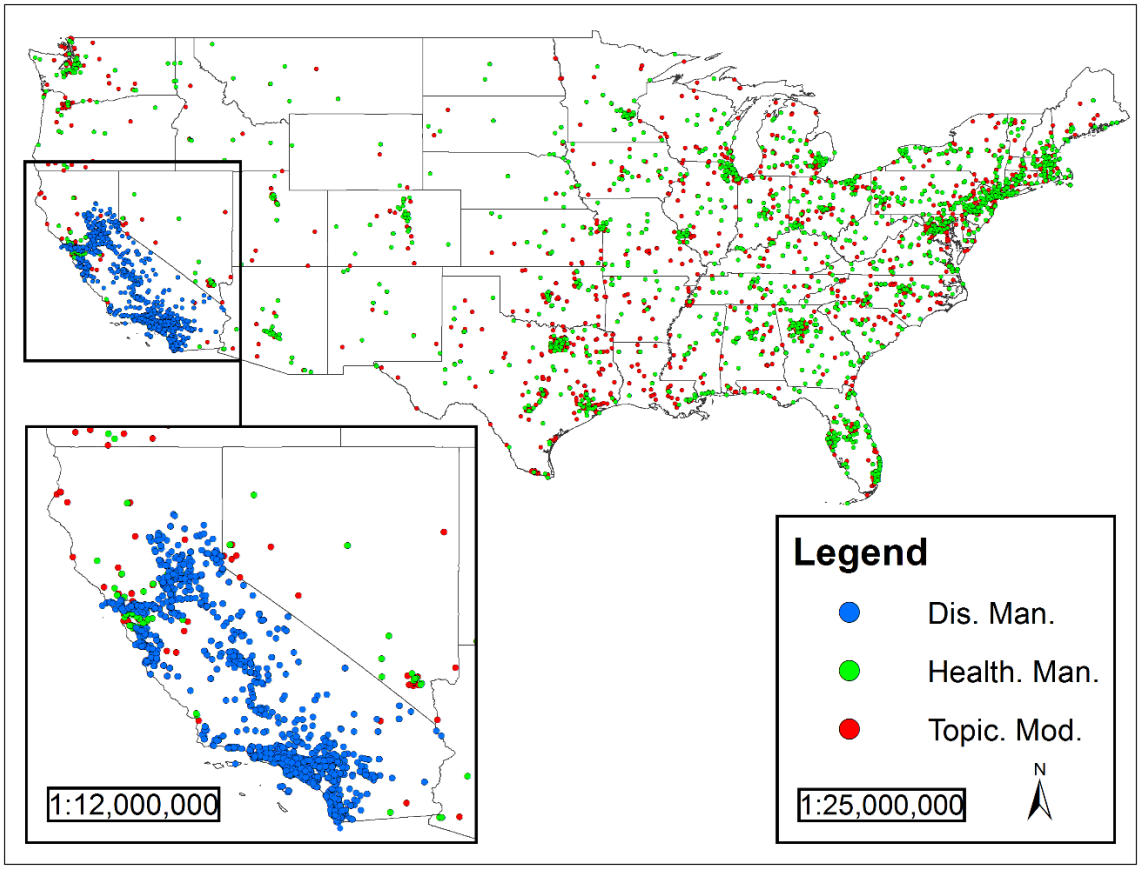

Figure 8.1: Geographical distribution of Twitter data sets used in thesis research

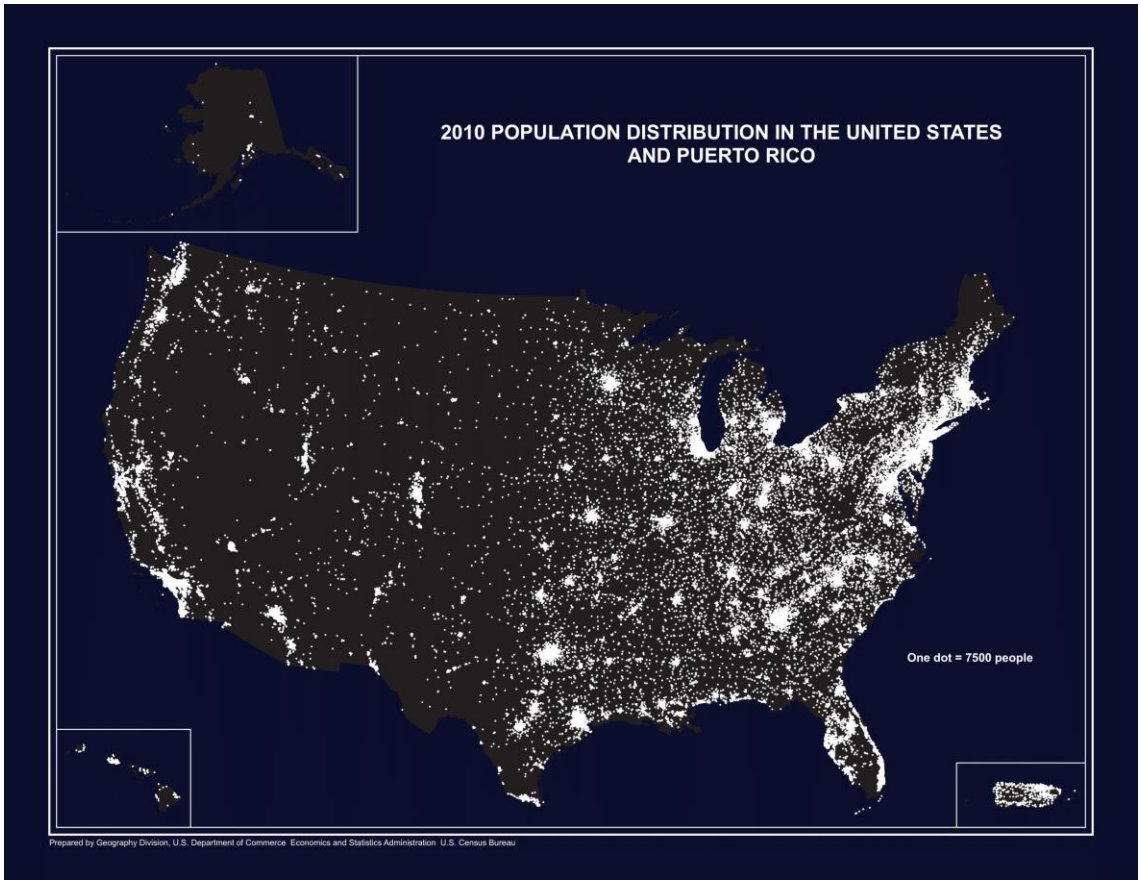

Figure 8.2: Geographical distribution of U.S. population in 2010 (Source: U.S. Census Bureau, 2010)

- While the Dec. (%)-values found for the standardizing step for the disaster management and health management data set are relatively close to each other with a difference of 4.3%, the same type of value found for the topic modelling data set is way higher. This means that in this respective data set a lot of user locations could not be standardized, possibly indicating that users do not use (well-known) toponyms as their user location.

- When clipping the data by catchment area, especially the Dec. (%)-value found for the disaster management data set is high. This either indicates that a lot of tweets were posted by users that have specified a user location outside of the disaster management research scenario study area or the bounding box from which the Twitter data was gathered contained more areas outside of the defined study area compared to health management and topic modelling research scenario. When performing analysis in ArcMap as specified in Appendix IV.2, it was found that the latter was the case.

- Comparing the final total number of rows in the data sets to the original number of rows, differences exist among the data sets when comparing the Tot(%)-values found. While the final data set used for the disaster management research scenario only contains 16.9% of the rows originally part of the data set, the data set used for the health management research scenario contains 42.8% of the rows originally part of the data set for example. These differences exist because some data sets are more heavily affected by certain pre-processing steps than others, as previously detailed in this paragraph and can be seen in Table 8.1 found earlier in this chapter.

- When comparing the geographical demographic distribution of the data sets to the actual geographical demographic distribution of the real population in 2010 by eye they seem similar, as illustrated in Figure 8.1 and Figure 8.2 on the previous page. This can especially be seen by the fact that both maps show dense populations on both the west coast and the east half of the contiguous United States.

For each GIS research scenario, the preferred and actual parameters of the data sets to be used in analysis have been defined earlier in paragraph 7.2. In Table 8.2 to 8.4 below and on the next page the preferred, final and actual parameters for each GIS research scenario part of the thesis research are put next to each other and discussed on the next page. The actual corpus sizes of the data sets have been calculated by dividing the Tot. (n)-values found for the "Filter by metadata" and "Derive users" pre-processing steps as presented in Table 8.1 respectively and multiply this value with the Tot. (n)-values found for the "Clip by area" pre-processing step presented in that table as well. By doing this the average number of tweets posted by each user part of the data set is determined, which is then used to estimate the number of tweets posted within the specified study area for each respective GIS research scenario.

Parameter	Preferred	Final	Actual
Application methodology	Event detection	Event detection	Event detection
Application domain	Disaster management	Disaster management	Disaster management
Real-time	Yes	No	No
Additional sources	Yes	Yes	Yes
Corpus size	128000 to 156000 tweets	128000 to 155000 tweets	≈ 87977 tweets
Period of gathering	7 to 9 days	7 days	7 days
Study area	Contiguous United States	Contiguous United States	California
Scale	Sub-national/City-level	Sub-national level	Sub-national level

Table 8.2: Preferred, final and actual parameters for disaster management research scenario

Parameter	Preferred	Final	Actual
Application methodology	Event detection	Event detection	Event detection
Application domain	Health management	Health management	Health management
Real-time	No	No	No
Additional sources	No	No	No
Corpus size	134000 to 164000 tweets	4000 to 5000 tweets	≈ 8665 tweets
Period of gathering	185 to 226 days	7 days	7 days
Study area	Contiguous United States	Contiguous United States	Contiguous United States
Scale	National-level	National-level	National-level

Table 8.3: Preferred, final and actual parameters for health management research scenario

Parameter	Preferred	Final	Actual
Application methodology	Event detection	Event detection	Event detection
Application domain	Topic modelling	Topic modelling	Topic modelling
Real-time	No	No	No
Additional sources	No	No	No
Corpus size	410000 to 502000 tweets	12000 to 14000 tweets	≈ 11087 tweets
Period of gathering	124 to 152 days	7 days	7 days
Study area	Contiguous United States	Contiguous United States	Contiguous United States
Scale	National-level	National-level	National-level

Table 8.4: Preferred, final and actual parameters for topic modelling research scenario

Some differences can be noted among Table 8.2 to 8.4 above and on the previous page when comparing the values found for the different groups of parameters. This goes especially for the corpus size of the data sets and periods of gathering. Note that the differences between the preferred and final parameters defined have been previously detailed earlier in paragraph 7.2 and will therefore not be discussed again here. For the disaster management and topic modelling scenario the actual amount of tweets part of the data sets is lower than the final number of tweets preferred (approximately 45% and 8% respectively compared to the minimum value), while for the health management research scenario this value is higher (approximately 73% compared to the maximum value). While this means that the actual corpus size values do not correspond very well with the final corpus sizes values preferred, this does not necessarily mean that these actual corpus sizes are not representative for any study within the same application domain as well. As detailed earlier in paragraph 4.5 the corpus sizes of studies within the same application domain vary heavily. Since the rest of the actual parameters are the same as the final parameters defined, the corpus sizes of the data sets are presumed to be representative majorly.

It has to be noted that for the analysis a random sample of 1000 entries of each data set has been used to evaluate and compare the GIMs among each other instead of the complete Twitter data sets as processed so far. The main reason for this decision is that the network-user method, as detailed earlier in paragraph 7.3.2, uses the geolocation of friends and followers to infer the geolocation of users part of the datasets.
The problem is that the rate limitations to gather data on the social network of Twitter users is very restricted compared to rate limitations to gather tweets through the Twitter API (Twitter, 2017q). As a result of this it takes approximately 1 minute to infer a user location using the network-user method. Since the user locations of thousands of users are meant to be inferred through these methods it would approximately take 6 weeks for this method to infer all user locations. Given that this means that the thesis research could not be finished before the preferred deadline (June 2017 as of writing) the choice has been made to analyse samples of the data sets rather than all entries of the data sets. Random samples of 1000 for each of the data sets used in the GIS research scenarios have been created using the R programming language, as defined in Appendix III.14. This specific code was written and run within the RStudio software package as detailed in Appendix IV.1. This means that even though the majority of the parameters of the data sets as defined earlier are still representative, the data set becomes less representative because the corpus sizes are not representative anymore due to them being sampled. This means that any results found or conclusions made should be interpreted indicative rather than axioms found. This line of thought will be taken into account in the upcoming parts of the thesis report.

8.3 Results

8.3.1 Regular analysis

With the data sets being described in the previous paragraph, the results found through analysis will be discussed now. The content-user and network-user method as defined previously in paragraph 7.3.1 and 7.3.2 respectively have been used on the data sets as described previously in paragraph 8.2 in the GIS research scenarios as defined earlier in paragraph 7.2. According to the methodology described in paragraph 7.4.4, 7.5 and 7.6 the output of these GIMs have been evaluated and compared in each GIS research scenario.
The results can be found in Table 8.5 and Table 8.6 on the next page, detailing the absolute observed values and normalized values for each evaluation metric respectively. A few observations considering the values presented in Table 8.5 and Table 8.6 can be made, as will be done on the next page as well. It has to be noted that the hybrid-user method has been excluded from this thesis research because the privacy policy of Twitter

changed during the analysis of this method (Twitter, 2017r). Due to this change of policy metadata necessarily to use the hybrid-user method could not be derived anymore because access was restricted. Given that the other two GIMs were not analysed under these new policies comparing the three methods would simply be unfair to the part of the hybrid-user method, being analysed within a more restricted context. To what extent this decision has impacted the validity of the thesis research will be discussed later in paragraph 11.3.

Metric	Sub-metric	Content-user method				Network-user method			
		D.M.	H.M.	T.M.	Avg.	D.M.	H.M.	T.M.	Avg.
Reliability	Spatial reliability	0.267	0.187	0.165	0.206	**0.493**	0.326	0.441	0.420
	Temporal reliability	**15.882**	17.337	16.405	16.541	16.713	24.080	24.438	21.743
Scale		**2.885**	2.782	2.791	2.819	2.711	2.209	2.643	2.521
Completeness		0.664	0.696	0.625	0.662	0.725	0.674	**0.792**	0.730
Effectiveness	Precision	**0.976**	0.975	0.966	0.972	0.636	0.544	0.474	0.551
	Recall	0.656	0.645	0.767	0.689	**0.801**	0.792	0.775	0.783
	F-measure	0.784	0.776	**0.855**	0.805	0.709	0.645	0.588	0.647
Speed		**4.499**	4.599	5.514	4.871	12.700	12.270	12.973	12.648

Table 8.5: Absolute observed values for evaluations metrics

Metric	Content-user method				Network-user method			
	D.M.	H.M.	T.M.	Avg.	D.M.	H.M.	T.M.	Avg.
Reliability	0.630	0.590	0.579	0.600	**0.743**	0.658	0.715	0.705
Scale	**0.721**	0.696	0.698	0.705	0.678	0.552	0.661	0.630
Completeness	0.664	0.696	0.625	0.662	0.725	0.674	**0.792**	0.730
Effectiveness	0.784	0.776	**0.855**	0.805	0.709	0.645	0.588	0.647
Speed	**0.948**	0.941	0.875	0.921	0.360	0.391	0.340	0.364
Overall	**0.750**	0.740	0.726	0.739	0.643	0.584	0.619	0.615

Table 8.6: Normalized observed values for evaluations metrics and totals

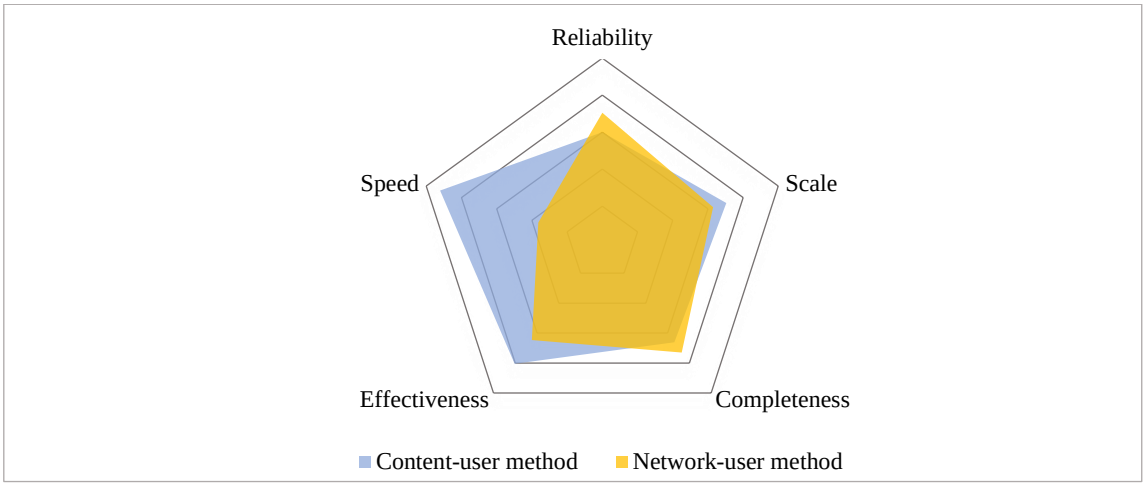

Figure 8.3: Normalized evaluation metric averages for GIMs examined in thesis research

Considering reliability, the network-user method's performance is the best overall as can be seen in Table 8.6 and Figure 8.3 above specifically. Differences can be observed when comparing both methods based on their spatial reliability and temporal reliability, however. While the network-user method has the best spatial reliability due to being able to infer 21.4% more user locations with an error distance less than 100 kilometres on average, the content-user method has the best temporal reliability with user locations being 5.2 days more topical on average as can be seen in Table 8.5 above specifically. A possible reason for this difference might be that the user locations inferred by the network-user method are based on social networks who tend to have a higher proximity overall, as detailed previously in paragraph 1.4, and therefore a higher spatial reliability is found compared to the content-user method as well. At the same time, with the content-user method being based on occurrences of toponyms in tweets primarily and the temporal reliability of the methods being based on the latest occurrence of the inferred user locations in tweets as well, the higher temporal reliability found for the content-user method can be explained as well. In the case of both the

content-user and network-user method, both the spatial and temporal reliability is best for the disaster management research scenario. It is difficult to determine the exact reason behind this is. A possible reason might be the fact that a different case study area was set for the disaster management scenario (California instead of the contiguous United States) compared to the other two GIS research scenarios and the behaviour of the Twitter users in that particular study area was different as well. If users in that particular area tweet more often and are more active on the social medium in general for example, this can considerably increase the amount of content or size of the social network from which the user locations can be derived from. This would then explain the higher reliability found for the disaster management research scenario compared to the reliability found for the to the other GIS research scenarios.

Considering scale, the content-user method performs best with an overall scale of 2.819, which is just above a city-level as previously explained in Table 7.8 in paragraph 7.4.2. While the overall normalized scale levels of the data output of the content-user method in each GIS research scenario do not differ that much (0.023 at max), bigger differences exist among the normalized values found for the network-user method (0.126 at max). This is especially the case when comparing the disaster management research scenario to the health management research scenario. The exact reason behind this is difficult to determine. A possible reason might be that the users part of the health management research scenario data set define user locations on a relatively higher scale level compared to users part of the disaster management and topic modelling research scenario, resulting in a higher overall scale level as well. When calculating the overall scale of the standardized user-specified user locations of each respective data set used through a Python script it is found that in fact this is true. This script is run within the Canopy software package as detailed in Appendix IV.1. The script itself has been detailed in Appendix III.15. While for the disaster management and topic modelling data sets an overall scale of 3.063 and 2.905 is found respectively, the overall scale for the health management data set is 2.757. While these relative differences are not as big as the scale level differences found for the overall scales in Table 8.5 and 8.6 on the previous page for the network-user method, this pattern is definitely interesting and possibly indicates a correlation explaining the differences among the GIS research scenarios found.

Considering completeness, the network-user method performs best. Differences between the different GIS research scenarios exists, with a value of 0.071 at max found for the content-user method and a value of 0.118 at max found for the network-user method. The differences can possibly be explained due the different content of the tweets posted by users part of the data sets and different structures of social networks these users are part of. This would then directly influence the number of toponyms that can be derived from this content and social network respectively.

Considering effectiveness, the content-user method performs best overall. When looking at the two measures that make up the effectiveness metric (precision and recall respectively), it is interesting to see that while the content-user method has the best precision, the network-user method has the best recall. In other words, when using the content-user method less false positives are found but when using the network-user method less false negatives are found, as reasoned through the formulas as detailed earlier in paragraph 7.4.2. A plausible reason is that whether a positive is true or false is primarily based on the fact whether the inferred user location is mentioned in the tweets of this user as well. Since the user locations are inferred through their occurrence in tweets primarily when using the content-user method, it is obvious that less false positives are found compared to the output of the network-user method. This has its downside, given that any toponym mentioned is taken into consideration for the possible user location to be inferred causing a bigger chance of these inferred user location to be false negatives. This is in a lesser extent the case for the network-user method because users tend to specify user locations that are true to their real geolocations, apparently, as can be reasoned through the higher spatial reliability found for the network-user method compared to the content-user method. While values found for the precision for the content-user method are similar (with a difference of 0.010 at max), the precisions of the network-user method are not (with a difference of 0.162 at max). A possible reason might be that due the different content of tweets posted by users that are part of the data sets differ and therefore result in a different number of false positives being found as well, given that the latter correlates whether the inferred user location is mentioned within tweets as previously explained in this paragraph. At the same time, the values found for the recall of the network-user method are similar (with a difference of 0.026 at max) while the values found for the content-user method are not (with a difference of 0.122 at max). This is possibly due the same reason as explained for the differences of the precision metric found for the network-user method as well.

Considering speed, the content-user method performs best with this particular method being approximately 7.8 seconds faster than the network-user method per user location inference. The differences among the GIS research scenarios are small, except the absolute value found for the topic modelling research scenario when using the content-user method as presented in Table 8.5 on the previous page. The reason for this might be that at the time the data was being inferred using this method the APIs used within this method responded more slowly due to unknown reasons.

8.3.2 Sensitivity analysis

As described in paragraph 7.4.3 earlier in this thesis report, two types of sensitivity analysis will be conducted. During these sensitivity analyses, values of parameters of the GIMs and weight scenarios respectively will be altered to determine whether one of these parameters or evaluation metrics has a big influence on the final value found for the total performance of the GIMs. Whether this is the case can then be taken into consideration when interpreting the results found and developing conclusions based on these results. The data set used for the sensitivity analysis has been created by taking all the three sample data sets for each respective GIS research scenario (as defined previously in paragraph 8.2) and get a random sample of 1000 from this newly compiled data set through the same method as described earlier in paragraph 8.2 as well. This way a "neutral" data set is created to determine the sensitivity of the GIM parameters and weight scenarios. The values found for the first sensitivity analysis have been presented in Table 8.7 and 8.8 below and illustrated in Figure 8.4 and 8.5 on the next page. A few observations considering the values presented in Table 8.7 and Table 8.8 can be made, as will be done below and on the next page as well.

Metric	Sub-metric	Content-user method			Network-user method		
		Low	Mid	High	Low	Mid	High
Reliability	Spatial reliability	**0.208**	0.191	0.155	**0.601**	0.592	0.574
	Temporal reliability	**15.104**	17.401	15.313	22.566	24.804	**21.064**
Scale		2.808	2.809	**2.864**	2.601	2.622	**2.655**
Completeness		**0.773**	0.645	0.394	0.640	**0.728**	0.555
Effectiveness	Precision	0.965	0.956	**0.972**	**0.587**	0.558	0.595
	Recall	**0.711**	0.660	0.634	**0.802**	0.785	0.758
	F-measure	**0.811**	0.780	0.768	**0.678**	0.652	0.667
Speed		5.787	5.174	**5.957**	13.204	**12.975**	13.106

Table 8.7: Absolute observed values for first sensitivity analysis

Metric	Content-user method			Network-user method		
	Low	Mid	High	Low	Mid	High
Reliability	**0.601**	0.592	0.574	0.677	0.703	0.656
Scale	0.702	0.703	**0.716**	0.650	0.656	0.664
Completeness	**0.773**	0.645	0.394	0.640	0.728	0.555
Effectiveness	**0.811**	0.780	0.768	0.678	0.652	0.667
Speed	0.856	**0.900**	0.844	0.328	0.340	0.331
Overall	**0.749**	0.724	0.659	0.594	0.616	0.575
Difference	+0.025		**-0.065**	-0.022		-0.041

Table 8.8: Normalized values for first sensitivity analysis

When altering the parameters within the GIMs, certain values found for the evaluation metrics change while others do not. The first interesting observation to be made is that when setting the parameters high, this does not necessarily result into a better performance of the GIMs. Especially the spatial reliability and completeness of the data output are affected by the alternations set. For both methods, the spatial reliability decreases when increasing the values used for the parameters as well. A potential reason might be that whether user locations will be inferred is determined by stricter parameters, less user locations will be inferred in general as well due to them not meeting the minimal standards (as proved by the completeness of the data sets in Table 8.7 and 8.8 on the previous page). Since the spatial reliability is based on the amount of inferred user locations with an error distance less than 100 kilometres, this value becomes lower due this reason as well. When increasing the values used for the parameters in the GIMs, different patterns are seen for each respective method. While the completeness of the data output of the content-user method decreases continuously, the completeness levels of the data output of the network-user method does not. Setting the parameter values for this method either lower or higher both results in a lower completeness as well.

This can be explained by the fact that when setting the parameters very high (at least 100 followers and friends are needed), less users are qualified to use the GIM on and are therefore their geolocation is not inferred as well. When setting all parameters very low, there is potentially not enough content available to derive a user location with certainty to which no geolocation is inferred as well. The scale, effectiveness and speed of the GIMs are hardly affected by the alterations set, with none of the normalized values decreasing or increasing above the 0.1 mark.

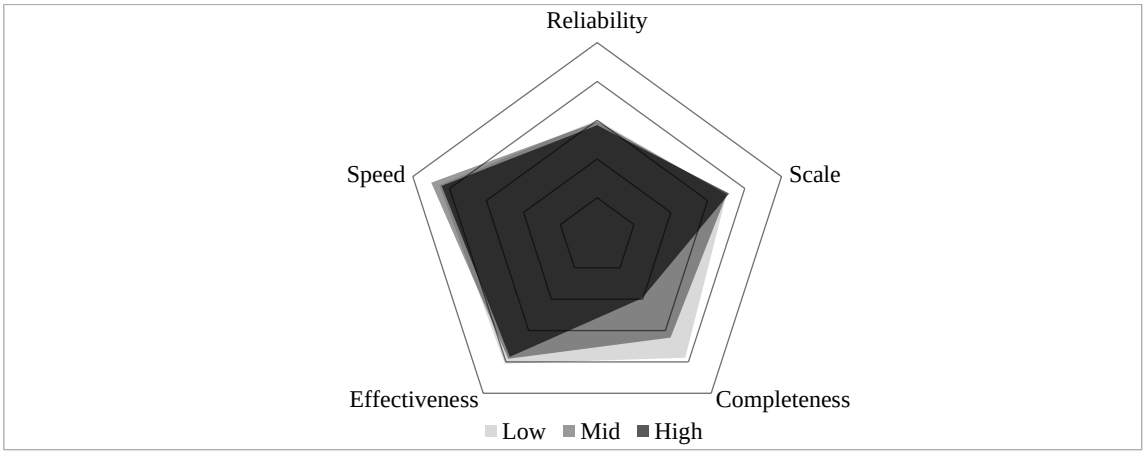

Figure 8.4: Normalized values found for content-user method in first sensitivity analysis

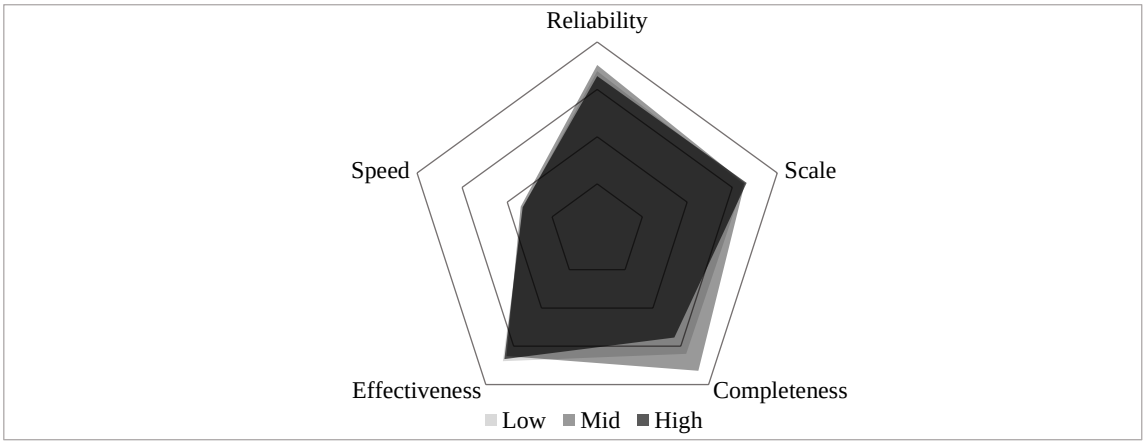

Figure 8.5: Normalized values found for network-user method in first sensitivity analysis

Metric	W.S. 1	W.S. 2	W.S. 3	W.S. 4	W.S. 5	W.S. 6
Reliability	0.592	1.184	0.444	0.444	0.444	0.444
Scale	0.702	0.527	1.405	0.527	0.527	0.527
Completeness	0.645	0.484	0.484	1.290	0.484	0.484
Effectiveness	0.780	0.585	0.585	0.585	1.560	0.585
Speed	0.900	0.675	0.675	0.675	0.675	1.799
Overall	0.724	0.691	0.718	0.704	0.738	0.768
Difference		-0.033	-0.006	-0.020	+0.014	**+0.044**

Table 8.9: Observed values for content-user method in second sensitivity analysis

Metric	W.S. 1	W.S. 2	W.S. 3	W.S. 4	W.S. 5	W.S. 6
Reliability	0.703	1.406	0.527	0.527	0.527	0.527
Scale	0.656	0.492	1.311	0.492	0.492	0.492
Completeness	0.728	0.546	0.546	1.456	0.546	0.546
Effectiveness	0.652	0.489	0.489	0.489	1.304	0.489
Speed	0.340	0.255	0.255	0.255	0.255	0.680
Overall	0.616	0.638	0.626	0.644	0.625	0.547
Difference		+ 0.022	+ 0.010	+ 0.028	+ 0.009	**-0.069**

Table 8.10: Observed values for network-user method in second sensitivity analysis

The values found for the second sensitivity analysis have been presented in Table 8.9 and 8.10 on the previous page. It has to be noted that for the second sensitivity analysis only the normalized values for the neutral data set as presented in Table 8.6 are used. A few observations considered the values presented in Table 8.9 and Table 8.10 can be made, as will be done below.

As can be seen in Table 8.9 and 8.10 on the previous page the differences between the values found for the overall performance in each weight scenarios are small, ranging from -0.069 to +0.044 when incorporating the overall performance values found for both methods. The biggest difference compared to the "neutral" overall performance value is perceived when the weight for the speed value is increased. The reason for this is the fact that the normalized value for the evaluation metric for the content-user and network-user method is the highest and lowest of all normalized values found for the evaluation metrics for those specific GIMs respectively. The influence of these values is therefore greater on the overall performance compared to the evaluation metrics because these values tend to be more "extreme". Given that the differences between the overall performance values are small as described earlier, the overall performance value as presented in Table 8.5 and 8.6 can be interpreted with trust.

8.4 Summary

With the GIMs being compared within each respective GIS research scenarios, a set of conclusions can be made considering the performance of these GIMs by taken the results found in paragraph 8.2 and 8.3 earlier in the thesis report into account:

- When pre-processing the data sets to be used in the GIS research scenarios, some similarities and differences exists on the amount these data sets are affected by the pre-processing steps. The data sets are similarly affected when filtering the data sets by language and the presence of a user-specified user location. The disaster management data set is especially affected when deriving unique users and clipping the data set by catchment area. The topic modelling data set is especially affected when standardizing the user-specified user locations. The spatial distribution of the users part of the data sets is similar to the spatial distribution of the real population of the contiguous United States. Concluding, to what degree the data sets are affected by each of the pre-processing steps seems to be dependent on the content of these data sets.

- The content-user method provides good temporal reliability, a low scale level, high precision and high speed in particular. At the same time the network-user method provides good spatial reliability, a high completeness of the data and a high recall.

- In some cases, differences in performance of the GIMs exists among the GIS research scenarios. In the case of the content-user method differences are especially apparent considering spatial reliability, recall and speed. In the case of the network-user method differences are especially apparent considering spatial reliability, temporal reliability, scale and precision.

- The sensitivity of both methods is low. When altering the parameters within the GIMs, especially the spatial reliability and completeness of the data output is affected even though these changes are small and cannot be considered considerably sensitive. Other evaluation metrics cannot be considered to be sensitive to these alternations. When altering the weight of each evaluation metric, the total performance is especially under the influence of the speed metric even though the difference compared to the original total performance is small and can therefore not be considered considerably sensitive.

9. Sub question 5: Evaluating and comparing GIMs to original data output

9.1 Introduction

The fifth sub question will be answered in this chapter by performing an analysis in which the performances of the GIMs as found in chapter 8 will be evaluated and compared to the unprocessed[55] Twitter API output in the GIS research scenarios using the evaluation and comparison metrics defined earlier in paragraph 7.2 and 7.4.2 respectively. This chapter has a similar structure as chapter 8, with a description of the data sets used in this analysis being given first. Secondly, the results found will be presented in tables and figures. Descriptive analysis will be performed where needed and arguments made will be supported with academic literature where needed as well. Next, a sensitivity analysis will be performed as defined earlier in paragraph 7.4.3. Finally, all findings will be summarized at the end of this chapter.

9.2 Data set descriptions

As described earlier in paragraph 7.2, three data sets based on an equal amount of GIS research scenarios have been created to evaluate and compare the performances of the GIMs presented earlier in paragraph 7.3. Before looking into the performances of the GIMs within these GIS research scenarios attention will be given on the way these data sets are structured and how they have been created In Table 9.1 below the corpus sizes as a result of each pre-processing step as defined earlier in paragraph 7.2.4 have been given for each application domain part of this thesis research respectively. It has to be noted that only the pre-processing steps that affected the corpus sizes have been included in the table below. In this table "Tot. (n)" represents the absolute number of rows in the data sets, "Dec. (%)" the relative number of rows decreased after conducting each pre-processing step and "Tot. (%)" the relative number of rows in the data sets after conducting each pre-processing step compared to the original data set corpus sizes. The value found for each pre-processing step in each column has been made bold. The geographical distribution of the data sets has been visualised in Figure 9.1 on the next page.

