GIMA

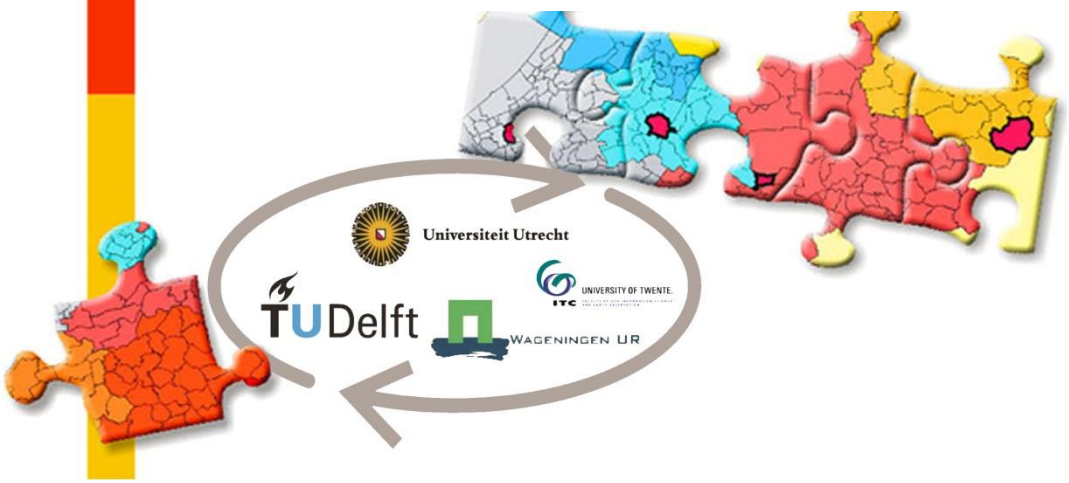
Geographical Information Management and Applications

Twitter as a spatio-temporal information source for traffic incident management

GIMA – Master Thesis

Roeland Steur

August, 2014



Twitter as a spatio-temporal information source for traffic incident management

Master Thesis – GIMA

Author:
Roeland. J. Steur
Ina Boudier-Bakkerlaan 141
3581 XV Utrecht
roelandsteur@gmail.com
06 44 345 444

Geographical Information Management and Applications (GIMA)

Supervisor: Dr. Ir. A. Ligtenberg
Professor: Prof. Dr. Ir. A. Bregt

August 25, 2014

Abstract

Although Dutch highways are monitored extensively, important details about incidents are sometimes lacking or arrive late in the traffic control center. Tweets sent by traffic participants about their experiences on the road could provide useful information in such cases. A difficult task however is to find relevant tweets in the thousands of tweets that are sent each second by Twitter users worldwide. Very interesting in this matter are tweets that are geographically localized by GPS-coordinates, so-called 'geotagged' tweets. It is expected that the spatio-temporal characteristics of geotagged tweets can be used to identify incident-related tweets that are sent on or around highways. This thesis addresses the question how useful Twitter is as a source of spatio-temporal information in the domain of incident management. For a period of 5 months geotagged tweets were harvested from the Twitter API and stored in a geographic database. Zonal regularity analysis was used in an attempt to detect traffic-related events in the area around Amsterdam from the database. It was found that geotagged Twitter data is lacking sufficient quantity and quality in order to be a valuable source of spatio-temporal information for incident management.

Table of Contents

List of Figures

List of tables

1. Introduction

1.1. Research Background and Motivation

1.1.1 Twitter as information medium in the Netherlands

Since its foundation in 2006, Twitter has become an increasingly popular micro-blogging service. With over 271 million monthly active users, sending about 500 billion 'tweets' a day as of 2012, Twitter has become one of the major social media today (Semiocast, 2012; Twitter, 2012c). Because of this massive number of users and messages, Twitter is already a well-known source of information within different countries.

Strikingly, the Netherlands score quite high in terms of Twitter accounts. Recent studies found that approximately 3,3 million Twitter users are from the Netherlands (Newcom Research & Consultancy B.V., 2013). Other studies revealed that the Netherlands is one of the countries that is most active on Twitter (Dawson, 2012; Lunden, 2012).

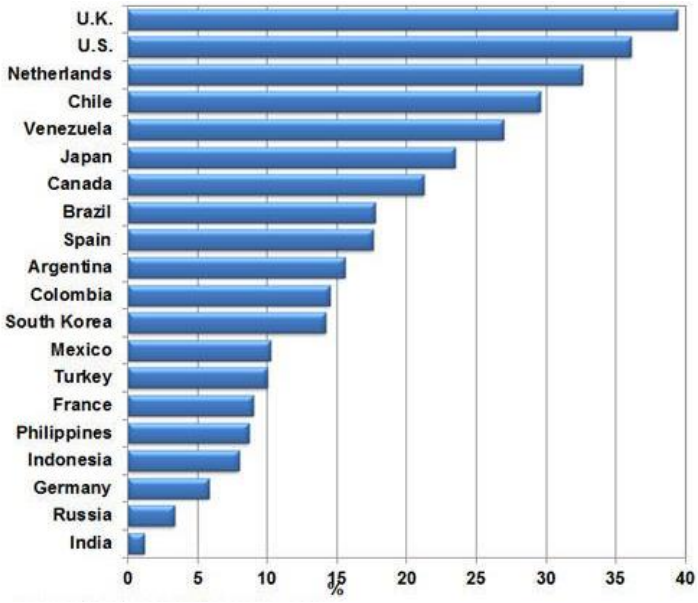


Figure 1 Ratio (in %) Twitter accounts/population (Dawson, 2012)

Figure 1 shows that the ratio Twitter accounts to population is one of the world's highest in the Netherlands. In Figure 2, a tweet density visualization of Europe shows the relative high density of tweets from the UK and the Netherlands compared to other areas.

Twitter appears to be not just a hype. In the Netherlands Twitter has become a serious source of information. In a large study among more than 13.000 Dutch subjects, 54% of participants responded that they think social media will become increasingly important in their way of gathering information (Newcom Research & Consultancy B.V., 2013). The importance of Twitter as a source of information manifests itself as well through numerous examples of public and commercial organizations that use Twitter to reach their public. Different news carriers like the Dutch public news channel NOS, newspapers, and governmental bodies each have their own Twitter channel. Also quite remarkably, some Dutch politicians use Twitter as their main means of communication with the public, rather than for instance press conferences.

Figure 2 Tweet-density map of Europe (Twitter, 2011)

1.1.2 Twitter as a source of geographical information and event detection

Twitter has become a promising source of information. Perhaps even more promising is the 'hidden content' of information that Twitter holds. Twitter isn't used only as a kind of news channel: most messages posted on Twitter are messages about people's daily experiences and opinions. It is for this reason that in several studies Twitter users have already been employed as potential valuable sensors for a range of things. For example, Zhao et al. (2011) discuss how to use Twitter's public messages for detecting football games throughout the United States in order to generate a real-time electronic program guide for these football events. Next to event detection, Twitter has been a topic of study for opinion mining. For example, Tumasjan et al. (2010) propose using Twitter content as a valuable indicator of the offline political landscape during election times.

In the framework of geosciences, it is even more interesting to discuss the geographical component of Twitter. Because Twitter offers GPS-enabled messaging, a small number of the tweets can be mapped (Figure 3). In these cases where tweets have GPS-coordinates, tweets can be seen as a source of geo-information, assuming that there is a relation between the user's location and the content of the message that they post. In the cases that no GPS-coordinates are sent with a tweet, these tweets can often be geo-located because the tweet holds some geographical description, for example the name of a train station. Although the level of detail of the geographical component of the tweet becomes smaller in these cases, it still can be seen as geographic information.

The idea to use Twitter as a valid source of geographical information is increasingly studied. Various attempts have been made to detect spatio-temporal events from Twitter. Sakaki et al. (2010) attempted to detect earthquake locations in Japan in real-time from Twitter messages. Unlike sensing for well-defined events like earthquakes, the event topic can also be completely unknown. Still, these unknown events can be detected just by searching for unusual regional Twitter activity (Lee et al., 2011). Hence, it appears that Twitter can be a valuable provider of geo-information in many different cases.

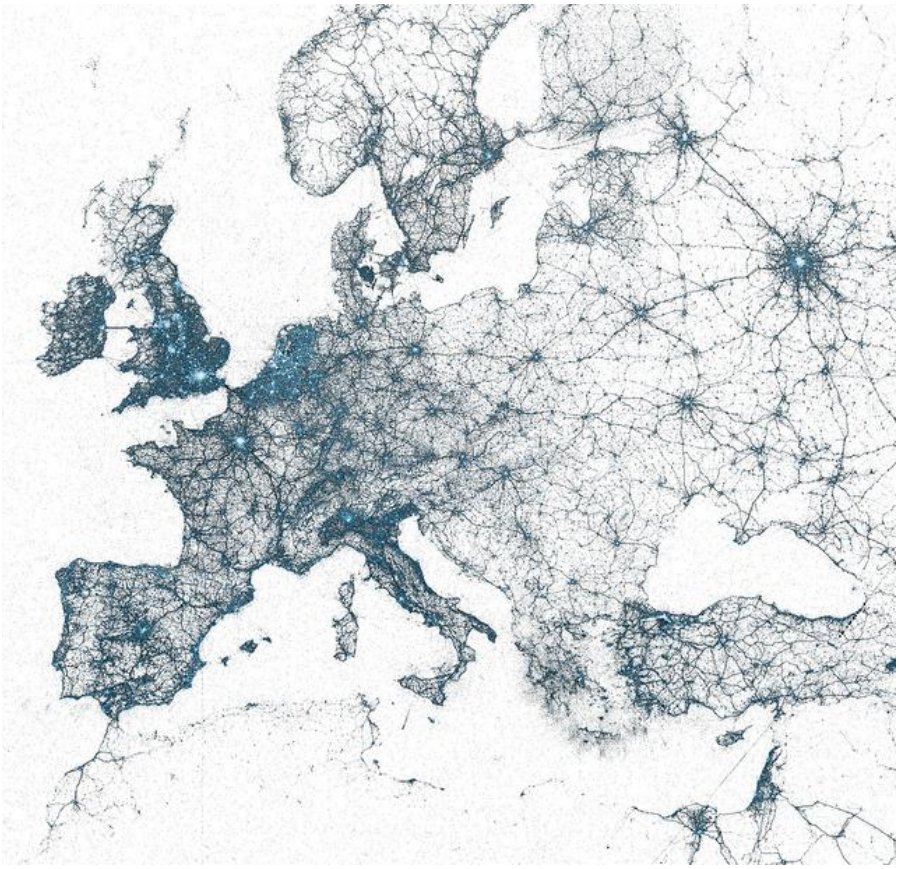


Figure 3 Visualization of all geotagged Tweets posted since 2009 (Rios, 2013)

1.1.3 Twitter and traffic management in the Netherlands

In traffic management, Twitter could be an especially useful additional source of information. The goal of traffic management is to ensure a smooth flow of traffic over the regional road network. In the Netherlands, Rijkswaterstaat is responsible for traffic management of the Dutch highways. In five traffic management centers, road traffic is monitored continuously using roadside technology. Operators in the traffic management centers are able to control the traffic flow remotely by using, for example, dynamic road signs, cameras, information panels and controlled access to motorways by traffic lights.

One of the main responsibilities of traffic management centers is giving support to incident management services. After an incident occurs on the road, it is extremely important that the traffic flow stagnates as little as possible. In order to take the right decisions and actions, detailed information about the incident is essential. It is important to verify the location and the type of the incident in order to take efficient measures. Traffic jams have various causes and therefore have different solutions differently as well. For example, if a big truck turns over and blocks the road, other actions will follow than in case a traffic jam is caused by a broken-down car or a small accident with only material damage.

The incident management process can be split up into different time phases, which together form the total incident duration (Figure 4). Rijkswaterstaat set the ambition in 2008 to decrease the average incident duration of 2015 with 25% compared to 2008 (Drolenga, 2011). The time between the incident occurring and the alert coming in at Rijkswaterstaat (detection time) is not counted as part of the incident duration as defined by Rijkswaterstaat. There is very little information about the average duration of this detection time, because it is hard to register and it can vary greatly for different types of incidents. Nevertheless, this detection time has an influence on the overall time it takes to normalize the road traffic. The sooner the traffic management centers are informed about the cause of a traffic jam and the precise location of an incident, the faster and more efficient this traffic jam can be solved again.

Unfortunately a relatively long delay exists between the incident happening and the moment traffic management centers are informed about the incident, most of the time by officials (for instance via the emergency calls of 1-1-2). There is a possibility that Twitter traffic from traffic participants could reveal more about the location and cause of the traffic jam or incident. Because Twitter is a real-time medium, information about incidents could arrive faster at the traffic management centers than the official information. Especially in cases where there's no need to call the emergency number (1-1-2), Twitter may offer valuable information to verify an incident's location and type.

TNO found that due to traffic jams, a financial loss of 400 million euros for cargo trade was inflicted in the year of 2010 (Zanten & Veth, 2011). If Twitter can be used as a fast source of information during incident management, resulting in a faster resolution of traffic jams, social and financial benefits could be high.