Processing step	Disaster management			Health management			Topic modelling		
	Tot. (n)	Dec. %	Tot. (%)	Tot. (n)	Dec. (%)	Tot. (%)	Tot. (n)	Dec. (%)	Tot. (%)
Gather data	**262226**		100	15536		100	30085		100
Derive geotagged	**7355**	-97.2	2.8	22	**-99.9**	0.14	8	**-99.9**	0.03
Clip by area	**5095**	**-30.7**	1.9	22	-0.0	0.14	8	-0.0	0.03
Filter by metadata	**4943**	-3.0	1.9	22	-0.0	0.14	6	**-25.0**	0.02

Table 9.1: Corpus sizes during pre-processing steps of Twitter data sets used in thesis research

When looking at the table above a few observations can be made concerning the similarities and differences in which the data sets are structured during each pre-processing stage of this thesis research, as described below and on the next page:

- Concerning all pre-processing steps, the corpus size of the disaster management data set is considerably higher than the ones used in the health management and topic modelling research scenarios and shows a similar hierarchy as found in academic literature. This point has previously been made in paragraph 8.2 as well.

- The number of geotagged tweets differs per data set but is low overall. While 2.8% of all tweets part of the disaster management data set is geotagged, only 0.14% and 0.03% of the tweets part of the health management and topic modelling data sets are geotagged respectively. This explains the high values found for Dec.(%) as well in the "Derive geotagged" processing step as well.
  These observations indicate that the amount of geotagged tweets part of data sets differs when the content of the tweets is different as well. It also shows that, would the health management and topic modelling research scenario be real, GIMs would be essential to be able to conduct the research successfully due to an otherwise lack of geographical data to base any conclusions on.

[55] Not that while deriving geotagged tweets from the original data sets is through processing the data, the data being "unprocessed" in this case means that the data has not been processed with the use of a GIM.

- When clipping the data by study area, only the corpus size of the disaster management data set is affected. Given that the corpus sizes of the health management and topic modelling data sets are very low it is difficult to determine whether this is due chance or due the fact that people outside of the (contiguous) United States do not (or hardly) geotag their tweets.

- When filtering the data based on metadata attributes (in particular language used in tweets and by users), only the corpus sizes of the disaster management and topic modelling data sets are affected. Due the low corpus sizes found for the health management and topic modelling data sets is difficult to determine whether this is due to different distribution of languages used among data sets or this is due chance.

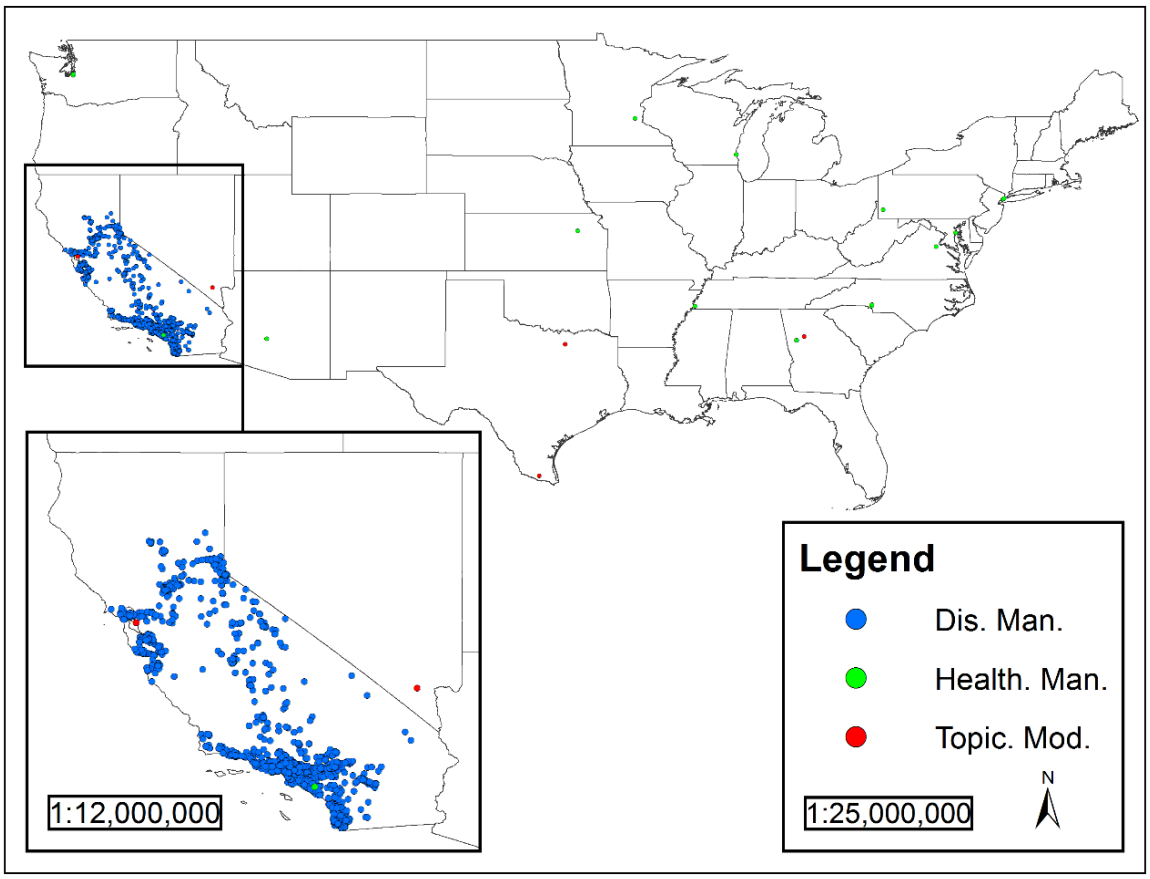

Figure 9.1: Geographical distribution of geotagged tweets in Twitter data sets used in thesis research

9.3 Results

9.3.1 Regular analysis

With the data sets being described in the previous paragraph, the results found through analysis will be discussed now. According to the methodology previously described in paragraph 7.4.3, 7.5 and 7.6 the output of these GIMs will be compared to the unprocessed Twitter API output. The results can be found in Table 9.2 and Table 9.3 on the next page, detailing the absolute observed values and normalized observed values for each evaluation metric respectively. It has to be noted that the values found for the content-user and network-user method have been excluded since they have already been detailed previously in Table 8.5 and Table 8.6 and illustrated in Figure 8.3. A few observations considered the values presented in Table 9.2 and Table 9.3 can be made, as will be done below on the next page as well.

When comparing the unprocessed data output with the output of the GIMs, it seems that the performance of the unprocessed data output is best. On all evaluation metrics the unprocessed data output scores best and the same among all GIS research scenarios, except for the completeness of the data. The reason why the scores for all GIS research scenarios are the same for most evaluation metrics is that the performance of these

evaluation metrics is based on the structure of the data and not the content of the data itself, which was the case for the GIMs evaluated in this thesis research.

Metric	Sub metric	Unprocessed data output			
		D.M.	H.M.	T.M.	Avg.
Reliability	Spatial reliability	**1.000**	**1.000**	**1.000**	**1.000**
	Temporal reliability	**0.000**	**0.000**	**0.000**	**0.000**
Scale		**4.000**	**4.000**	**4.000**	**4.000**
Completeness		0.019	0.001	0.000	0.006
Effectiveness	Precision	**1.000**	**1.000**	**1.000**	**1.000**
	Recall	**1.000**	**1.000**	**1.000**	**1.000**
	F-measure	**1.000**	**1.000**	**1.000**	**1.000**
Speed		**0.000**	**0.000**	**0.000**	**0.000**

Table 9.2: Absolute observed values for evaluation metrics

Metric	Unprocessed data output				Content-user method				Network-user method			
	D.M.	H.M.	T.M.	Avg.	D.M	H.M.	T.M.	Avg.	D.M.	H.M.	T.M.	Avg.
Reliability	**1.000**	**1.000**	**1.000**	**1.000**	0.630	0.590	0.579	0.600	0.743	0.658	0.716	0.706
Scale	**1.000**	**1.000**	**1.000**	**1.000**	0.721	0.696	0.698	0.705	0.678	0.552	0.660	0.630
Completeness	0.019	0.001	0.000	0.006	0.664	0.696	0.625	0.661	0.725	0.674	**0.792**	0.730
Effectiveness	**1.000**	**1.000**	**1.000**	**1.000**	0.784	0.776	0.855	0.805	0.709	0.645	0.588	0.647
Speed	**1.000**	**1.000**	**1.000**	**1.000**	0.746	0.740	0.689	0.725	0.283	0.308	0.268	0.286
Overall	**0.804**	**0.800**	**0.800**	**0.801**	0.709	0.700	0.689	0.699	0.628	0.567	0.604	0.600

Table 9.3: Normalized observed values for evaluations metrics and totals

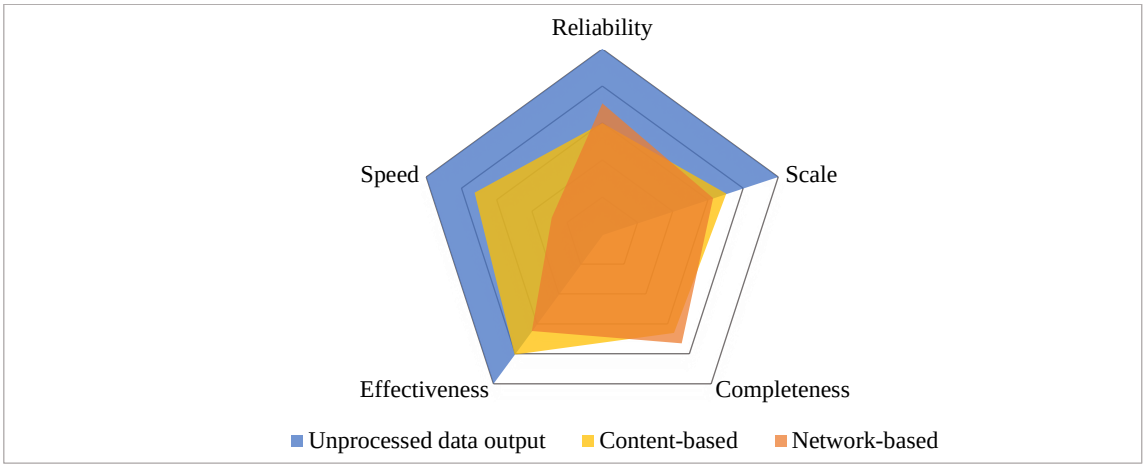

Figure 9.2: Normalized evaluation metric averages for GIMs and unprocessed Twitter data

The completeness of the data sets differs among the different GIS research scenarios, as also described earlier in paragraph 9.2. The reason for this is difficult to determine. A possible reason is that the phenomena subject to the disaster management research scenario typically occurs outside while the phenomena subject to the health management and topic modelling research scenarios do not. When a user feels a certain sentiment and wants to express it on Twitter (immediately), it will probably use a mobile phone over a desktop computer given that the latter is not easy to take outside. Given the wider availability of ways to use GPS on mobile phones compared to desktop computers, it is possible that this is the reason that more geotagged tweets are found in this specific data set as well. Another reason might be that a different case study area was specified for the disaster management research scenario and the users residing in this study area typically geotag their tweet more often than the users residing in the study area set for the other GIS research scenarios. This is probably not the case, given that for the health management and topic modelling research scenarios also hardly any tweets are found within the area specified for the disaster management research scenario.

As can be concluded so far, the unprocessed Twitter data performs best compared to the data output of both GIMs. With this in mind, one could conclude that using GIMs is useless because it does not increase the applicability of the Twitter data but actually decreases it overall. To determine whether this is actually the case a sensitivity analysis will be performed in the next paragraph to see whether the performance values found are sensitive to the weights added to these values or not.

9.3.2 Sensitivity analysis

To ensure whether the results found as a result of the analysis conducted earlier in paragraph 9.3.1 can be interpreted with trust a sensitivity analysis has been performed. It has to be noted that only the second sensitivity analysis as described earlier in paragraph 7.4.3 has been conducted due the fact deriving geotagged tweets is not met with setting any parameters. For this analysis, the average normalized values found for the evaluation metrics as detailed in Table 9.3 on the previous page have been used and differently weighted. It is assumed that these averages are representative for all GIS research scenarios as a whole. The results of the sensitivity analysis have been presented in Table 9.4 to 9.6 below and will be interpreted below and on the next page as well.

Metric	W.S. 1	W.S. 2	W.S. 3	W.S. 4	W.S. 5	W.S. 6
Reliability	1.000	2.000	0.750	0.750	0.750	0.750
Scale	1.000	0.750	2.000	0.750	0.750	0.750
Completeness	0.006	0.005	0.005	0.012	0.005	0.005
Effectiveness	1.000	0.750	0.750	0.750	2.000	0.750
Speed	1.000	0.750	0.750	0.750	0.750	2.000
Overall	0.801	0.851	0.851	0.602	0.851	0.851
Difference		+0.050	+0.050	**-0.200**	+0.050	+0.050

Table 9.4: Observed values for unprocessed Twitter data in sensitivity analysis

Metric	W.S. 1	W.S. 2	W.S. 3	W.S. 4	W.S. 5	W.S. 6
Reliability	0.600	1.200	0.450	0.450	0.450	0.450
Scale	0.705	0.528	1.410	0.528	0.528	0.528
Completeness	0.661	0.496	0.496	1.322	0.496	0.496
Effectiveness	0.805	0.604	0.604	0.604	1.610	0.604
Speed	0.725	0.544	0.544	0.544	0.544	1.450
Overall	0.699	0.674	0.701	0.690	0.626	0.706
Difference		-0.025	+0.002	-0.009	**-0.073**	+0.007

Table 9.5: Observed values for content-user method in sensitivity analysis

Metric	W.S. 1	W.S. 2	W.S. 3	W.S. 4	W.S. 5	W.S. 6
Reliability	0.706	1.412	0.530	0.530	0.530	0.530
Scale	0.630	0.473	1.260	0.473	0.473	0.473
Completeness	0.730	0.548	0.548	1.460	0.548	0.548
Effectiveness	0.647	0.485	0.485	0.485	1.294	0.485
Speed	0.286	0.215	0.215	0.215	0.215	0.572
Overall	0.600	0.623	0.608	0.633	0.612	0.522
Difference		+0.023	+0.008	+0.033	+0.012	**-0.078**

Table 9.6: Observed values for network-user method in sensitivity analysis

Some similarities and differences can be observed among the tables as presented above. In most cases, the sensitivity of the different weight scenarios is low (below the +/- 0.1 mark). In the case of the content-user and network-user method this goes for all weight scenarios, similar to the findings as presented in paragraph 8.3.3 earlier in the thesis report. In the case of the content-user method the biggest change is observed when weighting the effectiveness of the GIM more heavily, while in the case of the network-user method this is the case for the speed metric. The reason for this is the fact that the normalized value for these respective evaluation metrics is the highest and lowest of all normalized values found for the evaluation metrics for those specific GIMs respectively as well. The influence of these values is therefore greater on the overall performance compared to the evaluation metrics because these values tend to be more "extreme". However, it can be concluded that the sensitivity of the content-user and network-user method is low even if

the values for the evaluation metrics found are renormalized according to the unprocessed Twitter API output because these changes do not exceed the 0.1 mark as previously mentioned above.

When looking at the sensitivity of the different weights in Table 9.4 it can be seen that the sensitivity of the scenario in which the completeness of the data is weighted more heavily the overall performance of the data output decreases dramatically with 0.200 points, meaning that this particular evaluation metric is met with a big sensitivity. When comparing the values found for the content-user and network-user method for the same weight scenario it turns out that instead of the unprocessed Twitter API output performing best, it performs the worst. This fundamentally influences the way in which the overall performances found for the unprocessed Twitter API and GIMs part of the thesis research should be interpreted. They should be interpreted based on the aim of the research the data is used in rather than assuming that this performance is representative for all possible research scenarios. This argument will be further discussed in paragraph 11.1.2 later in this thesis report.

9.4 Summary

With the GIMs being compared to the unprocessed data output of the Twitter API within each respective GIS research scenarios, a set of conclusions can be made considering the performance of these GIMs by taken the results found in paragraph 9.2 and 9.3 earlier into account:

- The number of geotagged tweets found differs heavily per GIS research scenario, with the disaster management research scenario providing the most georeferenced tweets. Just 0.69% of all tweets is geotagged when taking the average percentages found for each GIS research scenario into account.

- When weighting all evaluation metrics equally, it is found that the performance of the unprocessed Twitter data performs best compared to the data output of both GIMs. This suggests that using GIMs is obsolete because they do not improve the overall performance of the data sets.

- When performing sensitivity analysis however, it turned out that the overall performance is heavily affected when weighting the completeness of the data sets more heavily. By doing this both GIMs are observed to have a better performance than the unprocessed data sets.

10. Conclusion

With the first and second part of the research as detailed and argued in paragraph 2.2 and paragraph 2.6. previously, being conducted according the methodologies presented in chapter 3 and 7 respectively and presented in chapter 4 to 6 and 8 to 9 respectively as well, the sub questions and central question can now be answered. This will be done in the current chapter of the thesis report. It has to be noted that these conclusions will not be discussed further in the current chapter but in paragraph 11.1 later in the thesis report.

10.1 Sub questions

To answer the central question, five additional sub questions will be answered. These sub questions have been previously detailed in paragraph 2.2 and will be answered below and on the next pages:

What are currently the most frequently and relevant used types of application of Twitter data in GIS research and how is this research structured?

Between 2013 and 2016 the most frequently used type of application methodology of Twitter data in GIS research was event detection. The application of the majority of the articles lied either within the disaster management, health management or topic modelling domain. The research structure varied among each application domain. Disaster management research primarily used real-time data, additional sources, had a short period of data gathering and was either set on a sub-national or city-level scale. Health management and topic modelling research had similar research structures. They primarily did not use real-time data or additional sources, had a relatively long period of data gathering and were set on a national scale level. Topic modelling research used relatively big data sets compared to disaster- and health management research, whose data sets used in research had similar corpus sizes. The only similarity between these three most often used application domains was that the contiguous United States was set as the (primary) study area.

What are the benefits and drawbacks of using Twitter data in these research applications?

For the usability of Twitter data in GIS research applications it was found that each benefit also automatically resulted in an interrelated drawback. These benefits and drawbacks are illustrated below in Figure 10.1:

Benefits		**Drawbacks**
Big data quantity	↔	Questionable data quality
Easy data access	↔	Rate limitations
Open nature of Twitter	↔	Privacy issues for Twitter users
Great academic interest	↔	Potential lack of technical expertise
Georeferencing potential	↔	Poor georeferencing in practice

Figure 10.1: Benefits and drawbacks of Twitter data usage in GIS research

What geolocation inference methodologies of Twitter data currently exist and how are their workflows structured?

The GIMs currently available to infer Twitter data are primarily either content-based, network-based or a hybrid of these two variations. Content-based GIMs are either focussed on inferring messages or user locations while network-based and hybrid GIMs are almost exclusively focussed on inferring user locations. Content-based GIMs primarily use text mining to infer the geolocation of users or tweets while network-

based GIMs primarily use tie-strength. Methodologies used for hybrid GIMs vary heavily, ranging from machine learning to GIS. The contiguous United States was most often chosen as the (primary) study area to test the GIMs as presented in academic research. Within the majority of the articles it was not specified what language was used by the users part of these data sets. The accuracy and effectiveness of the GIMs presented in academic literature was measured using different statistic measures and methods among these articles. Therefore, it was not possible to make overall statements on the differences in accuracy and effectiveness of the different types of GIMs through the selected academic literature.

What are the strengths and weaknesses concerning the applicability of these methodologies in event detection using Twitter data?

The two GIMs evaluated and compared in this thesis research both have their strengths and weaknesses, as illustrated in Figure 10.2 and 10.3 below:

Content-user method		**Network-user method**	
Strengths	*Weaknesses*	*Strengths*	*Weaknesses*
Temporal reliability	Spatial reliability	Spatial reliability	Temporal reliability
Scale	Completeness	Completeness	Scale
Precision	Recall	Recall	Precision
Speed			Speed

Figure 10.2: Strengths and weaknesses of GIMs part of the thesis research

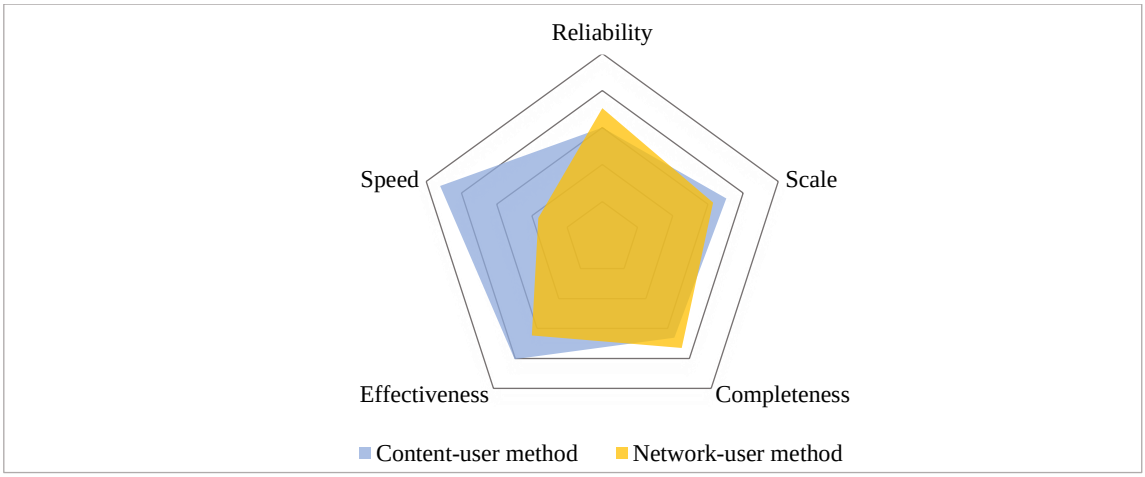

Figure 10.3: Normalized evaluation metric averages for GIMs examined in thesis research

Among the different GIS research scenarios different values for the evaluation metrics were found in some occasions as well. In the case of the content-user method differences were apparent considering spatial reliability, recall and speed while in the case of the network-user method differences were apparent considering the spatial reliability, temporal reliability, scale and precision in particular. The sensitivity of the parameters set within both GIMs and the weights attached to the evaluation metrics were low.
When comparing the overall performance of the GIMs tested in this thesis research, the content-user method performed best.

How does the geolocation inference methodologies data output compare to the geotagged Twitter data validity?

When comparing the unprocessed data output of the Twitter API to the output of the GIMs, the unprocessed data performs best when weighting all evaluation metrics equally. When weighting the completeness of the data more heavily however, both GIMs researched in this thesis research perform better. The average normalized values found for both GIMs and the unprocessed Twitter API output have been illustrated in Figure 10.4 below:

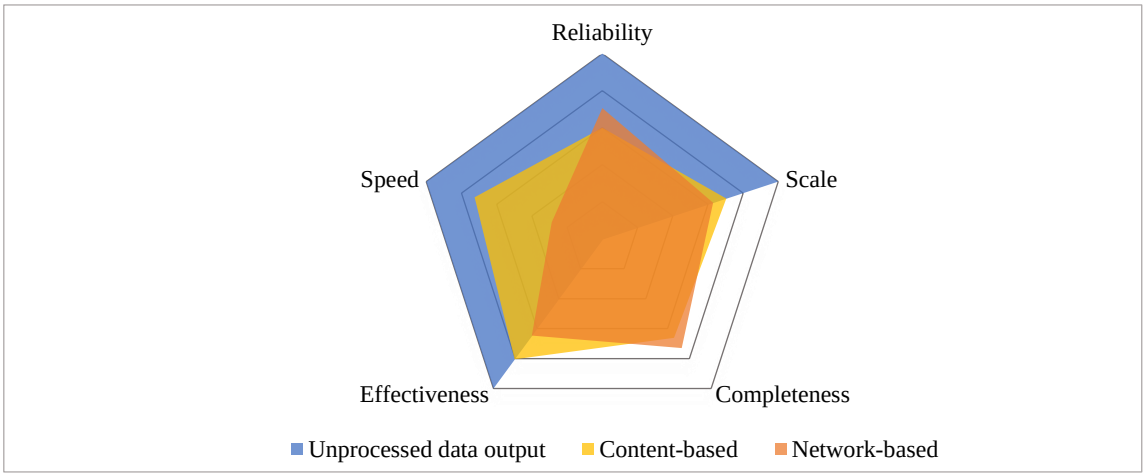

Figure 10.4: Normalized evaluation metric averages for GIMs and unprocessed Twitter data

10.2 Central question

The sub questions as answered in the previous paragraph will be used to answer the central question as specified in paragraph 2.2 earlier in this thesis report. This central question is the following:

To what extent can the usability of Twitter data in event detection research scenarios using this type of data be increased through the application of geolocation inference methodologies?

When using GIMs to increase the amount of geographical references in Twitter data sets to be used in GIS research, the increase of usability is a matter of compromise rather than an overall increase of data usability. While at the same time GIMs can potentially drastically increase the completeness of data sets, this goes at the cost of data quality characteristics such as reliability and truthfulness of the data. It has to be noted that the decrease of data quality does not rend the post-processed unusable but simply less accurate than the unprocessed Twitter data. Using GIMs is also met with a longer time needed to process Twitter data which might not be preferable when applying this type of data in a real-time application for example.
Concluding, the usability of Twitter data can especially be increased considering the completeness of the data at the cost of data quality to some degree. It depends on the aim of the research these methodologies are used in whether the data quality of the GIMs' output is sufficient or not, however.

11. Discussion

11.1 Interpretation of thesis research results and conclusions

In the previous chapter of the thesis report the sub questions and central question have been answered to which the thesis research has been concluded. These conclusions have not been interpreted yet, which will be done in the current chapter. Given that the thesis research consisted of two separate parts, these parts will be interpreted separately as well in paragraph 11.1.1 and 11.1.2 respectively. For each sub question the findings will be interpreted and compared to other academic research done with a similar aim. The similarities and differences between the thesis research and other academic research on the subject will be discussed in these paragraphs as well.

11.1.1 Research part 1

In the first part of the thesis research three literature studies were conducted to answer the first three sub questions respectively, as detailed in paragraph 2.2. previously. The main aim of the first sub question was to find out how Twitter data was used in GIS research and how this research was structured within the years 2013 to 2016. The findings have been presented previously in paragraph 10.1 in a concise way.
Similar literature studies have been performed by other academics in the field, even though these efforts are scarce. The literature study as performed by Steiger et al (2015) is of interest specifically because it is similar to the study conducted as part of this thesis research, given that it is also of a systematic nature.
Some similarities and differences exists when comparing the findings of Steiger et al with the findings of the literature study as discussed in chapter 4 earlier in this thesis report. These have been presented systematically below and discussed below and on the next page as well.

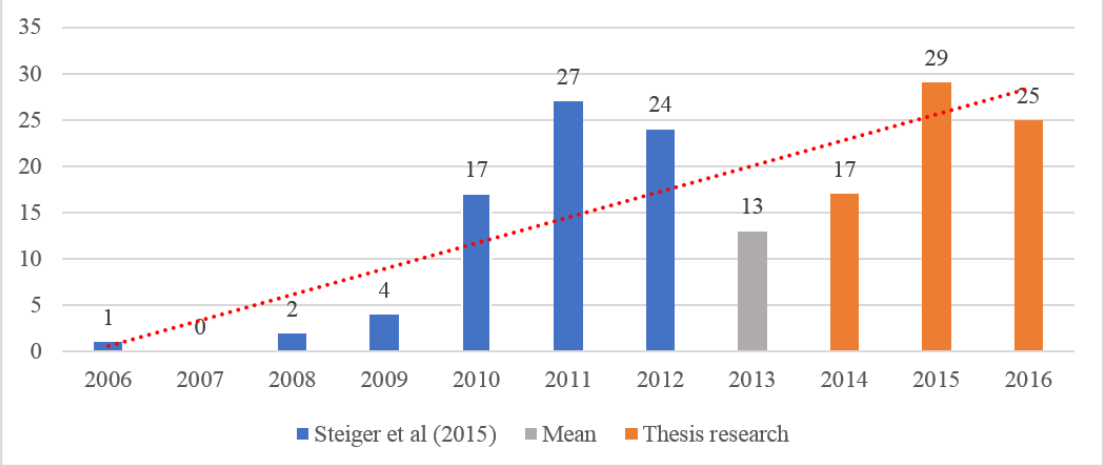

Figure 11.1: Number of articles combining Twitter data and GIS from 2006 to 2016

Even though the interest among the academic field to combine Twitter data and GIS seems to come in waves rather than grow continuously and constantly, the interest among the academic field seems to increase overall when combining the number of articles found per year for both literature studies. This has been illustrated effectively in Figure 11.1 above. Whether this trend will continue in the future is dependent on several factors. While the current user base of the social medium is still growing, this growth has decreased over the years (Statista, 2017b). It is plausible that at some point in the future Twitter's user base will decline to such extent that the service will not be able to remain active in its current form or even in any form.
Similar scenarios happened for other social media in the past such as MySpace, Digg and Friendster (see The Guardian, 2015; Financial Times, 2009; Bloomberg, 2011; Techradar, 2012; MIT Technology Review, 2012; Forbes, 2012; Mashable, 2014; Wired, 2013 for extensive journalistic writings on the subject). If this would be the case for Twitter as well, it would be natural for the academic interest in combining Twitter data and GIS would decrease as well given that this results primarily in a continuously decreasing amount of data to be derived from the social medium (among other negative effects). It does not necessarily mean that any knowledge gathered in research on Twitter data becomes obsolete, given that quite possibly a lot of this

knowledge can be applied on other social media as well. Whether the situation as illustrated here will actually occur in the future cannot be determined as of writing and will therefore not be done so.

The distribution of study areas specified in studies combining Twitter data and GIS differs. While Steiger et al (2015, p. 9) specified that New York City, NY, United States and Japan as a whole where among the most popular study areas in this field of research between the years 2006 and 2013, the conductor of the thesis research found that the most popular study areas where the (contiguous) United States and Greater London in research published between the years 2013 to 2016. The latter has been discussed earlier in paragraph 4.2 specifically. A plausible reason is due a more widespread popularity of using Twitter data in GIS research in recent years described in this thesis research compared to the period described by Steiger et al has caused a bigger diversity of the study areas specified in GIS research as well. Another difference perceived is that the geographical distribution of study areas specified in GIS research using Twitter data differs from the distribution of Twitter users worldwide. While recent academic literature on the subject focuses on either the (contiguous) United States or the Greater London area, big user bases also reside in countries scarcely set as a study area in academic research such as India, Indonesia and Japan (Statista, 2017d). This is problematic due the fact that knowledge found through research in which the (contiguous) United States or Greater London area is applied is not necessarily applicable for other study areas as well. This is the case because for example different languages are used in these other study areas while at the same time tweeting behaviour seems to differ per country, as detailed by Sloan and Morgan (2015) and previously discussed in paragraph 5.6 as well. To what extent this causes flaws in applications combining Twitter data in GIS is as of yet not researched and therefore difficult to determine as of writing.

Considering application methodologies, a similar hierarchy can be observed in both systematic literature studies in which event detection is the most popular method among academics. While in the literature study conducted by Steiger et al (2015, p. 12) the amount of studies on geolocation inference and SNA were equal, the conductor of the thesis research found that SNA was less popular than geolocation inference.
This difference is obvious given that SNA has been defined differently in the works of Steiger et al (2015, p. 16-17) and the thesis research. This has previously been argued in paragraph 4.3 earlier in this thesis report. When using the same definition as used by Steiger et al when classifying the articles used to answer the first sub question, it is found that the distribution found in this thesis research concerning application methodologies used in GIS research using Twitter data is very similar. This has effectively been illustrated in Figure 11.2 below. It can therefore be concluded that the different distribution of application methodologies among the article selection of the literature currently discussed are a result of a different definition of SNA rather than a decreasing interest in the field of SNA among academics.

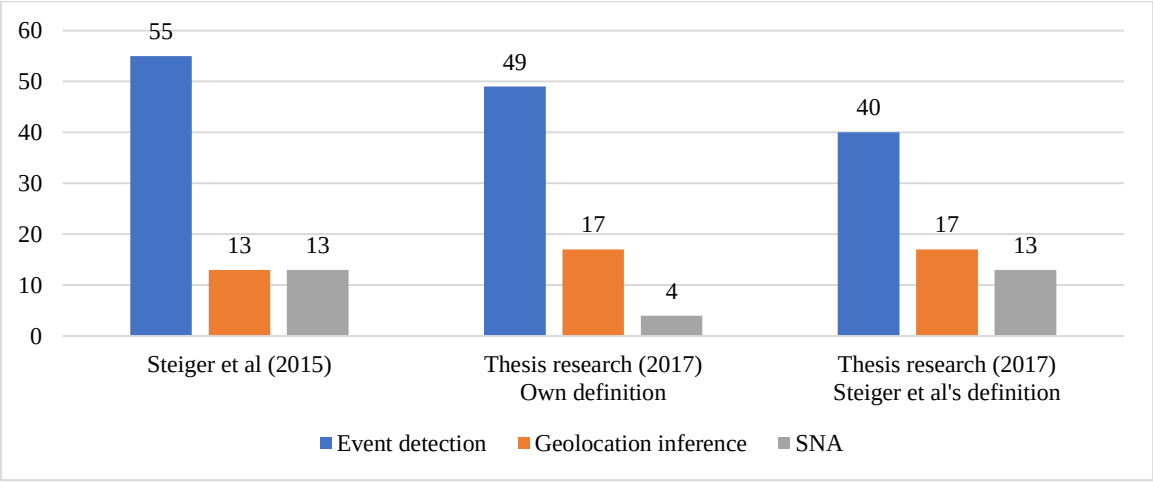

Figure 11.2: Distribution of application methodologies in academic literature part of literature studies

The distribution of the application domains among both literature studies differed as well. While Steiger et al (2015, p. 9, 13-16) found that disaster management, disease/health management and traffic management where among the most popular application domains in this field of research between the years 2006 and 2013 it was found in this thesis research that disaster management, health management and topic modelling were found to be the most popular application domains between the years 2013 and 2016 respectively. A possible

reason for this might be the different application domains definition used by the conductor of the thesis research and Steiger et al. In the case of the thesis research traffic management was not even considered as an application domain class, with traffic-related subject being classified in either crisis management (for example traffic accidents) and demographics (for example mobility). When classifying the final article selection for sub question 1 using the traffic management definition of Steiger et al, this application domain class still appears small compared to the top-3 most-often used application domains with only one study (by Gu et al, 2016) to be part of this application domain class. This implies that either the interest by academics to use Twitter data in combination with GIS in traffic management has plummeted or the article selection criteria caused this apparent lack of interest. The exact reason behind this can as of yet not be determined and should be explicitly researched before any arguments related can be made.

The main aim of the second sub question was to find out what the benefits and drawbacks of Twitter data in (GIS) research were. The findings have been presented in paragraph 10.1 in a concise way previously in the thesis report. In the past, no literature studies are known by other academics on the same subject in a similar nature as conducted as part of this thesis research. Therefore, the findings found through answering this sub question will not be compared to any academic literature as well but the findings themselves will be discussed here. Performing this literature study has been found to be important to the conductor of the thesis research because many of the knowledge gained from this respective literature study has either been implemented in the research design or become apparent during the conduction of the thesis research. Characteristics of Twitter data (gathering) such as rate limitations, assuring the privacy of the users part of the data sets and (lack of) technical expertise have been great factors in the research design as presented in this thesis report. In the future, it would be preferable to have a literature review available on the opportunities and limitations of Twitter data (in GIS research specifically) to enable academics to use these opportunities and work around the limitations as discussed in this thesis research as will be gone into in paragraph 11.2 later in this thesis report.

The main aim of the third sub question was to find out what GIMs are currently available and how these GIMs were structured. The findings have been presented in paragraph 10.1 in a concise way previously in the thesis report. Similar literature studies have been performed by other academics in the field, even though they are scarce and with a different nature as the literature study on the subject conducted as part of this thesis research. The literature studies as performed by Ajao et al (2015) and Jurgens et al (2015) are of interest to this thesis research specifically. Even though these literature studies are not of a systematic literature their main aim was to provide an overview on what GIMs where available at the time. In neither of the literature studies conducted by Ajao et al (2015) and Jurgens et al (2015) respectively, articles on GIMs have been classified in the way as has been done in this thesis research. Therefore, it is difficult to compare these three literature studies (including the one conducted as part of thesis research) with each other. What can be observed when comparing the GIMs discussed by either Ajao et al (2015, p.7), Jurgens et al (2015, p.3) and the conductor of the thesis research (in paragraph 6.5 earlier in this thesis report) is that the performance of GIMs presented in research have increased through time. This has effectively been illustrated in Table 11.1 below, in which statistics on the best-performing GIM discussed in each of the literature studies have been presented.

Author(s)	McGee et al	Ryoo & Moon	Laylavi et al
Year	2013	2014	2016
Discussed by	Jurgens et al (2015)	Ajao et al (2015)	Thesis research (2017)
GIM-type	Network-based	Other	Content-based
Inference subject	User location	User location	Message
General methodology	Tie-strength	Machine learning	Text mining
Amount inferred	100	100	87
Avg. E.D.	685	26.9	12.2

Table 11.1: Metadata on best-performing GIM in each respective literature study discussed

11.1.2 Research part 2

In the second part of the research several analyses were performed to answer the fourth and fifth sub question. The main aim of the fourth sub question was to determine the strengths and weaknesses of GIMs relevant to the thesis research. These were a content-based method and network-based method to infer users

specifically. It was found that each of these GIMs had specific strengths and weaknesses and behave differently among the GIS research scenarios they were evaluated by, as detailed earlier in paragraph 10.1. This knowledge is important to consider when determining what GIM will be most applicable for use in research and also depends on the aim and nature of the research. If within a certain research a big completeness and spatial reliability is preferred the network-user method would be the best option while if the aim would be different, say, a high processing speed would be preferred the content-user method would be the best option for example. This example shows that the performance of a GIM is also (highly) dependent on the nature and aim of the research it is used in and not necessarily of the performance of the GIM itself. Perhaps more interesting is the finding that the use of different data sets with different content also resulted in a different performance of the GIMs. In the case of the content-user method differences are especially apparent considering spatial reliability, recall and speed. In case of the network-user method differences are especially apparent considering spatial reliability, temporal reliability, scale and precision. This has previously been detailed in paragraph 8.4 as well. This means that using the same methods in different GIS research scenarios causes different performances as well. While the differences currently found in this thesis research are not very big within the application domains researched, potentially differences are bigger when comparing other application domains to each other not implemented in this thesis research. Therefore, further research on the subject is advised as will be done in paragraph 11.2 later.

The aim of the fifth sub question was to find out to what extent the output of the GIMs compared to the unprocessed data output of the Twitter API. It was found that overall the latter performed better but when weighting the completeness of the data more heavily the GIMs performed better. It was therefore concluded that the choice to use GIMs very much depends on the preferred data quantity and quality needs of the research these methods might be implemented in. This results somewhat in a Catch-22 in which researchers need to choose between either data quality at the cost of data quantity or vice versa. By the conductor of the thesis research it is argued that while data quality is incredibly important when performing research and ensure the validity of conclusions made, without a sufficient data quantity this research cannot be conducted since these conclusions cannot be made at all. Looking at the data sets used in the health management and topic modelling research scenarios for example, only 22 and 8 geotagged tweets were part of the unprocessed data sets respectively. This data quantity is so incredibly low that it would simply be impossible and irresponsible as an academic to make any conclusions from this type of data. This is especially the case for the health management research scenario given that one of the aims of this particular research is to increase the health of the population, which should not be done on doubtful conclusions. Using GIMs in these GIS research scenarios can be considered essential to be able to conduct these researches overall. According to the conductor of the thesis research it is therefore better to compare the unprocessed Twitter API output with the GIMs' output when weighting the completeness of the data heavier than weighting all evaluation metrics equally. Other researchers might come to different conclusions given that the previous point made is of a subjective nature and a matter of opinion.

11.2 Recommendations for future research

Within this thesis research a specific scope has been defined due various reasons as explained in paragraph 2.4 previously. For that reason, not all aspects of the usability of Twitter data within GIS research and the applicability of GIMs within this type of research have been researched within the thesis research context. Therefore, several recommendations for future research are discussed in this paragraph. They could serve as a starting point for other (geo)scientists interested in researching the use of GIMs in GIS research to determine the scope and focus of their research. The recommendations as made by the conductor of the thesis research are the following:

Research other types of social media data: Within this thesis research Twitter data has been researched exclusively. This choice was made because Twitter is currently a relatively popular social medium to use as a data source in GIS research compared to other social media platforms. A strong indication supporting this statement is that when using both the terms "GIS" and a certain type of social media data such as "Flickr data" in Google Scholar, way more results show up for Twitter data than data originating from any other social media platform. This does not necessarily mean that other social media cannot be used as a data source for GIS research and are not worth researching in the future. Other social media such as Foursquare and Flickr provide similar APIs to the one provided by Twitter which data output contains metadata attributes containing geographical information as well (Foursquare, 2017b; Flickr, 2017b). This geographical

information can then be used in a GIS for example. These social media data sources alternative to Twitter show potential but should be researched further to find out how and when these sources can and should be used in GIS research. Academic research on the inference of the geolocation of users or messages of other social media besides Twitter have been performed over the years (see Pontes et al, 2012; Jurgens, 2013; Lee et al, 2014; Friedland et al, 2011 for some examples). This academic framework is currently relatively small compared to the research done on Twitter data, however. Therefore, research on the applicability of social media data originating from other services as Twitter within GIS research scenarios is necessary to able geoscientists to evaluate whether these alternative sources of social media data can come handy within their own research as well.

Research other application methodologies: Within this thesis research event detection research scenarios have been researched exclusively. This choice has been made because the (vast) majority of GIS research using Twitter data as its main input uses this type of application methodology. Other application methodologies have been excluded from this thesis research due them either hardly being used within GIS research (SNA) or because the thesis research itself is already about this respective application methodology (geolocation inference). Another important reason to exclude certain application methodologies was the strict time-limit at which the thesis research could be conducted. Incorporating SNA into this thesis research would lead to several extra months of time needed to complete the thesis research for example. These arguments have previously been detailed in paragraph 4.4. This does not mean that other application methodologies should not be researched. Through SNA a better understanding on how social media networks work, develop and are structured can be derived for example. This knowledge can then be used to benefit society as a whole as previously explained in paragraph 2.3 in more detail. As previously explained in paragraph 1.4, hardly any research has been done on the differences of applicability of different types of GIMs within GIS research scenarios. Event detection has been researched specifically in this thesis research. This still means that limited research has been done on the applicability of different GIMs within SNA research scenarios, however. Therefore, research similar to this thesis research focussing on social network analysis specifically is needed to enable geoscientists using this specific application methodology within their research to evaluate what GIM is most applicable to use in their research and what is not.

Research other application domains: Within this thesis research the applicability of GIMs was researched within disaster management, health management and topic modelling research scenarios exclusively. Other application domains have been excluded from this thesis research due them hardly being used within GIS research compared to the three application domains as mentioned previously in paragraph 4.4. in more detail. Another important reason was the strict time-limit at which the thesis research could be conducted. Incorporating all application domains found and turning them into separate GIS research scenarios would lead to too much extra time needed for the thesis research to be conducted. This does not mean that the application domains not researched in this thesis research are not worth investigating scientifically. These application domains all have their purpose and the knowledge gained from research can be used to benefit society as a whole as previously explained in paragraph 2.3 in more detail. Only sparsely efforts have been made previously where the use of GIMs within specific application domains have been researched. A rare example comes from Laylavi et al (2016), who have researched geolocation inference techniques within the context of emergency response (which overlaps application domains such as disaster management and crisis management). Therefore, research similar to this thesis research focussing on application domains other than the ones researched in this thesis research is needed to able geoscientists performing GIS research within these specific domains to evaluate what GIM is most applicable in their research and what is not.

Research other GIM-types: Within this thesis research GIMs meant to infer users were researched exclusively. The main reason for this choice has been that within the time-limit at which the thesis research could be conducted no accurate message-inferring GIM could be designed. Another reason is that the majority of the GIMs as presented in academic research are meant to infer the geolocation of users as previously detailed in 6.3 in more detail. A hybrid-user method was considered but finally not implemented in the thesis research due to a change in the privacy policy of Twitter leading to a different amount of metadata that could be derived from users compared to the other GIMs that were already analysed and evaluated. This point has been detailed previously in paragraph 8.3 as well. This does not mean that other types of GIMs are not worth investigating scientifically. The fact that some GIM-types, such as network-based GIMs that infer messages, are not often presented in academic literature does not necessarily mean that thesis type of GIMs perform worse compared to more often presented types of GIMs. As previously

explained in paragraph 1.4, hardly any research has been done on the differences of applicability of different types of GIMs within GIS research scenarios. In this thesis research efforts have been made to evaluate the applicability of network-based and content-based focussing on inferring the geolocation of users. This still means that other types of GIMs have not been researched yet. Therefore, research similar to this thesis research focussing on GIM-types not discussed in this thesis research is needed to provide geoscientists working with Twitter data a wider set of GIMs to use in their research with certainty.

Research other GIM-workflows: Within this thesis research content-based GIMs that used text mining and network-based GIMs that used tie-strength were researched exclusively. The reason for this choice is that these types of GIM-workflows are most often presented in academic literature as mentioned previously in paragraph 6.4 in more detail. This does not mean that other types of GIM-workflows are not worth investigating scientifically. The fact that other GIM-workflows, such as workflows based on machine learning or GIS, are not often presented in academic literature does not necessarily mean that these GIMs perform worse compared to more often presented types of GIM-workflows. As previously has been explained in paragraph 1.4, hardly any research has been done on the applicability of GIMs within GIS research and differences in performance within GIS research scenarios. This means that limited research has been done on the applicability of GIMs with different workflows in GIS research scenarios as well. In this thesis research efforts have been made to evaluate the applicability of GIMs using either text mining or tie-strength. This still means that other GIM-workflows have not been researched yet. Therefore, research similar to this thesis research focussing on GIMs with workflows not discussed in this thesis research is needed to possibly provide geoscientists working with Twitter data a wider set of GIMs to use in their research with certainty.

Research other study areas: Within this thesis research both the GIS research scenarios and GIMs were researched within the context of the contiguous United States exclusively. Initially the reason for this was that this area has the most active Twitter users as to date in absolute value thus a bigger data set could be derived from this bounding box compared to the scenario where another study area is chosen, as detailed earlier in paragraph 2.4. During the conduction of the thesis research additional reasons supporting this choice surfaced. The first one was the fact that the majority of the GIS research using Twitter data as its main input defines the contiguous United States as their (primary) study area. This has previously been detailed in paragraph 4.2 in more detail. The same goes for GIMs, who have been primarily researched within the context of the contiguous United States as detailed in paragraph 6.2 previously in more detail. This does not mean that other study areas are not worth investigating scientifically. Beside the United States countries such as India, Indonesia and Japan could serve as interesting study areas because these countries have relatively large Twitter user bases as well (Statista, 2017d). So far academic research on GIMs with study areas outside the (contiguous) United States is relatively scarce as previously detailed in paragraph 6.2. Therefore, research is needed to determine whether GIMs used within the context of the (contiguous) United States are applicable on other study areas as well or work out different when applied on a different study area.

Research data using other languages: Within this thesis research Twitter data using the English language was researched exclusively. The first reason was that the researcher conducting this thesis research is not proficient with languages other than English and Dutch. The second reason was that most natural language processing packages have a general bias towards the English language partly due to the ease of tokenization of this particular language, as mentioned previously in paragraph 2.4 in more detail. It was later found that the majority of the Twitter data sets used to evaluate GIMs in academic research primarily used English data sets, as previously detailed in paragraph 6.4. This does not mean that Twitter data using other languages are not worth investigating scientifically. Given that countries such as India, Indonesia and Japan have relatively large Twitter user bases as previously mentioned in this paragraph, other languages are worth investigating as well due the native language of people from these countries is different than English. So far academic research on GIMs does not focus on the influence of language at all given that in most research the language used within the Twitter data used is not even specified anywhere in the article. This has been detailed earlier in paragraph 6.4. Therefore, research is needed to determine whether GIMs used within the context of the English language are applicable on Twitter data using other languages as well or work out different when applied on data using other languages.

Perform a new literature study on the use of Twitter data in GIS research within a couple of years: The main reason why a literature study had to be conducted on the use of Twitter data in GIS research is because no topical overview was available on the subject. There was only one literature review available specifically focussing on the use of Twitter data in GIS research as conducted by Steiger et al (2015) at the time of writing this thesis research, who have researched the use of Twitter data in GIS research from the years 2006 to 2013. Within a few years the literature study as performed on the subject as part of this thesis research cannot be considered topical anymore. New or different uses of Twitter data in GIS research might have emerged by then. It is therefore necessary to conduct a new literature study on the subject within a few years to be sure these new trends are summarized and geoscientists interested in the subjects are provided with a topical overview on the subject which can be referenced and used in their own research.

Perform a new literature study on GIMs for Twitter data within a couple of years: The main reason why a literature study had to be conducted on GIMs for Twitter data had to be conducted is because no topical overview was available on the subject. Some literature studies on the subject had already been conducted by the time of writing this thesis research. Ajao et al (2015) discussed several GIMs presented within the period 2010 to 2014 focussing on corpus sizes and granularity level while Jurgens et al (2015) discussed and compared GIMs from the period 2010 to 2014 as well but by workflow and performance. Within a few years the literature study as performed on the subject as part of this thesis research cannot be considered topical anymore. New or different types of GIMs to infer Twitter users and messages might have emerged by then. It is therefore necessary to conduct a new literature study on the subject within a few years to be sure these new trends are summarized and geoscientists interested in the subjects are provided with a topical overview on the subject which can be referenced and used in their own research.

Perform a more extensive literature study on the benefits and drawbacks of Twitter data usage in GIS research: As described previously in paragraph 11.1.1, a lack of knowledge on the benefits and drawbacks of Twitter data usage in GIS research might lead to various conflicts during the conduction of the research. While an effort in this thesis research has been made to examine the benefits and drawbacks, this overview is not very detailed. Therefore, future research on this specific subject is advised. This knowledge is considered vital by the conductor of the thesis research because researching the use of Twitter data in combination with a GIS is more difficult than is apparent on first eye. Any misjudgements on that part might result in radical research redesigns, time wasted and loss of interest by the conductors of the research due to frustration (among others). Given that these negative effects are relatively easy to tackle further research on the subject of the benefits and drawbacks of Twitter data usage in GIS research is advised.

11.3 Reflection

While the thesis research has been conducted successfully, it has not been performed perfectly necessarily. This may or may not be the fault of the conductor of the thesis research. It is important to consider what aspects of the thesis research as presented in this report could have been improved to prevent the same mistakes to be made and problems to occur in the future. These flaws in research will therefore be detailed and evaluated in this paragraph. The improvements as identified by the conductor of the thesis research are as followed:

Research constraints due to time-limit: A lot of research constraints arose because of the time-limit at which the thesis research could be conducted. The reason for this time-limit is that the conduction of this thesis research (or not finishing it by time) would conflict with other interests of the conductor of the thesis research such as taking part in extra courses within the master's programme, doing an internship and graduating from the master's programme. Because of these interests the thesis had to be finished before June 7th 2017, as specified by the course coordinator. This has led to certain choices considering the thesis research design during the conduction of the research, as described below and on the next page:

- An explicit focus has been set on event detection only, excluding SNA. The reasons behind this specific focus have been detailed and argued in paragraph 4.4 earlier in this thesis report.

- An explicit focus has been set on the most frequently used application domains in GIS research, excluding many other application domains who were less frequently used within academic research.

The reasons behind this specific focus have been detailed and argued in paragraph 4.5 earlier in this thesis report.

- An explicit focus has been set on inferring the geolocation of Twitter users exclusively, excluding GIMs that were meant to infer the geolocation of tweets. The reasons behind this specific focus have been detailed and argued in paragraph 6.4 earlier in this thesis report.

- The final data sets used to evaluate and compare the GIMs were sampled of 1000 entries rather than the full data sets. The reasons behind this research design choices have been detailed and argued in paragraph 8.3 earlier in this thesis report.

Another important reason why the focussed of the thesis research had to be increased is because some vital components of this thesis research, such as the systematic literature study, had hardly been performed previously by the conductor of the thesis research. Therefore, it was difficult to determine whether research focus initially set was realistic within the initial time planning or not.

The fact that during the conduction of the thesis research the amount of focus was increased or changed does necessarily not mean that the results and conclusions as presented in this thesis report are flawed. It merely means that the applicability of the results and conclusions is smaller than preferred at the beginning of the conduction of the thesis research. Without the time-limit as set now the focus could have been widened and the applicability of the results and conclusions presented in this thesis research increased. The time-limit could not be prevented however given that the other interests of the conductor of the thesis research as previously mentioned in this paragraph weight heavier than the thesis research itself. If anything, the aspects of GIMs within GIS research that could not be researched in this thesis research could be researched in the future. It can therefore be concluded that the decisions as described above concerning increasing the focus of the thesis research can be considered just. While previously it was difficult to determine the amount of time needed for a research with a similar structure as the one described in this thesis report, it is assumed that in the future the conductor of the thesis research is able to create a more realistic time planning and set a more realistic research focus based on the now gained experience.

Lack of technical expertise: Given that a certain amount of technical expertise is needed to be able to use GIMs, it is natural that a certain technical expertise is needed to research GIMs as well. Expertise is especially needed on programming languages such as Python and R and designing databases to store the data to-be-inferred in. The technical expertise of the conductor of this thesis research was sufficient to be able to finish the thesis research successfully but a better expertise would have been preferred. This is due two main reasons. The first reason is that due a lack of technical expertise designing the technical framework took longer than preferred. The second reason is that some of the scripts used in this thesis research are inefficient or designed illogical (while still serving their aim). In the future, it would be advisable to choose a research subject closer to the conductor's technical expertise or collaborate with academics from other relevant fields that have this technical expertise. Another way to increase the amount of technical expertise is to follow additional courses on for example programming languages and database design. This will be done by the conductor of the thesis research as of September 2017. After the completion of these courses it is assumed that the conductor of the thesis research has increased its technical expertise to such extent that the problems arising from the current technical expertise will not be occur or at least be less proportional in the future.

Lack of pre-luminary knowledge on thesis subject: GIMs are subject to a relatively small academic framework ranging multiple academic disciplines. While GIMs are meant to infer the geolocation of a certain subject, these methods are not necessarily part of the geosciences domain given that the techniques used to infer these geolocations originate from other disciplines such as statistics, computer science and information science. Therefore, even though the conductor of the research is an academically certified geoscientist, a lot of reading had to be done on GIMs and the techniques used within these GIMs when performing the thesis research. A part of this has also resulted in a literature study as detailed in chapter 6. The problem is that a lot of time was needed to understand the concepts subject to this thesis research. This was especially problematic given the time-limit, as previously mentioned multiple times. In the future, it would be advisable to choose a research subject closer to GIS research or collaborate with academics from other relevant academic fields to prevent an unnecessary amount of time to be put into understanding the research subject rather actually researching this subject. In the case of the subject of GIMs less preliminary research has to be

performed by the conductor of the thesis research in the future because plenty has already been done as part of this thesis research. If the conductor of the thesis research decides to perform research on GIMs in the future the problems arising from a lack of pre-luminary knowledge on the subject will are therefore assumed to be less proportional in the future.

Representativeness of GIS research scenarios: For each GIS research scenario used in this thesis research both preferred, final and actual parameters have been defined. The preferred parameters have been based on academic literature within the respective application domain the GIS research scenario is about while the final parameters are also set according to the scope of the thesis research as defined previously in paragraph 2.4. All parameter groups have been defined in paragraph 8.2 earlier. Some differences exist between the preferred, final and actual parameters which may have led to representativeness issues for the GIS research scenarios. For the disaster management scenario, initially real-time data was preferred but not used in the end. For the health management and topic modelling research scenarios initially a different corpus size and period of gathering was preferred compared to the ones used in the final GIS research scenarios. These choices have been argued previously in paragraph 7.2. While these GIS research scenarios have been kept as representative as possible there are still some issues concerning their representativeness for the whole application domains they are part of. When data is real-time criteria such as performance speed are much more important than when dealing with historic data in a disaster management research scenario.
The differences in corpus sizes and periods of gathering do not necessarily influence the representativeness of the GIS research scenarios given that the final parameters are based on the tweets per day found through the preferred parameters. Through design however, the health management and topic modelling research scenario are not truly representative even though the current differences do probably not influence the outcome of the results found in this thesis research necessarily. In future research, it would be worthwhile to research the relevant application domains again but now with complete representativeness.

Lack of use of word contexts when pre-processing data: The Twitter data used in this thesis research has been pre-processed using various techniques, such as filtering by attributes and clipping by area using GIS. The Twitter data has not been filtered by the word contexts of the tweets, however. When doing this possible disambiguation of certain words can be present in the data sets. This has been detailed earlier in paragraph 5.2 and used in the research of Yang and Mu (2015) in their research on depression among Twitter users. Within the GIS research scenarios keywords were used to gather data than can be misinterpreted and therefore not relevant to this GIS research scenario as well. Within the disaster management example keywords such as “rain” and “storm” were used which could be used in various different contexts not relevant to any disaster. The main reason why filtering the data by word contexts has not been part of the pre-processing steps as detailed earlier in paragraph 7.2.4 is because within the time-limit at which the thesis could be conducted this was not possible. In the future, it would be advisable to do that to create GIS research scenarios with greater representativeness.

Exclusion of hybrid-user method from thesis research: Due to changes in the privacy policy of Twitter the hybrid-user method has been excluded from the thesis research. The reason for this is that user metadata could not gathered with the same ease as done during the analysis of the content-user and hybrid-user method. Since comparing the methods under different policy contexts would be irresponsible from an academic point of view the hybrid-user method has therefore been excluded. Another option was to reanalyse the content-user and network-user method under the new privacy policies. This has not been done due the strict time-limit at which the thesis could be conducted, as mentioned multiple times in this chapter so far. Given that the conductor of the thesis research was met with a Catch-22 in this situation (either setting a new deadline or excluding the hybrid-user method), either decision would have conflicted with the research scope of the thesis research as previously defined in paragraph 2.4. Because finishing the thesis before the deadline weighted heavier than the inclusion of the hybrid-user method, as previously explained in this paragraph, the decision made by the conductor of the thesis research concerning the exclusion of the hybrid-user method is considered as just.

Appendix

Appendix I: References

Ahmed, A., L. Hong. & A. Smola (2013). Hierarchical geographical modelling of user locations from social media posts. In *Proceedings of the 22nd International Conference on World Wide Web*, 25-36. 10.1145/2488388.2488392

Aghabozorgi, S. (2016). *Analyze Social Media Data in Real Time*. https://bigdatauniversity.com/blog/analyze-social-media-data-real-time/. Retrieved on 30th May 2017.

Ajao, O., J. Hong & W. Liu (2015). A survey of location inference techniques on Twitter. *Journal of Information Science*, *41(6)*, 855-864. 10.1177/0165551515602847

Albuquerque, J., B. de Herbort, A. Brenning & A. Zipf (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. 10.1080/13658816.2014.996567

Anant, V. (2014). *Performance Analysis in Python*. https://vijayanant.github.io/blog/2014/11/02/Performance-analysis-in-Python. Retrieved on 31st May 2017.

Ao, J., P. Zhang & Y. Cao (2014). Estimating the locations of emergency events from Twitter streams. *Procedia Computer Science, 31*, 731-739. 10.1016/j.procs.2014.05.321

Apache (2017). *Welcome to Apache Hadoop!* https://hadoop.apache.org/. Retrieved on 29th May 2017.

Apreleva, S. & A. Cantarero (2015). Predicting the location of users on Twitter from low density graphs. In *Proceedings of the 2015 IEEE Conference on Big Data*, 976-983. 10.1109/BigData.2015.7363848

BBC (2016). *#BBCtrending: Poor Leo, Oscar selfies, and a Cumberbomb*. http://www.bbc.com/news/blogs-trending-26410106. Retrieved on 26th May 2017.

Bernard, H. (2012). *Social research methods: Qualitative and quantitative approaches*. Thousand Oaks: SAGE Publishing. 2nd edition.

Bloomberg (2011). *The Rise and Inglorious Fall of Myspace*. https://www.bloomberg.com/news/articles/2011-06-22/the-rise-and-inglorious-fall-of-myspace. Retrieved on 31st May 2017.

Bonzanini, M. (2015). *Mining Twitter Data With Python (and JS) – Part 7: Geolocation and Interactive* https://marcobonzanini.com/2015/06/16/mining-twitter-data-with-python-and-js-part-7-geolocation-and-interactive-maps/. Retrieved on 30th May 2017.

Bryman, A. (2012). *Social Research Methods*. Oxford: Oxford University Press. 4th edition.

Business Insider (2014). *"Instagram Rapture" Claims Millions of Celebrity Instagram Followers*. http://www.businessinsider.com/instagram-rapture-claims-millions-of-celebrity-instagram-followers-2014-12?international=true&r=US&IR=T. Retrieved on 29th May 2017.

Carr, C. & R. Hayes (2015). Social Media: Defining, Developing, and Divining. *Atlantic Journal of Communication*, *23(1)*, 46-65. 10.1080/15456870.2015.972282

Castells, M. (1997). An introduction to the information age. *City*, *2(7)*, 6-16. 10.1080/13604819708900050

CDC (2014). *Life expectancy*. https://www.cdc.gov/nchs/fastats/life-expectancy.htm. Retrieved on 31st May 2017.

Cervone, G., E. Sava, Q. Huang, E. Schnebele, J. Harrison & N. Waters (2016). Using Twitter for tasking remote-sensing data collection and damage assessment: 2013 Boulder flood case study. *International Journal of Remote Sensing*, *37(1)*, 100-124. 10.1080/01431161.2015.1117684.

Cha, M., Y. Gwon, H. Kung (2015). Twitter Geolocation and Regional Classification via Sparse Coding. In *Proceedings of the 9th International AAAI Conference on Web and Social Media*, 582-585. DOI Unknown.

Chae, J., D. Thom, Y. Jang, S. Kim, T. Ertl & D. Elbert (2014). Public behavior response analysis in disaster events utilizing visual analytics of microblog data. *Computers & Graphics*, *38*, 51-60. 10.1016/j.cag.2013.10.008

Chandra, S., L. Khan & F. Muhaya (2011). Estimating twitter user location using social interactions – A Content Based Approach. In *Proceedings of the 2011 IEEE Third International Conference on Privacy, Security Risk and Trust (PASSAT) and 2011 IEEE 3rd International Conference on Social Computing (SocialCom),* 838-843. 10.1109/PASSAT/SocialCom.2011.120

Chang, H., D. Lee, M. Eltaher & J. Lee (2012). @ Phillies tweeting from Philly? Predicting Twitter user locations with spatial word usage. In *Proceedings of the 2012 International Conference on Advances in Social Network Analysis and Mining*, 111-118. 10.1109/ASONAM.2012.29

Chen, J., Y. Liu & M. Zou (2016). Home location profiling for users in social media. *Information & Management*, *53(1)*, 135-143. 10.1016/j.im.2015.09.008

Cheng, Z., J. Caverlee & K. Lee (2013). A content-driven framework for geolocating microblog users. *ACM Transactions on Intelligent Systems and Technology*, *4(1)*, 2. 10.1145/2414425.2414427

Clarivate Analytics (2017). *Web of Science*. http://clarivate.com/scientific-and-academic-research/research-discovery/web-of-science/. Retrieved on 29th May 2017.

CNN (2017a). *More rain in store after 5 killed in California storms*. http://edition.cnn.com/2017/02/19/us/california-storm/index.html. Retrieved on 30th May 2017.

CNN (2017b). *Powerful storm takes aim at Southern California*. http://edition.cnn.com/2017/02/16/us/california-weather-storm/index.html. Retrieved on 30th May 2017.

CNN (2017c). *Updates: Deadly storm slams Southern California*. http://edition.cnn.com/2017/02/17/us/southern-california-storm-oroville-dam/index.html. Retrieved on 30th May 2017.

CNN (2017d). *Oroville Dam: Residents advised to remain vigilant as storm advances*. http://edition.cnn.com/2017/02/16/us/california-oroville-dam-storm-spillway/index.html. Retrieved on 30th May 2017.

CNN (2017e). *California braces for a new round of storms*. http://edition.cnn.com/2017/02/20/weather/weather-flooding-storms/index.html. Retrieved on 30th May 2017.

CNN (2017f). *Resident evacuate after levee breach in California*. http://edition.cnn.com/2017/02/21/weather/california-weather-flooding-storms/index.html. Retrieved on 30th May 2017.

CNN (2017g). *San Jose flooding: Thousands ordered to leave homes*. http://edition.cnn.com/2017/02/22/us/san-jose-flood/index.html. Retrieved on 30th May 2017.

CNN (2017h). *Flood evacuations underway in San Jose*. http://edition.cnn.com/2017/02/21/us/san-jose-flood/index.html. Retrieved on 30th May 2017.

CNN (2017i). *ABOUT CNN.COM*. http://edition.cnn.com/2014/01/17/cnn-info/about/index.html. Retrieved on 30th May 2017.

Compton, R., D. Jurgens & D. Allen (2014). Geotagging one hundred million twitter accounts with total variation minimization. In *Proceedings of the 2014 IEEE International Conference on Big Data*, 393-401. 10.1109/BigData.2014.7004256

comSysto (2012). *Real-time Twitter Heat Map With MongoDB*. https://comsysto.com/blog-post/real-time-twitter-heat-map-with-mongodb. Retrieved on 30th May 2017.

Conway, M. (2014). Ethical Issues in using Twitter for Public Health Surveillance and Research: Developing a Taxonomy of Ethical Concepts From the Research Literature. *Journal of Medical Internet Research*, *16(12)*, e290. 10.2196/jmir.3617

Crampton, J., M. Graham, A. Poorthuis, T. Shelton, M. Stephens, M. Wilson & M. Zook (2013). Beyond the geotag: situating 'big data' and leveraging the potential of the geoweb. *Cartography and Geographic Information Science*, *40(2)*, 130-139. 10.1080/15230406.2013.777137.

Cranshaw, J., R. Schwartz, J. Hong & N. Sadeh (2012). The Livehoods project: Utilizing Social Media to Understand the Dynamics of a City. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*, 58. 10.1.1.365.7792

Crawford, K. & M. Finn (2015). The limits of crisis data: analytical and ethical challenges of using social and mobile data to understand disasters. *GeoJournal*, *80(4)*, 491-502. 10.1007/s10708-014-9597-z

Crooks, A., A. Croitoru, A. Stefanidis & J. Radzikowski (2013). #Earthquake: Twitter as a distributed sensor system. *Transactions in GIS*, *17(1)*, 124-147. 10.1111/j.1467-9671.2012.01359.x

Daily Kos (2017). *#GOPDND trending on Twitter*. https://www.dailykos.com/story/2017/3/23/1646554/--GOPDND-trending-on-Twitter. Retrieved on 30th May 2017.

Davidov, D., O. Tsur, & A. Rappoport (2010). Semi-supervised recognition of sarcastic sentences in Twitter and Amazon. In *Proceeding of the 14th Conference on Computational Natural Language Learning*, 107-116. DOI Unknown.

Davis Jr, C., G. Pappa, D. de Oliveira & D. Arcanjo (2011). Inferring the location of twitter messages based on user relationship. *Transactions in GIS*, *15(6)*, 735-751. 10.111/j.1467-9671.2011.01297.x

Devillers, R., M. Gervais, M. Bédard & R. Jeansoulin (2002). Spatial data quality: from metadata to quality indicators and contextual end-user manual. *OEEPE/ISPRS Joint Workshop on Spatial Data Quality*, 45-55. DOI Unknown.

Devillers, R., A. Stein, Y. Bédard, N. Chrisman, P. Fisher & W. Shi (2010). Thirty years of research on spatial data quality: achievements, failures, and opportunities. *Transactions in GIS*, *14(4)*, 387-400. 10.1080/13658110600911879

Dorkly (2017). *15 Times #GOPDnD Proves That Republicans Would Be Bad At Dungeons & Dragons*. http://www.dorkly.com/post/82811/15-times-gopdnd-proves-that-republicans-would-be-bad-at-dungeons-dragons. Retrieved on 30th May 2017.

Duong-Trung, N., N. Schilling & L. Schmid-Thieme (2016). Near Real-time Geolocation Prediction in Twitter Streams via Matrix Factorization Based Regression. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, 1973-1976. 10.1145/2983323.2983887

Eisenstein, J., B. O'Connor, N. Smith & E. Xing (2010). A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 1277-1287. DOI Unknown.

Elite Daily (2017). *Women Are Sharing Photos of Self-Love With The Hashtag #BigThighTwitter*. http://elitedaily.com/women/plus-size-women-big-thigh-twitter-self-love-body-positivity-shaming-size-discrimination/1837243/. Retrieved on 30th May 2017.

Elsevier (2017). *Scopus*. https://www.elsevier.com/solutions/scopus. Retrieved on 29th May 2017.

ESRI (2017a). *About ArcGIS*. http://www.esri.com/arcgis/about-arcgis. Retrieved on 30th May 2017.

ESRI (2017b). *Normalization*. http://support.esri.com/other-resources/gis-dictionary/term/normalization. Retrieved on 30th May 2017.