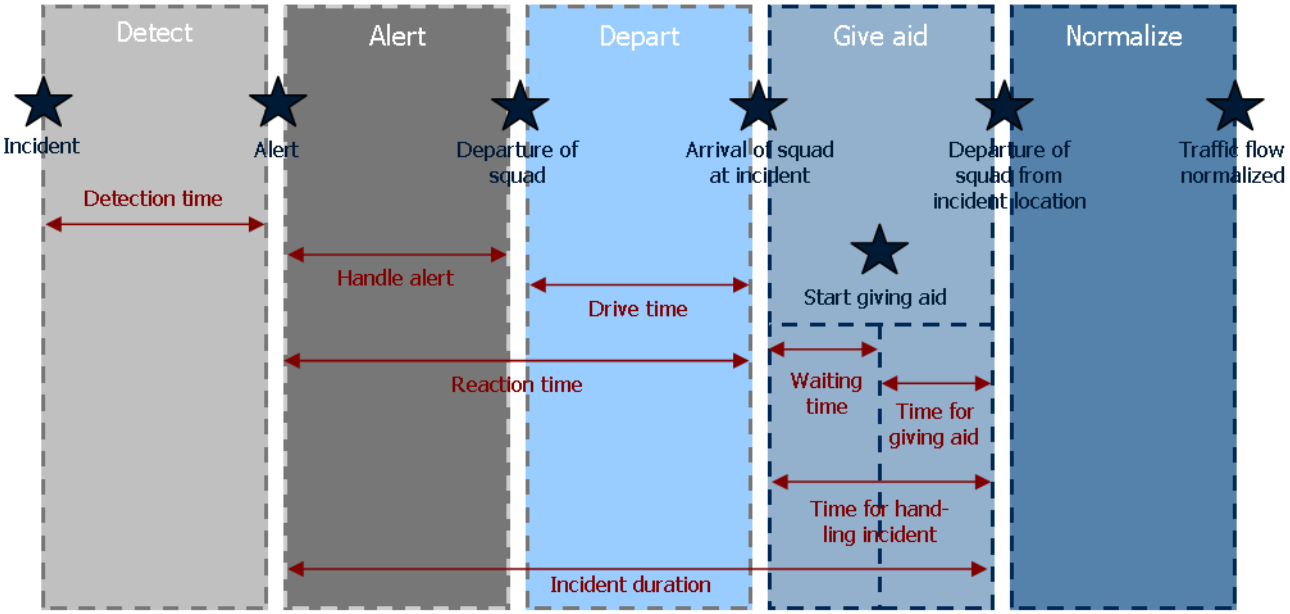


Figure 4 Definitions of incident duration and phases during the incident management process. (Drolenga, 2011)

1.1.4 Motivation for the use of Twitter

There are some critera that need to be met in order for social media to be useable for geographic research. Most importantly, messages that are posted in the social media have to be publicly accessible. Secondly, the messages should contain an indication of the location from which they are sent. Moreover, the social media should provide as much data as possible and preferably the data should be accessible real-time.

Comparing different social media that are used in the Netherlands, Twitter meets these criteria best. There are some advantages of using Twitter as a source of geo-information over other social media. The most important advantage is that messages on Twitter are publicly accessible via different REST APIs and streaming APIs. Not all social media are publicly accessible in this way. In the Netherlands specifically Twitter is a medium that is used by many active users. On an average day, Twitter is used by 1,6 million individuals in the Netherlands (Newcom Research & Consultancy B.V., 2013).

Other important advantages of Twitter are that it is accessible real-time and provides different methods for geo-tagging its messages. Chapter 3 will elaborate more on the characteristics of Twitter as a source of data.

1.2. Problem statement

Twitter appears to be a potential source of spatio-temporal information in different countries and for different domains. For the domain of incident management there are several indications from the literature that Twitter could provide spatio-temporal information, veiled in Tweet-content (Daly et al., 2013; Wanichayapong et al., 2011). However, very little is known about how Twitter can contribute to incident management. More insight is needed into the quality and quantity of potential information for incident management in tweets. How often do people tweet about situations in traffic? What kind of information do people provide about incident? Is this information really useful?

Secondly, more insight is needed into how to get information from Twitter for the purpose of incident management. In the literature many methods are described for knowledge discovery and event detection through Twitter data, but it is not clear if these methods provide the information that is needed for incident management. How do we identify relevant tweets from the huge number of tweets that are sent every day? How can we extract information from Twitter? Can we apply spatio-temporal techniques on the data in order to get useful geo-information?

1.3. Research Objectives

1.3.1. Objectives

The main objective of this thesis is to investigate the value of real-time Twitter data as a source of spatio-temporal information for traffic incident management. In advance, utility requirements for incident management applications need to be investigated and translated into criteria which Twitter data should meet. In an attempt to reach the main objective, Twitter data will be collected, prepared and analyzed using insights and techniques from state-of-the-art literature and other applications. During the analysis, twitter data must be tested on these criteria in order to assess their requirements on quality and practical usefulness for incident management applications.

In order to assess the requirements it will be necessary to evaluate the performance of a proof-of-concept model that will be set up as part of the research. Information derived from the Twitter data should be verified on different aspects, for example detail and correctness. The speed at which information can be extracted from Twitter data should be sufficiently high as well in order to be useful for traffic incident management.

An important objective of the thesis is to investigate the most important shortcomings and advantages of the Twitter data and the methodology that is used to derive useful information from the data. By highlighting these shortcomings and advantages, valuable recommendations can be given for further research and future application developments.

1.3.2. Research questions

The main question of the thesis is:

- *How useful is Twitter as a source of spatio-temporal information in the domain of incident management within the Netherlands?*

Sub-questions that need to be answered in the course of the research:

- *How is Twitter used as a source of (spatio-temporal) information in current applications and studies over the world?*
- *What are requirements of a Twitter-based incident management application in order to bring added value to the daily practice of incident management?*
- *What are the criteria for twitter data quality for it to be useful for incident management in the Netherlands, regarding accurateness, spatial and temporal scale, spatial and temporal density, topic, and completeness of information?*

- *How can (geo-)information be extracted from Twitter data?*
- *How can the requirements of Twitter data and a Twitter-based application be assessed and validated on the basis of their criteria for quality and usefulness for incident management?*

1.4. Thesis structure

First, in chapter 2, an overview will be given of important literature related to this thesis. The studies that are described can be grouped into three main research focuses: geolocalizing tweets, event detection from Twitter data, and traffic-related information extraction from Twitter data. The chapter finishes trying to place the thesis into perspective of the relevant literature by making use of a schema that summarizes objectives and methodological approaches for all of these studies. After reading chapter 2, the scope of the thesis in regard to other literature should be clear to the reader.

Chapter 3 will discuss the accessibility of Twitter data via APIs, characteristics and structure of the data and the way data have been gathered for usage in this thesis. A set of criteria for the data and the gathering process is formulated, and these criteria will eventually motivate the decision on which API and gathering techniques are used to collect the necessary data.

Chapter 4 will describe the methodology of this thesis. First, criteria concerning the information quality are defined that need to be met in order for the data to be useful for incident management. Next, the data collection methods are described. A motivation will be given for the study area that is chosen to apply the analyses in the methodology on. The different analyses, correlation statistics and geographic irregularity analysis that are used to reach the research objectives will be described in more detail.

In chapter 5 the results of the analyses are given. In chapter 6 these results are interpreted and discussed. Chapter 7 contains the final conclusions and recommendations.

2. Background

2.1. Twitter as an information source

The relatively open character of Twitter has resulted in many studies that made use of this social medium as a source of data. In addition to the wide range of academic fields of study that Twitter data have been used for, there are several examples of commercial applications that rely on the real-time data from Twitter streams. This chapter will give an overview of the related research and commercial applications that made use of Twitter data. In section 2.1.3 examples of research in the specific field of information extraction for traffic management will be discussed.

In the literature, there are different approaches of studying tweet content. Concerning the relevance of the literature for this thesis, an important distinction could be made between studies that focus on the geographic component of Twitter data and studies that do not. One thing that all studies on Twitter data have in common is that they should classify or cluster tweets one way or another in order to extract information from the data. In many cases, in order to extract useful geo-information from Twitter data, tweets should be classified both on the messages' content as well as on their geographical component. For this reason, this literature review will not discuss examples from the literature for which the geographical component of Twitter data is not relevant. The work of Bontcheva & Rout (2012) offers a comprehensive meta-review of semantic analysis for mining and information extraction of social media streams.

2.1.1. Geolocalizing tweets

A major challenge that needs to be dealt with when using Twitter data as a source of geographic information is the scarcity of tweets that are geo-referenced. Approximately 1% of all tweets are explicitly geotagged (Schulz et al., 2013). In an extensive literature review, Schulz et al. (2013) summarized twenty studies that dealt with this challenge of geolocating tweets or Twitter users. In these studies, different spatial indicators were used in order to geolocate a tweet or Twitter user. Most of the time, the tweet's text was used by applying natural language processing techniques on the terms in the text. An alternative for using natural language processing, that was used in the remainder of the studies is matching the terms in a tweet with a database of geographic locations, using a Gazetteer. An advantage of this approach is that it does not require training data and is much simpler (Schulz et al., 2013).

Instead of using the message text, some studies focus on using the location information that is sent with a tweet. However, an extensive study by Hecht et al. (2011) on the location field in tweets showed that the location field is not a very good spatial indicator on its own. Schulz et al. (2013) are the first that used a multi-indicator approach for geolocalizing tweets and Twitter users. They designed this multi-indicator approach because this method should be less vulnerable to missing or incomplete data. The work of Schulz and his colleagues is one of the best on geolocalizing tweets in recent literature. They managed to geotag 92% of all tweets with an average distance error of less than 30 kilometers. Hence, it seems that geo-tagging tweets without a GPS-coordinate is a very challenging job. In the scope of this thesis, tweets should be geo-tagged on a very small scale in order to be suitable for usage in incident management. For this reason, geo-referenced tweets are most useful for the objectives of this thesis.

2.1.2. Event detection with Twitter data

One of the leading articles in the field of real-time event detection from Twitter data is the work of Sakaki et al. (2010; 2012). Sakaki succeeded in detecting the locations of earthquakes with a seismic intensity scale of 3 or more from tweet content with a probability of 96%. The detection system that was developed by Sakaki is able to send earthquake notifications much faster than the official announcements that are broadcasted by the Japanese Meteorological Agency (JMA). The event detection mechanism works on the basis of classifying tweets and probabilistic spatio-temporal modeling. First, tweets are classified based on different features like the keywords in a tweet and the number of words that a tweet consists of. Next, the center and trajectory of

the event location are estimated on the basis of Kalman filtering and particle filtering, using the location information. An important assumption that is made in the analysis of Sakaki, is that only one event (earthquake) takes place in Japan at a time. Within the context of this thesis, multiple events (traffic incidents) could take place simultaneously and therefore this might be an additional challenge for the thesis.

Lee et al. (2011) discussed the possibility of detecting events by only making use of the geographical coordinates of geotagged tweets. Tweet content was not used in their event detection approach. Lee et al. attempted to identify geographical irregularities in the patterns of tweet traffic in their analysis. In advance, they constructed frameworks for geographical regularity of specific towns in Japan based on geo-tagged messages from microblogs, like Twitter. The twitter traffic within these frameworks of geographical regularity were monitored. If the geographical pattern of tweets deviated from the regular pattern, then most likely an event took place within the framework. Deviant patterns could for example be a sudden increase or drop in the number of tweets in a certain region. A sudden increase of unique Twitter users in a certain region for a short period could also be an indicator for an event. The authors succeeded in finding many announced, as well as unexpected events in the experiment. Depending on the number of available geo-tagged tweets in the Netherlands, this may be a useful methodology to apply in order to reach the objectives of this thesis.

A different approach to detecting events based on their spatio-temporal pattern has been developed by Sugitani et al. (2013). In this study local events, regardless of size and type, were identified by using spatio-temporal clustering techniques. First, the authors filtered out noise from their data. Secondly, clusters of tweets that were sent within a short time period and spatial distance from each other were identified. Within these clusters co-occurrences of keywords were searched for, which can indicate a relation between these tweets and a possible event that caused the formation of these tweet-clusters.

Next to the work of Sakaki et al. (2010; 2012) where focus lies on detecting events that occur on a large scale, there are similar studies that try to detect smaller scale events. Walther and Kaisser (2013) built an algorithm that was used to identify places in a given geographic region which showed high amounts of Twitter traffic. In case any high Twitter traffic was discovered as being an event, they used Machine Learning in order to classify this event as being a real-world event or not. The goal of Walther and Kaisser's work was not only to detect real-world events, but also to know the precise location of these events. In this way, events could be presented to a potential user on a map. The authors aimed at designing an event-detection system for the following customer groups and use cases:

- Police forces, fire departments and governmental organizations, which could use the system to become more aware about situations that could happen in the service area for which they are responsible for
- Journalists and news agencies, which could be informed about the latest breaking events
- Private customers, who want to know what is going on in their neighborhood

It's important to notice that many of the events that the authors aim to detect in these cases are covered by just a few tweets. This scale of event-detection requires a much different approach to large scale event detection such as earthquake detection. In order to detect small-scale events, Walther and Kaisser created a "ClusterCreator" which checks if a certain number of tweets (three) are issued within a certain timespan (30 minutes) and within a certain geographic radius (200 meter). The work of Walther and Kaisser shows that the detection of undefined events, instead of searching for specific pre-defined events such as traffic jams or forest fires, is a lot more complex.

Twitcident (Abel et al., 2012a; Abel et al., 2012b; Terpstra et al., 2012) is an application that makes use of real-time Twitter data in order to detect all kinds of real-world incidents by semantic analysis. The application is able to automatically filter information from social web streams that is relevant to any real-world event. Users of Twitcident are able to analyze a particular situation as reported on social media. In this way, Twitcident can support short-term decision making in case of incidents. The work of Abel et al. is highly

relevant to the topic of this thesis. Twitcident is not only a state-of-the-art piece of work in scientific literature; it is put into practice in the Netherlands too in various pilot cases.

For example, Twitcident was brought into action during a Dutch music festival which attracts around 55.000 visitors (Twitcident, 2013b). It was expected that Twitcident could support the festival workers in different crowd management tasks during the festival by early detection of increased risk of incidents on the festival ground. In this particular case Twitcident proved its power by informing the organizers about very long queues in front of tap water facilities. Relevant to this thesis is that Twitcident can also make use of the geographic component in social media. This is both shown in an analysis of Terpstra (2012) and in a second pilot case where Twitcident was used by ProRail weather service (Twitcident, 2013a). ProRail weather service used the Twitcident weather map in order to identify locations in the Netherlands where the weather could hinder the train traffic. Next to the traditional weather forecasts, the Twitcident weather map could verify real-time weather conditions by updates from Twitter users. In this way, Twitcident could detect locations where the weather was worse than expected, and so the organizers at ProRail could take appropriate action.

The work of Abel et al. provides several relevant insights into the topic of information extraction and event detection from social media and Twitter in particular. The architecture of Twitcident emphasizes the need for noise elimination, which entails extracting only useful and relevant information from the enormous amounts of data. Sophisticated semantic filtering is necessary in order to reduce the noise of Twitter data. Finally, the challenges that come with building real-time applications become clearer from the studies of Abel et al. Next to automatic information filtering, it is essential that this information is accessible and findable in a given incident context (Abel et al., 2012b). Twitcident succeeded in this by providing a user-friendly interface.

2.1.3. Twitter as information source in traffic management

Next to the most important literature on the topics of knowledge discovery and event detection from Twitter, some studies discussed the potential value of real-time Twitter data for usage in traffic management. Daly et al. (2013) motivated that there is a need for real-time information about the underlying reasons of traffic conditions. Traffic congestions can have different causes, such as broken traffic lights, road-works, accidents or large events like music concerts. Access to real-time traffic information is increasing; however, for citizens and traffic operators it is important to know what reason is behind a traffic jam. Daly et al. (2013) suggest that social media can be used to capture information and highlight the causes of traffic conditions. In order to substantiate this suggestion, Daly et al. (2013) built a user application called Dub-STAR (Dublin's Semantic Traffic Annotator and Reasoner). In Dub-Star both static data from event planners and dynamic data derived from social media are combined in order to bring users updates about traffic conditions. Input from social media is scanned for possible causes of a traffic congestion on the basis of the time-window and spatial relationship with a specific congestion event. In order to define this spatial relationship, first messages are geo-coded on basis of the user-generated text. In a geo-coding evaluation it was discovered that 50% of the geo-coded locations of messages where accurate with an error range of 500 meters, and 100% were accurate within 2 kilometers. Because the radius of the city of Dublin spans nearly 30 kilometers, these are promising results.

In line with the work of Daly et al. (2013), Mai and Hranac (2013) attempted to derive information about causes of traffic congestions from social media as well. The purpose of Mai en Hranac's work however differs from the work of Daly et al. The objective of Mai and Hranac is to collect external data on traffic on roadways over time to use in traffic performance analyses. Questions that could be answered with the external data from social media are for example "Why are speeds low in this area?" or "How much delay would a rainstorm at this time and location likely cause?" (Mai & Hranac, 2013). In order to discuss the value of Twitter data in answering these questions, they compared incident records from the California Highway Patrol with tweets related to roadway events over the same time period. The authors expected that more traffic-related tweets are sent in case of an incident, and so they tried to discover this pattern in a big dataset of incident-related tweets. Moreover, Mai and Hranac (2013) expected to find that tweet content of traffic-related tweets is more

incident-related near the time and location of an incident. In order to find this pattern in the data, they applied a semantic analysis. The authors found that Twitter use appeared to correlate with the California Highway Patrol, but that sophisticated filtering based on content and location is needed to maximize their conclusions.

Similar, but more sophisticated compared to the work of Mai and Hranac (2013), is the study of Schulz et al. (2012). In order to increase the situational awareness in case of incidents, the authors provided a solution for real-time identification of small-scale incidents using Twitter. Similar to the application of Twitcident (Abel et al., 2012b), Schulz et al. used text classification and semantic enrichment of tweets in order to detect events. A main difference however, is that in the work of Schulz et al. more focus lies on geo-locating incidents on a small scale. Whereas other approaches in the literature rely mostly on city-level precision, Schulz et al. managed to extract precise location information on street level for 87% of their tweets. The event detection approach of Schulz et al. was tested on its ability to detect car crashes. The authors compared fifteen car crash incident logs from open government data with incidents that were detected in their database of tweets. All fifteen car crashes could be identified based on an average of ten related tweets, and a minimum of three related tweets. This result is very promising regarding the research objectives of this thesis.

2.1.4. Research scope definition

From the preceding sections, the major challenges that come with using Twitter data as a source of geographical information are described. It became clear from literature that different research goals require different approaches of processing Twitter data to be useful for event detection. First, a distinction can be made in literature on the type of events that are tried to be detected out of Twitter data (Table 1):

- Geographic scale – For example, an earthquake is a larger scale event than a traffic jam.
- Well-defined or unclear – In some studies it is tried to seek for specific events like traffic jams or earthquakes, whereas in other studies Twitter data is mined in order to identify unknown events.

	Event to detect in Twitter data				
	Scale			Well-defined	
	Large	Medium	Small	yes	no
Sakaki et al. (2010)	x			x	
Lee et al. (2011)		x			x
Walther and Kaisser (2013)			x		x
Abel et al. (2012): Twitcident		x		x	
Daly et al. (2013)			x		x
Mai and Hranac (2013)		x	(x)		x
Schulz et al. (2012)			x		x
Thesis methodology			x	x	

Table 1 Overview of type of events that are aimed to detect in related studies.

Another distinction that can be made in literature is on the detection method of events (Table 2):

- The detection process can rely solely on Twitter data, or external data can be used next to Twitter data to support the detection process.
- The detection method can be based on semantic analysis, and/or can be based on spatio-temporal clustering.
- Different spatial indicators can be used in the detection process – For example a tweet's text, location field or coordinates can be used.
- The detection method could support real-time detection, or not.

	Detection method								
	Based on external data		Event identification based on		Spatial indicator in tweet			Real-time dection	
	yes	no	semantics	spatio-temporal cluster	text	location field	coordinates	yes	no
Sakaki et al. (2010)		x	x	x				x	
Lee et al. (2011)		x		x	x		x		x
Walther and Kaisser (2013)		x	x	x			x	x	
Abel et al. (2012): Twitcident	x		x		x			x	
Daly et al. (2013)		x		x	x			x	
Mai and Hranac (2013)	x		x	x			x		x
Schulz et al. (2012)		x	x		x	x	x	x	
Thesis methodology		x		x			x		x

Table 2 An overview of the type of detection methods that are used in all related studies

Within the diversity of related literature, the scope of this thesis can be further narrowed down. In both of the tables, Table 1 and Table 2, it is listed how the research is related to the state-of-the-art. Regarding the related literature, the scope can be defined as follows:

- The geographic scale of the events that will be studied in this thesis is small. Traffic incidents are one of the smallest scale events that have been studied in literature.
- The events are well-defined: we are only interested in traffic incidents on the Dutch roads.
- The detection method will mainly rely on Twitter data. Logs from Rijkswaterstaat on traffic incidents will be used for validation of the found results.
- The detection method will be based mainly on spatio-temporal clustering of Twitter data. No semantic techniques are applied.
- Only coordinates will be used as spatial indicators of tweets.
- The detection method as applied in this thesis, will not aim at real-time detection of traffic incidents. In the first place, gathered Twitter data will be used for analysis. On the other hand, the potential of Twitter streams for real-time incident detection will be discussed based on the thesis' results.

2.2. Twitter data characteristics and gathering options

This chapter will give an overview of the data characteristics of Twitter data and the way this data is accessible through the internet. In order to make use of Twitter data successfully, insight in these characteristics of Twitter data is essential. Section 2.1.1 will discuss the different ways of how Tweets can be filtered and retrieved from the Twitter APIs. In section 2.2.2 a motivation will be given for the decision to gather data from Twitter's streaming API. 2.2.3 will discuss the data structure of the 'raw' tweets that can be retrieved from Twitter's streaming API. Finally it is discussed in section 2.2.4 how the raw data can be stored and exported into databases that are usable in a GIS.

2.2.1. Twitter data accessibility

Twitter provides real-time access to their data through different APIs: REST (Representational State Transfer) API and Streaming API. There are some important differences between the REST API and Streaming API. The REST API is mainly used in applications and websites that need to make use of Twitter's functionality like searching for tweets based on a keyword, search within a Twitter user's timeline (all tweets that are ever sent by someone), or search for tweets based on a hashtag. Figure 5 shows the process that starts when an application or web site is making use of Twitter's functionality. A user can make a request to a website, which is issued by a server to Twitter's REST API. The API will generate some response from the server's request which is printed to the user via the server. This whole process requires keeping a persistent HTTP connection open. Moreover, there are some limitations to the number of request that the user can make in a certain time span, and the number of tweets that are generated as a response from the API. This all doesn't make the REST API suitable for harvesting great numbers of tweets for analysis.

Establishing a connection with the streaming API works differently (Figure 6). The processes that are necessary for maintaining the streaming connection and handling HTTP requests, run separately. Once a connection is made, access is provided to tweets as they occur in the real-time stream. Eventually, the streamed tweets can be stored.

The streaming API is most suitable for data mining and research purposes (Megally, 2012). The streaming API offers low latency (i.e. near real-time) access to all publicly available tweets in Twitter's global stream (Twitter, 2012b). There are three different streaming endpoints available, each customized to different use cases:

- Public stream – this stream contains all public Twitter data
- User stream – this stream contains all data corresponding with a single Twitter user
- Site stream – Multi-user version of the user stream

According to Twitter (2012b) the Public stream is best suitable for data mining. Once a connection is established to the public streaming endpoint, a feed of Tweets is delivered which has no rate limits. The 'unlimited' access to public Twitter data via the public streaming endpoint offers the opportunity to store great amounts of data for further analysis. For this reason, the public streaming endpoint is used for capturing all the data that is needed for the analyses in this thesis.

The only limitations when making use of the streaming APIs, is that only public tweets can be requested, and that only a small fraction of the complete stream is accessible. There are 3 different public streaming endpoints to choose from:

- POST statuses/filter – returns public tweets that match one or more filter predicates. There are 3 predicate parameters, from which at least one needs to be specified in order to receive tweets from the stream:
    - Follow: – returns public tweets that are sent by specific Twitter users

- Track: – returns public tweets that contain specific keywords
- Location: – returns public tweets that are sent within a certain geographic region (specified by a bounding box)

- GET statuses/sample – returns a small random sample of all public tweets.
- GET statuses/firehose – returns all public tweets, however this requires special authorization. There are only few applications that require this level of access.

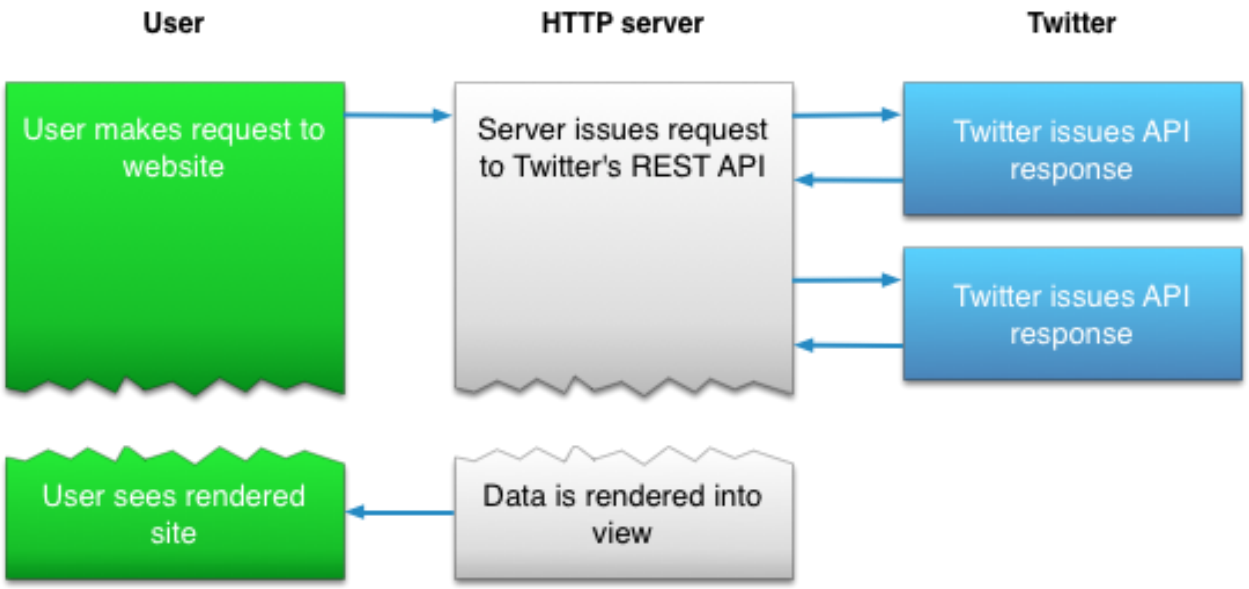


Figure 5 The process of making a connection to the Twitter REST API

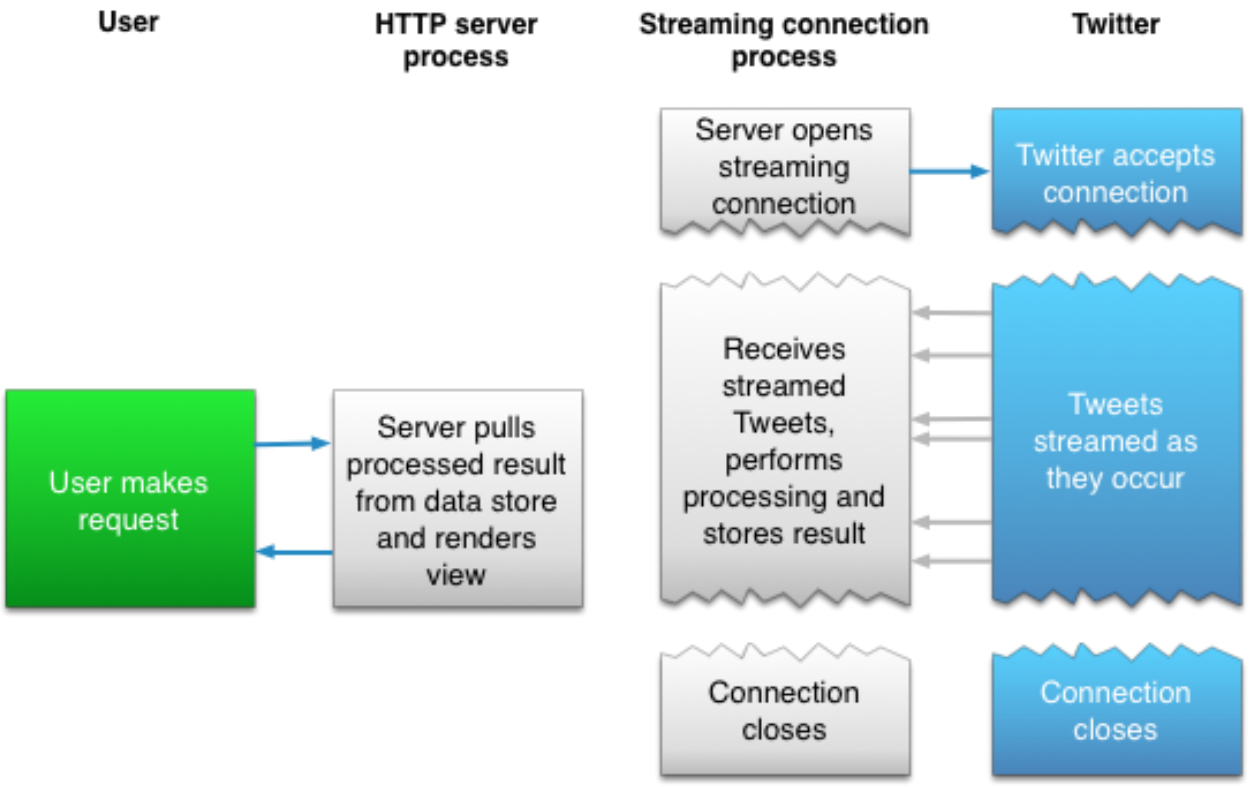


Figure 6 The process of making a connection to the Twitter Streaming API

2.2.2. Motivation for using the streaming API

In order to reach the thesis' objectives, analyses need to be done on Twitter's data. Hence, the first step of the research is building up a database of tweets. In order to do so, a decision needs to be made how this is done and which API should be used for data gathering.