Facebook (2016). *FORM 10-K*. https://d1lge852tjjqow.cloudfront.net/CIK-0001326801/0c796377-fba3-4e40-b18a-3503901af1a4.pdf?noexit=true. Retrieved on 29th May 2017.

FEMA (2017). *About the Agency*. https://www.fema.gov/about-agency. Retrieved on 29th May 2017.

Financial Times (2009). *The rise and fall of MySpace*. https://www.ft.com/content/fd9ffd9c-dee5-11de-adff-00144feab49a. Retrieved on 31st May 2017.

FiveThirtyEight (2015). *How Many Times Does The Average Person Move?* https://fivethirtyeight.com/datalab/how-many-times-the-average-person-moves/. Retrieved on 31st May 2017.

Flickr (2017a). *About Flickr*. https://www.flickr.com/about. Retrieved on 29th May 2017.

Flickr (2017b). *The App Garden*. https://www.flickr.com/services/api// Retrieved on 31st May 2017.

Forbes (2012). *Facebook Didn't Kill Digg, Reddit Did*. https://www.forbes.com/sites/insertcoin/2012/07/13/facebook-didnt-kill-digg-reddit-did/. Retrieved on 31st May 2017.

Foursquare (2017a). *About Us*. https://foursquare.com/about. Retrieved on 29th May 2017.

Foursquare (2017b). *The Foursquare API*. https://developer.foursquare.com/overview/. Retrieved on 31st May 2017.

Friedland, G., J. Choi & A. Janin (2011). Video2GPS: A Demo of Multimodel Location Estimation on Flick Videos. In *Proceedings of the 19th ACM International Conference on Multimedia*, 833-834. 10.1145/2072298.2072482

Garnett, R. & R. Stewart (2015). Comparison of GPS units and mobile Apple GPS capabilities in an urban landscape. *Cartography and Geographic Information Science, 42(1)*, 1-8. 10.1080/15230406.2014.974074.

GeoPy (2017). *Welcome to GeoPy's documentation! – GeoPy 1.10.0 documentation*. https://geopy.readthedocs.io/en/1.10.0/. Retrieved on 30th May 2017.

GeoText (2017). *About GeoText*. http://www.geotext.com/about-geotext. Retrieved on 30th May 2017.

Gizmodo (2017). *#GOPDnD Has The Best Dungeons & Dragons Game Ever After Trumpcare Failed*. https://www.gizmodo.com.au/2017/03/gopdnd-had-the-best-dungeons-dragons-game-ever-after-trumpcare-failed/. Retrieved on 30th May 2017.

Gnip (2017). *About Gnip*. https://gnip.com/about/. Retrieved on 29th May 2017.

Google Maps (2017). *About*. https://www.google.com/maps/about/. Retrieved on 31st May 2017.

Google Plus (2017). *Google+*. https://plus.google.com/+googleplus. Retrieved on 29th May 2017.

Google Scholar (2017). *About Google Scholar*. https://scholar.google.com/intl/en/scholar/about.html. Retrieved on 29th May 2017.

Gu, H., H. Hang, Q. Lv & D. Grunwald (2012). Fusing Text and Friendships for Location Inference in Online Social Networks. In *Proceedings of the 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology*, 158-165. 10.1109/WI-IAT.2012.243

Gu, Y., Z. Qian & F. Chen (2016). From Twitter to detector: Real-time traffic incident detection using social media data. *Transportation Research Part C: Emerging Technologies*, *67*, 321-342. 10.1016/j.trc.2016.02.011

Guan, X. & C. Chen (2014). Using social media data to understand and assess disasters. *Natural Hazards*, *74(2)*, 837-850. 10.1007/s11069-014-1217-1.

Guardian, The (2015). *MySpace – what went wrong: "The site was a massive spaghetti-ball mess'.* https://www.theguardian.com/technology/2015/mar/06/myspace-what-went-wrong-sean-percival-spotify. Retrieved on 31st May 2017.

Gurajala, S., J. White, B. Hudson & J. Matthews (2015). Fake Twitter accounts: profile characteristics obtained using an activity-based pattern detection approach. In *Proceedings of the 2015 International Conference on Social Media & Society*, 9-15. 10.1145/2789187.2789206

Han, B., P. Cook & T. Baldwin (2014). Text-Based Twitter User Geolocation Prediction. *Journal of Artificial Intelligence Research*, *49*, 451-500. 10.1613/jair.4200

Henderson, M., N. Johnson & G. Auld (2013). Silences of ethical practice: dilemmas for researchers using social media. *Educational Research and Evaluation*, *19(6)*, 546-560. 10.1080/13803611.2013.805656

Himelboim, I. & J. Han (2014). Cancer talk on twitter: community structure and information sources in breast and prostate cancer social networks. *Journal of Health Communication*, *19(2)*, 210-225. 10.1080/10810730.2013.811321

Huang, Q., C. Guofeng & C. Wang (2014). From where do tweets originate?: a GIS approach for user location inference. In *Proceedings of the 7th ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, 1-8. 10.1145/2755492.2755494

Huang, Q. & D. Wong (2016). Activity patterns, socioeconomic status and urban spatial structure: what can social media data tell us? *International Journal of Geographical Information Science*, *30(9)*, 1873-1898. 10.1080/13658816.2016.1145225

Huffington Post (2017). *Twitter's Latest Body Positive Hashtag Celebrates The Goodness of Thick Thighs*. http://www.huffingtonpost.ca/2017/03/24/big-thigh-twitter_n_15581544.html. Retrieved on 30th May 2017.

Jurgens, D. (2013). That's What Friends Are For: Inferring Location in Online Social Media Platforms Based on Social Relationships. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, 273-282. DOI Unknown.

Jurgens, D., T. Finnethy, J. McCorriston, Y. Xu & D. Ruths (2015). Geolocation Prediction in Twitter Using Social Networks: A Critical Analysis and Review of Current Practice. In *Proceeding of the 9th International AAAI Conference on Weblogs and Social Media (ICWSM)*, pp. 188-197. DOI Unknown.

Kamath, K., Z. Caverlee & D. Sui (2012). Spatial influence vs. community influence: modelling the global spread of social media. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, 962-971. 10.1145/2396761.2396883

Kamath, K., J. Caverlee, K. Lee, Z. Cheng (2013a). Spatio-temporal dynamics of online memes: a study of geo-tagged tweets. In *Proceedings of the 22nd International Conference on World Wide Web*, 667-678. 10.1145/2488388.2488447

Kamath, K. & J. Caverlee (2013). Spatio-temporal meme prediction: learning what hashtags will be popular where. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, 1341-1350. 10.1145/2505515.2505579

Katsuki, T., T. Mackey & R. Cuomo (2015). Establishing a Link Between Prescription Drug Abuse and Illicit Online Pharmacies: Analysis of Twitter Data. *Journal of Medical Internet Research*, *17(12)*, 1-12. 10.2196/jmir.5144

KFF (2017). *Population Distribution by Age: Timeframe 2015*. http://kff.org/other/state-indicator/distribution-by-age/?currentTimeframe=0. Retrieved on 29th May 2017.

Kim, K., I. Kojima & H. Ogawa (2016). Discovery of local topics by using latent spatio-temporal relationships in geo-social media. *International Journal of Geographical Information Science, 30*, 1899-1922. 10.1080/13658816.2016.1146956

Kinsella, S., V. Murdock & N. O'Hare (2016). "I'm Eating a Sandwich in Glasgow": Modeling Locations with Tweets. In *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents*, 61-68. 10.1145/2065023.2065039

Kong, L., Z. Liu & Y. Huang (2014). Spot: Locating social media users based on social network context. *Proceeding of the VLDB Endowment, 7(13)*, 1681-1684. 10.14778/2733004.2733060.

Kontaxis, G., I. Polakis, S. Ioannidis & E. Markatos (2011). Detecting social network profile cloning. In *Proceedings of the 2011 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, 295-300. 10.1109/PERCOMW.2011.5766886

Kos, S., D. Brčić, I. Musulin (2013). Smartphone application GPS performance during various space weather conditions: a preliminary study. In *Proceedings of the 21st International Symposium on Electronics in Transport*. pp. Unknown. DOI Unknown.

Kotzias, D., T. Lappas & D. Gunopulos (2016). Home is where your friends are: Utilizing the social graph to locate twitter users in a city. *Information Systems*, *57*, 77-87. 10.1016/j.is.2015.10.011

Krishnamurthy, R., P. Kapanipathi, A. Sheth & K. Thirunarayan (2015). Knowledge enabled approach to predict the location of twitter users. In *Proceedings of the European Semantic Web Conference*, 187-201. 10.1007/978-3-319-18818-8_12

Instagram (2017). *About Us*. https://www.instagram.com/about/us/. Retrieved on 29th May 2017.

ITV (2017). *London unites on Twitter #prayforlondon #wearenotafraid*. http://www.itv.com/news/london/2017-03-23/london-unites-on-twitter-prayforlondon-wearenotafraid/. Retrieved on 30th May 2017.

Jung, J. (2014). Code clouds: Qualitative geovisualization of geotweets. *The Canadian Geographer / Le Géographe canadien*, *59(1)*, 52-68. 10.1111/cag.12133

Lansley, G. & P. Longley (2016). The geography of Twitter topic in London. *Computers, Environment and Urban Systems*, *58*, 85-96. 10.1016/j.compenvurbsys.2016.04.002

Laylavi, F., A. Rajabifard & M. Kalantari (2016). A multi-element approach to location inference of twitter: A case for emergency response. *ISPRS International Journal of Geo-Information*, *5(5)*, 56. 10.3390/ijgi5050056

Lee, K., A. Agrawal & A. Choudhary (2013). Real-time disease surveillance using twitter data: demonstration on flue and cancer. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1474-1477. 10.1145/2487575.2487709

Lee, K., R. Ganti, M. Srivatsa & L. Liu (2014). When twitter meets foursquare: tweet location prediction using foursquare. In *Proceedings of the 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, 1*98-207. 10.4108/icst.mobiquitous.2014.258092

Li, L., M. Goodchild & B. Xu (2013). Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartography and Geographic Information Science*, *40(2)*, 61-77. 10.1080/15230406.2013777139

Li, R., S. Wang & K. Chang (2012a). Multiple locations profiling for users and relationships from social network and content. In *Proceedings of the VLDB Endowment*, *5(11)*, 1603-1614. 10.14778/2350229.2350273

Li, R., S. Wang, H. Deng, R. Wang & K. Chang (2012b). Towards social user profiling: unified and discriminative influence model for inferring home locations. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1023-1031. 10.1145/2339530.2339692

Lingad, J., S. Karimi & J. Yin (2013). Location extraction from disaster-related microblogs. In *Proceedings of the 22nd International Conference on World Wide Web*, 1017-1020. 10.1145/2487788.2488108

Liu, Z. & Y. Huang (2016). Closeness and structure of friends help to estimate user locations. In *Proceedings of the International Conference on Database Systems for Advanced Applications*, 33-48. 10.1007/978-3-319-32049-6_3

Liu, J. & D. Inkpen (2015). Estimating user location in social media with stacked denoising auto-encoders. In *Proceedings of NAACL-HLT*, 201-2010. DOI Unknown.

Lwin, K., K. Sugiura & K. Zettsu (2016). Space-time multiple regression model for grid-based population estimation in urban areas. *International Journal of Geographical Information Science*, *30(8)*, 1579-1593. 10.1080/13658816.2016.1143099

Mahmud, J., J. Nichols & C. Drews (2014). Home location identification of twitter users. *ACM Transactions on Intelligent Systems and Technology*, *5(3)*, 47. 10.1145/2528548

Malleson, N. & M. Andresen (2015). The impact of using social media data in crime calculations: crime hot spots and changing spatial patterns. *Cartography and Geographic Information Science*, *42(2)*, 112-121. 10.1080/15230406.2014.905756

Mary Sue, The (2017). *#GOPDnD: Trump's America Makes for a Hilariously Disastrous D&D Campaign*. https://www.themarysue.com/gop-dnd/. Retrieved on 30th May 2017.

Mashable (2014). *Friendster Founder Tells His Side of the Story, 10 Years After Facebook*. http://mashable.com/2014/02/03/jonathan-abrams-friendster-facebook/#7aSq8SVw6aqx. Retrieved on 31st May 2017.

May, T. (2015). *Social Research*. New York: McGraw-Hill Education. 4th edition.

Maynard, D. & M. Greenwood (2014). Who cares about Sarcastic Tweets? Investigating the Impact of Sarcasm on Sentiment Analysis. In *LREC*, pp. Unknown. DOI Unknown.

McGee, J., J. Caverlee & Z. Cheng (2013). Location prediction in social media based on tie-strength. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, 459-468. 10.1145/2505515.2505544

Metro (2017). *Pray For London and We Are Not Afraid Being Shared after London terror attack*. http://metro.co.uk/2017/03/22/people-are-sharing-posts-with-prayforlondon-and-wearenotafraid-in-wake-of-westminster-attack-6527320/. Retrieved on 30th May 2017.

Mislove, A., S. Lehmann, Y. Ahn, J. Onnela & J. Rosenquist (2011). Understanding the Demographics of Twitter Users. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media,* 5-9. DOI Unknown.

MIT Technology Review (2012). *What Did Reddit Succeed Where Digg Failed?* https://www.technologyreview.com/s/428520/why-did-reddit-succeed-where-digg-failed/. Retrieved on 31st May 2017.

Musulin, I., D. Brčić & S. Kos (2014). A study of smartphone satellite positioning performance at sea using GPS and GLONASS systems. In *Proceedings of the ISEP 2014 ITS for Seamless and Energy Smart Transport*. pp. Unknown. DOI Unknown.

Nagel, A., M. Tsou, B. Spitzberg, L. An, M. Gawron, D. Gupta, J. Yang, S. Han, M. Peddecord, S. Lindsay & M. Sawyer (2013). The Complex Relationship of Realspace Events and Messages in Cyberspace: Case Study of Influenza and Pertussis Using Tweets. *Journal of Medical Research*, *15(10)*, e237. 10.2196/jmir.2705.

Nelson, J., S. Quinn, B. Swedberg, W. Chu & A. MacEachren (2015). Geovisual Analytics Approach to Exploring Public Political Discourse on Twitter. *ISPRS International Journal of Geo-Informati*on, *4(1)*, 337-366. 10.3390/ijgi401337.

news.com.au (2017). *Pauline Hanson starts horrific new hashtag attacking Muslims in light of London terror attack*. http://www.news.com.au/national/politics/pm-malcolm-turnbull-has-responded-to-the-london-attacks-with-a-powerful-message-from-australia/news-story/62b8d06f7d64df9d835c8e6852dde813. Retrieved on 30th May 2017.

NGA (2014). *Office of Geomatics: World Geodetic System 1984 (WGS84)*. http://earth-info.nga.mil/GandG/wgs84/index.html. Retrieved on 29th May 2017.

Nguyen, H. (2013a). *A guide to analysing Python performance*. https://www.huyng.com/posts/python-performance-analysis. Retrieved on 31st May 2017.

Nguyen, H. (2013b). *About Huy*. https://www.huyng.com/about. Retrieved on 31st May 2017.

Nguyen, Q., S. Kath, H. Meng, D. Li, K. Smith, J. VanDerslice, M. Wen & F. Li (2016a). Leveraging geotagged Twitter data to examine neighborhood happiness, diet, and physical activity. *Applied Geography*, *73*, 77-88. 10.1016/j.apgeog.2016.06.003.

Nguyen, Q., D. Li, H. Meng, S. Kath, E. Nsoesie, F. Li & M. Wen (2016b). Building a National Neighborhood Dataset From Geotagged Twitter Data for Indicators of Happiness, Diet, and Physical Activity. *JMIR Public Health and Surveillance*, *2(2)*. 10/2196/publichealth.5869

OpenStreetMap (2017). *About*. https://www.openstreetmap.org/about. Retrieved on 30th May 2017.

Orita, A. & H. Hisakazu (2009). "Is that really you?: an approach to assure identity without revealing real-name online. In *Proceedings of the 5th ACM workshop on Digital identity management*, 17-20. 10.1145/1655028.1655034

OSM Foundation (2017). *Nominatim Usage Policy*. https://operations.osmfoundation.org/policies/nominatim/. Retrieved on 30th May 2017.

Ozdikis, O., H. Oguztuzun & P. Karagoz (2013). Evidential location estimation for events detected in Twitter. In *Proceedings of the 7th Workshop on Geographic Information Retrieval*, 9-16. 10.1145/2533888.2533929

Paraskevopoulos, P. & T. Palpanas (2016). Where has this tweet come from? Fast and fine-grained geolocalization of non-geotagged tweets. *Social Network Analysis and Mining*, *6(1)*, 89. 10.1007/s13278-016-0400-7

Park, M., H. Kim, S. Lee & K. Bae (2014). Performance evaluation of Android location service at the urban canyon. In *Proceedings of the 16th International Conference on Advanced Communication Technology,* 662-665. 10.1109/ICACT.2014.6779045

Paul, M. & M. Dredze (2011). You are what you Tweet: Analyzing Twitter for public health. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, 265-272. DOI Unknown.

Pavalanathan, U. & J. Eisenstein (2015). Confounds and Consequences in Geotagged Twitter Data. *arXiv preprint arXiv:1506.02275.*

Pickles, J. (1995). *Ground truth: the social implications of geographic information systems*. New York: Guilford Press. 1st edition.

Pontes, T., M. Vasconcelos, J. Almeida, P. Kumaraguru & V. Almeida (2012). We Know Where You Live: Privacy Characterization of Foursquare Behaviour. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, 898-905. 10.1145/2370216.2370419

Poorazizi, M., A. Hunter & S. Steiniger (2015). A Volunteered Geographic Information Framework to Enable Bottom-Up Disaster Management Platforms. *ISPRS International Journal of Geo-Information*, *4(3)*, 1389-1422. 10.3390/ijgi4031389

Popsugar (2017). *Why Women on Twitter Are Sharing Pictures of Their Glorious Thick Thighs*. https://www.popsugar.com/fitness/BigThighTwitter-Body-Positive-Movement-43349682. Retrieved on 30th May 2017.

Ptácek, T., I. Habernal & J. Hong (2014). Sarcasm Detection on Czech and English Twitter. In *Proceedings of the 25th International Conference on Computational Linguistics*, 213-223. DOI Unknown.

Punch, K. (2014). *Introduction to social research: Quantitative and qualitative approaches*. Thousand Oaks: SAGE Publishing. 3rd edition.

Python Software Foundation (2017). *What is Python? Extensive Summary*. https://www.python.org/doc/essays/blurb/. Retrieved on 29th May 2017.

R Foundation, The (2017). *What is R?* https://www.r-project.org/about.html. Retrieved on 29th May 2017.

Radzikowski, J., A. Stefanidis, K. Jacobsen, A. Croitoru, A. Crooks & P. Delamater (2016). The Measles Vaccination Narrative in Twitter: A Quantitative Analysis. J*MIR Public Health and Surveillance*, *2(1)*, e1. 10.2196/publichealth.5059.

Reddit (2017a). *welcome to /r/IAmA.* https://www.reddit.com/r/IAmA/. Retrieved on 26th May 2017.

Reddit (2017b). *I am Barack Obama, President of the United States – AMA*. https://www.reddit.com/r/IAmA/comments/z1c9z/i_am_barack_obama_president_of_the_united_states/ Retrieved on 26th May 2017.

Reddit (2017c). *I'm Bill Gates, co-chair of the Bill & Melinda Gates Foundation. Ask Me Anything*. https://www.reddit.com/r/IAmA/comments/5whpqs/im_bill_gates_cochair_of_the_bill_melinda_gates/ Retrieved on 26th May 2017.

Reddit (2017d). *I am Gordon Ramsay. AMA*. https://www.reddit.com/r/IAmA/comments/334wcy/i_am_gordon_ramsay_ama/ . Retrieved on 26th May 2017.

Reddit (2017e). *Reddit*. https://about.reddit.com/. Retrieved on May 26th 2017.

Refinery29 (2017). *This Trending Hashtag Is The Body-Positive Movement We Need*. http://www.refinery29.com/2017/03/146919/big-thigh-twitter-body-positivity. Retrieved on 30th May 2017.

Ren, K., S. Zhang & H. Lin (2012). Where Are You Settling Down: Geo-locating Twitter Users Based on Tweets and Social Networks. In *Proceedings of the Asia Information Retrieval Symposium*, 150-161. 10.1007/978-3-642-35341-3_13

Rodrigues, M. & A. Teixeira (2015). *Advanced Application of Natural Language Processing for Performing Information Extraction*. New York: Springer. 1st edition.

Rodrigues, E., R. Assanção, G. Pappa, D. Renno & W. Meira Jr. (2016). Exploring multiple evidence to infer users' location in Twitter. *Neurocomputing*, *171*, 30-38. 10.1016/j.neucom.2015.05.066

Roller, S., M. Speriosu, S. Rallapalli, B. Wing & J. Baldridge (2012). Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 1500-1510. DOI Unknown.

Rossant, C. (2012). *Profiling and optimizing Python code*. http://cyrille.rossant.net/profiling-and-optimizing-python-code/. Retrieved on 31st May 2017.

Russell, M. (2014). *Mining the Social Web.* Sebastopol: O'Reilly Media. 2nd Edition.

Seventeen (2017). *#BigThighTwitter Fights Back Against Unrealistic Beauty Standards by Proving Legs of All Sizes Are Sexy*. http://www.seventeen.com/beauty/news/a45935/big-thigh-twitter/. Retrieved on 30th May 2017.

Shelton, T., A. Poorthuis, M. Graham & M. Zook (2014). Mapping the data shadows of Hurricane Sandy: Unconvering the sociospatial dimensions of 'big data'. *Geoforum*, *52*, 167-179. 10.1016/j.geoforum.2014.01.006

Shook, E. & V. Turner (2016). The socio-environmental data explorer (SEDE): a social media-enhanced decision support system to explore risk perception to hazard events. *Cartography and Geographic Information Science*, *43(5)*, 427-441. 10.1080/15230406.2015.1131627

Slate (2017). *#GOPDnD Uses Dungeons & Dragons to Process Republicans' Cartoon Villainy*. http://www.slate.com/blogs/browbeat/2017/03/23/_gopdnd_uses_dungeons_dragons_to_explain_the_republican_party_s_villainy.html. Retrieved on 30th May 2017.

Sloan, L. & J. Morgan (2015). Who Tweets with Their Location? Understanding the Relationship between Demographic Characteristics and the Use of Geoservices and Geotagging on Twitter. *PLoS ONE*, *10(11)*, e0142209. 10.1371/journal.pone.0142209

Statista (2017a). *Most famous social network sites worldwide as of April 2017, ranked by number of active user (in millions)*. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/. Retrieved on 26th May 2017.

Statista (2017b). *Number of monthly active Twitter users worldwide from 1st quarter 2010 to 1st quarter 2017 (in millions)*. https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/. Retrieved on 26th May 2017.

Statista (2017c). *Number of social media users worldwide from 2010 to 2020 (in billions)*. https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/. Retrieved on 26th May 2017.

Statista (2017d). *Number of active Twitter user in leading markets as of May 2016 (in millions)*. https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/. Retrieved on 29th May 2017.

Statista (2017e). *Distribution of Twitter users in the United States as of December 2015, by age group*. https://www.statista.com/statistics/192703/age-distribution-of-users-on-twitter-in-the-united-states/. Retrieved on 29th May 2017.

Steiger, E., J. Porto de Albuquereque & A. Zipf (2015). An Advanced Systematic Literature Review on Spatiotemporal Analyses of Twitter Data. *Transactions in GIS*, *19(6)*, 809-834. 10.1111/tgis.12132

Sugawara, Y. (2012). Cancer patients on Twitter: a novel patient community on social media. *BMC Research Notes, 5(1)*, 699. 10.1186/1756-0500-5-699

Tashakkori, A. & C. Teddlie (2010). *Sage Handbook of Mixed Methods in Social & Behavioral Research.* Thousand Oaks: SAGE Publishing. 2nd edition.

Techradar (2012). *Whatever happened to Digg?* http://www.techradar.com/news/internet/web/whatever-happened-to-digg-1093422. Retrieved on 31st May 2017.

TeenVogue (2017). *#BigThighTwitter Is the New Body Positive Movement We Need*. http://www.teenvogue.com/story/bigthightwitter-body-positive-hashtag. Retrieved on 30th May 2017.

Tsou, M. (2015). Research challenges and opportunities in mapping social media and Big Data. *Cartography and Geographic Information Science*, *42(1)*, 70-74. 10.1080/15230406.2015.1059251

Twitter (2015). *Coordinate system user by Streaming API*. https://twittercommunity.com/t/coordinate-system-used-by-streaming-api/45039. Retrieved on 26th May 2017.

Twitter (2016). *#ThisHappened in 2016*. https://blog.twitter.com/2016/thishappened-in-2016. Retrieved on 26th May 2017.

Twitter (2017a). *Milestones*. https://about.twitter.com/en/company/press/milestones. Retrieved on 26th May 2017.

Twitter (2017b). *The Twitter Rules*. https://support.twitter.com/articles/18311#. Retrieved on 26th May 2017.

Twitter (2017c). *Posting photos or GIFs on Twitter*. https://support.twitter.com/articles/20156423#. Retrieved on 26th May 2017.

Twitter (2017d). *Sharing and watching videos on Twitter*. https://support.twitter.com/articles/20172128#. Retrieved on 26th May 2017.

Twitter (2017e). *Posting links in a Tweet*. https://support.twitter.com/articles/78124#. Retrieved on 26th May 2017.

Twitter (2017f). *Posting a Tweet*. https://support.twitter.com/articles/15367. Retrieved on 26th May 2017.

Twitter (2017g). *About public and protected Tweets*. https://support.twitter.com/articles/14016. Retrieved on 26th May 2017.

Twitter (2017h). *API Overview*. https://dev.twitter.com/overview/api. Retrieved on 26th May 2017.

Twitter (2017i). *Adding your location to a Tweet*. Retrieved on 10th January 2017 from https://support.twitter.com/articles/122236

Twitter (2017j). *Public streams*. Retrieved on 3rd November 2016 from https://dev.twitter.com/streaming/public

Twitter (2017k). *The Search API*. Retrieved on 10th January 2017 from https://dev.twitter.com/rest/public/search

Twitter (2017l). *Using hashtags on Twitter*. https://support.twitter.com/articles/20169394. Retrieved on 29th May 2017.

Twitter (2017m). *Streaming APIs*. https://dev.twitter.com/streaming/overview. Retrieved on 29th May 2017.

Twitter (2017n). *Tweets*. https://dev.twitter.com/overview/api/tweets. Retrieved on 29th May 2017.

Twitter (2017o). *Users*. https://dev.twitter.com/overview/api/users. Retrieved on 29th May 2017.

Twitter (2017p). *Places*. https://dev.twitter.com/overview/api/places. Retrieved on 29th May 2017.

Twitter (2017q). *Rate Limits: Chart*. https://dev.twitter.com/rest/public/rate-limits. Retrieved on 30th May 2017.

Twitter (2017r). *Privacy*. https://twitter.com/privacy. Retrieved on 30th May 2017.

Unicode Consortium, The (2017). *About the Unicode® Standard*. http://unicode.org/standard/standard.html. Retrieved on 30th May 2017.

USA Today (2017). *London terror attack spurs online #PrayforLondon effort*. http://www.usatoday.com/story/news/nation-now/2017/03/22/london-terror-attack-prayforlondon/99500378/. Retrieved on 30th May 2017.

Vincenty, T. (1975). Direct and Inverse Solutions of Geodesics on the Ellipsoid with application of nested equations. *Survey Review, XXIII (176)*, 88-93. 10.1179/sre.1975.23.176.88.

Washington Post, The (2017). *'I cast Repeal Obamacare': Twitter reimagines the Trump era as a game of 'Dungeons & Dragons'*. https://www.washingtonpost.com/news/the-intersect/wp/2017/03/26/trolls-are-reimagining-the-trump-era-as-a-game-of-dungeons-dragons/?utm_term=.2337da125224. Retrieved on 30th May 2017.

Widener, M. & L. Wenwen (2014). Using geolocated Twitter data to monitor the prevalence of healthy and unhealthy food references across the US. *Applied Geography*, *54*, 189-197. 10.1016/j.apgeog.2014.07.017

Wikipedia (2017a). *Selfie*. https://en.wikipedia.org/wiki/Selfie. Retrieved on 29th May 2017.

Wikipedia (2017b). *Application programming interface*. https://en.wikipedia.org/wiki/Application_programming_interface. Retrieved on 29th May 2017.

Wikipedia (2017c). *Metadata*. https://en.wikipedia.org/wiki/Metadata. Retrieved on 29th May 2017.

Wikipedia (2017d). *Minimum bounding rectangle*. https://en.wikipedia.org/wiki/Minimum_bounding_rectangle. Retrieved on 29th May 2017.

Wikipedia (2017e). *Toponomy*. https://en.wikipedia.org/wiki/Toponymy. Retrieved on 29th May 2017.

Wikipedia (2017f). *Geographic data and information*. https://en.wikipedia.org/wiki/Geographic_data_and_information. Retrieved on 29th May 2017.

Wikipedia (2017g). *Contiguous United States*. https://en.wikipedia.org/wiki/Contiguous_United_States Retrieved on 29th May 2017.

Wikipedia (2017h). *Comma-separated values*. https://en.wikipedia.org/wiki/Comma-separated_values. Retrieved on 29th May 2017.

Wikipedia (2017i). *Big data*. https://en.wikipedia.org/wiki/Big_data. Retrieved on 29th May 2017.

Wikipedia (2017j). *Gazetteer*. https://en.wikipedia.org/wiki/Gazetteer. Retrieved on 29th May 2017.

Wired (2013). *The Friendster Autopsy: How a Social Network Dies*. https://www.wired.com/2013/02/friendster-autopsy/. Retrieved on 31th May 2017.

WMO (2017). *Tropical Cyclone Naming*. http://www.wmo.int/pages/prog/www/tcp/Storm-naming.html. Retrieved on 30th May 2017.

Xie, Y., Y. Cheng, A. Agrawal & A. Choudhary (2014). Estimating online user location distribution without GPS location. In *Proceedings of the 2014 IEEE International Conference on Data Mining Workshop (ICDMW)*, 936-943. 10.1109/ICDMW.2014.30

Yahoo! Beauty (2017). *#BigThighTwitter Fights Back Against Unrealistic Beauty Standards by Proving Legs of All Sizes Are Sexy*. https://www.yahoo.com/beauty/bigthightwitter-fights-back-against-unrealistic-160253894.html. Retrieved on 30th May 2017.

Yamaguchi, Y., T. Amagasa & H. Kitagawa (2013). Landmark-based user location inference in social media. In *Proceedings of the 1st ACM conference on Online social networks*, 223-234. 10.1145/2512938.2512941.

Yang, W. & L. Mu (2015). GIS analysis of depression among Twitter users. *Applied Geography*, *60*, 217-223. 10.1016/j.apgeog.2014.10.016

Young, S., C. Rivers & B. Lewis (2014). Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes. *Preventive Medicine*, *63*, 112-115. 10.1016/j.ypmed.2014.01.024

Zandbergen, P. (2009). Accuracy of iPhone Locations: A Comparison of Assisted GPS, WiFi and Cellular Positioning. *Transactions in GIS*, *13(1)*, 5-26. 10.1111.j.1467-9671.2009.01152.x

Zandbergen, P. (2012). Comparison of WiFi Positioning on two mobile devices. *Journal of Location Based Services*, *6(1)*, 35-50. 10/1080/17489725.2011.630038

Zandbergen, P. & S. Barbeau (2011). Positional accuracy of assisted GPS data from high-sensitivity GPS-enabled mobile phones. *Journal of Navigation*, *64(3)*, 381-399. 10.1017/S0373463311000051

Zhang, W. & J. Gelernter (2014). Geocoding location expressions in Twitter messages: A preference learning method. *Journal of Spatial Information Science*, *2014(9)*, 37-70. 10.5311/JOSIS.2014.9.170

Zhang, J., J. Sun, R. Zhang & Y. Zhang (2015). Your Actions Tell Where You Are: Uncovering Twitter Users in a Metropolitan Area. In *Proceedings of the 2015 IEEE Conference on Communications Network Security*, 424-432. 10.1109/CNS.2015.7346854

Zhuang, Y., Z. Syed, Y. Li & N. El-Sheimy (2016). Evaluation of two WiFi positioning systems based on autonomous crowdsourcing of handheld devices for indoor navigation. *IEEE Transactions on Mobile Computing, 15(8)*, 1982-1995. 10.1109/TMC.2015.2451641

Zimmer, M. & N. Proferes (2014). A topology of Twitter research: disciplines, methods, and ethics. *Aslib Journal of Information Management*, *66(3)*, 250-261. 10.1108/AJIM-09-2013-0083

Appendix II: Article selection

III.1. Sub question 1

Citation			
Albuquerque, J. de, B. Herbort, A. Brenning & A. Zipf (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. 10.1080/13658816.2014.996567			
Methodology	Event detection	Domain	Disaster management
Study area	Germany	Cited by	34
Real-time	No	Additional sources	Authoritative
Corpus size	60524	Period of gathering	8th June 2013 – 10th June 2013

Citation			
Allen, C., M. Tsou, A. Aslam, A. Nagel & J. Gawron (2016). Applying GIS and Machine Learning Methods to Twitter Data for Multiscale Surveillance of Influenza. *PloS One*, *11(7)*, e0157734. 10.1371/journal.pone.0157734			
Methodology	Event detection	Domain	Health management
Study area	United States	Cited by	0
Real-time	No	Additional sources	No
Corpus size	Not specified	Period of gathering	Not specified

Citation			
Alowibdi, J., U. Buy, S. Philip, S. Ghani & M. Mokbel (2015). Deception detection in Twitter. *Social Network Analysis and Mining*, *5(1)*, 1-16. 10.1007/s13278-015-0273-1			
Methodology	Geolocation inference	Domain	Not specified
Study area	Saudi Arabia	Cited by	4
Real-time	No	Additional sources	No
Corpus size	Not specified	Period of gathering	Not specified

Citation			
Andrienko, G., N. Andrienko, H. Bosch, T. Ertl, G. Fuchs, P. Jankowksi & D. Thom (2013). Thematic patterns in georeferenced tweets through space-time visual analytics. *Computing in Science & Engineering*, *15(3)*, 72-82. 10.1109/MCSE.2013.70			
Methodology	Event detection	Domain	Demographics
Study area	Seattle, WA, United States	Cited by	45
Real-time	No	Additional sources	No
Corpus size	306326	Period of gathering	8th August 2011 – 8th October 2011

Citation			
Bakillah, M., R. Li, & S. Liang (2015). Geo-located community detection in Twitter with enhanced fast-greedy optimization of modularity: the case study of typhoon Haiyan. *International Journal of Geographical Information Science*, *29(2)*, 258-279. 10.1080/13658816.2014.964247			
Methodology	Geolocation inference	Domain	Not specified
Study area	Philippines	Cited by	13
Real-time	No	Additional sources	No
Corpus size	Not specified	Period of gathering	Not specified

Citation			
Banerjee, S., B. Hosack, B. Lim & J. Kostelnick (2013). Social media widget for emergency response. *Issues in Information Systems*, *14(2)*, 289-297. DOI Unknown.			
Methodology	Event detection	Domain	Crisis management
Study area	Illinois, United States	Cited by	0
Real-time	Yes	Additional sources	No
Corpus size	Not specified	Period of gathering	Not specified

Citation			
Cao, G., S. Wang, M. Hwang, A. Padmanabhan, Z. Zhang & K. Soltani (2015). A scalable framework for spatiotemporal analysis of location-based social media data. *Computers, Environment and Urban Systems*, *51*, 70-82. 10.1016/j.compenvurbsys.2015.01.002			
Methodology	Multiple	Domain	None specified
Study area	North America	Cited by	17
Real-time	Yes	Additional sources	No
Corpus size	Not specified	Period of gathering	Not specified

Citation			
Cervone, G., E. Sava, Q. Huang, E. Schnebele, J. Harrison & N. Waters (2016). Using Twitter for tasking remote-sensing data collection and damage assessment: 2013 Boulder flood case study. *International Journal of Remote Sensing*, *37(1)*, 100-124. 10.1080/01431161.2015.1117684.			
Methodology	Event detection	Domain	Disaster management
Study area	Boulder County, CO / Boulder, CO/ Longmont, CO, United States	Cited by	1
Real-time	Yes	Additional sources	Flickr, Commercial, Authoritative
Corpus size	148379	Period of gathering	11th September 2013 – 18th September 2013

Citation			
Chae, J., D. Thom, Y. Jang, S. Kim, T. Ertl & D. Elbert (2014). Public behaviour response analysis in disaster events utilizing visual analytics of microblog data. *Computers & Graphics*, *38*, 51-60. 10.1016/j.cag.2013.10.008			
Methodology	Event detection	Domain	Disaster management
Study area	NY / NJ / North-East US Coast / Moore, OK, United States	Cited by	46
Real-time	Yes	Additional sources	No
Corpus size	Not specified	Period of gathering	26th October 2012 – 30th October 2012

Citation			
Chen, X., G. Elmes, X. Ye & J. Chang (2016). Implementing a real-time Twitter-based system for resource dispatch in disaster management. *GeoJournal*, *81(6)*, 863-873. 10.1007/s10708-016-9745-8			
Methodology	Event detection	Domain	Disaster management
Study area	Not specified	Cited by	0
Real-time	Yes	Additional sources	No
Corpus size	Not specified	Period of gathering	Not specified

Citation			
Cheng, T. & T. Wicks (2014). Event detection using Twitter: a spatio-temporal approach. *PloS One*, *9(6)*, e97807. 10.1371/journal.pone.0097807			
Methodology	Event detection	Domain	None specified
Study area	London, United Kingdom	Cited by	17
Real-time	Yes	Additional sources	No
Corpus size	183731	Period of gathering	7th January 2013 – 18th January 2013

Citation			
Chorianopoulos, K. & K. Talvis (2016). Flutrack.org: Open-source and linked data for epidemiology. *Health informatics journal*, *22(4)*, 962-974. 10.1177/1460458215599822			
Methodology	Event detection	Domain	Health management
Study area	United States	Cited by	0
Real-time	Yes	Additional sources	No
Corpus size	Not specified	Period of gathering	2nd December 2012 – 7th April 2013

Citation			
Crampton, J., M. Graham, A. Poorthuis, T. Shelton, M. Stephens, M. Wilson & M. Zook (2013). Beyond the geotag: situating 'big data' and leveraging the potential of the geoweb. *Cartography and Geographic Information Science*, *40(2)*, 130-139. 10.1080/15230406.2013.777137.			
Methodology	Multiple	Domain	Not specified
Study area	Lexington, KY, United States	Cited by	162
Real-time	No	Additional sources	No
Corpus size	25	Period of gathering	Not specified

Citation			
Croitoru, A., A. Crooks, J. Radzikowski & A. Stefanidis (2013). Geosocial gauge: a system prototype for knowledge discovery from social media. *International Journal of Geographical Information Science*, *27(12)*, 2483-2508. 10.1080/13658816.2013.825724			
Methodology	Multiple	Domain	Not specified
Study area	United States	Cited by	41
Real-time	Yes	Additional sources	Yes
Corpus size	854	Period of gathering	Not specified

Citation			
Croitoru, A., N. Wayant, A. Crooks, J. Radzikowski & A. Stefanidis (2015). Linking cyber and physical spaces through community detection and clustering in social media feeds. *Computer, Environment and Urban Systems*, *53*, 47-64. 10.1016/j.compenvurbsys.2014.11.002			
Methodology	Multiple	Domain	Not specified
Study area	New York, NY / Boston, MA, United States	Cited by	12
Real-time	No	Additional sources	No
Corpus size	1110594	Period of gathering	Not specified

Citation			
Crooks, A., A. Croitoru, A. Stefanidis & J. Radzikowski (2013). #Earthquake: Twitter as a distributed sensor system. *Transactions in GIS*, *17(1)*, 124-147. 10.1111/j.1467-9671.2012.01359.x			
Methodology	Event detection	Domain	Disaster management
Study area	Mineral, VA, United States	Cited by	157
Real-time	No	Additional sources	No
Corpus size	21364	Period of gathering	23rd August 2011

Citation			
Crooks, A., D. Masad, A. Croitoru, A. Cotnoir, A. Stefanidis & J. Radzikowski (2014). International relations: State-driven and citizen-driven networks. *Social Science Computer Review*, *32(2)*, 205-220. 10.1177/0894439313506851			
Methodology	Social network analysis	Domain	Not specified
Study area	World	Cited by	6
Real-time	No	Additional sources	Yes
Corpus size	Not specified	Period of gathering	Not specified

Citation			
Fohringer, J., D. Dransch, H. Kreibich & K. Schröter (2015). Social media as an information source for rapid flood inundation mapping. *Natural Hazards and Earth System Sciences*, *15(12)*, 2725-2738. 10.5194/nhess-15-2725-2015			
Methodology	Event detection	Domain	Disaster management
Study area	Dresden, Germany	Cited by	11
Real-time	Yes	Additional sources	Flickr
Corpus size	Not specified	Period of gathering	5th June to 7th June 2013

Citation			
Ghosh, D. & E. Guha (2013). What are we 'tweeting' about obesity? Mapping tweets with topic modelling and Geographic Information System. *Cartography and geographic information science*, *40(2)*, 90-102. 10.1080/15230406.2013.776210			
Methodology	Event detection	Domain	Topic modelling
Study area	United States	Cited by	52
Real-time	No	Additional sources	No
Corpus size	455981	Period of gathering	1st October 2011 – 31th March 2012

Citation			
Gu, Y., Z. Qian & F. Chen (2016). From Twitter to detector: Real-time traffic incident detection using social media data. *Transportation Research Part C: Emerging Technologies*, *67*, 321-342. 10.1016/j.trc.2016.02.011			
Methodology	Event detection	Domain	Crisis management
Study area	Pittsburgh, PA / Philadelphia, PA, United States	Cited by	3
Real-time	No	Additional sources	No
Corpus size	973	Period of gathering	1st September 2014 – 30th September 2014

Citation			
Guan, X. & C. Chen (2014). Using social media data to understand and assess disasters. *Natural Hazards*, *74(2)*, 837-850. 10.1007/s11069-014-1217-1.			
Methodology	Event detection	Domain	Disaster management
Study area	North-East US Coast	Cited by	21
Real-time	No	Additional sources	Authoritative
Corpus size	1577402	Period of gathering	26th October 2012 – 7th November 2012

Citation			
Guo, D. & C. Chen (2014). Detecting Non-personal and Spam Users on Geo-tagged Twitter Network. *Transactions in GIS*, *18(3)*, 370-384. 10.1111/tgis.12101			
Methodology	Social network analysis	Domain	Not specified
Study area	North-East US Coast	Cited by	15
Real-time	No	Additional sources	No
Corpus size	Not specified	Period of gathering	Not specified

Citation			
Han, S., Y. Tsou & K. Clarke (2015). Do global cities enable global views? Using Twitter to quantify the level of geographical awareness of US cities. *PloS One*, *10(7)*, e0132464. 10.1371/journal.pone.0132464			
Methodology	Event detection	Domain	Topic modelling
Study area	United States	Cited by	4
Real-time	No	Additional sources	No
Corpus size	5013608	Period of gathering	1st December 2013 – 28th February 2014 / 10th December 2013 – 10th February 2014

Citation			
Hong, I. (2016). Python-based Integrated Architecture for Geotweet Analysis. *International Journal of Software Engineering and Its Applications*, *10(2)*, 247-256. 10.14257/ijseia.2016.10.2.20			
Methodology	Event detection	Domain	Not specified
Study area	South-Korea	Cited by	0
Real-time	No	Additional sources	No
Corpus size	47000	Period of gathering	1st July 2015 – 31st July 2015

Citation			
Hua, T., F. Chen, L. Zhao, C. Lu & N. Ramakrishnan (2016). Automatic targeted-domain spatiotemporal event detection in Twitter. *GeoInformatica*, *20(4)*, 765-795. 10.1007/s10707-016-0263-0			
Methodology	Event detection	Domain	Not specified
Study area	Mexico	Cited by	0
Real-time	Yes	Additional sources	No
Corpus size	Not specified	Period of gathering	1st July 2012 – 31st May 2013

Citation			
Huang, Q. (2017). Mining online footprints to predict user's next location. *International Journal of Geographical Information Science*, *31(3)*, 523-541. 10.1080/13658816.2016.1209506			
Methodology	Geolocation inference	Domain	Not specified
Study area	Washington D.C., United States	Cited by	0
Real-time	No	Additional sources	No
Corpus size	Not specified	Period of gathering	Not specified

Citation			
Huang, Q. & C. Xu (2014). A data-driven framework for archiving and exploring social media data. *Annals of GIS*, *20(4)*, 265-277. 10.1080/19475683.2014.942697			
Methodology	Methodology	Domain	Not specified
Study area	Not specified	Cited by	10
Real-time	Yes	Additional sources	Yes
Corpus size	Not specified	Period of gathering	Not specified

Citation			
Huang, Q. & D. Wong (2016). Activity patterns, socioeconomic status and urban spatial structure: what can social media data tell us? *International Journal of Geographical Information Science*, *30(9)*, 1873-1898. 10.1080/13658816.2016.1145225			
Methodology	Geolocation inference	Domain	Not specified
Study area	Washington D.C., United States	Cited by	6
Real-time	No	Additional sources	No
Corpus size	Not specified	Period of gathering	Not specified

Citation			
Huck, J., D. Whyatt & P. Coulton (2015). Visualizing patterns in spatially ambiguous point data. *Journal of Spatial Information Science*, *2015(10)*, 47-66. 10.5311/JOSIS.2015.10.211			
Methodology	Event detection	Domain	Not specified
Study area	Great Britain	Cited by	1
Real-time	No	Additional sources	No
Corpus size	550171	Period of gathering	29th April 2011

Citation			
Jian, B., D. Ma, J. Yin & M. Sandberg (2016). Spatial Distribution of City Tweets and Their Densities. *Geographical Analysis*, *48*, 337-351. 10.1111/gean.12096			
Methodology	Geolocation inference	Domain	Not specified
Study area	Paris, France / Toulouse, France / Berlin, Germany / Munich, Germany / London, United Kingdom / Birmingham, United Kingdom	Cited by	0
Real-time	No	Additional sources	No
Corpus size	Not specified	Period of gathering	Not specified

Citation			
Jovanovic, V. & D. Vukelic (2015). Using Geosocial Analysis for Real-time Monitoring the Marine Environments. *Journal of Environmental Protection and Ecology*, *16(4)*, 1344-1352. DOI Unknown.			
Methodology	Geolocation inference	Domain	Not specified
Study area	Malta	Cited by	0
Real-time	Yes	Additional sources	Yes
Corpus size	Not specified	Period of gathering	Not specified

Citation			
Jung, J. (2014). Code clouds: Qualitative geovisualization of geotweets. *The Canadian Geographer / Le Géographe canadien*, *59(1)*, 52-68. 10.1111/cag.12133			
Methodology	Event Detection	Domain	Not specified
Study area	King County, WA, United States	Cited by	4
Real-time	Yes	Additional sources	No
Corpus size	14858	Period of gathering	Not specified

Citation			
Kang, Y., J. Park & A. Kang (2015). An analysis on the spatial characteristics of satisfaction on the residential environment using tweets. *International Journal of Geospatial and Environmental Research*, *1(2)*, 5. DOI Unknown.			
Methodology	Event Detection	Domain	Topic modelling
Study area	South Korea	Cited by	1
Real-time	No	Additional sources	No
Corpus size	516	Period of gathering	1st November 2012 – 31st January 2013

Citation			
Kim, K., I. Kojima & H. Ogawa (2016). Discovery of local topics by using latent spatio-temporal relationships in geo-social media. *International Journal of Geographical Information Science*, *30(9)*, 1899-1922. 10.1080/13658816.2016.1146956			
Methodology	Event Detection	Domain	Topic modelling
Study area	United States	Cited by	2
Real-time	Yes	Additional sources	No
Corpus size	1928937	Period of gathering	10th June 2015 – 28th June 2015

Citation			
Kirilenko, A. & S. Stepchenkova (2014). Public microblogging on climate change: One year of Twitter worldwide. *Global Environmental Change*, *26*, 171-182. 10.1007/s11069-014-1217-1			
Methodology	Event Detection	Domain	Topic modelling
Study area	World	Cited by	24
Real-time	No	Additional sources	No
Corpus size	1872263	Period of gathering	Not specified

Citation			
Lampoltshammer, T., O. Kounadi, I. Sitko & B. Hawelka (2014). Sensing the public's reaction to crime news using the 'Links Correspondence Method". *Applied Geography*, *52*, 57-66. 10.1016/j.apgeog.2014.04.016			
Methodology	Event Detection	Domain	Topic modelling
Study area	London, United Kingdom	Cited by	7
Real-time	No	Additional sources	No
Corpus size	1821	Period of gathering	1st January 2012 – 31th December 2012

Citation			
Lansley, G. & P. Longley (2016). The geography of Twitter topic in London. *Computers, Environment and Urban Systems*, *58*, 85-96. 10.1016/j.compenvurbsys.2016.04.002			
Methodology	Event Detection	Domain	Topic modelling
Study area	London, United Kingdom	Cited by	6
Real-time	No	Additional sources	Authoritative
Corpus size	1301004	Period of gathering	1st January 2013 – 31th December 2013

Citation			
Lenormand, M., M. Picornell, O. Cantu-Ros, A. Tugores, T. Louail, R. Herranz & J. Ramasco (2014). Cross-checking different sources of mobility information. *PloS One*, *9(8)*, e105184. 10.1371/journal.pone.0105184			
Methodology	Geolocation inference	Domain	Not specified
Study area	Barcelona / Madrid, Spain	Cited by	29
Real-time	No	Additional sources	No
Corpus size	Not specified	Period of gathering	Not specified

Citation			
Li, L., M. Goodchild & B. Xu (2013). Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartography and Geographic Information Science*, *40(2)*, 61-77. 10.1080/15230406.2013777139			
Methodology	Event detection	Domain	Demographics
Study area	Los Angeles, United States	Cited by	111
Real-time	No	Additional sources	Flickr
Corpus size	19758954	Period of gathering	21st January 2011 – 7th March 2011

Citation			
Lin, J. & R. Cromley (2015). Evaluating geo-located Twitter data as a control layer for areal interpolation of population. *Applied Geography*, *58*, 41-47. 10.1016/j.apgeog.2015.01.006			
Methodology	Geolocation inference	Domain	Not specified
Study area	Hartford County, CO, United States	Cited by	4
Real-time	Yes	Additional sources	Yes
Corpus size	Not specified	Period of gathering	Not specified

Citation			
Lloyd, A. & J. Cheshire (2017). Deriving retail centre locations and catchments from geo-tagged Twitter data. *Computers, Environment and Urban Systems*, *61*, 108-118. 10.1016/j.compenvurbsys.2016.09.006			
Methodology	Not specified	Domain	Not specified
Study area	Not specified	Cited by	0
Real-time	Not specified	Additional sources	Not specified
Corpus size	Not specified	Period of gathering	Not specified

Citation			
Longley, P. & M. Adnan (2016). Geo-temporal Twitter demographics. *International Journal of Geographical Information Science*, *30(2)*, 369-289. 10.1080/13658816.2015.1089441			
Methodology	Geolocation inference	Domain	Not specified
Study area	London, United Kingdom	Cited by	7
Real-time	No	Additional sources	No
Corpus size	Not specified	Period of gathering	Not specified

Citation			
Longley, P., M. Adnan & G. Lansley (2015). The geotemporal demographics of Twitter usage. *Environment and Planning A*, *47(2)*, 465-484. 10.1068/a130122p			
Methodology	Geolocation inference	Domain	Not specified
Study area	London, United Kingdom	Cited by	16
Real-time	No	Additional sources	Yes
Corpus size	Not specified	Period of gathering	Not specified

Citation			
Lwin, K., K. Sugiura & K. Zettsu (2016). Space-time multiple regression model for grid-based population estimation in urban areas. *International Journal of Geographical Information Science*, *30(8)*, 1579-1593. 10.1080/13658816.2016.1143099			
Methodology	Geolocation inference	Domain	Not specified
Study area	Kobe City, Japan	Cited by	1
Real-time	No	Additional sources	Yes
Corpus size	Not specified	Period of gathering	Not specified

Citation			
Malleson, N. & M. Andresen (2015). Exploring the impact of ambient population measures on London crime hotspots. *Journal of Criminal Justice*, *46*, 52-63.10.1016/j.jcrimjus.2016.03.002			
Methodology	Event detection	Domain	Crime management
Study area	London, United Kingdom	Cited by	3
Real-time	No	Additional sources	Authoritative/Commercial
Corpus size	204159	Period of gathering	1st September 2013 – 30th September 2013

Citation			
Malleson,. N. & M. Andresen (2015). The impact of using social media data in crime rate calculations: shifting hot spots and changing spatial patterns. *Cartography and Geographic Information Science*, *42(2)*, 112-121. 10.1080/15230406.2014.905756			
Methodology	Event detection	Domain	Crime management
Study area	Leeds, United Kingdom	Cited by	17
Real-time	No	Additional sources	Authoritative
Corpus size	1955655	Period of gathering	22nd June 2011 – 14th April 2013

Citation			
Mcardle, G., E. Furey, A. Lawlor & A. Pozdnoukhow (2014). Using digital footprints for a city-scale traffic simulation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *5(3)*, 41. 10.1145/2517028			
Methodology	Geolocation inference	Domain	Not specified
Study area	Dublin, Ireland	Cited by	5
Real-time	No	Additional sources	Yes
Corpus size	Not specified	Period of gathering	Not specified

Citation			
Mearns, G., R. Simmonds, R. Richardson, M. Turner, P. Watson & P. Missier (2014). Tweet my street: a cross-disciplinary collaboration for the analysis of local Twitter data. *Future Internet*, *6(2)*, 378-396. 10.3390/fi6020378			
Methodology	Multiple	Domain	Not specified
Study area	Newcastle, United Kingdom	Cited by	5
Real-time	Yes	Additional sources	No
Corpus size	Not specified	Period of gathering	Not specified

Citation			
Middleton, S., L. Middleton & S. Modaferri (2014). Real-time crisis mapping of natural disasters using social media. *IEEE Intelligent Systems*, *29(2)*. 9-17. 10.1109/MIS.2013.126			
Methodology	Multiple	Domain	Disaster management
Study area	NY / NJ / Moore, OK, United States	Cited by	52
Real-time	Yes	Additional sources	No
Corpus size	Not specified	Period of gathering	Not specified

Citation			
Nagel, A., M. Tsou, B. Spitzberg, L. An, M. Gawron, D. Gupta, J. Yang, S. Han, M. Peddecord, S. Lindsay & M. Sawyer (2013). The Complex Relationship of Realspace Events and Messages in Cyberspace: Case Study of Influenza and Pertussis Using Tweets. *Journal of Medical Research*, *15(10)*, e237. 10.2196/jmir.2705.			
Methodology	Event detection	Domain	Health management
Study area	United States	Cited by	29
Real-time	No	Additional sources	No
Corpus size	169322	Period of gathering	31th August 2012 – 4th March 2013

Citation			
Nelson, J., S. Quinn, B. Swedberg, W. Chu & A. MacEachren (2015). Geovisual Analytics Approach to Exploring Public Political Discourse on Twitter. *ISPRS International Journal of Geo-Informatio*n, *4(1)*, 337-366. 10.3390/ijgi401337.			
Methodology	Event detection	Domain	Topic modelling
Study area	United States	Cited by	3
Real-time	No	Additional sources	No
Corpus size	70000	Period of gathering	1st September 2013 – 27th October 2013

Citation			
Nguyen, Q., D. Li, H. Meng, S. Kath, E. Nsoesie, F. Li & M. Wen (2016b). Building a National Neighborhood Dataset From Geotagged Twitter Data for Indicators of Happiness, Diet, and Physical Activity. *JMIR Public Health and Surveillance*, *2(2)*. 10/2196/publichealth.5869			
Methodology	Event detection	Domain	Health management
Study area	United States	Cited by	0
Real-time	Yes	Additional sources	Authoritative
Corpus size	79848992	Period of gathering	1st February 2015 – 31th March 2016

Citation			
Nguyen, Q., S. Kath, H. Meng, D. Li, K. Smith, J. VanDerslice, M. Wen & F. Li (2016a). Leveraging geotagged Twitter data to examine neighborhood happiness, diet, and physical activity. *Applied Geography*, *73*, 77-88. 10.1016/j.apgeog.2016.06.003.			
Methodology	Event detection	Domain	Health management
Study area	Salt Lake City, UT / San Fransisco, CA / New York, NY, United States	Cited by	2
Real-time	No	Additional sources	Authoritative
Corpus size	2848900	Period of gathering	1st February 2015 – 31th August 2015

Citation			
Oleksiak, P. (2014). Analysing and processing of geotagged social media. *Information Systems in Management*, *4(3)*, 250-260. DOI Unknown.			
Methodology	Event detection	Domain	Not specified
Study area	World	Cited by	0
Real-time	Yes	Additional sources	Authoritative
Corpus size	Not specified	Period of gathering	Not specified

Citation			
Padmanabhan, A., S. Wang, G. Cao, M. Hwang, Z. Zhang, Y. Gao & Y. Liu (2014). A cyberGIS application for interactive analysis of massive location-based social media. *Concurrency and Computation: Practice and Experience*, *26(13)*, 2253-2265. 10.1002/cpe.3287			
Methodology	Event detection	Domain	Health Management
Study area	Not specified	Cited by	19
Real-time	Yes	Additional sources	No
Corpus size	Not specified	Period of gathering	Not specified

Citation			
Panteras, G., S. Wise, X. Lu, A. Croitoru, A. Crooks & A. Stefanidis (2015). Triangulating social multimedia content for event localization using Flickr and Twitter. *Transactions in GIS*, *19(5)*, 694-715. 10.1111/tgis.12122			
Methodology	Event detection	Domain	Not specified
Study area	Colorado Springs, CO, United States	Cited by	10
Real-time	No	Additional sources	Flickr
Corpus size	97866	Period of gathering	Not specified

Citation			
Patel, N., F. Stevens, Z. Huang, A. Gaughan, I. Elyazar & A. Tatem (2016). Improving large area population mapping using geotweet densities. *Transactions in GIS*, *21(2)*, 317-331. 10.1111/tgis.12214			
Methodology	Geolocation inference	Domain	Not specified
Study area	Indonesia	Cited by	0
Real-time	No	Additional sources	Yes
Corpus size	Not specified	Period of gathering	Not specified

Citation			
Poorazizi, M., A. Hunter & S. Steiniger (2015). A Volunteered Geographic Information Framework to Enable Bottom-Up Disaster Management Platforms. *ISPRS International Journal of Geo-Information*, *4(3)*, 1389-1422. 10.3390/ijgi4031389			
Methodology	Event detection	Domain	Disaster management
Study area	World	Cited by	0
Real-time	Yes	Additional sources	Flickr, Google Plus, Instagram
Corpus size	440	Period of gathering	4th December 2014 – 17th December 2014

Citation			
Radzikowski, J., A. Stefanidis, K. Jacobsen, A. Croitoru, A. Crooks & P. Delamater (2016). The Measles Vaccination Narrative in Twitter: A Quantitative Analysis. J*MIR Public Health and Surveillance*, *2(1)*, e1. 10.2196/publichealth.5059.			
Methodology	Multiple	Domain	Health management
Study area	United States	Cited by	6
Real-time	No	Additional sources	No
Corpus size	Not specified	Period of gathering	Not specified

Citation			
Shelton, T. (2016). Spatialities of data: Mapping social media 'beyond the geotag'. *GeoJournal*, 1-14. 10.1007/s10708-016-9713-3			
Methodology	Event detection	Domain	Not specified
Study area	World	Cited by	1
Real-time	No	Additional sources	No
Corpus size	Not specified	Period of gathering	Not specified

Citation			
Shelton, T., A. Poorthuis & M. Zook (2015). Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information. *Landscape and Urban Planning*, *142*, 198-211. 10.1016/j.landurbplan.2015.02.020			
Methodology	Event detection	Domain	Demographics
Study area	Louisville, KY, United States	Cited by	26
Real-time	No	Additional sources	No
Corpus size	101399	Period of gathering	Late June 2012 - Early July 2014

Citation			
Shook, E. & V. Turner (2016). The socio-environmental data explorer (SEDE): a social media-enhanced decision support system to explore risk perception to hazard events. *Cartography and Geographic Information Science*, *43(5)*, 427-441. 10.1080/15230406.2015.1131627			
Methodology	Event detection	Domain	Disaster management
Study area	US East coast	Cited by	0
Real-time	Yes	Additional sources	Authoritative
Corpus size	8000000	Period of gathering	19th January 2015 – 23th February 2015

Citation			
Stefanidis, A., A. Crooks & J. Radzikowski (2013). Harvesting ambient geospatial information from social media feeds. *GeoJournal*, *78(2)*, 319-338. 10.1007/s10708-011-9438-2			
Methodology	Multiple	Domain	Not specified
Study area	Cairo, Egypt / Tokyo, Japan	Cited by	156
Real-time	No	Additional sources	Yes
Corpus size	Not specified	Period of gathering	Not specified

Citation			
Steiger, E., T. Ellersiek, B. Resch & A. Zipf (2015). Uncovering latent mobility patterns from twitter during mass events. *GI_Forum*, *1*, 525-534. 10.1553/giscience2015			
Methodology	Event detection	Domain	Crisis management
Study area	Boston, MA, United States	Cited by	2
Real-time	No	Additional sources	No
Corpus size	251771	Period of gathering	25th October 2013 – 5th November 2013

Citation			
Steiger, E., B. Resch & A. Zipf (2016). Exploration of spatiotemporal and semantic clusters of Twitter data using unsupervised neural networks. *International Journal of Geographical Information Science*, *30(9)*, 1694-1716. 10.1080/13658816.2015.1099658			
Methodology	Event detection	Domain	Not specified
Study area	London, United Kingdom	Cited by	8
Real-time	No	Additional sources	No
Corpus size	41200000	Period of gathering	1st January 2014 – 31th December 2014

Citation			
Steiger, E., R. Westerholt, B. Resch & A. Zipf (2016). Twitter as an indicator for whereabouts of people? Correlating Twitter with UK census data. *Computers, Environment and Urban Systems*, *54*, 255-265. 10.1007/s11069-014-1217-1			
Methodology	Geolocation inference	Domain	Not specified
Study area	London, United Kingdom	Cited by	11
Real-time	No	Additional sources	Yes
Corpus size	Not specified	Period of gathering	Not specified

Citation			
Shelton, T., A. Poorthuis, M. Graham & M. Zook (2014). Mapping the data shadows of Hurricane Sandy: Unconvering the sociospatial dimensions of 'big data'. *Geoforum*, *52*, 167-179. 10.1016/j.geoforum.2014.01.006			
Methodology	Event detection	Domain	Disaster management
Study area	New York, NY / Los Angeles, CA, United States	Cited by	56
Real-time	No	Additional sources	No
Corpus size	141909	Period of gathering	24th October 2012 – 31th October 2012

Citation			
Wachowicz, M., M. Arteaga, S. Cha & Y. Bourgeois (2016). Developing a streaming data processing workflow for querying space-time activites from geotagged tweet. *Computers, Environment and Urban Systems*, *59*, 256-268. 10.1016/j.compenvurbsys.2015.12.001			
Methodology	Event detection	Domain	Not specified
Study area	Canada	Cited by	0
Real-time	Yes	Additional sources	No
Corpus size	Not specified	Period of gathering	1st February 2014 – 31th July 2014

Citation			
Wang, Q. & J. Taylor (2015). Process Map for Urban-Human Mobility and Civil Infrastructure Data Collection Using Geosocial Networking Platforms. *Journal of Computing in Civil Engineering*, *30(2)*, 04015004. 10.1061/(ASCE)CP.1943-5487.0000469			
Methodology	Geolocation inference	Domain	Not specified
Study area	New York, NY, United States	Cited by	7
Real-time	No	Additional sources	No
Corpus size	Not specified	Period of gathering	Not specified

Citation			
Wang, F., E. Mack & R. Maciewjeski (2017). Analyzing Entrepreneurial Social Network with Big Data. *Annals of the American Association of Geographers*, *107(1)*, 130-150. 10.1080/24694452.2016.1222263			
Methodology	Social network analysis	Domain	Not specified
Study area	United States	Cited by	0
Real-time	No	Additional sources	No
Corpus size	Not specified	Period of gathering	Not specified

Citation			
Widener, M. & L. Wenwen (2014). Using geolocated Twitter data to monitor the prevalence of healthy and unhealthy food references across the US. *Applied Geography*, *54*, 189-197. 10.1016/j.apgeog.2014.07.017			
Methodology	Event detection	Domain	Health management
Study area	United States	Cited by	29
Real-time	No	Additional sources	No
Corpus size	128914	Period of gathering	26th June 2013 – 22nd July 2014

Citation			
Xu, C., H. Qin & M. Yu (2015). Visualising spatiotemporal trajectories of mobile social media users using space-time cube. *Cartography and Geographic Information Science*, *42(sup1)*, 75-83. 10.1080/15230406.2015.1059253			
Methodology	Event detection	Domain	Not specified
Study area	United States	Cited by	0
Real-time	Yes	Additional sources	No
Corpus size	Not specified	Period of gathering	Not specified

Citation			
Xu, C., D. Wong & C. Yang (2013). Evaluating the 'geographical awareness' of individuals: an exploratory analysis of Twitter data. *Cartography and Geographic Information Science*, *40(2)*, 103-115. 10.1080/15230406.2013.776212			
Methodology	Geolocation inference	Domain	Not specified
Study area	United States	Cited by	29
Real-time	Yes	Additional sources	No
Corpus size	Not specified	Period of gathering	Not specified

Citation			
Yang, W. & L. Mu (2015). GIS analysis of depression among Twitter users. *Applied Geography*, *60*, 217-223. 10.1016/j.apgeog.2014.10.016			
Methodology	Event detection	Domain	Health management
Study area	New York, NY, United States	Cited by	8
Real-time	No	Additional sources	No
Corpus size	402	Period of gathering	5th September 2013 – 5th March 2014

Citation			
Yang, W., L. Mu & Y. Shen (2015). Effect of climate and seasonality on depressed mood among twitter users. *Applied Geography*, *63*, 184-191. 10.1016/j.apgeog.2015.06.017			
Methodology	Event detection	Domain	Health management
Study area	United States	Cited by	1
Real-time	No	Additional sources	No
Corpus size	Not specified	Period of gathering	5th September 2013 – 3rd September 2014

Citation			
Yang, J., M. Tsou, C. Jung, C. Allen, B. Spitzberg, J. Gawron & S. Han (2016). Social media analysis and research testbed (SMART): Exploring spatiotemporal patterns of human dynamics with geo-targeted social media messages. *Big Data & Society*, *3(1)*, 2053951716652914. 10.1177/2053951716652914			
Methodology	Event detection	Domain	Not specified
Study area	United States / West-Africa	Cited by	0
Real-time	Yes	Additional sources	No
Corpus size	Not specified	Period of gathering	Not specified

Citation			
Yang, C., I. Jensen & P. Rosen (2014). A multiscale approach to network event identification using geolocated twitter data. *Computing*, *96(1)*, 3-13. 10.1007/s00607-013-0285-5			
Methodology	Multiple	Domain	Not specified
Study area	Salt Lake City, UT, United States	Cited by	4
Real-time	No	Additional sources	No
Corpus size	Not specified	Period of gathering	Not specified

Citation			
Young, S., C. Rivers & B. Lewis (2014). Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes. *Preventive Medicine*, *63*, 112-115. 10.1016/j.ypmed.2014.01.024			
Methodology	Event detection	Domain	Health management
Study area	United States	Cited by	36
Real-time	No	Additional sources	No
Corpus size	9800	Period of gathering	26th May 2012 – 9th December 2012

Citation			
Zhang, S. & R. Feick (2016). Understanding Public Opinions from Geosocial Media. *ISPRS International Journal of Geo-information*, *5(6)*, 74. 10.3390/ijgi5060074			
Methodology	Event detection	Domain	Topic modelling
Study area	Waterloo, ON, United States	Cited by	0
Real-time	No	Additional sources	No
Corpus size	4889	Period of gathering	1st March 2014 – 31st July 2015

Citation			
Zhou, X. & L. Zhang (2016). Crowdsourcing functions of the living city from Twitter and Foursquare data. *Cartography and Geographic Information Science*, *43(5)*, 393-404. 10.1080/15230406.2015.1128852			
Methodology	Multiple	Domain	Demographics
Study area	Boston, MA / Chicago, IL, United States	Cited by	1
Real-time	Yes	Additional sources	Yes
Corpus size	Not specified	Period of gathering	Not specified

III.2. Sub question 3

Citation			
Ahmed, A., L. Hong. & A. Smola (2013). Hierarchical geographical modelling of user locations from social media posts. In *Proceedings of the 22nd International Conference on World Wide Web*, 25-36. 10.1145/2488388.2488392			
GIM-type	Content-based	Inference subject	User, static
Study area	Not specified	Cited by	43
Language	English	General methodology	Text mining
Methods mentioned	Content Analysis and indexing; Learning - Parameter Learning; NLP; Non-parametric Bayesian Models; Chinese Restaurant Process; Generative model		