It is decided that the public streaming API is used for filling a database with tweets. There are several reasons for choosing this API for building a tweet database. The most important reason is that the streaming API is designed for long lasting connections, which makes this API very useful for automated harvesting of tweets. The REST API has many limits and restrictions (Twitter, 2013c) which makes automated storage of tweets much more complex. It would for example require more than one user account and IP address to make request to the APIs (Stronkman, 2011).

The architecture of the streaming API is designed in such a way that only one request is needed in order to make a long-lasting connection which receives tweets real-time as they occur. Though, there is a rate limited set for using the streaming API. The public streaming APIs cap the number of tweets that are sent to a client to a small fraction of the total volume of the tweet stream (Twitter, 2013a). In many cases however, the public streaming cap is not reached because the stream is already filtered on one or more filter predicates. A message will be streamed to the user if more messages are found, matching the filter predicate, then the streaming cap would allow to pass through.

Another decision that needed to be taken, is which stream API endpoint should be used. It is decided that the POST statuses/filter is the only usable endpoint. The GET statuses/firehose isn't accessible with default authorization, so this endpoint can't be used. The GET statuses/sample would receive a random sample of the complete data stream which is not very useful since we focus the research on Twitter traffic in the Netherlands. The POST statuses/filter provides many ways of filtering the twitter stream which makes it possible to fill a database with tweets representing Dutch Twitter traffic in the best possible way.

2.2.3. Structure of raw Twitter data

When making a request to the POST statuses/filter endpoint, the API will give response. This response consists of all tweets from the public stream that match one or more filter predicates. The response format is JSON (JavaScript Object Notation).

A tweet that is responded in JSON format is built up by many fields (Appendix 2). Next to the tweet text, many metadata values are included with a tweet like a timestamp, the place of the Twitter user, the source from which the tweet is sent etc. On Twitter's developers website (2013f), an extensive overview and explanation is given of all fields that are included with a tweet.

2.2.4. Twitter data gathering options

Making requests to the Twitter streaming API can be done in many different ways. On the internet, different scripts in different languages are available for free re-use (Barbera, 2013; Cantino, 2013; Graser, 2012; Haslam, 2012; McCarroll, 2012). Whether the scripts are useful or not, depends on a couple of criteria:

- All attribute fields of a tweet can be gathered from the stream
- The script should be able to make a connection through OAuth authorization (Twitter, 2013b). Simple authorization via username and password isn't supported anymore by all Twitter APIs since May 2013.
- The script can handle connection time-outs in order to realize long-lasting connections.
- The script can write the responded tweets to an output file which can be used to transport tweets into a geo-database.

In order to fulfill all criteria, a Python script (McCarroll, 2012) is used for the data gathering process. Appendix 3 contains McCarrol's full original script.

2.3. Traffic incident management background

2.3.1. Practices of incident management

Incident management is defined as all the actions that aim at stabilizing the traffic flow on the road after an incident has happened (Eurlings, 2010). In the Netherlands different parties are involved in the practice of incident management. The police, fire brigade, ambulance, Rijkswaterstaat, ANWB and salvage workers are cooperating in incident management on the Dutch roads.

The main goal of incident management is to reduce social costs inflicted by traffic jams. This is realized by making good arrangements between all partners that are involved in incident management. The aim of these arrangements is to resolve an incident as quickly as possible while taking good care of incident victims and investigate the incident cause. When handling an incident, priorities are set as follows (The Netherlands Traffic Management Centre (VCNL), 2005):

- The emergency worker's own safety
- Traffic safety
- Treatment of casualties
- Maintaining the flow of traffic
- Vehicle / cargo salvaging

Regarding these priorities, the incident management process is started as soon as an incident is called in or detected. The incident management process can be split up into 4 phases:

- Detection and notification phase
- Getting to the scene
- Action phase
- Normalization phase

Detection and notification phase

Incidents can be identified via various ways (Figure 7). In most cases, an incident is called in by an incident-involved person or some other road user. Often, more than one call about the same incident is received by the recipient emergency center. Next to these incident calls, incidents can be detected as well. Incidents can be detected automatically by detection systems of one of the regional traffic control centers of Rijkswaterstaat. Otherwise, incidents can be detected by officials on patrol like social workers, road inspectors or ANWB-employees.

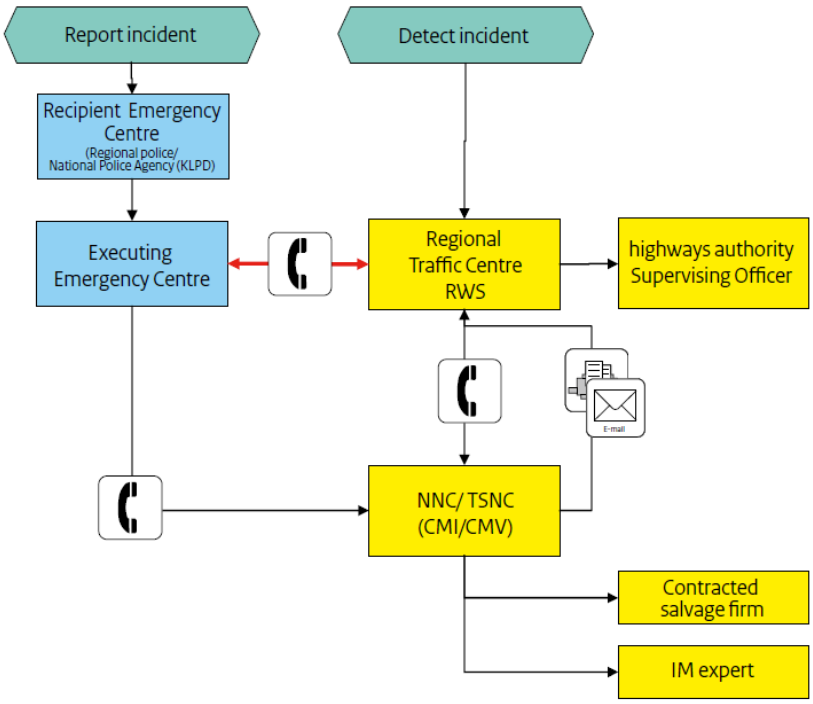


Figure 7 Incident notification chart (Rijkswaterstaat, 2011)

The recipient emergency center verifies the incident and asks which emergency services (police, fire brigade or ambulance) are needed to give aid. Subsequently, the executing emergency center will send the required emergency services towards the incident. In the meantime, the regional traffic center of RWS and the National Notification Centre Incidents (NNCI) and/or Truck Salvage Notification Centre (TSNC) are called and informed about the incident. The regional traffic center of RWS sends out a road inspector to the location of the incident. The road inspector can verify the type of incident and can assist in clearing the road. If possible and necessary the traffic control center takes safety measures and takes actions to stimulate the traffic flow.

The NNCI/TSNC central calls in a salvage worker in order to tow away the incident involved cars or trucks. If a truck is involved in the incident, an incident-management expert is called in as well in order to estimate the damage of the truck and its load.

Getting to the scene, action and normalization phase

After the detection and notification phase (see Figure 4 for all phases), all emergency services make their way to the incident location. In the meantime information about the incident is verified by the emergency center, regional traffic center and via any possible assisting people at the incident location. A precise location and description of the incident situation is essential for the emergency services to arrive fast and work efficiently. The more complete and nuanced information about the incident situation is available, the better the emergency services are prepared when they arrive at the incident location. For this reason, the emergency services keep in contact with the emergency center and/or regional control center until they arrive at the incident location.

In the action phase, different disciplines have their own responsibilities. In Table 3 all actions that need to be taken by various emergency services are summarized in order of priority. In the normalization phase, focus lies on re-enabling a safe traffic flow. All traffic and safety measures taken in order to deal with the incident are withdrawn.

Work process	Highways authority	Police	Fire brigade	Ambulance
Concern for safety	X	X	X	X
Concern for victims			X	X
Investigation		X		
Traffic flow	X	X		
Salvage	X			
Recover/clear damage	X			

Table 3 Incident management work processes for the various emergency services. Work processes that are higher in the tale, have higher priority. (Rijkswaterstaat, 2011)

2.3.2. Improvement goals on incident management

It is extremely important that an incident is dealt with very fast. The time traffic flow is stagnated due to an incident, has to be as short as possible. TNO estimated that the annual total financial loss to transportation of goods and logistic processes due to traffic jams is between 954 million to 1,240 billion euros per year. In different studies for the Netherlands, United States and England, it is estimated that about 20% of the total lost hours by traffic jams is caused by incidents (Knibbe, 2007). Hence, a reasonable amount of the total loss due to traffic jams is directly inflicted due to incidents. The faster an incident is solved, the less financial damage is done.

The effectiveness of incident management is proven by calculations of TNO. In 2003, the Netherlands would have been affected with 65% more hours lost by vehicles as a consequence of incidents, if no incident management had been applied (Immers, 2007). These numbers give every reason to start improving the incident management process in order to enhance the benefits that come with it. In order to improve the application of incident management a set of ambitions and 'SMART' (specific, measurable, acceptable, realistic, time-dependent) aims have been set up in a professionalizing program for the period of 2008 to 2015.

One of the program items aims at shortening the total incident duration with 25% from 2008 till 2015 for the different categories of incidents. Three categories of incidents are distinguished:

- Category 1: car breakdown
- Category 2: truck breakdown, or car accident (only material damage done)
- Category 3: car accident (injury done), truck accident

Each year, the average incident duration for the whole year is calculated and compared with the average duration as it should be according to the enhancement program. The results of this comparison can be seen in Figure 8. Developments on the duration of incidents of category 1 and 3 are ahead of schedule. Developments on the duration of incidents category 2 are behind schedule.

Although the incident duration of category 3 incidents is ahead of schedule, attention to this category is needed in order to keep the development on schedule towards achieving the goal for 2015. The management of incidents of category 1 does not require any additional measures in order to achieve the 2015 incident duration development goal. Management of incidents of category 2 do need additional measures in order to reach the goals.

Another goal of the professionalizing program is to enhance the information provision to the traveler and media. Information about the incident, the expected normalization time and advises for alternative routing should be sent within five minutes from the incident report or detection (Immers, 2007).

It is for these two goals, reduction of the incident management process and information provision to the public and media, that Twitter could play a supportive role. The next chapter will elaborate more on the expectations of Twitter as an information source to support different incident management practices.