Citation			
Apreleva, S. & A. Cantarero (2015). Predicting the location of users on Twitter from low density graphs. In *Proceedings of the 2015 IEEE Conference on Big Data*, 976-983. 10.1109/BigData.2015.7363848			
GIM-type	Network-based	Inference subject	User, static
Study area	Not specified	Cited by	1
Language	English	General methodology	Tie-strength
Methods mentioned	Graph theory; Gaussian distributions; Diffusion processes; Data models		

Citation			
Berggren, M., J. Karlgren, R. Östling & M. Parkvall (2016). Inferring the location of authors from words in their text. *arXiv preprint arXiv: 1612.06671*.			
GIM-type	Content-based	Inference subject	User, static
Study area	Not specified	Cited by	1
Language	Swedish	General methodology	Not specified
Methods mentioned	Not specified		

Citation			
Cha, M., Y. Gwon, H. Kung (2015). Twitter Geolocation and Regional Classification via Sparse Coding. In *Proceedings of the 9th International AAAI Conference on Web and Social Media*, 582-585. DOI Unknown.			
GIM-type	Hybrid	Inference subject	User, static
Study area	Not specified	Cited by	10
Language	Not specified	General methodology	Text mining
Methods mentioned	Sparse coding; Dictionary learning; Pattern recognition		

Citation			
Chandra, S., L. Khan & F. Muhaya (2011). Estimating twitter user location using social interactions – A Content Based Approach. In *Proceedings of the 2011 IEEE Third International Conference on Privacy, Security Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom),* 838-843. 10.1109/PASSAT/SocialCom.2011.120			
GIM-type	Content-based	Inference subject	User, static
Study area	Not specified	Cited by	68
Language	Not specified	General methodology	Text mining + Tie-strength
Methods mentioned	Data mining		

Citation			
Chang, H., D. Lee, M. Eltaher & J. Lee (2012). @ Phillies tweeting from Philly? Predicting Twitter user locations with spatial word usage. In *Proceedings of the 2012 International Conference on Advances in Social Network Analysis and Mining*, 111-118. 10.1109/ASONAM.2012.29			
GIM-type	Content-based	Inference subject	User, static
Study area	Not specified	Cited by	47
Language	Not specified	General methodology	Machine learning
Methods mentioned	Probability models; Gaussian Mixture Model; Maximum Likelihood; Non-Localness; Geometric-Localness		

Citation			
Chauhan, A., K. Kummamuru & D. Toshniwal (2017). Prediction of places of visit using tweets. *Knowledge and Information Systems*, *50(1)*, 145-166. 10.1107/s10115-016-0936-x			
GIM-type	Content-based	Inference subject	User, mobile
Study area	New York, NY, United States	Cited by	0
Language	Not specified	General methodology	Not specified
Methods mentioned	Not specified		

Citation			
Chen, J., Y. Liu & M. .Zou (2016). Home location profiling for users in social media. *Information & Management*, *53(1)*, 135-143. 10.1016/j.im.2015.09.008			
GIM-type	Network-user	Inference subject	User, home location, static
Study area	Not specified	Cited by	2
Language	Not specified	General methodology	Tie-strength
Methods mentioned	Tie-strength; Social tie		

Citation			
Cheng, Z., J. Caverlee, & K. Lee (2013). A content-driven framework for geolocating microblog users. *ACM Transactions on Intelligent Systems and Technology*, *4(1)*, 2. 10.1145/2414425.2414427			
GIM-type	Content-based	Inference subject	User, static
Study area	Not specified	Cited by	25
Language	Not specified	General methodology	Text mining
Methods mentioned	Data mining; Spatial data mining; Text mining		

Citation			
Compton, R., D. Jurgens & D. Allen (2014). Geotagging one hundred million twitter accounts with total variation minimization. In *Proceedings of the 2014 IEEE International Conference on Big Data*, 393-401. 10.1109/BigData.2014.7004256			
GIM-type	Network-based	Inference subject	User, static
Study area	Not specified	Cited by	44
Language	Not specified	General methodology	Tie-strength
Methods mentioned	Data mining		

Citation			
Davis Jr, C., G. Pappa, D. de Oliveira & D. Arcanjo (2011). Inferring the location of twitter messages based on user relationship. *Transactions in GIS*, *15(6)*, 735-751. 10.111/j.1467-9671.2011.01297.x			
GIM-type	Network-based	Inference subject	Message, static
Study area	Not specified	Cited by	100
Language	Multiple	General methodology	Not specified
Methods mentioned	Not specified		

Citation			
Duong-Trung, N., N. Schilling & L. Schmid-Thieme (2016). Near Real-time Geolocation Prediction in Twitter Streams via Matrix Factorization Based Regression. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, 1973-1976. 10.1145/2983323.2983887			
GIM-type	Content-based	Inference subject	User, static
Study area	Contiguous US / North-America / World	Cited by	0
Language	Multiple	General methodology	GIS
Methods mentioned	GIS; Models of learning; Matrix Factorization; Regression		

Citation			
Gonzalez, R., G. Figueroa & Y. Chen (2012). Tweolocator: a non-intrusive geographical locator system for twitter. In *Proceedings of the 5th ACM SIGSPATIAL International Workshop on Location-based social networks*, 24-31. 10.1145/2442796.2442804			
GIM-type	Content-based	Inference subject	User, static
Study area	Not specified	Cited by	11
Language	Multiple	General methodology	Not specified
Methods mentioned	Not specified		

Citation			
Gu, H., H. Hang, Q. Lv & D. Grunwald (2012). Fusing Text and Friendships for Location Inference in Online Social Networks. In *Proceedings of the 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology*, 158-165. 10.1109/WI-IAT.2012.243			
GIM-type	Hybrid	Inference subject	User, static
Study area	Not specified	Cited by	20
Language	Not specified	General methodology	Tie-strength + text mining
Methods mentioned	Not specified		

Citation			
Han, B., P. Cook & T. Baldwin (2014). Text-Based Twitter User Geolocation Prediction. *Journal of Artificial Intelligence Research*, *49*, 451-500. 10.1613/jair.4200			
GIM-type	Content-based	Inference subject	User, static
Study area	Contiguous US / World	Cited by	69
Language	Multiple	General methodology	Text mining
Methods mentioned	Not specified		

Citation			
Hecht, B., L. Hong, B. Suh & E. Chi (2011). Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. In *Proceedings of the SIGCHI conference on human factors in computing systems*, 237-246. 10.1145/1978942.1978976			
GIM-type	Content-based	Inference subject	User, static
Study area	Not specified	Cited by	301
Language	English	General methodology	Text mining
Methods mentioned	Machine learning		

Citation			
Huang, Q., C. Guofeng & C. Wang (2014). From where do tweets originate?: a GIS approach for user location inference. In *Proceedings of the 7th ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, 1-8. 10.1145/2755492.2755494			
GIM-type	Content-based	Inference subject	Message, static
Study area	St. Louis, MO, United States	Cited by	7
Language	Not specified	General methodology	GIS
Methods mentioned	GIS, Spatial clustering, spatiotemporal clustering, Big Data		

Citation			
Huang, Q. & D. Wong (2016). Activity patterns, socioeconomic status and urban spatial structure: what can social media data tell us? *International Journal of Geographical Information Science*, *30(9)*, 1873-1898. 10.1080/13658816.2016.1145225			
GIM-type	Content-based	Inference subject	User, mobile
Study area	Washington D.C., United States	Cited by	6
Language	Not specified	General methodology	Not specified
Methods mentioned	Not specified		

Citation			
Huang, Q. (2017). Mining online footprints to predict user's next location. *International Journal of Geographical Information Science*, *31(3)*, 523-541. 10.1080/13658816.2016.1209506			
GIM-type	Content-based	Inference subject	User, mobile
Study area	Washington D.C., United States	Cited by	0
Language	Not specified	General methodology	Not specified
Methods mentioned	Not specified		

Citation			
Hulden, M., M. Silfverberg & J. Francom (2015). Kernel Density Estimation for Text-based geolocation. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, 145-150. DOI Unknown.			
GIM-type	Content-based	Inference subject	Message, static
Study area	Contiguous US / World	Cited by	3
Language	Multiple	General methodology	Not specified
Methods mentioned	Not specified		

Citation			
Ikawa, Y., M. Enoki & M. Tatsubori (2012). Location inference using microblog messages. In *Proceedings of the 21st International Conference on World Wide Web*, 687-690. 10.1145/2187980.2188181			
GIM-type	Content-based	Inference subject	Message, static
Study area	Not specified	Cited by	58
Language	Not specified	General methodology	Not specified
Methods mentioned	Not specified		

Citation			
Intagorn, S. & K. Lerman(2014). Placing user-generated content on the map with confidence. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advanced in Geographic Information Systems*, 413-416. 10.1145/2666310.2666433			
GIM-type	Content-based	Inference subject	Message, static
Study area	Not specified	Cited by	1
Language	Not specified	General methodology	Not specified
Methods mentioned	Not specified		

Citation			
Ishida, K. (2015). Estimation of User Location and Local Topics Based on Geographical Distribution of Microblogging. *Information Engineering Express*, 1(4), 33-42. DOI Unknown.			
GIM-type	Content-based	Inference subject	User, static
Study area	Not specified	Cited by	0
Language	Not specified	General methodology	Not specified
Methods mentioned	Not specified		

Citation			
Jurgens, D. (2013). That's What Friends Are For: Inferring Location in Online Social Media Platforms Based on Social Relationships. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media, 13*, 273-282. DOI Unknown.			
GIM-type	Network-based	Inference subject	User, static
Study area	Not specified	Cited by	85
Language	Not specified	General methodology	Tie-strength
Methods mentioned	Not specified		

Citation			
Katragadda, S., M. Jin & V. Raghavan (2014). An unsupervised approach to identify location based on the content of user's tweet history. In *Proceedings of the International Conference on Acitive Media Technology*, 311-323. 10.1107/978-3-319-09912-5_26			
GIM-type	Content-based	Inference subject	User, static
Study area	Not specified	Cited by	5
Language	Not specified	General methodology	Not specified
Methods mentioned	Not specified		

Citation			
Kawano, M. & K. Ueda (2016). Where Are You Talking From?: Estimating the Location of tweets Using Recurrent Neural Networks. In *Proceedings of the Second International Conference on IoT in Urban Space*, 57-60. 10.1145/2962735.2962759			
GIM-type	Content-based	Inference subject	Message, static
Study area	Not specified	Cited by	0
Language	Not specified	General methodology	Not specified
Methods mentioned	Not specified		

Citation			
Kinsella, S., V. Murdock & N. O'Hare (2016). "I'm Eating a Sandwich in Glasgow": Modeling Locations with Tweets. In *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents*, 61-68. 10.1145/2065023.2065039			
GIM-type	Content-based	Inference subject	Multiple, static
Study area	Jakarta, Indonesia / New York, NY / Chicago, IL / San Fransisco, CA / Houston, TX, United States / London, United Kingdom / Toronto, Canada / Amsterdam, Netherlands / Sydney, Australia / Santiago, Chile	Cited by	139
Language	Not specified	General methodology	Not specified
Methods mentioned	Not specified		

Citation			
Kong, L., Z. Liu & Y. Huang (2014). Spot: Locating social media users based on social network context. In *Proceedings of the VLDB Endowment, 7(13)*, 1681-1684. 10.14778/2733004.2733060.			
GIM-type	Network-based	Inference subject	User, static
Study area	Not specified	Cited by	9
Language	Not specified	General methodology	Tie-strength
Methods mentioned	Social closeness, local social coefficient		

Citation			
Kotzias, D., T. Lappas & D. Gunopulos (2016). Home is where your friends are: Utilizing the social graph to locate twitter users in a city. *Information Systems*, *57*, 77-87. 10.1016/j.is.2015.10.011			
GIM-type	Hybrid	Inference subject	User, static
Study area	Dublin, Ireland / Manchester, United Kingdom, Boston, MA, United States	Cited by	2
Language	Not specified	General methodology	Tie-strength
Methods mentioned	Social graph		

Citation			
Krishnamurthy, R., P. Kapanipathi, A. Sheth & K. Thirunarayan (2015). Knowledge enabled approach to predict the location of twitter users. In *Proceedings of the European Semantic Web Conference*, 187-201. 10.1007/978-3-319-18818-8_12			
GIM-type	Content-based	Inference subject	User, static
Study area	Not specified	Cited by	4
Language	Not specified	General methodology	Text mining
Methods mentioned	Knowledge based approach		

Citation			
Laylavi, F., A. Rajabifard & M. Kalantari (2016). A multi-element approach to location inference of twitter: A case for emergency response. *ISPRS International Journal of Geo-Information*, *5(5)*, 56. 10.3390/ijgi5050056			
GIM-type	Content-based	Inference subject	Message, static
Study area	Sydney and major regional centres of NSW, Australia	Cited by	2
Language	English	General methodology	Text mining
Methods mentioned	Multi-elemental location inference method		

Citation			
Li, R., S. Wang, H. Deng, R. Wang & K. Chang (2012b). Towards social user profiling: unified and discriminative influence model for inferring home locations. In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1023-1031. 10.1145/2339530.2339692			
GIM-type	Hybrid	Inference subject	User, home location, static
Study area	100 cities with most Twitter user worldwide	Cited by	176
Language	Not specified	General methodology	Tie strength + Text mining
Methods mentioned	Data Mining; Influence Model; Unified discriminative influence model		

Citation			
Li, R., S. Wang & K. Chang (2012a). Multiple locations profiling for users and relationships from social network and content. In *Proceedings of the VLDB Endowment*, *5(11)*, 1603-1614. 10.14778/2350229.2350273			
GIM-type	Hybrid	Inference subject	User, mobile
Study area	Not specified	Cited by	39
Language	Not specified	General methodology	Not specified
Methods mentioned	Not specified		

Citation			
Lingad, J., S. Karimi & J. Yin (2013). Location extraction from disaster-related microblogs. In *Proceedings of the 22nd International Conference on World Wide Web*, 1017-1020. 10.1145/2487788.2488108			
GIM-type	Content-based	Inference subject	Message, static
Study area	Not specified	Cited by	43
Language	English	General methodology	Text mining
Methods mentioned	Text Analysis; Named Entity Recognition, Social Media Mining		

Citation			
Liu, Z. & Y. Huang (2016). Closeness and structure of friends help to estimate user locations. In *Proceedings of the International Conference on Database Systems for Advanced Applications*, 33-48. 10.1007/978-3-319-32049-6_3			
GIM-type	Network-based	Inference subject	User, static
Study area	Not specified	Cited by	2
Language	Not specified	General methodology	Tie-strength
Methods mentioned	Social closeness, confidence iteration method		

Citation			
Liu, J. & D. Inkpen (2015). Estimating user location in social media with stacked denoising auto-encoders. In *Proceedings of NAACL-HLT*, 201-210. DOI Unknown.			
GIM-type	Content-based	Inference subject	User, static
Study area	Contiguous US / North-America	Cited by	5
Language	Not specified	General methodology	Deep learning
Methods mentioned	Deep learning		

Citation			
Liu, Z. & Y. Huang (2016). Where are You Tweeting?: A Context and User Movement Based Approach. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management*, 1949-1952. 10.1145/2983323.2983881			
GIM-type	Content-based	Inference subject	User, static
Study area	Not specified	Cited by	0
Language	Not specified	General methodology	Not specified
Methods mentioned	Not specified		

Citation			
Mahmud, J., J. Nichols & C. Drews (2014). Home location identification of twitter users. *ACM Transactions on Intelligent Systems and Technology*, *5(3)*, 47. 10.1145/2528548			
GIM-type	Content-based	Inference subject	User, static
Study area	100 cities with most Twitter user worldwide	Cited by	66
Language	Not specified	General methodology	Text mining
Methods mentioned	Data Mining; Gazetteer		

Citation			
McGee, J., J. Caverlee, Z. Cheng (2013). Location prediction in social media based on tie-strength. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, 459-468. 10.1145/2505515.2505544			
GIM-type	Network-user	Inference subject	User, static
Study area	Not specified	Cited by	41
Language	Not specified	General methodology	Tie-strength
Methods mentioned	Data Mining, Spatial data mining; Social tie strength; Maximum likelihood estimator		

Citation			
Melo, F. & B. Martins (2015). Geocoding textual documents through theusage of hierarchical classifiers. In *Proceedings of the 9th Workshop on Geographic Information Retrieval*, 7. 10.1145/2837689.2837690			
GIM-type	Content-based	Inference subject	Message, static
Study area	Contiguous US / World	Cited by	0
Language	Multiple	General methodology	Not specified
Methods mentioned	Not specified		

Citation			
Pang, J. & Y. Zhang (2015). Location prediction: communities speak louder than friends. In *Proceedings of the 2015 ACM on Conference on Online Social Networks*, 161-171. 10.1145/2817946.2817954			
GIM-type	Network-based	Inference subject	User, mobile
Study area	New York, NY / San Francisco, CA, United States	Cited by	4
Language	Not specified	General methodology	Not specified
Methods mentioned	Not specified		

Citation			
Paraskevopoulos, P. & T. Palpanas (2016). Where has this tweet come from? Fast and fine-grained geolocalization of non-geotagged tweets. *Social Network Analysis and Mining*, *6(1)*, 89. 10.1007/s13278-016-0400-7.			
GIM-type	Content-based	Inference subject	Message, static
Study area	Rome / Milan / Naples / Bologna / Venice / Turin, Italy / Berlin, Germany /Amsterdam, Netherlands	Cited by	0
Language	Not specified	General methodology	Not specified
Methods mentioned	Not specified		

Citation			
Priedhorsky, R., A. Culotta & S. Del Valle (2014). Inferring the origin locations of tweets with quantitative confidence. In *Proceedings of the 17th ACM Confernece on Computer supported cooperative work and social computing*, 1523-1536. 10.1145/2531602.2531607			
GIM-type	Content-based	Inference subject	Message, static
Study area	Not specified	Cited by	31
Language	Not specified	General methodology	Not specified
Methods mentioned	Not specified		

Citation			
Rahimi, A., T. Cohn & T. Baldwin (2016). pigeo: A python geotagging tool. In *Proceedings of ACL-2016 System Demonstrations*, 127-132. DOI Unknown.			
GIM-type	Hybrid	Inference subject	Multiple, static
Study area	Not specified	Cited by	2
Language	English	General methodology	Not specified
Methods mentioned	Not specified		

Citation			
Ren, K., S. Zhang & H. Lin (2012). Where Are You Settling Down: Geo-locatin Twitter Users Based on Tweets and Social Networks. In *Proceedings of the 2012 Asia Information Retrieval Symposium*, 150-161. 10.1007/978-3-642-35341-3_13			
GIM-type	Hybrid	Inference subject	User, static
Study area	Not specified	Cited by	3
Language	Not specified	General methodology	Text mining
Methods mentioned	Text mining		

Citation			
Rodrigues, E., R. Assanção, G. Pappa, D. Renno & W. Meira Jr. (2016). Exploring multiple evidence to infer users' location in Twitter. *Neurocomputing*, *171*, 30-38. 10.1016/j.neucom.2015.05.066			
GIM-type	Hybrid	Inference subject	User, static
Study area	Belo Horizonte / Rio De Janeiro / Sao Paula / Brasilia / Curitiba / Fortaleza / Manaus / Porto Alegre / Recife Vitoria, Brazil	Cited by	1
Language	Not specified	General methodology	Not specified
Methods mentioned	Not specified		

Citation			
Roller, S., M. Speriosu, S. Rallapalli, B. Wing & J. Baldridge (2012). Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 1500-1510. DOI Unknown.			
GIM-type	Content-based	Inference subject	Message, static
Study area	Contiguous US / World	Cited by	69
Language	English	General methodology	k-trees/Grid
Methods mentioned	k-d trees		

Citation			
Ryoo, K. & S. Moon (2014). Inferring twitter user locations with 10 km accuracy. In *Proceedings of the 23rd International Conference on World Wide Web*, 643-648. 10.1145/2567948.2579236			
GIM-type	Content-based	Inference subject	User, home location
Study area	Not specified	Cited by	20
Language	Korean	General methodology	Not specified
Methods mentioned	Not specified		

Citation			
Schultz, A., A. Hadjakos, H. Paulheim, J. Nachtwey & M. Mühlhäuser (2013). A Multi-Indicator Approach for Geolocalization of Tweets. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media,* 573-582. DOI Unknown.			
GIM-type	Content-based	Inference subject	Multiple, static
Study area	U.S. East Coast region	Cited by	56
Language	Not specified	General methodology	Not specified
Methods mentioned	Not specified		

Citation			
Ueda, S., Y. Yamaguchi, H. Kitagawa & T. Amagasa (2015). Tweet Location Inference Based on Contents and Temporal Association. In *Proceedings of the International Conference on Web Information Systems Engineering*, 259-266. 10.1007/978-3-319-26187-4_22			
GIM-type	Content-based	Inference subject	Message, static
Study area	Not specified	Cited by	1
Language	Japanese	General methodology	Not specified
Methods mentioned	Not specified		

Citation			
Wing, B. & J. Baldridge (2014). Hierarchical Discriminative Classification for Text-based Geolocation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 336-348. DOI Unknown.			
GIM-type	Content-based	Inference subject	Message, static
Study area	Contiguous US/ World	Cited by	13
Language	Multiple	General methodology	Not specified
Methods mentioned	Not specified		

Citation			
Xie, Y., Y. Cheng, A. Agrawal & A. Choudhary (2014). Estimating online user location distribution without GPS location. In *Proceedings of the 2014 IEEE International Conference on Data Mining Workshop (ICDMW)*, 936-943. 10.1109/ICDMW.2014.30			
GIM-type	Content-based	Inference subject	User, static
Study area	Not specified	Cited by	0
Language	Not specified	General methodology	Text mining
Methods mentioned	Neural language model		

Citation			
Yamaguchi, Y., T. Amagasa & H. Kitagawa (2013). Landmark-based user location inference in social media. In *Proceedings of the first ACM conference on Online social networks*, 223-234. 10.1145/2512938.2512941.			
GIM-type	Network-based	Inference subject	User, home location, static
Study area	Not specified	Cited by	10
Language	Not specified	General methodology	Tie-strength
Methods mentioned	Data Mining; Social graphs, landmarks		

Citation			
Yamaguchi, Y., T. Amagasa, H. Kitagawa & Y. Ikawa (2014). Online user location inference exploiting spatiotemporal correlations in social streams. In *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management*, 1139-1148. 10.1145/2661829.2662039			
GIM-type	Content-based	Inference subject	User, static
Study area	Not specified	Cited by	6
Language	Japanese	General methodology	Not specified
Methods mentioned	Not specified		

Citation			
Yuan, Q., G. Cong, Z. Ma, A. Sun & N. Thalmann (2013). Who, where, when and what: discover spatio-temporal topics for twitter users. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge discovery and data mining*, 605-613. 10.1145/2487575.2487576			
GIM-type	Content-based	Inference subject	Message, mobile
Study area	Contiguous US / World	Cited by	83
Language	Not specified	General methodology	Not specified
Methods mentioned	Not specified		

Citation			
Zhang, W. & J. Gelernter (2014). Geocoding location expressions in Twitter messages: A preference learning method. *Journal of Spatial Information Science*, *2014(9)*, 37-70. 10.5311/JOSIS.2014.9.170			
GIM-type	Content-based	Inference subject	Message, static
Study area	Not specified	Cited by	16
Language	English	General methodology	Text mining
Methods mentioned	Geocoding; Toponym resolution; Named entity disambiguation; Machine learning; Gazetteer		

Citation			
Zhang, J., J. Sun, R. Zhang & Y. Zhang (2015). Your Actions Tell Where You Are: Uncovering Twitter Users in a Metropolitan Area. In *Proceedings of the 2015 IEEE Conference on Communications Network Security*, 424-432. 10.1109/CNS.2015.7346854			
GIM-type	Hybrid	Inference subject	User, static
Study area	Tuscon, TX / Philadelphia, PY / Chicago, IL / Los Angeles, California, United States	Cited by	5
Language	Not specified	General methodology	Tie-strength
Methods mentioned	Tie-strength		

Citation			
Zhang, Y., C. Szabo & Q. Sheng (2015). Sense and focus: towards effective location inference and event detection on Twitter. In *Proceedings of the International Conference on Web Information Systems Engineering*, 463-477. 10.1007/978-3-319-26190-4_31			
GIM-type	Content-based	Inference subject	Message, static
Study area	Not specified	Cited by	1
Language	English	General methodology	Text mining
Methods mentioned	Microblog content classification		

Appendix III: Scripts used

III.1 Gathering Twitter data in Python

Metadata	Description
Original code by	Alexander Galea
Retrieved from	https://github.com/agalea91/twitter_search
Edited	Yes
Parameters	**consumer_key, consumer_secret, access_token, access_secret, search_phrases, time_limit, max_tweets, min_days_old, max_days_old, USA**
Notes	Parameters used to authorize API use have been anonymized. All values given for the parameters are merely examples of values that can be used and are not necessarily representative for the actual values used in research.

```
import tweepy
from tweepy import OAuthHandler
import json
import datetime as dt
import time
import os
import sys

def load_api():
	consumer_key = [redacted]
	consumer_secret = [redacted]
	access_token = [redacted]
	access_secret = [redacted]
	auth = OAuthHandler(consumer_key, consumer_secret)
	auth.set_access_token(access_token, access_secret)
	return tweepy.API(auth)

def tweet_search(api, query, max_tweets, max_id, since_id, geocode):
	searched_tweets = []
	while len(searched_tweets) < max_tweets:
		remaining_tweets = max_tweets - len(searched_tweets)
		try:
			new_tweets = api.search(q=query, count=remaining_tweets,
						since_id=str(since_id),
						max_id=str(max_id-1),
						geocode=geocode)
			print('found',len(new_tweets),'tweets')
			if not new_tweets:
				print('no tweets found')
				break
			searched_tweets.extend(new_tweets)
			max_id = new_tweets[-1].id
		except tweepy.TweepError:
			print('exception raised, waiting 15 minutes')
			print('(until:', dt.datetime.now()+dt.timedelta(minutes=15), ')')
			time.sleep(15*60)
			break
	return searched_tweets, max_id

def get_tweet_id(api, date='', days_ago=9, query='a'):
	if date:
		td = date + dt.timedelta(days=1)
		tweet_date = '{0}-{1:0>2}-{2:0>2}'.format(td.year, td.month, td.day)
		tweet = api.search(q=query, count=1, until=tweet_date)
	else:
		td = dt.datetime.now() - dt.timedelta(days=days_ago)
		tweet_date = '{0}-{1:0>2}-{2:0>2}'.format(td.year, td.month, td.day)
		tweet = api.search(q=query, count=10, until=tweet_date)
		print('search limit (start/stop):',tweet[0].created_at)
		return tweet[0].id
```

```
def write_tweets(tweets, filename):
	with open(filename, 'a') as f:
		for tweet in tweets:
			json.dump(tweet._json, f)
			f.write('\n')

def main():
	search_phrases = ["search_phrase1”, “search_phrase2”, “search_phrase3”]
	time_limit = 24.0
	max_tweets = 100
	min_days_old, max_days_old = 0, 7
	USA = '39.8,-97.4,2600km'
	for search_phrase in search_phrases:
		print('Search phrase =', search_phrase)
		name = search_phrase.split()[0]
		json_file_root = name + '/' + name
		os.makedirs(os.path.dirname(json_file_root), exist_ok=True)
		read_IDs = False
		if max_days_old - min_days_old == 1:
			d = dt.datetime.now() - dt.timedelta(days=min_days_old)
			day = '{0}-{1:0>2}-{2:0>2}'.format(d.year, d.month, d.day)
		else:
			d1 = dt.datetime.now() - dt.timedelta(days=max_days_old-1)
			d2 = dt.datetime.now() - dt.timedelta(days=min_days_old)
			day = '{0}-{1:0>2}-{2:0>2}_to_{3}-{4:0>2}-{5:0>2}'.format(
				d1.year, d1.month, d1.day, d2.year, d2.month, d2.day)
		json_file = json_file_root + '_' + day + '.json'
		if os.path.isfile(json_file):
			print('Appending tweets to file named: ',json_file)
			read_IDs = True
		api = load_api()
		if read_IDs:
			with open(json_file, 'r') as f:
				lines = f.readlines()
				max_id = json.loads(lines[-1])['id']
				print('Searching from the bottom ID in file')
		else:
			if min_days_old == 0:
				max_id = -1
			else:
				max_id = get_tweet_id(api, days_ago=(min_days_old-1))
		since_id = get_tweet_id(api, days_ago=(max_days_old-1))
		print('max id (starting point) =', max_id)
		print('since id (ending point) =', since_id)
		start = dt.datetime.now()
		end = start + dt.timedelta(hours=time_limit)
		count, exitcount = 0, 0
		while dt.datetime.now() < end:
			count += 1
			print('count =',count)
			tweets, max_id = tweet_search(api, search_phrase, max_tweets,
						max_id=max_id, since_id=since_id,
						geocode=USA)
			if tweets:
				write_tweets(tweets, json_file)
				exitcount = 0
			else:
			exitcount += 1
			if exitcount == 3:
				if search_phrase == search_phrases[-1]:
					sys.exit('Maximum number of empty tweet strings reached
				- exiting')
				else:
					print('Maximum number of empty tweet strings reached -
				breaking')
					break

if __name__ == "__main__":
main()
```

III.2 Converting JSON file to CSV file in Python

Metadata	Description
Original code by	Michal Migurski
Retrieved from	http://mike.teczno.com/notes/streaming-data-from-twitter.html
Edited	Yes
Parameters	**json_file, csv_file**
Notes	All values given for the parameters are merely examples of values that can be used and are not necessarily representative for the actual values used in research.

```
import json

tweets = []

json_file = 'data.json'
csv_file = 'data.csv'

for line in open(json_file):
	try:
		tweets.append(json.loads(line))
	except:
		pass

tweet = tweets[0]

ids_all = [tweet['user']['id_str'] for tweet in tweets]
lang_user = [tweet['user']['lang'] for tweet in tweets]
lang_tweet = [tweet['lang'] for tweet in tweets]
location = [tweet['user']['location'] for tweet in tweets]

out = open(csv_file, 'wb')

print >> out, 'id_str,lang_user,lang_tweet,location'

rows = zip(ids_all,lang_user,lang_tweet,location)

from csv import writer
csv = writer(out)

for row in rows:
	values = [(value.encode('utf8') if hasattr(value, 'encode') else value) for value
	in row]
	csv.writerow(values)

out.close()
```

III.3 Merging CSV files to new CSV file in Python

Metadata	Description
Original code by	wisty
Retrieved from	http://stackoverflow.com/a/2512572
Edited	Yes
Parameters	**input, output, range**
Notes	All values given for the parameters are merely examples of values that can be used and are not necessarily representative for the actual values used in research.

```
input = "data"
output = "csv_merged.csv"

fout=open(output,"a")
for line in open(input + 1 + ".csv"):
	fout.write(line)
for num in range(2,12):
	f = open(input + str(num)+ ".csv")
	f.next()
	for line in f:
		fout.write(line)
	f.close()
fout.close()
```

III.4 Create subset based on attributes, delete id_str duplicates, export to CSV in R

Metadata	Description
Original code by	Joe d'Hont
Retrieved from	Original code
Edited	No
Parameters	**csv_file**
Notes	All values given for the parameters are merely examples of values that can be used and are not necessarily representative for the actual values used in research.

```
csv_file <- "csv_merged.csv")

library(readr)
data_all <- read_csv("csv_merged.csv")
View(data_all)

data_en <- subset(data_all, lang_user == "en")
data_en <- subset(data_en, lang_tweet == "en")
data_loc <- na.omit(data_en)
data_unique <- unique(data_loc)
write.csv(data_unique, file = "data_filtered.csv", row.names=FALSE)
```

III.5 Standardizing and adding coordinates to user-specified user locations in Python

Metadata	Description
Original code by	Joe d'Hont
Retrieved from	Original code
Edited	No
Parameters	**input, output**
Notes	All values given for the parameters are merely examples of values that can be used and are not necessarily representative for the actual values used in research.

```
import pandas as pd
from geopy.geocoders import Nominatim
from geopy.exc import GeocoderQueryError

input = "data_filtered.csv"
output = "data_standardized_coord.csv"

data = pd.read_csv(input)

geolocator = Nominatim(country_bias = "United States")

location_st_names = []
location_lats = []
location_lons = []

for index, row in data.iterrows():
	location_st = geolocator.geocode(row["location"])
	try:
		try:
			location_st_name = str(location_st.raw["display_name"])
			location_st_names.append(location_st_name)
			location_lat = location_st.latitude
			location_lats.append(location_lat)
			location_lon = location_st.longitude
			location_lons.append(location_lon)
		except Exception:
			location_st_name = None
			location_st_names.append(location_st_name)
			location_lat = None
			location_lats.append(location_lat)
			location_lon = None
			location_lons.append(location_lon)
	except GeocoderQueryError as e:
		location_st_name = None
		location_st_names.append(location_st_name)
	print location_st_name
	time.sleep(1.1)
```

```
data["location_st"] = location_st_names
data["latitude"] = location_lats
data["longitude"] = location_lons

data.to_csv(output, sep=',', encoding='utf-8', index=False)
```

III.6 Omit users whose location could not be standardized in R

Metadata	Description
Original code by	Joe d'Hont
Retrieved from	Original code
Edited	No
Parameters	**input, output**
Notes	All values given for the parameters are merely examples of values that can be used and are not necessarily representative for the actual values used in research.

```
input <- "data_standardardized_coord.csv"
output <- "data_st_loc.csv"

library(readr)
data_st <- read_csv(input)
View(data_all)

data_st_loc <- na.omit(data_st)
write.csv(data_unique, file = output, row.names=FALSE)
```

III.7 Divide data sets in pieces of 1000 entries in Python

Metadata	Description
Original code by	Rudziankou
Retrieved from	http://stackoverflow.com/a/36445821
Edited	Yes
Parameters	**delimiter, row_limit, input, output_name, output_path**
Notes	All values given for the parameters are merely examples of values that can be used and are not necessarily representative for the actual values used in research.

```
input = 'data_filtered.csv'
output_name = 'data_norm%s.csv'
output_path = '.'

import os

def split(filehandler, delimiter=',', row_limit=1000,
        output_name = output_name, output_path=output_path, keep_headers=True):
        import csv
        reader = csv.reader(filehandler, delimiter=delimiter)
        current_piece = 1
        current_out_path = os.path.join(
                output_path,
                output_name_template % current_piece
        )
        current_out_writer = csv.writer(open(current_out_path, 'wb'), delimiter=delimiter)
        current_limit = row_limit
        if keep_headers:
                headers = reader.next()
                current_out_writer.writerow(headers)
        for i, row in enumerate(reader):
                if i + 1 > current_limit:
                        current_piece += 1
                        current_limit = row_limit * current_piece
                        current_out_path = os.path.join(
                                output_path,
                                output_name_template % current_piece
                        )
                        current_out_writer = csv.writer(open(current_out_path, 'w'),
                        delimiter=delimiter)
                        if keep_headers:
                                current_out_writer.writerow(headers)
                current_out_writer.writerow(row)
split(open(input, 'r'));
```