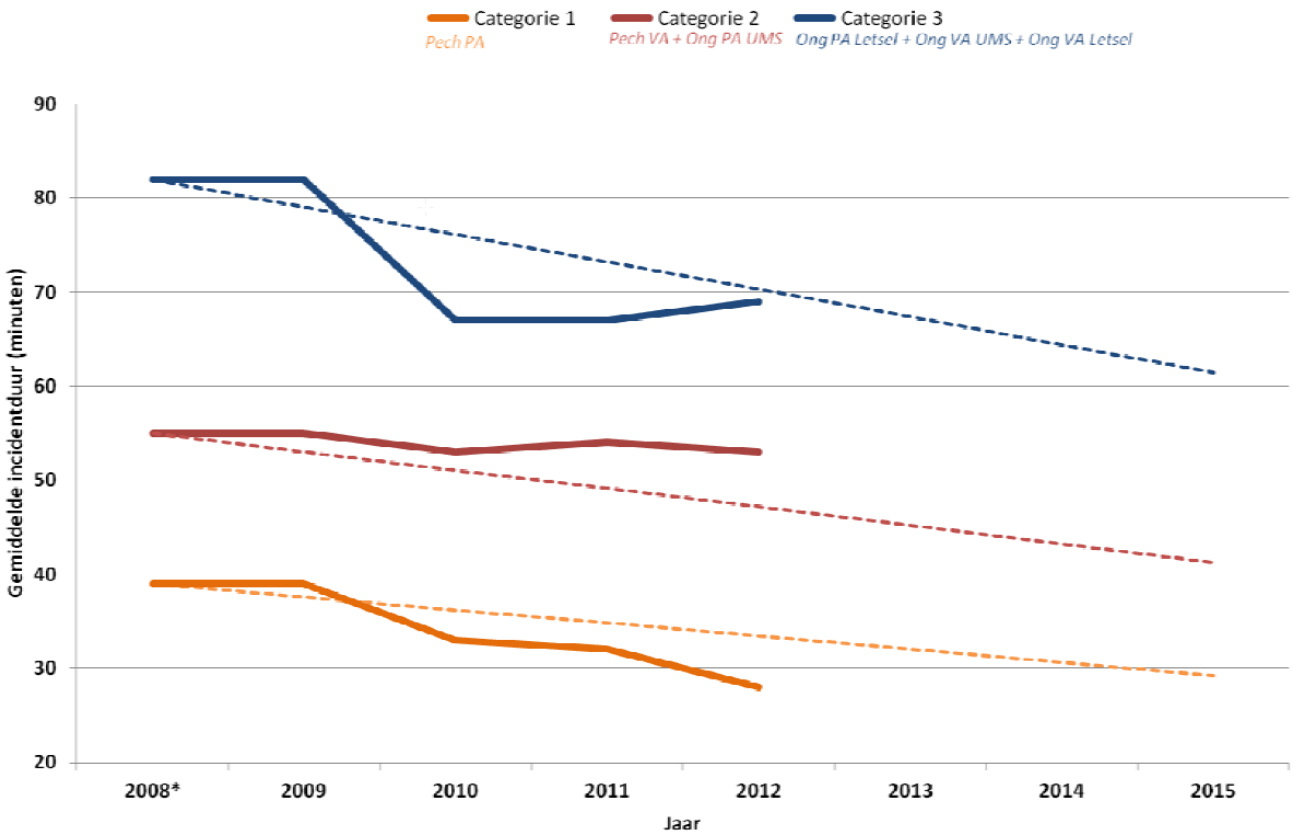


Figure 8 Development of incident duration regarding the ambition of the incident management professionalizing program (Ammerlaan et al., 2013)

3. Use cases: Twitter as information source for IM

The preceding sections of this chapter made clear that incident information is essential in the incident management process. The faster and completer an image of the incident scene can be drawn, the faster actions and efficient measures can be taken by all IM-involved disciplines. In addition, the chance of a follow-up accident will be smaller if the situation is under control due to good information provision (The Netherlands Traffic Management Centre (VCNL), 2005). Twitter, as a potential source of (geo-)information, could possibly provide support in different practices within incident management process:

- Incident report verification
- Incident detection
- Incident communication to road users

3.1. Incident report verification and enrichment

The first practice for which Twitter is expected to be useful, is the verification and enrichment of incident reports. Incidents that are called in are verified first by asking the person reporting the incident for details. After the verification, emergency services are sent out to the incident scene. After the emergency services have departed, the report is enriched, if possible, with complementing or more nuanced information. The emergency service first arriving at the scene will provide feedback about the incident to the control centers. All responsibilities of the control center are subdivided and structured as (The Netherlands Traffic Management Centre (VCNL), 2005):

- ASK: asking the person reporting the incident for information / details
- DISPATCH: sending the service to the scene
- CONSULT: consulting the emergency worker at the scene of the incident
- COMMUNICATE: communicating all information to the other control centers involved
- RECORD & EVALUATE: recording and evaluating the data of the handling of the incident

If people around the incident scene are using Twitter to report about the incident, then this information could be used for verification of the incident, and enrichment of the incident details. Immers (2007) found out from discussions with emergency workers that often there is misunderstanding about the incident location or the nature of the incident. These misunderstandings lead to loss of speed and quality of the emergency assistance. Ideally, Twitter could give both indications on the location of an incident as well as the nature of an incident. This information could perhaps be used as verification material and so reduce misunderstandings about the incident scene. Aside from the hypothesis that Twitter could provide information that is usable during the incident handling, Twitter could maybe provide useful information for the RECORD & EVALUATE-task of the control center as well.

Figure 9 shows an example of an incident-related tweet which is suitable for incident report verification. From the text in this tweet it becomes clear that the person who sent the tweet witnessed a car accident and called for emergency service immediately. Thereafter, he posted a tweet with a picture of the scene included. Because the Twitter user posted information about its location (#Kerkstraat, #Hogezand), this tweet can be linked to the emergency call that he made earlier. Very favorable in this case, a picture is included in the tweet that gives a clear overview of the incident. The picture makes immediately clear that one of the cars is total loss and should be towed away.

A second example of tweet that could potentially be used for incident report verification and enrichment can be seen in Figure 10. The Twitter user describes his location on the road, and uploads a photo of the incident scene with his tweet. Information from both examples of tweets and pictures in Figure 9 and Figure 10 can be verified with the information that emergency callers provided, and additionally enriches the image of the accident scene.

Figure 9 Example of an incident-related tweet, suitable for incident report verification. Free translation of tweet: quite a heavy crash on the #Kerkstraat in #Hoogezand, I got out immediately and run to, called 112.

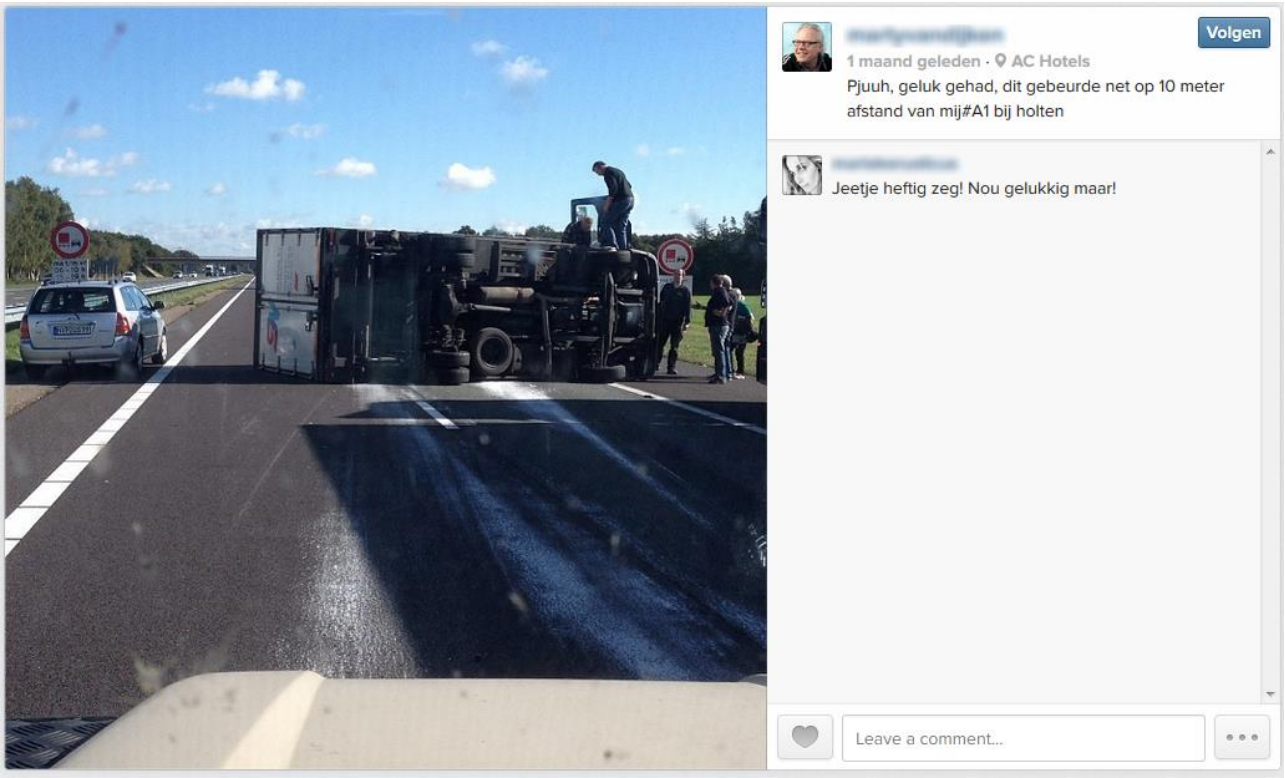


Figure 10 Example of an incident-related tweet, suitable for incident report verification. Free translation of tweet: I was lucky. This just happened 10 meters in front of me. #A1 near Holten.

3.2.Incident detection

Next to incident verification, Twitter could be used as a first-line, stand-alone information source, from which first indications could be received for possible incidents on the road. Twitter as base for event detection is a topic for many examples of studies (see Table 2 for an overview). Also, there are some examples in literature that aim at detecting traffic incidents from Twitter data (see section 2.1.3). Incident detection is therefore a high potential study topic, however, a highly challenging topic as well. Because Dutch Traffic Centres have much technology on board for traffic monitoring and incident detection, Twitter will face high 'competition' regarding fast incident detection. Maybe past research already proved Twitter's ability for incident detection purposes, but in order to be actually useful Twitter should perform better than traditional incident detection technologies.

The fastest way of traditional incident detection is if someone involved in an accident makes an emergency call. Information about the accident is provided directly from first hand to the emergency services in these cases. It is highly improbable that someone involved in an accident will post information on Twitter first, before calling for emergency service. A higher potential for usefulness of Twitter in incident management is maybe more likely to be found in incident cases where no emergency services need to be called. 85 percent of the incidents concern material damage only (Immers, 2007). Approximately 270 incidents a day occur in the Netherlands (Steenbruggen et al., 2013), so there are many occasions in which an incident does not require emergency service. For these cases, Twitter could be a faster detector of the incident than other means.

In section 2.1.4 it is summarized that in literature event detection from Twitter data can be based on semantic analysis and/or spatio-temporal clustering. In this thesis, more focus lies on the use of GIS to cluster tweets based on their spatio-temporal relation in order to detect incident-related events in the Twitter data. Semantic analysis is beyond the scope of the thesis and will only be applied in the form of queries on data attributes. It is expected that spatio-temporal clustering of tweets could be used for incident detection if sufficient tweets, which can be accurately geotagged, are sent around an incident scene.

Figure 11 Example of a tweet that provides information about the cause of a traffic jam. Free translation of tweet: Chaos on Capelseplein in the direction of Rotterdam, because of a delivery van in the crashbarrier.

3.3.Incident communication to road users

A final practice of incident management for which Twitter could be useful, is the communication of incidents to road users. Fast communication about road incidents to the public could reduce the growth of traffic jams due to the incident. The smaller the traffic jam, the faster the traffic flow is normalized. Moreover, there's always the risk that cars on the motorway crash into the end of a tailback. Hence, the shorter the period in which traffic is stagnated, the less chance exist on secondary incidents due to traffic jams. Therefore, communication to the public about incidents and traffic jams is an essential task within incident management.

Traffic information however is not provided directly to the road user by the regional traffic centers of RWS (see Figure 12). Communication of traffic information is done via a service provider (ANWB, VID and TomTom are the largest service providers). The service providers buy a license from RWS in order to receive real-time basic traffic information like travel times, raw data and video images. This data is provided by the National Traffic Centre of RWS (VCNL). In special incident cases, the regional traffic center informs the VCNL of the cause of an incident by phone. This information can again be passed through to the service providers. Sometimes, service providers receive additional traffic information from in-house informants or measuring instruments. Through various media, service providers inform the road user about traffic congestions and possible causes of these traffic congestions. The only direct communication between RWS and road users about traffic, is through road-side electronic displays (DRIPs).

A current issue with modern traffic information provision in this context is stressed by Daly et al. (2013). Although Real-time information about traffic congestion has become easily accessible and is very detailed, real-time information about the underlying reason for traffic congestions is much harder to provide. Nevertheless, information about the underlying reason of traffic congestion is essential for road users. If road users are aware of the underlying reason of a traffic congestion, they can make better routing decisions.

It can be expected that Twitter could be an additional source of information during these situations if people in traffic jams write messages about the traffic jams that they are in. Examples of this type of information sharing can already be discovered in current Twitter traffic. It often happens for example, that the ANWB has detected a traffic jam, but is unaware of the cause. Via Twitter, the ANWB asks their followers to keep their eyes open and post information about the cause of the traffic jam on Twitter as soon as they notice (and are able to safely send a tweet while driving). The need for this kind of information sharing comes forward as well in the work of Fernandes et al. (2012). They built an app in which road users could share their problematic experiences on the road in order to prevent traffic jams from happening.

In the Netherlands, Twitter is maybe one of the most suitable media to spread this information and to send and receive updates about the incident progress in a fast way. It is expected that in many cases of traffic congestion Twitter can complement traditional traffic information provisioning if road users share their experiences on the road.

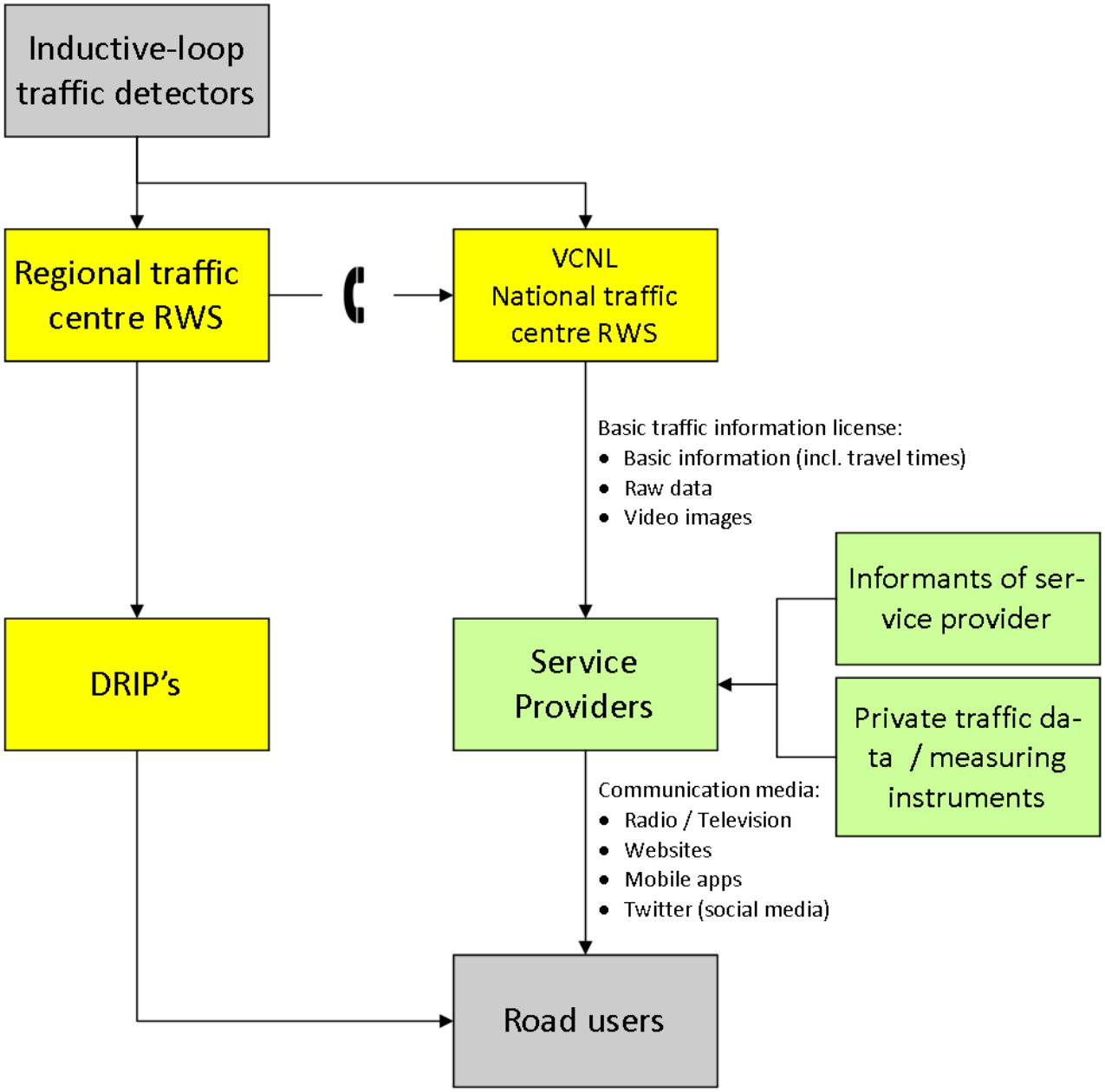


Figure 12 Process of traffic information provision to road users, based on Coëmet (2006)

4. Methodology – Identification of relevant tweets

For the sake of geographic research it is very advantageous that tweets can be georeferenced accurately. In cases when Twitter users share their geographic location, a coordinate pair is sent with the tweets which enables the processing of Twitter data into geographic data. Using the coordinate pair and the timestamp of the tweet, tweets can be located in both space and time as spatio-temporal point data. This spatio-temporal point data can be made ready for different spatio-temporal analysis techniques.

As this thesis tries to investigate the potential value of Twitter data as being geographic data for the application of incident management in particular, it is key to test the 'performance' of this spatio-temporal component of Tweets. The performance of the spatio-temporal component could be defined as the data's ability to be used in spatio-temporal analyses and the usefulness of these analyses outcomes.

In this and upcoming sections, the main goal is to apply spatio-temporal analysis on the Twitter data in order to identify traffic- and/or incident-related tweets. Only when incident-related tweets are found, the information content of these tweets could be qualified on its value for the incident management process. There are many ways to identify relevant tweets. Most often the identification of relevant tweets is done through language processing, as is the case in Stronkman (2011). In this thesis however more focus lies on the identification of relevant tweets by spatio-temporal analysis. In the end, spatio-temporal analysis should bring more insight into the performance of Twitter data as geographical data.

The main question that is tried to be answered in this thesis is whether or not Twitter data can be used as a potential valuable source of geo-information in the field of incident management. In this methodological chapter, the thesis' approach of answering this main question is explained. Answering the main question will be done by answering several sub questions.

Figure 13 shows an image of the 'pipeline' of the research approach. The pipeline shows that the methodology consists of three phases. In the first phase, all data is collected and criteria are defined on the Twitter data quality in order to be useful for incident management. In the second phase, the pre-analysis phase, the study area is defined by doing some analysis on the tweets and incident data. Secondly, correlation calculations are made in order to find a relation between incident happenings and twitter activity.

In the third phase of the methodology the zonal regularity analysis is used to detect traffic-related tweet clusters. First a pre-analysis is carried out in order to test the applicability of the regularity analysis on the Twitter data in the study area. Thereafter, the regularity analysis in the form of a sensitivity analysis is applied on zones around highways in the study area. This sensitivity analysis will bring results, which will be interpreted and evaluated in the final phase. From the interpretation and discussion, a conclusion will be drawn and recommendations for further study will be given.

The first subquestion that will first be answered is about the relevancy and potential value of geo-information for incident management. What information should be extracted from tweets and when (in which situations) is this information relevant and potential valuable for incident management? In section 4.1, the criteria are defined on geo-information in order to be potential valuable for the incident management process.

The next subquestion is about the identification of relevant information from tweets. How can we 'mine' and find relevant information in geographic Twitter data? To answer this question it is key to identify tweets that could meet the definition of being potential valuable for incident management as of the informative content that can be extracted from these tweets. For this data mining of relevant tweets out of the Twitter databases, different approaches and techniques are applied using GIS.

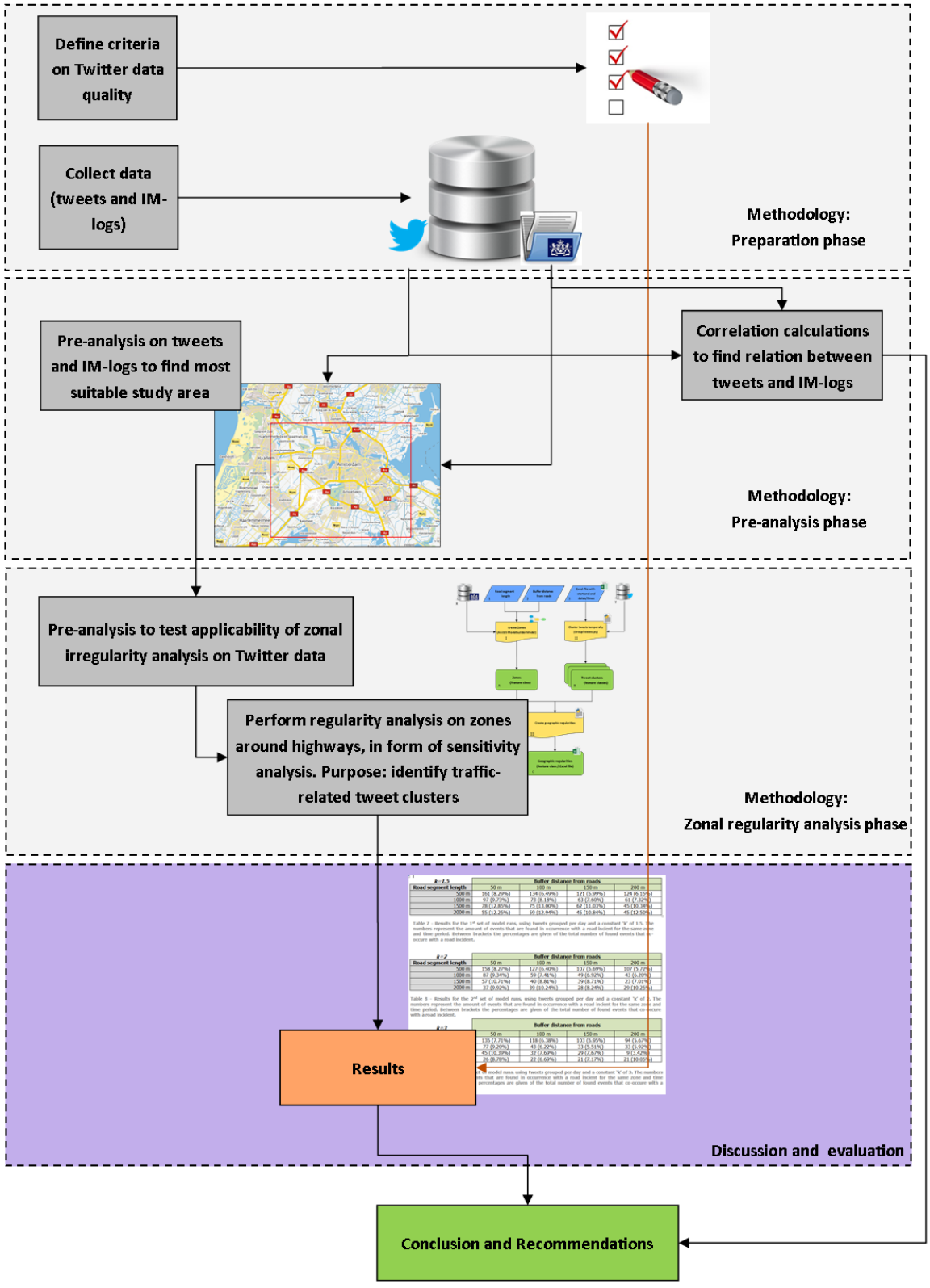


Figure 13 Schematic overview of the 'pipeline' of the research approach

4.1.Criteria (what information is relevant for traffic incidents?)

The preceding chapter described different practices of incident management for which it could be expected that Twitter can give informative support. Whether Twitter really can be of any value in one of these practices or not, depends on different sets of criteria. Further on in Chapter 6.1, the value of Twitter data for different practices of incident management will be investigated based on these criteria.

Criteria on Twitter data for their potential value for incident report verification and enrichment

In order to be useful for incident report verification and enrichment, it is essential that Tweets can be linked to specific incidents on the road based on the location from which a tweet is sent. The usability of tweets for incident management would be minimal if no indications could be found that a Twitter user was really a witness at the location of the incident scene. Therefore a first criterion can be defined as:

(criterion 1) A tweet should be sent in the neighborhood of an incident, and therefore a tweet should hold sufficient indications or evidence about the location from which it was sent.

A tweet can be linked to an incident in different ways. First, a tweet can be linked to a specific incident based on its proximity to this incident both in space and time. In other words, tweets that are sent near an incident's location and shortly after the happening of this incident, could be related to this specific incident. Secondly, tweets can be linked to specific incidents because they hold incident-related information. Twitter users can give details about the incident in the tweet's text, or upload photos of the incident.

Next to the ability to verify an incident report using tweets, an incident report could as well be enriched by tweets that hold detailed information about the incident situation. Because time plays a very important role in incident management, information from tweets should arrive in time. For example, if very useful information from a tweet comes available just after the emergency services arrived, then this information comes too late for the emergency services and is useless. The moment on which relevant tweets can be accessed and the time it takes to extract information from these tweets, in other words the recency of the information, is essential for their potential value for incident report enrichment. A second criterion could therefore be defined as:

(criterion 2) Tweets should be sent within relevant time limits after an incident happened.

A final criterion for the potential value of tweets for incident report enrichment, is regarding to which extent tweets hold complementary information about the incident situation. If tweets hold information that was not known by authorities through other media or information sources, then the potential value of these tweets for incident report enrichment would increase significantly. Hence, the last criterion will be defined as:

(criterion 3) Tweets should bring detailed information updates about the incident situation

Criteria on Twitter data for their potential value for incident detection

The ability to detect events from Twitter data has been studied for a wide range of event types. One perhaps obvious but nonetheless important criterion for the ability to detect events from Twitter data, is that the events particularly influence the daily life of people. Only if people are influenced by an event, people are induced to tweet about this event (Sakaki et al., 2012). For both large and small scale events it has been shown that these events could be identified based on spatio-temporal characteristics of tweets (Lee et al., 2011; Sakaki et al., 2012; Sugitani et al., 2013).

In order to detect events based on the spatio-temporal components of Twitter data, a spatial and/or temporal relation should exist between the events and the Twitter data. In other words, it is only possible to identify

events from tweets if these tweets can be located near the event location and when the tweets are sent within a short period after the event has happened.

(criterion 4) A spatio-temporal relation should exist between road traffic conditions and patterns in Twitter data traffic.

Criteria on Twitter data for their potential value for incident communication to road users

A key criterion for the potential value of relevant tweets for incident communication to road users, is that information from tweets will arrive in time. The road traffic condition is very dynamic. This means that the development and solution of traffic congestions (due to a road incident) is a matter of minutes. It would be most beneficial for communication purposes if road users would respond as quickly as possible to traffic jams via Twitter. For this reason, criterion 2 will also apply for the potential value of tweets for incident communication to road users.

Another important criterion for the potential value of tweets for incident communication to road users, is that tweet should hold detailed information about the incident situation. Moreover, in the case of incident communication to road users, not only information about the incident is relevant. It could be the case that tweets provide insight into road traffic conditions which are caused by an incident. For this reason criterion 3 is extended to:

(criterion 3B) Tweets should contain detailed information updates about the incident situation or road traffic conditions.

Next to the recency of tweets and the detail of information that they provide about the incident situation, the ability to geolocate a tweet is also important for the application of incident communication to road users. Tweets have to be linkable to incidents or at least it should be clear on which part of the highway the information in a tweet applies. It is for this reason that criterion 1 and criterion 4 will apply to the application of incident communication to road users as well.

4.2.Data acquisition

The following chapter will describe the data that is used in this thesis, and the way it was gathered from different sources.

Criteria 1 and 4 from section 4.1 stated that Twitter data can only be potential valuable for incident management practices if a relation exists between patterns in the Twitter data, and events on the road. Additionally, indications or evidence for this relation have to be traceable in the data and must be verifiable with other data sources.

Road incident-related tweets will have to be searched for in a Twitter database. In order to support the tracing of incident-related tweets, and the verification of an existing relation between road events and patterns in the Twitter database, external sources from Rijkswaterstaat are used. Incident loggings are available for analysis and validation purposes. More details about the content of these databases will be given in the coming sections of this chapter.