III.8 Add geometry to database in SQL

Metadata	Description
Original code by	Joe d'Hont
Retrieved from	Original code
Edited	No
Parameters	None
Notes	Examplar tablenames have been given in the code below. All values given for the parameters are merely examples of values that can be used and are not necessarily representative for the actual values used in research.

```
create table rs_data(
	id_str text,
	location text,
	location_st text,
	latitude float,
	longitude float
);

create extension postgis

SELECT AddGeometryColumn ('public','rs_data','geometry',4326,'POINT',2);
UPDATE rs_data SET geometry = ST_SetSRID(ST_MakePoint(longitude, latitude), 4326)
```

III.9 Content-user method in Python

Metadata	Description
Original code by	Joe d'Hont
Retrieved from	Original code
Edited	No
Parameters	**consumer_key, consumer_secret, access_token, access_secret, country_bias, input, output, statuses_count, count, topos_freq**
Notes	All values given for the parameters are merely examples of values that can be used and are not necessarily representative for the actual values used in research.

```
import time
import tweepy
from collections import Counter
from geotext import GeoText
from geopy.geocoders import Nominatim
import pandas as pd
from geopy.exc import GeocoderQueryError
from geopy import distance
import winsound

consumer_key = [redacted]
consumer_secret = [redacted]
access_token = [redacted]
access_secret = [redacted]
auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_secret)

api = tweepy.API(auth)
geolocator = Nominatim(country_bias = "United States")

input = "data_norm.csv"
output = "data_analysis.csv"
data = pd.read_csv(input, dtype={"id_str": "str", "location": "str", "location_s": "str",
"latitude": "float", "longitude": "float"})

inf_locs1, inf_locs2, inf_occs1, inf_occs2, obsv_times1, obsv_times2, tot_times, scales1,
err_dists1, scales2, err_dists2 = [], [], [], [], [], [], [], [], [], [], []

for index, row in data.iterrows():
	start = time.time()
	try:
		id_str= row["id_str"]
		loc_st_lat=row["latitude"]
		loc_st_lon=row["longitude"]
```

```
        user_meta = api.get_user(id_str)
except tweepy.TweepError:
        inf_loc1, inf_loc2, inf_occ1, inf_occ2, obsv_time1, obsv_time2, tot_time,
        scale1, err_dist1, scale2, err_dist2 = None, None, None, None, None, None,
        None, None, None, None, None
        inf_locs1.append(inf_loc1)
        inf_occs1.append(inf_occ1)
        inf_locs2.append(inf_loc2)
        inf_occs2.append(inf_occ2)
        obsv_times1.append(obsv_time1)
        obsv_times2.append(obsv_time2)
        tot_times.append(tot_time)
        scales1.append(scale1)
        scales2.append(scale2)
        err_dists1.append(err_dist1)
        err_dists2.append(err_dist2)
        print inf_loc1
        continue
if user_meta.statuses_count <10:
        inf_loc1, inf_loc2, inf_occ1, inf_occ2, obsv_time1, obsv_time2, tot_time,
        scale1, err_dist1, scale2, err_dist2 = None, None, None, None, None, None,
        None, None, None, None, None
        inf_locs1.append(inf_loc1)
        inf_occs1.append(inf_occ1)
        inf_locs2.append(inf_loc2)
        inf_occs2.append(inf_occ2)
        obsv_times1.append(obsv_time1)
        obsv_times2.append(obsv_time2)
        tot_times.append(tot_time)
        scales1.append(scale1)
        scales2.append(scale2)
        err_dists1.append(err_dist1)
        err_dists2.append(err_dist2)
        print inf_loc1
        continue

try:
        user_tweets = api.user_timeline(user_id=id_str, count=200)
except tweepy.TweepError:
        inf_loc1, inf_loc2, inf_occ1, inf_occ2, obsv_time1, obsv_time2, tot_time,
        scale1, err_dist1, scale2, err_dist2 = None, None, None, None, None, None,
        None, None, None, None, None
        inf_locs1.append(inf_loc1)
        inf_occs1.append(inf_occ1)
        inf_locs2.append(inf_loc2)
        inf_occs2.append(inf_occ2)
        obsv_times1.append(obsv_time1)
        obsv_times2.append(obsv_time2)
        tot_times.append(tot_time)
        scales1.append(scale1)
        scales2.append(scale2)
        err_dists1.append(err_dist1)
        err_dists2.append(err_dist2)
        print inf_loc1
        continue
tweet_text = [tweet.text for tweet in user_tweets]
tweet_texts = ''.join(tweet_text)
tweet_topos = GeoText(tweet_texts)
tweet_places_meta = [(tweet.place.full_name if tweet.place else None)for tweet in
user_tweets]
tweet_places_cities = [city for city in tweet_places_meta if city is not None]

user_descr = user_meta.description
descr_topos = GeoText(user_descr)

topos_occ = Counter(tweet_topos.cities + tweet_places_cities + descr_topos.cities)
topos_df = pd.DataFrame.from_dict(topos_occ, orient='index').reset_index()
topos_df= topos_df.rename(columns={'index':'place', 0:'count'})
try:
        topos_freq = topos_df["count"].sum(axis=0)
```

```
except Exception:
	inf_loc1, inf_loc2, inf_occ1, inf_occ2, obsv_time1, obsv_time2, tot_time,
	scale1, err_dist1, scale2, err_dist2 = None, None, None, None, None, None,
	None, None, None, None, None
	inf_locs1.append(inf_loc1)
	inf_occs1.append(inf_occ1)
	inf_locs2.append(inf_loc2)
	inf_occs2.append(inf_occ2)
	obsv_times1.append(obsv_time1)
	obsv_times2.append(obsv_time2)
	tot_times.append(tot_time)
	scales1.append(scale1)
	scales2.append(scale2)
	err_dists1.append(err_dist1)
	err_dists2.append(err_dist2)
	print inf_loc1
	continue
topos_df_top5= topos_df.nlargest(5,"count")
topo_st_names = []
topos_df_top2 = topos_df_top5.head(2)
for index, row in topos_df_top2.iterrows():
	topo_st = geolocator.geocode(row["place"], timeout=10000)
	time.sleep(1.1)
	try:
		try:
			topo_st_name = str(topo_st.raw["display_name"])
			topo_st_names.append(topo_st_name)
		except Exception:
			topo_st_name = None
			topo_st_names.append(topo_st_name)
	except GeocoderQueryError as e:
		topo_st_name = None
		topo_st_names.append(topo_st_name)
topos_df_top2["place"] = topo_st_names
topos_df_top2 = topos_df_top2.dropna()
try:
	top_1_occ = topos_df_top2.iloc[0,1]
except Exception:
	inf_loc1, inf_loc2, inf_occ1, inf_occ2, obsv_time1, obsv_time2, tot_time,
	scale1, err_dist1, scale2, err_dist2 = None, None, None, None, None, None,
	None, None, None, None, None
	inf_locs1.append(inf_loc1)
	inf_occs1.append(inf_occ1)
	inf_locs2.append(inf_loc2)
	inf_occs2.append(inf_occ2)
	obsv_times1.append(obsv_time1)
	obsv_times2.append(obsv_time2)
	tot_times.append(tot_time)
	scales1.append(scale1)
	scales2.append(scale2)
	err_dists1.append(err_dist1)
	err_dists2.append(err_dist2)
	print inf_loc1
	continue
try:
	top_2_occ = topos_df_top2.iloc[1,1]
except Exception:
	top_2_occ = 0
inf_loc1 = topos_df_top2.iloc[0,0]
print inf_loc1
try:
	inf_loc2 = topos_df_top2.iloc[1,0]
except Exception:
	inf_loc2 = None
if ((top_1_occ == top_2_occ) or (topos_freq < 10)):
	inf_loc1, inf_loc2, inf_occ1, inf_occ2, obsv_time1, obsv_time2, tot_time,
	scale1, err_dist1, scale2, err_dist2 = None, None, None, None, None, None,
	None, None, None, None, None
	inf_locs1.append(inf_loc1)
	inf_occs1.append(inf_occ1)
	inf_locs2.append(inf_loc2)
```

```
        inf_occs2.append(inf_occ2)
        obsv_times1.append(obsv_time1)
        obsv_times2.append(obsv_time2)
        tot_times.append(tot_time)
        scales1.append(scale1)
        scales2.append(scale2)
        err_dists1.append(err_dist1)
        err_dists2.append(err_dist2)
        print inf_loc1
        continue

inf_locs1.append(inf_loc1)
inf_locs2.append(inf_loc2)
inf_occ1 = top_1_occ
inf_occ2 = top_2_occ
inf_occs1.append(inf_occ1)
inf_occs2.append(inf_occ2)

tweet_time = [tweet.created_at for tweet in user_tweets]
tweet_times= pd.DataFrame({'tweet_text': tweet_text, 'tweet_time': tweet_time})
obsv_text_meta1 = topos_df_top2.ix[topos_df_top2['count'].idxmax()]
obsv_text_full1 = obsv_text_meta1['place']
obsv_text_compact1 = obsv_text_full1.split(',', 1)[0]
tweet_times_obsv1 =
tweet_times[tweet_times['tweet_text'].str.contains(obsv_text_compact1)==True]
place_times = pd.DataFrame({'tweet_place': tweet_places_meta, 'tweet_time':
tweet_time})
place_times_obsv1 =
place_times[place_times['tweet_place'].str.contains(obsv_text_compact1)==True]
try:
        obsv_time_all1 = tweet_times_obsv1.append(place_times_obsv1)
        obsv_time1 = obsv_time_all1.iloc[0]["tweet_time"]
        obsv_times1.append(obsv_time1)
except Exception:
        obsv_time1 = None
        obsv_times1.append(obsv_time1)

obsv_text_meta2 = topos_df_top2.ix[topos_df_top2['count'].idxmin()]
obsv_text_full2 = obsv_text_meta2['place']
obsv_text_compact2 = obsv_text_full2.split(',', 1)[0]
tweet_times_obsv2 =
tweet_times[tweet_times['tweet_text'].str.contains(obsv_text_compact2)==True]
place_times_obsv2 =
place_times[place_times['tweet_place'].str.contains(obsv_text_compact2)==True]
try:
        obsv_time_all2 = tweet_times_obsv2.append(place_times_obsv2)
        obsv_time2 = obsv_time_all2.iloc[0]["tweet_time"]
        obsv_times2.append(obsv_time2)
except Exception:
        obsv_time2 = None
        obsv_times2.append(obsv_time2)

tot_time = time.time()-start
tot_times.append(tot_time)

try:
        scale1 = inf_loc1.count(',')
        scales1.append(scale1)
except Exception:
        scale1 = 0
        scales1.append(scale1)
try:
        scale2 = inf_loc2.count(',')
        scales2.append(scale2)
except Exception:
        scale2 = 0
        scales2.append(scale2)
```

```
        try:
                loc_st_coord = (loc_st_lat,loc_st_lon)
                inf_st1 = geolocator.geocode(inf_loc1, timeout=10000)
                time.sleep(1.1)
                inf_st_coord1 = (inf_st1.latitude,inf_st1.longitude)
                err_dist1 = distance.vincenty(loc_st_coord, inf_st_coord1, ellipsoid="WGS-
                84").kilometers
                time.sleep(1.1)
                err_dists1.append(err_dist1)
        except Exception:
                err_dist1 = None
                err_dists1.append(err_dist1)
        try:
                inf_st2 = geolocator.geocode(inf_loc2, timeout=10000)
                time.sleep(1.1)
                inf_st_coord2 = (inf_st2.latitude,inf_st2.longitude)
                err_dist2 = distance.vincenty(loc_st_coord, inf_st_coord2, ellipsoid="WGS-
                84").kilometers
                time.sleep(1.1)
                err_dists2.append(err_dist2)
        except Exception:
                err_dist2 = None
                err_dists2.append(err_dist2)

data["inf1_cu"] = inf_locs1
data["occ1_cu"] = inf_occs1
data["obsv1_cu"] = obsv_times1
data["scale1_cu"] = scales1
data["err_dist1_cu"] = err_dists1
data["inf2_cu"] = inf_locs2
data["occ2_cu"] = inf_occs2
data["obsv2_cu"] = obsv_times2
data["scale2_cu"] = scales2
data["err_dist2_cu"] = err_dists2
data["time_cu"] = tot_times

data.to_csv(output, sep=',', encoding='utf-8', index=False)
winsound.Beep(300,2000)
```

III.10 Network-user method in Python

Metadata	Description
Original code by	Joe d'Hont
Retrieved from	Original code
Edited	No
Parameters	**consumer_key, consumer_secret, access_token, access_secret, country_bias, input, output, followers_count, , friends_count, count, topos_freq**
Notes	All values given for the parameters are merely examples of values that can be used and are not necessarily representative for the actual values used in research.

```
import time
import tweepy
from collections import Counter
from geopy.geocoders import Nominatim
import pandas as pd
from geopy.exc import GeocoderQueryError
from geopy import distance
import winsound

all_start = time.time()

consumer_key = [redacted]
consumer_secret = [redacted]
access_token = [redacted]
access_secret = [redacted]
auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_secret)
```

```
api = tweepy.API(auth)
geolocator = Nominatim(country_bias = "United States")

input = "data_norm.csv"
output = "data_nu.csv"
data = pd.read_csv(csv_file, dtype={"id_str": "str",,"location": "str", "location_s":
"str", "latitude": "float", "longitude": "float"})

inf_locs1, inf_locs2, inf_occs1, inf_occs2, obsv_times1, obsv_times2, tot_times, scales1,
err_dists1, scales2, err_dists2 = [], [], [], [], [], [], [], [], [], [], []

for index, row in data.iterrows():
       start = time.time()
       try:
              id_str = row["id_str"]
              loc_st_lat=row["latitude"]
              loc_st_lon=row["longitude"]
              user_meta = api.get_user(id_str)
       except tweepy.TweepError:
              print "User_meta could not be accessed!"
              inf_loc1, inf_loc2, inf_occ1, inf_occ2, obsv_time1, obsv_time2, tot_time,
              scale1, err_dist1, scale2, err_dist2 = None, None, None, None, None, None,
              None, None, None, None, None
              inf_locs1.append(inf_loc1)
              inf_occs1.append(inf_occ1)
              inf_locs2.append(inf_loc2)
              inf_occs2.append(inf_occ2)
              obsv_times1.append(obsv_time1)
              obsv_times2.append(obsv_time2)
              tot_times.append(tot_time)
              scales1.append(scale1)
              scales2.append(scale2)
              err_dists1.append(err_dist1)
              err_dists2.append(err_dist2)
              continue
       if user_meta.followers_count <1:
              print "Followers_count too low!"
              inf_loc1, inf_loc2, inf_occ1, inf_occ2, obsv_time1, obsv_time2, tot_time,
              scale1, err_dist1, scale2, err_dist2 = None, None, None, None, None, None,
              None, None, None, None, None
              inf_locs1.append(inf_loc1)
              inf_occs1.append(inf_occ1)
              inf_locs2.append(inf_loc2)
              inf_occs2.append(inf_occ2)
              obsv_times1.append(obsv_time1)
              obsv_times2.append(obsv_time2)
              tot_times.append(tot_time)
              scales1.append(scale1)
              scales2.append(scale2)
              err_dists1.append(err_dist1)
              err_dists2.append(err_dist2)
              continue
       if user_meta.friends_count <1:
              print "Friends_count too low!"
              inf_loc1, inf_loc2, inf_occ1, inf_occ2, obsv_time1, obsv_time2, tot_time,
              scale1, err_dist1, scale2, err_dist2 = None, None, None, None, None, None,
              None, None, None, None, None
              inf_locs1.append(inf_loc1)
              inf_occs1.append(inf_occ1)
              inf_locs2.append(inf_loc2)
              inf_occs2.append(inf_occ2)
              obsv_times1.append(obsv_time1)
              obsv_times2.append(obsv_time2)
              tot_times.append(tot_time)
              scales1.append(scale1)
              scales2.append(scale2)
              err_dists1.append(err_dist1)
              err_dists2.append(err_dist2)
              continue
```

```
try:
	followers_ids, friends_ids, mutuals_ids = [], [], []
	followers_ids = api.followers_ids(user_id=id_str, count=100)
	friends_ids = api.friends_ids(user_id=id_str, count=100)
	mutuals_ids = set(followers_ids).intersection(friends_ids)
except tweepy.RateLimitError:
	try:
		print "Rate limit reached.. sleeping for about 15 minutes!"
		time.sleep(915)
		followers_ids, friends_ids, mutuals_ids = [], [], []
		followers_ids = api.followers_ids(user_id=id_str, count=100)
		friends_ids = api.friends_ids(user_id=id_str, count=100)
		mutuals_ids = set(followers_ids).intersection(friends_ids)
	except tweepy.TweepError:
		print "For some reason metadata from followers or friends could not
		be accessed!"
		inf_loc1, inf_loc2, inf_occ1, inf_occ2, obsv_time1, obsv_time2,
		tot_time, scale1, err_dist1, scale2, err_dist2 = None, None, None,
		None, None, None, None, None, None, None, None
		inf_locs1.append(inf_loc1)
		inf_occs1.append(inf_occ1)
		inf_locs2.append(inf_loc2)
		inf_occs2.append(inf_occ2)
		obsv_times1.append(obsv_time1)
		obsv_times2.append(obsv_time2)
		tot_times.append(tot_time)
		scales1.append(scale1)
		scales2.append(scale2)
		err_dists1.append(err_dist1)
		err_dists2.append(err_dist2)
		continue
except tweepy.TweepError:
	print "For some reason metadata could not be accessed!"
	inf_loc1, inf_loc2, inf_occ1, inf_occ2, obsv_time1, obsv_time2, tot_time,
	scale1, err_dist1, scale2, err_dist2 = None, None, None, None, None, None,
	None, None, None, None, None
	inf_locs1.append(inf_loc1)
	inf_occs1.append(inf_occ1)
	inf_locs2.append(inf_loc2)
	inf_occs2.append(inf_occ2)
	obsv_times1.append(obsv_time1)
	obsv_times2.append(obsv_time2)
	tot_times.append(tot_time)
	scales1.append(scale1)
	scales2.append(scale2)
	err_dists1.append(err_dist1)
	err_dists2.append(err_dist2)
	continue
try:
	followers_meta = api.lookup_users(user_ids=followers_ids)
	friends_meta = api.lookup_users(user_ids=followers_ids)
	mutuals_meta = api.lookup_users(user_ids=mutuals_ids)
except Exception:
	print "Followers_meta, friends_meta or mutuals_meta is not sufficient!"
	inf_loc1, inf_loc2, inf_occ1, inf_occ2, obsv_time1, obsv_time2, tot_time,
	scale1, err_dist1, scale2, err_dist2 = None, None, None, None, None, None,
	None, None, None, None, None
	inf_locs1.append(inf_loc1)
	inf_occs1.append(inf_occ1)
	inf_locs2.append(inf_loc2)
	inf_occs2.append(inf_occ2)
	obsv_times1.append(obsv_time1)
	obsv_times2.append(obsv_time2)
	tot_times.append(tot_time)
	scales1.append(scale1)
	scales2.append(scale2)
	err_dists1.append(err_dist1)
	err_dists2.append(err_dist2)
	continue
```

```
followers_locs = filter(None, [u.location for u in followers_meta])
friends_locs = filter(None, [u.location for u in friends_meta])
mutuals_locs = filter(None, [u.location for u in mutuals_meta])
topos_occ = Counter(followers_locs + friends_locs)
topos_df = pd.DataFrame.from_dict(topos_occ, orient='index').reset_index()
topos_df= topos_df.rename(columns={'index':'place', 0:'count'})
try:
	topos_freq = topos_df["count"].sum(axis=0)
except Exception:
	print "Topos_freq could not be counted!"
	inf_loc1, inf_loc2, inf_occ1, inf_occ2, obsv_time1, obsv_time2, tot_time,
	scale1, err_dist1, scale2, err_dist2 = None, None, None, None, None, None,
	None, None, None, None, None
	inf_locs1.append(inf_loc1)
	inf_occs1.append(inf_occ1)
	inf_locs2.append(inf_loc2)
	inf_occs2.append(inf_occ2)
	obsv_times1.append(obsv_time1)
	obsv_times2.append(obsv_time2)
	tot_times.append(tot_time)
	scales1.append(scale1)
	scales2.append(scale2)
	err_dists1.append(err_dist1)
	err_dists2.append(err_dist2)
	continue

topos_df_top5= topos_df.nlargest(5,"count")
topo_st_names = []

try:
	for index, row in topos_df_top5.iterrows():
		topo_st = geolocator.geocode(row["place"], timeout=10000)
		time.sleep(1.1)
		try:
			try:
				topo_st_name = str(topo_st.raw["display_name"])
				topo_st_names.append(topo_st_name)
			except Exception:
				topo_st_name = None
				topo_st_names.append(topo_st_name)
		except GeocoderQueryError as e:
			topo_st_name = None
			topo_st_names.append(topo_st_name)
			topos_df_top5["place"] = topo_st_names
			topos_df_top2 = topos_df_top5.dropna()
	except Exception:
		print "Geoserviceerror!"
		inf_loc1, inf_loc2, inf_occ1, inf_occ2, obsv_time1, obsv_time2,
		tot_time, scale1, err_dist1, scale2, err_dist2 = None, None, None,
		None, None, None, None, None, None, None, None
		inf_locs1.append(inf_loc1)
		inf_occs1.append(inf_occ1)
		inf_locs2.append(inf_loc2)
		inf_occs2.append(inf_occ2)
		obsv_times1.append(obsv_time1)
		obsv_times2.append(obsv_time2)
		tot_times.append(tot_time)
		scales1.append(scale1)
		scales2.append(scale2)
		err_dists1.append(err_dist1)
		err_dists2.append(err_dist2)
		time.sleep(5)
		continue
try:
	top_1_occ = topos_df_top2.iloc[0,1]
except Exception:
	print "No toponyms found!"
	inf_loc1, inf_loc2, inf_occ1, inf_occ2, obsv_time1, obsv_time2, tot_time,
	scale1, err_dist1, scale2, err_dist2 = None, None, None, None, None, None,
	None, None, None, None, None
	inf_locs1.append(inf_loc1)
```

```
        inf_occs1.append(inf_occ1)
        inf_locs2.append(inf_loc2)
        inf_occs2.append(inf_occ2)
        obsv_times1.append(obsv_time1)
        obsv_times2.append(obsv_time2)
        tot_times.append(tot_time)
        scales1.append(scale1)
        scales2.append(scale2)
        err_dists1.append(err_dist1)
        err_dists2.append(err_dist2)
        continue
try:
        top_2_occ = topos_df_top2.iloc[1,1]
except Exception:
        top_2_occ = 0
inf_loc1 = topos_df_top2.iloc[0,0]
print inf_loc1
try:
        inf_loc2 = topos_df_top2.iloc[1,0]
except Exception:
        inf_loc2 = None
if ((top_1_occ == top_2_occ) or (topos_freq < 10)):
        print "Top_1_occ == top_2_occ or topos_freq is not sufficient!"
        inf_loc1, inf_loc2, inf_occ1, inf_occ2, obsv_time1, obsv_time2, tot_time,
        scale1, err_dist1, scale2, err_dist2 = None, None, None, None, None, None,
        None, None, None, None, None
        inf_locs1.append(inf_loc1)
        inf_occs1.append(inf_occ1)
        inf_locs2.append(inf_loc2)
        inf_occs2.append(inf_occ2)
        obsv_times1.append(obsv_time1)
        obsv_times2.append(obsv_time2)
        tot_times.append(tot_time)
        scales1.append(scale1)
        scales2.append(scale2)
        err_dists1.append(err_dist1)
        err_dists2.append(err_dist2)
        continue

inf_locs1.append(inf_loc1)
inf_locs2.append(inf_loc2)

inf_occ1 = top_1_occ
inf_occ2 = top_2_occ
inf_occs1.append(inf_occ1)
inf_occs2.append(inf_occ2)

try:
        user_tweets = api.user_timeline(user_id=id_str, count=200)
        tweet_text = [tweet.text for tweet in user_tweets]
        tweet_time = [tweet.created_at for tweet in user_tweets]
        tweet_times= pd.DataFrame({'tweet_text': tweet_text, 'tweet_time':
        tweet_time})
        obsv_text_meta1 = topos_df_top2.ix[topos_df_top2['count'].idxmax()]
        obsv_text_full1 = obsv_text_meta1['place']
        obsv_text_compact1 = obsv_text_full1.split(',', 1)[0]
        tweet_places_meta = [(tweet.place.full_name if tweet.place else None)for
        tweet in user_tweets]
        tweet_times_obsv1 =
        tweet_times[tweet_times['tweet_text'].str.contains(obsv_text_compact1)==True
        ]
        place_times = pd.DataFrame({'tweet_place': tweet_places_meta, 'tweet_time':
        tweet_time})
        place_times_obsv1 =
        place_times[place_times['tweet_place'].str.contains(obsv_text_compact1)==Tru
        e]
        obsv_time_all1 = tweet_times_obsv1.append(place_times_obsv1)
        obsv_time1 = obsv_time_all1.iloc[0]["tweet_time"]
        obsv_times1.append(obsv_time1)
        except Exception:
        obsv_time1 = None
```

```
            obsv_times1.append(obsv_time1)

        try:
            obsv_text_meta2 = topos_df_top2.ix[topos_df_top2['count'].idxmin()]
            obsv_text_full2 = obsv_text_meta2['place']
            obsv_text_compact2 = obsv_text_full2.split(',', 1)[0]
            tweet_times_obsv2 =
            tweet_times[tweet_times['tweet_text'].str.contains(obsv_text_compact2)==True
            ]
            place_times_obsv2 =
            place_times[place_times['tweet_place'].str.contains(obsv_text_compact2)==Tru
            e]
            obsv_time_all2 = tweet_times_obsv2.append(place_times_obsv2)
            obsv_time2 = obsv_time_all2.iloc[0]["tweet_time"]
            obsv_times2.append(obsv_time2)
        except Exception:
            obsv_time2 = None
            obsv_times2.append(obsv_time2)

        tot_time = time.time()-start
        tot_times.append(tot_time)

        try:
            scale1 = inf_loc1.count(',')
            scales1.append(scale1)
        except Exception:
            scale1 = 0
            scales1.append(scale1)
        try:
            scale2 = inf_loc2.count(',')
            scales2.append(scale2)
        except Exception:
            scale2 = 0
            scales2.append(scale2)

        try:
            loc_st_coord = (loc_st_lat,loc_st_lon)
            inf_st1 = geolocator.geocode(inf_loc1, timeout=10000)
            time.sleep(1.1)
            inf_st_coord1 = (inf_st1.latitude,inf_st1.longitude)
            err_dist1 = distance.vincenty(loc_st_coord, inf_st_coord1, ellipsoid="WGS-
            84").kilometers
            time.sleep(1.1)
            err_dists1.append(err_dist1)
        except Exception:
            err_dist1 = None
            err_dists1.append(err_dist1)
        try:
            inf_st2 = geolocator.geocode(inf_loc2, timeout=10000)
            time.sleep(1.1)
            inf_st_coord2 = (inf_st2.latitude,inf_st2.longitude)
            err_dist2 = distance.vincenty(loc_st_coord, inf_st_coord2, ellipsoid="WGS-
            84").kilometers
            time.sleep(1.1)
            err_dists2.append(err_dist2)
        except Exception:
            err_dist2 = None
            err_dists2.append(err_dist2)

data["inf1_cu"] = inf_locs1
data["occ1_cu"] = inf_occs1
data["obsv1_cu"] = obsv_times1
data["scale1_cu"] = scales1
data["err_dist1_cu"] = err_dists1
data["inf2_cu"] = inf_locs2
data["occ2_cu"] = inf_occs2
data["obsv2_cu"] = obsv_times2
data["scale2_cu"] = scales2
data["err_dist2_cu"] = err_dists2
data["time_cu"] = tot_times
```

```
data.to_csv(output, sep=',', encoding='utf-8', index=False)
all_tot_time = time.time()-all_start
print all_tot_time
winsound.Beep(300,2000)
```