Overview of datasets
The following main datasets were used for the analyses in this thesis:

- *Twitter data (Geotagged Twitter database):* This database contains over a million of tweets that can be accurately located on a map by coordinates that were sent with the tweets.

- *Incident loggings from Rijkswaterstaat*

  Hundreds of logs are available from Rijkswaterstaat Noord-Holland. These logs contain details about all incidents that take place on the highways and roads that are monitored by Rijkswaterstaat Noord-Holland. The logs contain the following information:

  - Date and time of incident log entry
  - Estimation of date and time of incident happening
  - Date and time of last mutation of the log
  - Log type
  - Description of the incident
  - Incident type (category)
  - Location of incident
  - Traffic lane information

Twitter Data harvesting setup

For specific periods, raw tweets are gathered constantly via Twitter's streaming API using different filters. The streamed raw tweets are automatically stored as JSON-formatted strings1. Tweets that flow through the streaming API are filtered in two ways (Figure 15). The first filter was set to filter all tweets with a geographic coordinate pair that could be placed within a bounding box around the Netherlands, drawn between WGS-coordinates in decimal degrees: 3, 50, 7 and 54 (see Figure 14) . According to Twitter (2013d), the following heuristic is used by the streaming API to determine whether a tweet is sent within the borders of this bounding box:

- If the field "coordinates" is populated, the coordinate pair in this field is tested against the bounding box.
- In cases where the field "coordinates" is empty, but "place" is populated, it is checked whether overlap exists between the bounding box of this place and the bounding box that is defined by the user. The tweets matches the user defined bounding box if any overlap is found.

If none of these two rules exist then the tweet does not match the filter defined by the bounding box and will not be stored.

On a second computer, tweets are streamed through a filter of specific traffic-related keywords in the tweet text. A comma-separated list of phrases is used as a composition of different queries in order to filter out possible relevant tweets. Some rules are defined by Twitter to use this comma-separated list as a query (2013e) . The following rules are the most relevant for filtering incident-related tweets:

- Commas in the comma-separated list are equivalent to logical OR operators
- Space in the comma-sperated list are equivalent to logical AND operator
- The following entity fields of a tweet are checked for matches with the comma-separated list:
    - text
    - expanded_url
    - display_url
    - screen_name
- exact matching of phrases is not supported
- punctuation and special characters are considered part of the term they are adjacent to

Data were collected in the year 2013 from the beginning of July until the end of December.

[1] JSON stands for JavaScript Object Notation. It is an open standard format that uses human-readable text to transmit data objects consisting of attribute-value pairs. JSON is used primarily to transmit data between a server and a web application (The Basics | JSON - JavaScript Object Notation [online].; JSON - Wikipedia, the free encyclopedia [online].2013).

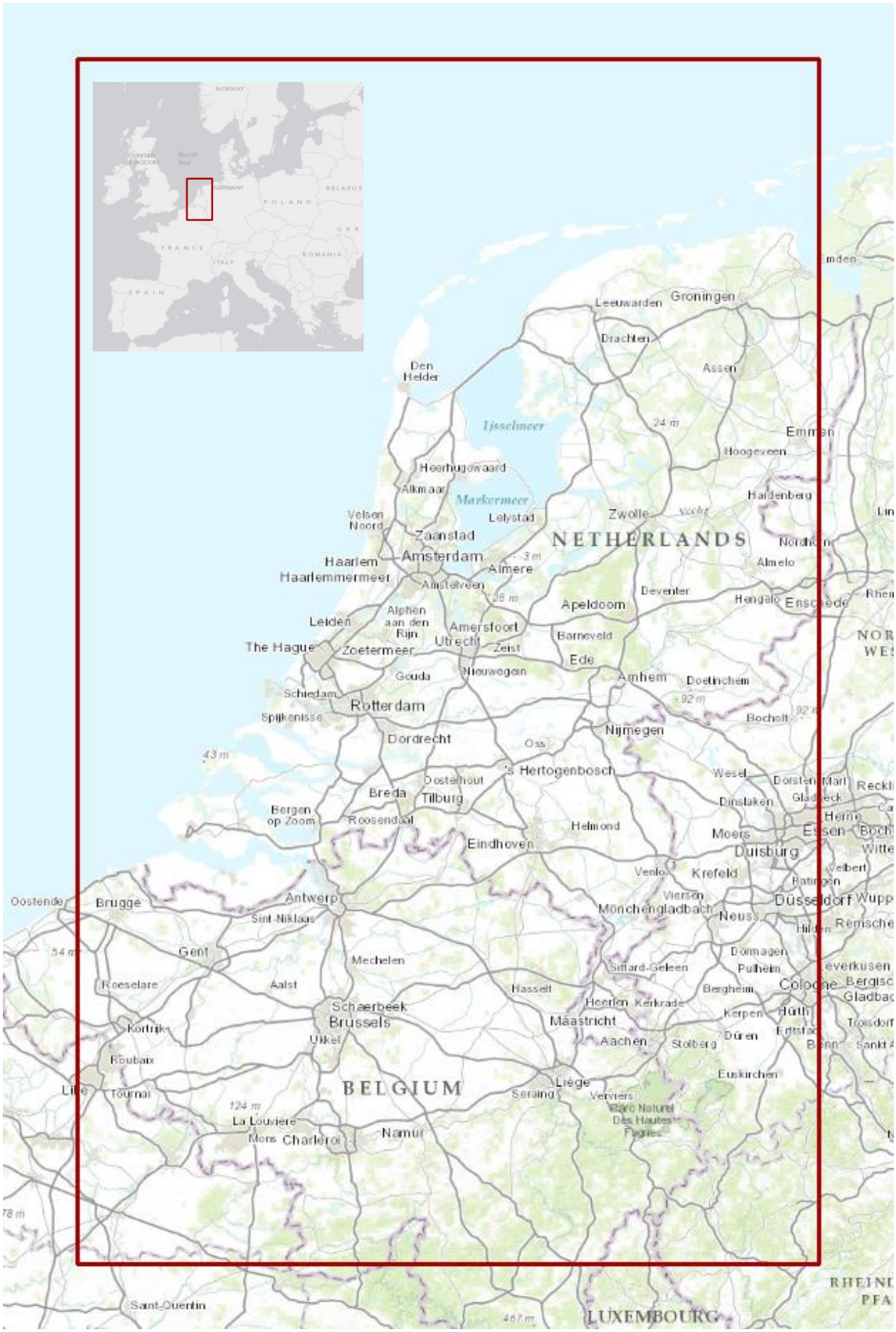


Figure 14 Bounding box which is used for filtering geotagged tweets from the Twitter stream

Data preparation process

The json strings that are generated from the streaming API are prepared for usage in GIS applications. The json data preparation had three main goals:

- *reduce the data volume of the raw json strings.* Tweets that are stored in raw json format contain many entities of metadata (Twitter, 2013f). Some of these metadata entities are considered irrelevant for the analyses in this thesis and can therefore be removed from the data. A conversion from the raw json strings to comma-separated values containing only the relevant data entities, reduced data to 10,5% of the total volume.

- *convert the json data to tables that are readable and have good performance in GIS and other administrative software packages.* Because JSON data isn't structured as a straightforward administrative table, GIS applications as well as other administrative software packages cannot read JSON data. In order to enable queries and analyses on the Twitter data using GIS or other software, the data needed to be converted to tabular data.

- *enable temporal and spatial analysis on the data.* The notation of the coordinate pair and the timestamp attributes of the tweets was not supported by the used GIS software. In order to geographically reference the tweets and to make the data time aware, the coordinate pair and timestamp attributes were converted into a different notation.

In order to reach these goals a process of data preparation has been set up. This data preparation process consists of the following steps (Figure 15):

1: Convert data from JSON format to CSV format.
Using GO programming language[2], only the relevant field values from the JSON lines were extracted and written to a CSV file. This reduced the volume of the data of the tweets drastically. From the 71 attribute fields, only 16 attribute fields were assumed to have potential valuable information for this study. The attribute fields listed in Table 4 were extracted from the JSON lines.

2: Convert data from CSV to File Geodatabase[3].
Although CSV files are readable in ArGIS, performance of analyses will increase substantially when data is stored in a File Geodatabase. Moreover, a File Geodatabase has additional advantages. For example it is able to store up to 1 terabyte of data which can be of different formats (like geographic data in feature classes or non-geographic data in database tables). Because of these high benefits for data storage and analysis performances, a File Geodatabse is founded most useful for data storage.

In order to convert all tweets stored in CSV-lines to data in a File Geodatabse, ArcGIS modelbuilder models[4] are used. Tweets that were filtered from the streaming API based on location and contained a longitude-latitude pair, were stored in point feature classes[5] within the File Geodatabse. All tweets that were filtered from the streaming API based on location, but did not contained a longitude-latitude pair were removed from the data. Tweets that were filtered from the streaming API based on keywords by the second computer, were stored as non-geographic tables within the file geodatabase.

[2] http://golang.org/
[3] http://resources.arcgis.com/en/help/main/10.2/index.html#//003n00000007000000
[4] http://resources.arcgis.com/en/help/main/10.2/index.html#//002w00000001000000
[5] http://resources.arcgis.com/en/help/main/10.2/index.html#//003n00000005000000

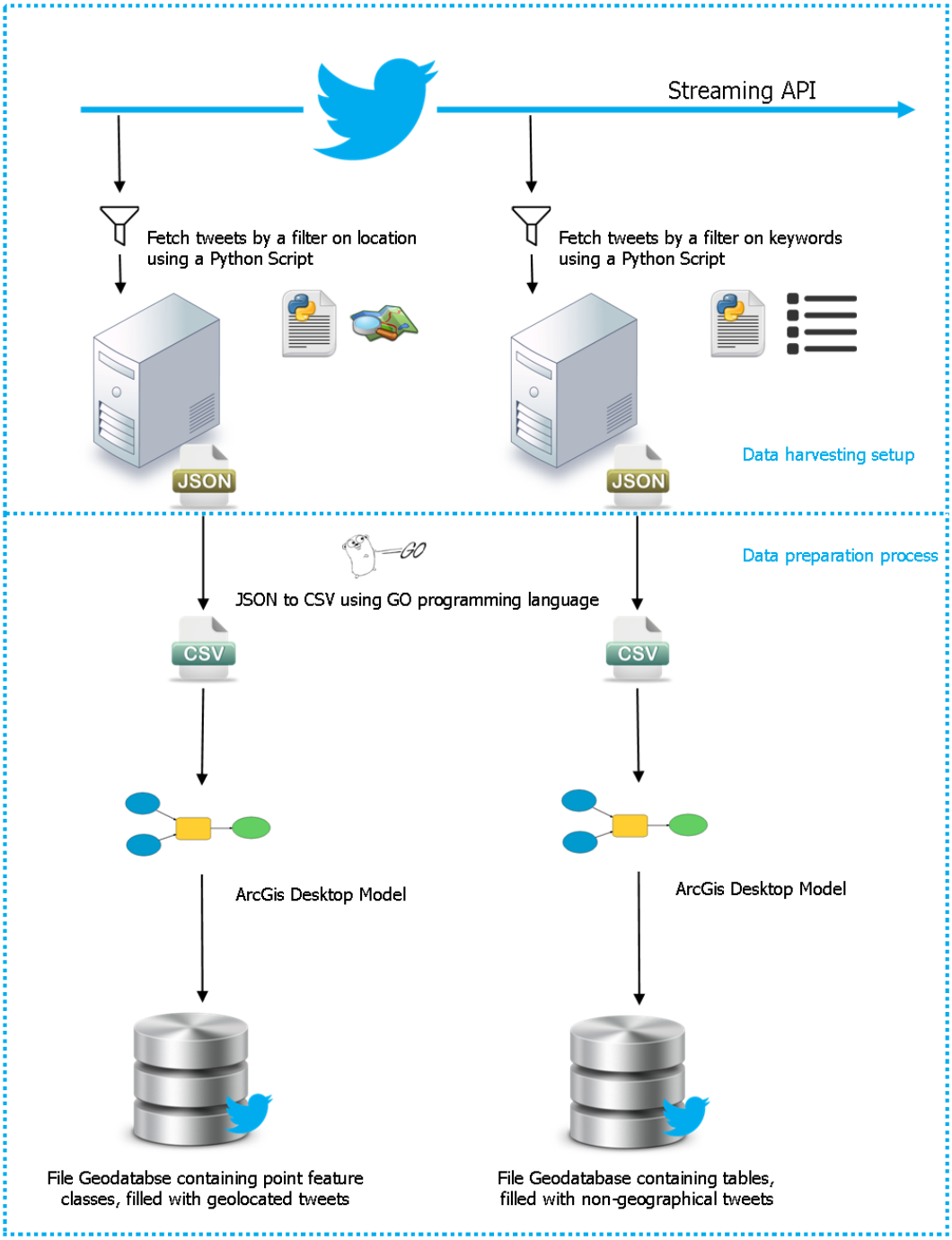


Figure 15 Schematic overview of data harvesting setup and data preparation process