III.10 Hybrid-user method in Python

Metadata	Description
Original code by	Joe d'Hont
Retrieved from	Original code
Edited	No
Parameters	**consumer_key, consumer_secret, access_token, access_secret, country_bias, input, output, statuses_count, followers_count, , friends_count, count, topos_freq**
Notes	All values given for the parameters are merely examples of values that can be used and are not necessarily representative for the actual values used in research.

```
import time
import tweepy
from collections import Counter
from geotext import GeoText
from geopy.geocoders import Nominatim
import pandas as pd
from geopy.exc import GeocoderQueryError
from geopy import distance
import winsound

all_start = time.time()
consumer_key = [redacted]
consumer_secret = [redacted]
access_token = [redacted]
access_secret = [redacted]
auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_secret)

api = tweepy.API(auth)
geolocator = Nominatim(country_bias = "United States)

input = "data_norm.csv"
output = "data_hu.csv"
data = pd.read_csv(csv_file, dtype={"id_str": "str", "location": "str", "location_s":
"str", "latitude": "float", "longitude": "float"})

inf_locs1, inf_locs2, inf_occs1, inf_occs2, obsv_times1, obsv_times2, tot_times, scales1,
err_dists1, scales2, err_dists2 = [], [], [], [], [], [], [], [], [], [], []

for index, row in data.iterrows():
       start = time.time()
       try:
              id_str= row["id_str"]
              loc_st_lat=row["latitude"]
              loc_st_lon=row["longitude"]
              user_meta = api.get_user(id_str)
       except tweepy.TweepError:
              print "TweepError!"
              inf_loc1, inf_loc2, inf_occ1, inf_occ2, obsv_time1, obsv_time2, tot_time,
              scale1, err_dist1, scale2, err_dist2 = None, None, None, None, None, None,
              None, None, None, None, None
              inf_locs1.append(inf_loc1)
              inf_occs1.append(inf_occ1)
              inf_locs2.append(inf_loc2)
              inf_occs2.append(inf_occ2)
              obsv_times1.append(obsv_time1)
              obsv_times2.append(obsv_time2)
              tot_times.append(tot_time)
              scales1.append(scale1)
              scales2.append(scale2)
              err_dists1.append(err_dist1)
              err_dists2.append(err_dist2)
              print inf_loc1
              continue
```

```
if user_meta.statuses_count <10:
        print "Statuses_count too low!"
        inf_loc1, inf_loc2, inf_occ1, inf_occ2, obsv_time1, obsv_time2, tot_time,
        scale1, err_dist1, scale2, err_dist2 = None, None, None, None, None, None,
        None, None, None, None, None
        inf_locs1.append(inf_loc1)
        inf_occs1.append(inf_occ1)
        inf_locs2.append(inf_loc2)
        inf_occs2.append(inf_occ2)
        obsv_times1.append(obsv_time1)
        obsv_times2.append(obsv_time2)
        tot_times.append(tot_time)
        scales1.append(scale1)
        scales2.append(scale2)
        err_dists1.append(err_dist1)
        err_dists2.append(err_dist2)
        print inf_loc1
        continue
if user_meta.followers_count <10:
        print "Followers_count too low!"
        inf_loc1, inf_loc2, inf_occ1, inf_occ2, obsv_time1, obsv_time2, tot_time,
        scale1, err_dist1, scale2, err_dist2 = None, None, None, None, None, None,
        None, None, None, None, None
        inf_locs1.append(inf_loc1)
        inf_occs1.append(inf_occ1)
        inf_locs2.append(inf_loc2)
        inf_occs2.append(inf_occ2)
        obsv_times1.append(obsv_time1)
        obsv_times2.append(obsv_time2)
        tot_times.append(tot_time)
        scales1.append(scale1)
        scales2.append(scale2)
        err_dists1.append(err_dist1)
        err_dists2.append(err_dist2)
        continue
if user_meta.friends_count <10:
        print "Friends_count too low!"
        inf_loc1, inf_loc2, inf_occ1, inf_occ2, obsv_time1, obsv_time2, tot_time,
        scale1, err_dist1, scale2, err_dist2 = None, None, None, None, None, None,
        None, None, None, None, None
        inf_locs1.append(inf_loc1)
        inf_occs1.append(inf_occ1)
        inf_locs2.append(inf_loc2)
        inf_occs2.append(inf_occ2)
        obsv_times1.append(obsv_time1)
        obsv_times2.append(obsv_time2)
        tot_times.append(tot_time)
        scales1.append(scale1)
        scales2.append(scale2)
        err_dists1.append(err_dist1)
        err_dists2.append(err_dist2)
        continue
try:
        user_tweets = api.user_timeline(user_id=id_str, count=200)
except tweepy.TweepError:
        print "Tweeperror!"
        inf_loc1, inf_loc2, inf_occ1, inf_occ2, obsv_time1, obsv_time2, tot_time,
        scale1, err_dist1, scale2, err_dist2 = None, None, None, None, None, None,
        None, None, None, None, None
        inf_locs1.append(inf_loc1)
        inf_occs1.append(inf_occ1)
        inf_locs2.append(inf_loc2)
        inf_occs2.append(inf_occ2)
        obsv_times1.append(obsv_time1)
        obsv_times2.append(obsv_time2)
        tot_times.append(tot_time)
        scales1.append(scale1)
        scales2.append(scale2)
        err_dists1.append(err_dist1)
        err_dists2.append(err_dist2)
```

```
        print inf_loc1
        continue

tweet_text = [tweet.text for tweet in user_tweets]
tweet_texts = ''.join(tweet_text)
tweet_topos = GeoText(tweet_texts)
tweet_places_meta = [(tweet.place.full_name if tweet.place else None)for tweet in
user_tweets]
tweet_places_cities = [city for city in tweet_places_meta if city is not None]

user_descr = user_meta.description
descr_topos = GeoText(user_descr)

try:
        followers_ids, friends_ids, mutuals_ids = [], [], []
        followers_ids = api.followers_ids(user_id=id_str, count=100)
        friends_ids = api.friends_ids(user_id=id_str, count=100)
        mutuals_ids = set(followers_ids).intersection(friends_ids)
except tweepy.RateLimitError:
        try:
                print "Rate limit reached.. sleeping for about 15 minutes!"
                time.sleep(915)
                followers_ids, friends_ids, mutuals_ids = [], [], []
                followers_ids = api.followers_ids(user_id=id_str, count=100)
                friends_ids = api.friends_ids(user_id=id_str, count=100)
                mutuals_ids = set(followers_ids).intersection(friends_ids)
        except tweepy.TweepError:
                print "For some reason metadata from followers or friends could not
                be accessed!"
                inf_loc1, inf_loc2, inf_occ1, inf_occ2, obsv_time1, obsv_time2,
                tot_time, scale1, err_dist1, scale2, err_dist2 = None, None, None,
                None, None, None, None, None, None, None, None
                inf_locs1.append(inf_loc1)
                inf_occs1.append(inf_occ1)
                inf_locs2.append(inf_loc2)
                inf_occs2.append(inf_occ2)
                obsv_times1.append(obsv_time1)
                obsv_times2.append(obsv_time2)
                tot_times.append(tot_time)
                scales1.append(scale1)
                scales2.append(scale2)
                err_dists1.append(err_dist1)
                err_dists2.append(err_dist2)
                continue
except tweepy.TweepError:
        print "For some reason metadata could not be accessed!"
        inf_loc1, inf_loc2, inf_occ1, inf_occ2, obsv_time1, obsv_time2, tot_time,
        scale1, err_dist1, scale2, err_dist2 = None, None, None, None, None, None,
        None, None, None, None, None
        inf_locs1.append(inf_loc1)
        inf_occs1.append(inf_occ1)
        inf_locs2.append(inf_loc2)
        inf_occs2.append(inf_occ2)
        obsv_times1.append(obsv_time1)
        obsv_times2.append(obsv_time2)
        tot_times.append(tot_time)
        scales1.append(scale1)
        scales2.append(scale2)
        err_dists1.append(err_dist1)
        err_dists2.append(err_dist2)
        continue
try:
        followers_meta = api.lookup_users(user_ids=followers_ids)
        friends_meta = api.lookup_users(user_ids=followers_ids)
        mutuals_meta = api.lookup_users(user_ids=mutuals_ids)
except Exception:
        print "Followers_meta, friends_meta or mutuals_meta is not sufficient!"
        inf_loc1, inf_loc2, inf_occ1, inf_occ2, obsv_time1, obsv_time2, tot_time,
        scale1, err_dist1, scale2, err_dist2 = None, None, None, None, None, None,
        None, None, None, None, None
        inf_locs1.append(inf_loc1)
```

```
        inf_occs1.append(inf_occ1)
        inf_locs2.append(inf_loc2)
        inf_occs2.append(inf_occ2)
        obsv_times1.append(obsv_time1)
        obsv_times2.append(obsv_time2)
        tot_times.append(tot_time)
        scales1.append(scale1)
        scales2.append(scale2)
        err_dists1.append(err_dist1)
        err_dists2.append(err_dist2)
        continue
followers_locs = filter(None, [u.location for u in followers_meta])
friends_locs = filter(None, [u.location for u in friends_meta])
mutuals_locs = filter(None, [u.location for u in mutuals_meta])

topos_occ = Counter(tweet_topos.cities + tweet_places_cities + descr_topos.cities
+ followers_locs + friends_locs + mutuals_locs)
topos_df = pd.DataFrame.from_dict(topos_occ, orient='index').reset_index()
topos_df= topos_df.rename(columns={'index':'place', 0:'count'})
try:
        topos_freq = topos_df["count"].sum(axis=0)
except Exception:
        print "Topos_freq could not be counted!"
        inf_loc1, inf_loc2, inf_occ1, inf_occ2, obsv_time1, obsv_time2, tot_time,
        scale1, err_dist1, scale2, err_dist2 = None, None, None, None, None, None,
        None, None, None, None, None
        inf_locs1.append(inf_loc1)
        inf_occs1.append(inf_occ1)
        inf_locs2.append(inf_loc2)
        inf_occs2.append(inf_occ2)
        obsv_times1.append(obsv_time1)
        obsv_times2.append(obsv_time2)
        tot_times.append(tot_time)
        scales1.append(scale1)
        scales2.append(scale2)
        err_dists1.append(err_dist1)
        err_dists2.append(err_dist2)
        print inf_loc1
        continue
        topos_df_top5= topos_df.nlargest(5,"count")
        topo_st_names = []

try:
        for index, row in topos_df_top5.iterrows():
        topo_st = geolocator.geocode(row["place"], timeout=10000)
        time.sleep(1.1)
        try:
                try:
                        topo_st_name = str(topo_st.raw["display_name"])
                        topo_st_names.append(topo_st_name)
                except Exception:
                        topo_st_name = None
                        topo_st_names.append(topo_st_name)
        except GeocoderQueryError as e:
                topo_st_name = None
                topo_st_names.append(topo_st_name)
                topos_df_top5["place"] = topo_st_names
                topos_df_top2 = topos_df_top5.dropna()
except Exception:
        print "Geoserviceerror!"
        inf_loc1, inf_loc2, inf_occ1, inf_occ2, obsv_time1, obsv_time2, tot_time,
        scale1, err_dist1, scale2, err_dist2 = None, None, None, None, None, None,
        None, None, None, None, None
        inf_locs1.append(inf_loc1)
        inf_occs1.append(inf_occ1)
        inf_locs2.append(inf_loc2)
        inf_occs2.append(inf_occ2)
        obsv_times1.append(obsv_time1)
        obsv_times2.append(obsv_time2)
        tot_times.append(tot_time)
        scales1.append(scale1)
```

```
        scales2.append(scale2)
        err_dists1.append(err_dist1)
        err_dists2.append(err_dist2)
        time.sleep(5)
        continue
try:
        top_1_occ = topos_df_top2.iloc[0,1]
except Exception:
        print "No toponyms found!"
        inf_loc1, inf_loc2, inf_occ1, inf_occ2, obsv_time1, obsv_time2, tot_time,
        scale1, err_dist1, scale2, err_dist2 = None, None, None, None, None, None,
        None, None, None, None, None
        inf_locs1.append(inf_loc1)
        inf_occs1.append(inf_occ1)
        inf_locs2.append(inf_loc2)
        inf_occs2.append(inf_occ2)
        obsv_times1.append(obsv_time1)
        obsv_times2.append(obsv_time2)
        tot_times.append(tot_time)
        scales1.append(scale1)
        scales2.append(scale2)
        err_dists1.append(err_dist1)
        err_dists2.append(err_dist2)
        continue
try:
        top_2_occ = topos_df_top2.iloc[1,1]
except Exception:
        top_2_occ = 0
inf_loc1 = topos_df_top2.iloc[0,0]
print inf_loc1
try:
        inf_loc2 = topos_df_top2.iloc[1,0]
except Exception:
        inf_loc2 = None
if ((top_1_occ == top_2_occ) or (topos_freq < 10)):
        print "Top_1_occ == top_2_occ or topos_freq is not sufficient!"
        inf_loc1, inf_loc2, inf_occ1, inf_occ2, obsv_time1, obsv_time2, tot_time,
        scale1, err_dist1, scale2, err_dist2 = None, None, None, None, None, None,
        None, None, None, None, None
        inf_locs1.append(inf_loc1)
        inf_occs1.append(inf_occ1)
        inf_locs2.append(inf_loc2)
        inf_occs2.append(inf_occ2)
        obsv_times1.append(obsv_time1)
        obsv_times2.append(obsv_time2)
        tot_times.append(tot_time)
        scales1.append(scale1)
        scales2.append(scale2)
        err_dists1.append(err_dist1)
        err_dists2.append(err_dist2)
        continue

inf_locs1.append(inf_loc1)
inf_locs2.append(inf_loc2)

inf_occ1 = top_1_occ
inf_occ2 = top_2_occ
inf_occs1.append(inf_occ1)
inf_occs2.append(inf_occ2)

try:
        user_tweets = api.user_timeline(user_id=id_str, count=200)
        tweet_text = [tweet.text for tweet in user_tweets]
        tweet_time = [tweet.created_at for tweet in user_tweets]
        tweet_times= pd.DataFrame({'tweet_text': tweet_text, 'tweet_time':
        tweet_time})
        obsv_text_meta1 = topos_df_top2.ix[topos_df_top2['count'].idxmax()]
        obsv_text_full1 = obsv_text_meta1['place']
        obsv_text_compact1 = obsv_text_full1.split(',', 1)[0]
        tweet_places_meta = [(tweet.place.full_name if tweet.place else None)for
        tweet in user_tweets]
```

```
        tweet_times_obsv1 =
        tweet_times[tweet_times['tweet_text'].str.contains(obsv_text_compact1)==True
        ]
        place_times = pd.DataFrame({'tweet_place': tweet_places_meta, 'tweet_time':
        tweet_time})
        place_times_obsv1 =
        place_times[place_times['tweet_place'].str.contains(obsv_text_compact1)==Tru
        e]
        obsv_time_all1 = tweet_times_obsv1.append(place_times_obsv1)
        obsv_time1 = obsv_time_all1.iloc[0]["tweet_time"]
        obsv_times1.append(obsv_time1)
except Exception:
        obsv_time1 = None
        obsv_times1.append(obsv_time1)

try:
        obsv_text_meta2 = topos_df_top2.ix[topos_df_top2['count'].idxmin()]
        obsv_text_full2 = obsv_text_meta2['place']
        obsv_text_compact2 = obsv_text_full2.split(',', 1)[0]
        tweet_times_obsv2 =
        tweet_times[tweet_times['tweet_text'].str.contains(obsv_text_compact2)==True
        ]
        place_times_obsv2 =
        place_times[place_times['tweet_place'].str.contains(obsv_text_compact2)==Tru
        e]
        obsv_time_all2 = tweet_times_obsv2.append(place_times_obsv2)
        obsv_time2 = obsv_time_all2.iloc[0]["tweet_time"]
        obsv_times2.append(obsv_time2)
except Exception:
        obsv_time2 = None
        obsv_times2.append(obsv_time2)

tot_time = time.time()-start
tot_times.append(tot_time)

try:
        scale1 = inf_loc1.count(',')
        scales1.append(scale1)
except Exception:
        scale1 = 0
        scales1.append(scale1)
try:
        scale2 = inf_loc2.count(',')
        scales2.append(scale2)
except Exception:
        scale2 = 0
        scales2.append(scale2)

try:
        loc_st_coord = (loc_st_lat,loc_st_lon)
        inf_st1 = geolocator.geocode(inf_loc1, timeout=10000)
        time.sleep(1.1)
        inf_st_coord1 = (inf_st1.latitude,inf_st1.longitude)
        err_dist1 = distance.vincenty(loc_st_coord, inf_st_coord1, ellipsoid="WGS-
        84").kilometers
        time.sleep(1.1)
        err_dists1.append(err_dist1)
except Exception:
        err_dist1 = None
        err_dists1.append(err_dist1)
try:
        inf_st2 = geolocator.geocode(inf_loc2, timeout=10000)
        time.sleep(1.1)
        inf_st_coord2 = (inf_st2.latitude,inf_st2.longitude)
        err_dist2 = distance.vincenty(loc_st_coord, inf_st_coord2, ellipsoid="WGS-
        84").kilometers
        time.sleep(1.1)
        err_dists2.append(err_dist2)
except Exception:
        err_dist2 = None
        err_dists2.append(err_dist2)
```

```
data["inf1_cu"] = inf_locs1
data["occ1_cu"] = inf_occs1
data["obsv1_cu"] = obsv_times1
data["scale1_cu"] = scales1
data["err_dist1_cu"] = err_dists1
data["inf2_cu"] = inf_locs2
data["occ2_cu"] = inf_occs2
data["obsv2_cu"] = obsv_times2
data["scale2_cu"] = scales2
data["err_dist2_cu"] = err_dists2
data["time_cu"] = tot_times

data.to_csv(output, sep=',', encoding='utf-8', index=False)
all_tot_time = time.time()-all_start
print all_tot_time
winsound.Beep(300,2000)
```

III.12 Calculating evaluation metrics using Python

Metadata	Description
Original code by	Joe d'Hont
Retrieved from	Original code
Edited	No
Parameters	**input, output**
Notes	All values given for the parameters are merely examples of values that can be used and are not necessarily representative for the actual values used in research.

```
import pandas as pd
import numpy
import tweepy
from datetime import datetime
import winsound

input = "data_analysis.csv"
output = "data_eval.csv:
data = pd.read_csv(csv_file)

consumer_key = [redacted]
consumer_secret = [redacted]
access_token = [redacted]
access_secret = [redacted]
auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_secret)

api = tweepy.API(auth)


err_dists = []

for index, row in data.iterrows():
	err_dist1 = row["err_dist1_cu"]
	err_dist2 = row["err_dist2_cu"]
	err_dists_list = [err_dist1, err_dist2]
	err_dists_min = min(err_dists_list)
	if err_dists_min < 100:
		err_dists.append(err_dists_min)
	else:
		err_dists.append('nan')
		err_dists_clean = [x for x in err_dists if str(x) != 'nan']
		spatial_rel = float(len(err_dists_clean))/1000
		print spatial_rel

ages = []
date_format = "%Y-%m-%d %H:%M:%S"
end_time = datetime.strptime('2017-05-04 00:00:00', date_format)

for index, row in data.iterrows():
	err_dist1 = row["err_dist1_cu"]
	err_dist2 = row["err_dist2_cu"]
	if (err_dist1 < err_dist2):
```

```
                date = str(row["obsv1_cu"])
                try:
                        date_formatted = datetime.strptime(date, date_format)
                        date_diff = end_time - date_formatted
                        age = date_diff.days
                        ages.append(age)
                except:
                        pass
        elif (err_dist1 > err_dist2):
                date = str(row["obsv2_cu"])
                try:
                        date_formatted = datetime.strptime(date, date_format)
                        date_diff = end_time - date_formatted
                        age = date_diff.days
                        ages.append(age)
                except:
                        pass
        elif (err_dist1 == err_dist2):
                date = str(row["obsv1_cu"])
                try:
                        date_formatted = datetime.strptime(date, date_format)
                        date_diff = end_time - date_formatted
                        age = date_diff.days
                        ages.append(age)
                except:
                        pass
                else:
                        pass

age_avg = numpy.mean(ages)
age_med = numpy.median(numpy.array(ages))
age = (age_avg + age_med)/2

scales = []

for index, row in data.iterrows():
        err_dist1 = row["err_dist1_cu"]
        err_dist2 = row["err_dist2_cu"]
        if (err_dist1 < err_dist2):
                scale = row["scale1_cu"]
                scales.append(scale)
        elif (err_dist1 > err_dist2):
                scale = row["scale2_cu"]
                scales.append(scale)
        elif (err_dist1 == err_dist2):
                scale = row["scale1_cu"]
                scales.append(scale)
        else:
                scale = None
                scales.append(scale)

scales_clean = [x for x in scales if x is not None]
scales_avg = numpy.mean(scales_clean)
scales_med = numpy.median(numpy.array(scales_clean))
scale = (scales_avg + scales_med)/2

inf_locs = data["inf1_cu"]
err_dists_clean = [x for x in err_dists if str(x) != 'nan']
completeness = (float(len(err_dists_clean)))/1000

results = []
true_pos = []
false_pos = []
false_neg = []
location_s_occs = []

for index, row in data.iterrows():
        err_dist1 = row["err_dist1_cu"]
        err_dist2 = row["err_dist2_cu"]
        location_s = row["location_s"]
        location_s_compact = str(location_s.split(',', 1)[0])
```

```
inf_loc1 = str(row["inf1_cu"])
inf_loc2 = str(row["inf2_cu"])
obsv_time1 = row["obsv1_cu"]
obsv_time2 = row["obsv2_cu"]
inf_occ1 = row["occ1_cu"]
inf_occ2 = row["occ2_cu"]
scale1 = row["scale1_cu"]
scale2 = row["scale2_cu"]

try:
       id_str= row["id_str"]
       user_meta = api.get_user(id_str)
       user_tweets = api.user_timeline(user_id=id_str, count=200)
       tweet_text = [tweet.text for tweet in user_tweets]
       tweet_texts = ''.join(tweet_text)
       tweet_place = [(tweet.place.full_name if tweet.place else None)for tweet in
       user_tweets]
       tweet_places_clean = [x for x in tweet_place if x is not None]
       tweet_places = ''.join(tweet_places_clean)
except tweepy.TweepError:
       location_s_occ = 0
       location_s_occs.append(location_s_occ)
       result = "TN"
       print id_str, ", ", location_s_compact, ",", location_s_occ, result
       results.append(result)
       continue
location_s_occ = tweet_texts.count(location_s_compact) +
tweet_places.count(location_s_compact)
if (err_dist1 < err_dist2):
       if ((((err_dist1 < 0.01) and (scale1 > 2)) or (inf_loc1 == location_s) or
       (((inf_loc1 in location_s_compact)==True) and (scale1 > 2))) and
       (pd.isnull(obsv_time1)==False)):
              result = "TP"
              print id_str, ", ", location_s_compact, ",", location_s_occ, result
              true_pos.append(inf_loc1)
              results.append(result)
              continue
       elif ((((err_dist1 < 0.01) and (scale1 > 2)) or (inf_loc1 == location_s) or
       (((inf_loc1 in location_s_compact)==True) and (scale1 > 2))) and
       (pd.isnull(obsv_time1)==True)):
              result = "FP"
              print id_str, ", ", location_s_compact, ",", location_s_occ, result
              false_pos.append(inf_loc1)
              results.append(result)
              continue
       elif ((((err_dist1 > 0.01) and (scale1 > 2)) or (inf_loc1 != location_s) or
       (((inf_loc1 in location_s_compact)==False) and (scale > 2))) and
       (pd.isnull(obsv_time1)==False)):
              if ((inf_occ1 > location_s_occ) and (inf_occ1 > 10)):
              result = "FN"
              print id_str, ", ", location_s_compact, ",", location_s_occ, result
              false_neg.append(inf_loc1)
              results.append(result)
              continue
       else:
              result = "TN"
              print id_str, ", ", location_s_compact, ",", location_s_occ, result
              results.append(result)
              continue
else:
       result = "TN"
       print id_str, ", ", location_s_compact, ",", location_s_occ, result
       results.append(result)
       continue

elif (err_dist1 > err_dist2):
       if ((((err_dist2 < 0.01) and (scale2 > 2)) or (inf_loc2 == location_s) or
       (((inf_loc2 in location_s_compact)==True) and (scale2 > 2))) and
       (pd.isnull(obsv_time2)==False)):
              result = "TP"
              print id_str, ", ", location_s_compact, ",", location_s_occ, result
```

```
                        true_pos.append(inf_loc2)
                        results.append(result)
                        continue
                elif ((((err_dist2 < 0.01) and (scale2 > 2)) or (inf_loc2 == location_s) or
                (((inf_loc2 in location_s_compact)==True) and (scale2 > 2))) and
                (pd.isnull(obsv_time2)==True)):
                        result = "FP"
                        print id_str, ", ", location_s_compact, ",", location_s_occ, result
                        false_pos.append(inf_loc2)
                        results.append(result)
                        continue
                elif ((((err_dist2 > 0.01) and (scale2 > 2)) or (inf_loc2 != location_s) or
                (((inf_loc2 in location_s_compact)==False) and (scale > 2))) and
                (pd.isnull(obsv_time2)==False)):
                        if ((inf_occ2 > location_s_occ) and (inf_occ2 > 10)):
                        result = "FN"
                        print id_str, ", ", location_s_compact, ",", location_s_occ, result
                        false_neg.append(inf_loc2)
                        results.append(result)
                        continue
                else:
                        result = "TN"
                        print id_str, ", ", location_s_compact, ",", location_s_occ, result
                        results.append(result)
                        continue
        else:
                result = "TN"
                print id_str, ", ", location_s_compact, ",", location_s_occ, result
                results.append(result)
                continue

if (err_dist1 == err_dist2):
        if ((((err_dist1 < 0.01) and (scale1 > 2)) or (inf_loc1 == location_s) or
        (((inf_loc1 in location_s_compact)==True) and (scale1 > 2))) and
        (pd.isnull(obsv_time1)==False)):
                result = "TP"
                print id_str, ", ", location_s_compact, ",", location_s_occ, result
                true_pos.append(inf_loc1)
                results.append(result)
                continue
        elif ((((err_dist1 < 0.01) and (scale1 > 2)) or (inf_loc1 == location_s) or
        (((inf_loc1 in location_s_compact)==True) and (scale1 > 2))) and
        (pd.isnull(obsv_time1)==True)):
                result = "FP"
                print id_str, ", ", location_s_compact, ",", location_s_occ, result
                false_pos.append(inf_loc1)
                results.append(result)
                continue
        elif ((((err_dist1 > 0.01) and (scale1 > 2)) or (inf_loc1 != location_s) or
        (((inf_loc1 in location_s_compact)==False) and (scale > 2))) and
        (pd.isnull(obsv_time1)==False)):
                if ((inf_occ1 > location_s_occ) and (inf_occ1 > 10)):
                        result = "FN"
                        print id_str, ", ", location_s_compact, ",", location_s_occ, result
                        false_neg.append(inf_loc1)
                        results.append(result)
                        continue
                else:
                        result = "TN"
                        print id_str, ", ", location_s_compact, ",", location_s_occ, result
                        results.append(result)
                        continue
        else:
                result = "TN"
                print id_str, ", ", location_s_compact, ",", location_s_occ, result
                results.append(result)
                continue
else:
result = "TN"
print id_str, ", ", location_s_compact, ",", location_s_occ, result
results.append(result)
```

```
continue
data["result"] = results
true_pos_no = len(true_pos)
false_pos_no = len(false_pos)
false_neg_no = len(false_neg)
precision = (float(true_pos_no)/(false_pos_no + true_pos_no))
recall = (float(true_pos_no)/(true_pos_no + false_neg_no))
fmeasure = (2*float(precision)*recall)/(precision+recall)

speed_all = data["time_cu"]
speed_clean = [x for x in speed_all if ((str(x) != ('nan')) and (x < 20))]
speed_avg = numpy.mean(speed_clean)
speed_med = numpy.median(numpy.array(speed_clean))
speed = (speed_avg + speed_med)/2

print "For", csv_file, "the values for the evaluation metrics are:"
print "Spatial reliability: ", spatial_rel, "km"
print "Temporal reliability: ", age, "days"
print "Scale: ", scale
print "Completeness: ", completeness
print "Precision: ", precision
print "Recall: ", recall
print "F-measure: ", fmeasure
print "Programming speed: ", speed, "sec"

data.to_csv(output, sep=',', encoding='utf-8', index=False)
winsound.Beep(300,2000)
```

III.13 Normalize evaluation metrics and perform second sensitivity analysis using Python

Metadata	Description
Original code by	Joe d'Hont
Retrieved from	Original code
Edited	No
Parameters	**csv_file**
Notes	All values given for the parameters are merely examples of values that can be used and are not necessarily representative for the actual values used in research.

```
import pandas as pd

csv_file = "abs.csv"
data = pd.read_csv(csv_file, sep=",")
data.index = data["METRIC"]
data = data.iloc[:,1:7]
index = ["SPAT_REL", "TEMP_REL", "SCALE", "COMP", "PRECISION", "RECALL", "FMEASURE",
"SPEED"]
index = ["REL", "SCALE", "COMP", "FMEASURE", "SPEED", "TOT"]
columns = ["LOW", "MID", "HIGH"]
columns_w = ["WS1", "WS2", "WS3", "WS4", "WS5", "WS6"]

print "\nABSOLUTE VALUES:\n\n", data, "\n"

data_n = pd.DataFrame(index = index, columns= columns)
data_mid = pd.DataFrame (index = index, columns = columns_w)
data_w = pd.DataFrame(index = index, columns = columns_w)

spat_rel = data.iloc[0,0:3]

temp_rel = data.iloc[1,0:3]
temp_max = 0
temp_min = 2547
temp_n = ((temp_rel-temp_min)/(temp_max-temp_min))

rel_n = (spat_rel + temp_n)/2
data_n.loc["REL"] = rel_n

scales = data.iloc[2,0:3]
scales_min = 0
scales_max = 4
```

```
scales_n = ((scales-scales_min)/(scales_max-scales_min))
data_n.loc["SCALE"] = scales_n

comp = data.iloc[3,0:3]
data_n.loc["COMP"] = comp

fmeasure=data.iloc[6,0:3]
data_n.loc["FMEASURE"] = fmeasure

speed=data.iloc[7,0:3]
speed_max = 3.775
speed_min = 17.7195
speed_n = ((speed-speed_min)/(speed_max-speed_min))
data_n.loc["SPEED"] = speed_n

totals = data_n.mean()
data_n.loc["TOT"] = totals

print "\nNORMALIZED VALUES:\n\n", data_n, "\n"

sens_mid = data_n["MID"]
data_mid.loc[0:6,"WS1"] = sens_mid
data_mid.loc[0:6,"WS2"] = sens_mid
data_mid.loc[0:6,"WS3"] = sens_mid
data_mid.loc[0:6,"WS4"] = sens_mid
data_mid.loc[0:6,"WS5"] = sens_mid
data_mid.loc[0:6,"WS6"] = sens_mid

ws1 = [1.0,1.0,1.0,1.0,1.0,1.0]
ws2 = [2,0.75,0.75,0.75,0.75,1.0]
ws3 = [0.75,2.0,0.75,0.75,0.75,1.0]
ws4 = [0.75,0.75,2.0,0.75,0.75,1.0]
ws5 = [0.75,0.75,0.75,2.0,0.75,1.0]
ws6 = [0.75,0.75,0.75,0.75,2.0,1.0]

data_w.loc[0:6,"WS1"] = ws1
data_w.loc[0:6,"WS2"] = ws2
data_w.loc[0:6,"WS3"] = ws3
data_w.loc[0:6,"WS4"] = ws4
data_w.loc[0:6,"WS5"] = ws5
data_w.loc[0:6,"WS6"] = ws6

v1 = data_mid.reindex(columns=columns_w).values
v2 = data_w.reindex(columns=columns_w).values
data_weighted = pd.DataFrame(v1 * v2, index=index, columns=columns_w)
data_weighted = data_weighted.iloc[0:5,:]
w_totals = data_weighted.mean()
data_weighted.loc["TOT"] = w_totals

print "\nWEIGHTED VALUES:\n\n", data_weighted, "\n"
```

III.14 Create random samples from CSV files using R

Metadata	Description
Original code by	Joe d'Hont
Retrieved from	Original code
Edited	No
Parameters	**input**, **output**
Notes	All values given for the parameters are merely examples of values that can be used and are not necessarily representative for the actual values used in research.

```
input <- "data_norm.csv"
output <- "data_random.csv"

library(readr)
data_st <- read_csv(input)
View(data_all)

data_sample = data_all[sample(nrow(data_all), 1000), ]
```

```
write.csv(data_sample, file = output, row.names=FALSE)
```

III.15 Calculating scale differences using Python

Metadata	Description
Original code by	Joe d'Hont
Retrieved from	Original code
Edited	No
Parameters	**input**
Notes	All values given for the parameters are merely examples of values that can be used and are not necessarily representative for the actual values used in research.

```
import pandas as pd
import numpy

input = "data_analysis.csv"
data = pd.read_csv(csv_file)

scales_s = []



for index, row in data.iterrows():
       location_s = row["location_s"]
       scale = location_s.count(',')
       scales_s.append(scale)

scales_avg = numpy.mean(scales_s) #to calculate avg
scales_med = numpy.median(numpy.array(scales_s)) #to calculate median
scale = (scales_avg + scales_med)/2
print scale
```

Appendix IV. Software used

IV.1 Main packages

Name	Version	Developer(s)
ArcMap	10.5.6491	ESRI
Canopy	1.7.4.3348	Enthought, Inc.
Excel	16.0.7766.2060	Microsoft
pgAdmin III	1.22	The pgAdmin Development Team
QGIS	2.18.1 Las Palmas	QGIS Development Team
RStudio	1.0.136	RStudio

IV.2 Sub packages

Name	Version	Developer(s)	Used in
GeoPy	1.11.0	GeoPy Contributors	Canopy
GeoText	0.3.0	Yaser Martinez Pelenzuela	Canopy
KuTools	16.00	ExtentOffice	Excel
Python	2.7.12	Python Software Foundation	ArcMap, Canopy, QGIS
PostgreSQL	9.4	PostgreSQL Global Development Group	pgAdmin III
PostGIS Bundle	2.3.2	Various	pgAdmin III
R for Windows	3.3.3	R Core Team	RStudio
Time		Python Software Foundation	Canopy
Tweepy	3.5.0	Joshua Roesslein	Canopy

Appendix V. Shapefiles used

Screen name	Folder name	Source	Year	Size
Cartographic Boundary Shapefiles - States	cb_2015_us_state_500k	U.S. Census Bureau	2015	4.57 MB
Cartographic Boundary Shapefiles - Counties	cb_2015_us_county_500k	U.S. Census Bureau	2015	16.6 MB
World Borders Dataset	TM_WORLD_BORDERS-0.3	Bjorn Sandvik	2009	6.2 MB

Appendix VI: USB-content

Folder	Sub-folder	File-name	Format	Size	Description
Data	rs1	rs1_analysis_cu	CSV	245KB	rs1_random processed with content-user method.
		rs1_analysis_cu_eval	CSV	248KB	rs1_analysis_cu evaluated.
		rs1_analysis_nu	CSV	250KB	rs1_random processed with network-user method.
		rs1_analysis_nu_eval	CSV	253KB	rs1_analysis_nu evaluated.
		rs1_clipped	CSV	6.78MB	rs1_st_loc clipped by study area
		rs1_filtered	CSV	4.12MB	Users filtered by metadata attributes.
		rs1_random	CSV	121KB	Random sample of rs1_clipped.
		rs1_st_loc	CSV	9.85MB	rs1_filtered with standardized user locations.
		dam	JSON	61.86MB	Tweets containing “dam”-keyword.
		downpour	JSON	2.79MB	Tweets containing “downpour”-keyword.
		evacuation	JSON	30.44MB	Tweets containing “evacuation”-keyword.
		flood	JSON	152.24MB	Tweets containing “flood”-keyword.
		landslide	JSON	6.50MB	Tweets containing “landslide”-keyword.
		mud	JSON	31.11MB	Tweets containing “mud”-keyword.
		rain	JSON	589.1MB	Tweets containing “rain”-keyword.
		sinkholes	JSON	3.74MB	Tweets containing “sinkholes”-keyword.
		snow	JSON	130.3MB	Tweets containing “snow”-keyword.
		spillway	JSON	45.93MB	Tweets containing “spillway”-keyword.
		storm	JSON	251.85MB	Tweets containing “storm”-keyword.
		dam_geojson	JSON	72KB	Tweets containing “dam”-keyword.
		downpour_geojson	JSON	9KB	Geotagged tweets containing “downpour”-keyword.
		evacuation_geojson	JSON	16KB	Geotagged tweets containing “evacuation”-keyword.
		flood_geojson	JSON	123KB	Geotagged tweets containing “flood”-keyword.
		landslide_geojson	JSON	5KB	Geotagged tweets containing “landslide”-keyword.
		mud_geojson	JSON	32KB	Geotagged tweets containing “mud”-keyword.
		rain_geojson	JSON	2.17MB	Geotagged tweets containing “rain”-keyword.
		sinkholes_geojson	JSON	2KB	Geotagged tweets containing “sinkholes”-keyword.
		snow_geojson	JSON	399KB	Geotagged tweets containing “snow”-keyword.
		spillway_geojson	JSON	15KB	Geotagged tweets containing “spillway”-keyword.
		storm_geojson	JSON	422KB	Geotagged tweets containing “storm”-keyword.
	rs2	rs2_analysis_cu	CSV	236KB	rs2_random processed with content-user method.
		rs2_analysis_cu_eval	CSV	259KB	rs2_analysis_cu evaluated.
		rs2_analysis_nu	CSV	224KB	rs2_random processed with network-user method.
		rs2_analysis_nu_eval	CSV	227KB	rs2_analysis_nu evaluated.
		rs2_clipped	CSV	739KB	rs2_st_loc clipped by study area
		rs2_filtered	CSV	403KB	Users filtered by metadata attributes.
		rs2_random	CSV	113KB	Random sample of rs2_clipped.
		rs2_st_loc	CSV	950KB	rs2_filtered with standardized user locations.
		breast_cancer	JSON	75.60MB	Tweets containing “breast cancer”-keywords.
		breast_cancer_geojson	JSON	10KB	Geotagged tweets containing “breast cancer”-keywords
	rs3	rs3_analysis_cu	CSV	225KB	rs3_random processed with content-user method.
		rs3_analysis_cu_eval	CSV	194KB	rs3_analysis_cu evaluated.
		rs3_analysis_nu	CSV	249KB	rs3_random processed with network-user method.
		rs3_analysis_nu_eval	CSV	252KB	rs3_analysis_nu evaluated.
		rs3_clipped	CSV	969KB	rs3_st_loc clipped by study area
		rs3_filtered	CSV	698KB	Users filtered by metadata attributes.
		rs3_random	CSV	117KB	Random sample of rs3_clipped.
		rs3_st_loc	CSV	1.34MB	rs3_filtered with standardized user locations.
		bigthightwitter	JSON	187.77MB	Tweets containing “#bigthightwitter”-keyword.
		gopdnd	JSON	52.11MB	Tweets containing “#gopdnd”-keyword.
		bigthightwitter_geojson	JSON	4KB	Geotagged tweets with “#bigthightwitter”-keyword.
		gopdnd_geojson	JSON	1KB	Geotagged tweets containing “#gopdnd”-keyword.
	sens	sens_high_cu	CSV	194KB	sens_high processed with content-user method
		sens_high_cu_eval	CSV	197KB	sens_high evaluated.
		sens_high_nu	CSV	215KB	sens_high processed with network-user method
		sens_high_nu_eval	CSV	218KB	sens_high evaluated.
		sens_mid_cu	CSV	223KB	sens_mid processed with content-user method
		sens_mid_cu_eval	CSV	236KB	sens_mid evaluated.
		sens_mid_nu	CSV	241KB	sens_mid processed with network-user method
		sens_mid_nu_eval	CSV	244KB	sens_mid evaluated.
		sens_low_cu	CSV	250KB	sens_low processed with content-user method
		sens_low_cu_eval	CSV	253KB	sens_low evaluated.

		sens_low_nu	CSV	226KB	sens_low processed with network-user method
		sens_low_nu_eval	CSV	229KB	sens_low evaluated.
		sens_data	CSV	118KB	Sensitivity analysis data set.
		sq2_data	JSON	246.58MB	Test data set used for sub question 2.
Scripts		gather_twitter_data	PY	8KB	Used to gather Twitter data.
		csv_convert	PY	4KB	Used to convert JSON to CSV.
		merge_csv	PY	1KB	Used to merge CSV files.
		standardize_names	PY	2KB	Used to standardize user-specified user locations.
		divide_csv	PY	2KB	Used to split CSV in equal parts.
		content_user_method	PY	12KB	The GIM as described in paragraph 7.3.1.
		network_user_method	PY	16KB	The GIM as described in paragraph 7.3.2.
		hybrid_user_method	PY	18KB	The GIM as described in paragraph 7.3.3.
		evaluation	PY	12KB	Used to evaluate GIM data output.
		normalizing_sens	PY	3KB	Used to normalize metrics and perform second sensitivity analysis.
		scale_diff	PY	1KB	Used to calculate original and new average scale.