Attribute field	Type	Description	Potential usefulness of the attribute field for analysis
created_at	String	UTC time when Tweet was created	Making data time aware
text	String	UTF-8 text of the status update	This attribute field can contain situational information
retweeted	Boolean	Indicates whether the tweet has been retweeted by the authenticating user.	This attribute field could be an indicator that a tweet does not contain new information (Schulz et al., 2012).
in_reply_to_user_id_str	String (*Nullable)*	If the tweet is a reply, this field will contain the integer representation of the original Tweet's author ID.	This attribute field is an indicator that a tweet is part of a conversation. This field could therefore indicate the usefulness of this tweet in particular situations. It could for instance be likely that Twitter users do not communicate in dialogues during incident situations.
in_reply_to_status_id_str	String (*Nullable)*	If the tweet is a reply, this field will contain the string representation of the original Tweet's ID.	This attribute field is an indicator that a tweet is part of a conversation. If it would for some reason be valuable to track conversations, then it would be possible to link different tweets of a conversation to each other using this attribute field.
source	String	Utility used to post the Tweet, as an HTML-formatted string. Tweets from the Twitter website can be identified by the value "web".	It could be useful to know whether a tweet is sent from a desktop computer or a smartphone for example.
place_full_name	String	Full human-readable representation of the place's name. Tweets associated with places are not necessarily issued from that location but could also potentially be about that location.	This attribute field could be used to determine the spatial context of a tweet. It may be an alternative to geotag tweets that do not contain a coordinate pair.
place_place_type	String	The type of location represented by this place.	This attribute field could be an indication of the usable scale of the place field.
lang	String (*Nullable)*	When present, indicates a BCP47 language identifier corresponding to the machine-detected language of the tweet, or "und" if no language could be detected.	The language of a tweet may be used as a noise filter, if you are for example only interested in Dutch tweets.
id_str	String	A unique identifier for a tweet.	Could be used to link tweets within conversations, or save lists of interesting tweets.
user_id_str	String	A unique identifier for the user that sent a tweet.	Could be used to identify users and to link users to conversations. It could as well be used as noise filter when blocking tweets of a particular user.
user_name	String	The name of the user, as they've defined it. Not necessarily a person's name. Typically capped at 20 characters, but subject to change.	This field is better human readable as the user_id_str field. But, because the field is subject to change and not unique this field should never be used to identify a particular user.
user_location	String (*Nullable)*	The user-defined location for this account's profile. Not necessarily a location nor parseable.	Altough this field does not represent a location with high accuracy of certainty, it may be used to identify the geography of a tweet when no other options are available to geolocate this tweet.
user_followers_count	Int	The number of followers that the user has a t the moment of sending the tweet.	In different studies this field is used as an indicator of reliability of the tweet's information, assuming that someone with many followers will face more social control on their tweets than followers with few followers (Morris et al., 2012).
user_statuses_count	Int	The number of tweets (including retweets) issued by the user.	This field could as well be used as an indicator of reliability for the tweets. It could also be used to filter out tweets of 'institutions' like news channels which send many tweets but may not be useful for certain purposes.
user_utc_offset	Int *(Nullable)*	The offset from GMT/UTC in seconds.	Because this field defines the time zone, it is used as a geographic indicator of tweets in different studies (Krishnamurthy et al., 2008).
coordinates_coordinates	Collection of Float	The longitude and latitude of the tweet's location as a collection in the form of [longitude, latitude].	This field provides very accurate information about the tweet's location. It is therefore very useful for geographic analysis in this thesis.

Table 4 Overview of all attribute fields that are extracted from the streamed JSON tweets during the data preparation process. For each attribute field the data type, a description and a motivation for potential usefulness of the specified attribute field are given. Information is based on (Twitter, 2012a; Twitter, 2013f; Twitter, 2013g).

4.3.Study area

During the data harvesting period relatively large amounts of data were collected. Because it is assumed that it would be too computational heavy to process all data for the whole of the harvesting time and spatial extent, it is decided to focus on a smaller area and a shorter time extent for which analyses will be performed and the research questions will be answered. In order to increase chances on positive results from the analyses, the study area should be defined with caution to both the collected data as well as the road and incident characteristics in this area. In the current section a motivation will be given for the choice of the study area that is used in this thesis.

In order to define a suitable study area, a few criteria were defined on forehand. These criteria were defined regarding the chance to find as many as possible incident-related tweets. The criteria were defined as:

- The study area should contain a high density of state highways and secondary roads
- The study area should contain a high density of geotagged Twitter data
- Many incidents should occur in the study area on a daily basis

The criteria stated above are investigated in a few short pre-analyses. The following sections will describe the results of these pre-analyses, and the final motivation for the choice of the study area.

Spatial distribution of state highways and secondary roads in the Netherlands

The first criterion that should be met by the study area, is that it contains a high density of state highways and secondary roads. It is only on state highways and secondary roads that incidents are registered by Rijkswaterstaat in a standardized and structured way. Rijkswaterstaat made incident loggings available for the research purposes of this thesis.

An essential part of this research will be the validation of incident-related information from tweets using incident loggings from Rijkswaterstaat. It would therefore be advantageous to define a study area that has a high density of state highways and secondary roads for which incident loggings are abound. In order to investigate Dutch areas with high densities of state highways and secondary roads, a line density analysis was performed based on the Dutch National Road File (in Dutch: Nationaal Wegenbestand - NWB). This spatial line density analysis resulted in a 1 by 1 kilometer grid map, representing for each grid cell the density of state highways and roads that were found within a search area of 10 square kilometer from each grid cell (Figure 16).

Overview of spatial distribution of Twitter data in the Netherlands

The second criterion that should be met by the study area is that a high density of geotagged Twitter data should be available in the study area. In order to increase chances on finding incident related tweets, the number of tweets available in the study area should be as high as possible. The exact number of tweets that are sent within a particular area cannot be calculated because approximately 1% of all tweets have coordinates available for geotagging. Nevertheless, all geotagged tweets available in a particular area could be seen as a representative sample for the Twitter activity. It is assumed that the higher the Twitter activity is in a region, the higher the chances are on finding relevant tweets there.

In order to get some insight into the Twitter activity in the Netherlands, a spatial point density was performed based on all collected geotagged tweets. First, all available tweets were projected based on their coordinate values. Thereafter, a point density analysis was performed resulting in a 1 by 1 kilometer grid map, representing for each grid cell the density of points that was found in a search area of 10 square kilometers around each grid cell (Figure 17).

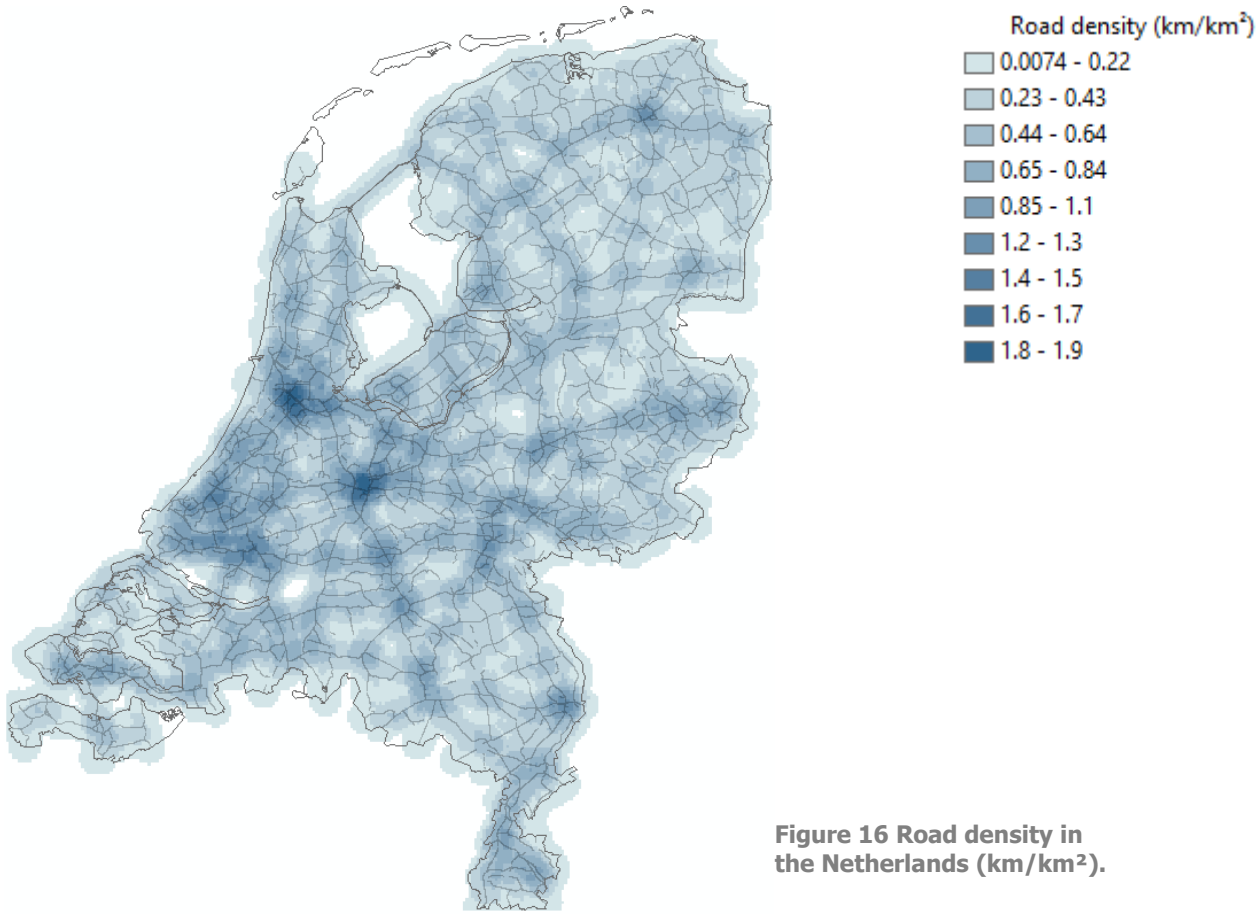


Figure 16 Road density in the Netherlands (km/km²).

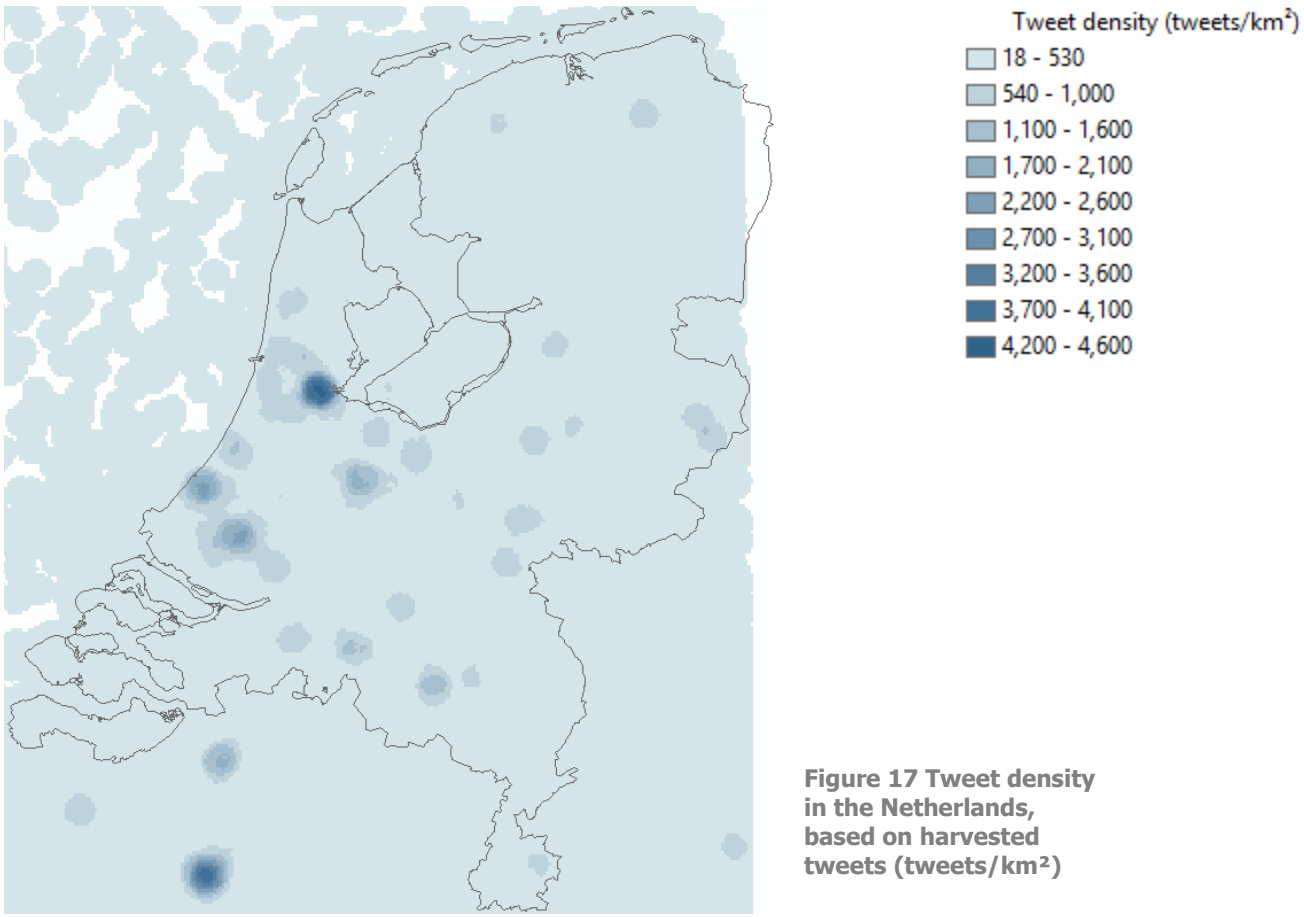


Figure 17 Tweet density in the Netherlands, based on harvested tweets (tweets/km²)

Spatial distribution of incidents in the Netherlands

A final criterion that should be met by the study area is the number of incidents that happen in the study area. It is assumed that in an area with a high density of incident occurrences, a higher chance exist that incident-related tweets will be sent from that area.

Because especially incidents on state highways and secondary roads are studied in this thesis, the study area should have a high density of incidents that occurred on these roads. A database from Rijkswaterstaat containing registrations of incidents occurrences in the Netherlands was used to get insight into the spatial distribution of incidents in the Netherlands. The database only contained incidents in which road users were injured, hence no registrations were available of incidents that only led to material damage.

First, a selection was made on the incidents based on the speed limit of the road on which the incidents occurred. Incidents that happened on roads with speed limits from 80 to 130 kilometers per hour were selected, as these incidents happened on state highways or secondary roads. A density raster was created from the selection of incident points (Figure 18).

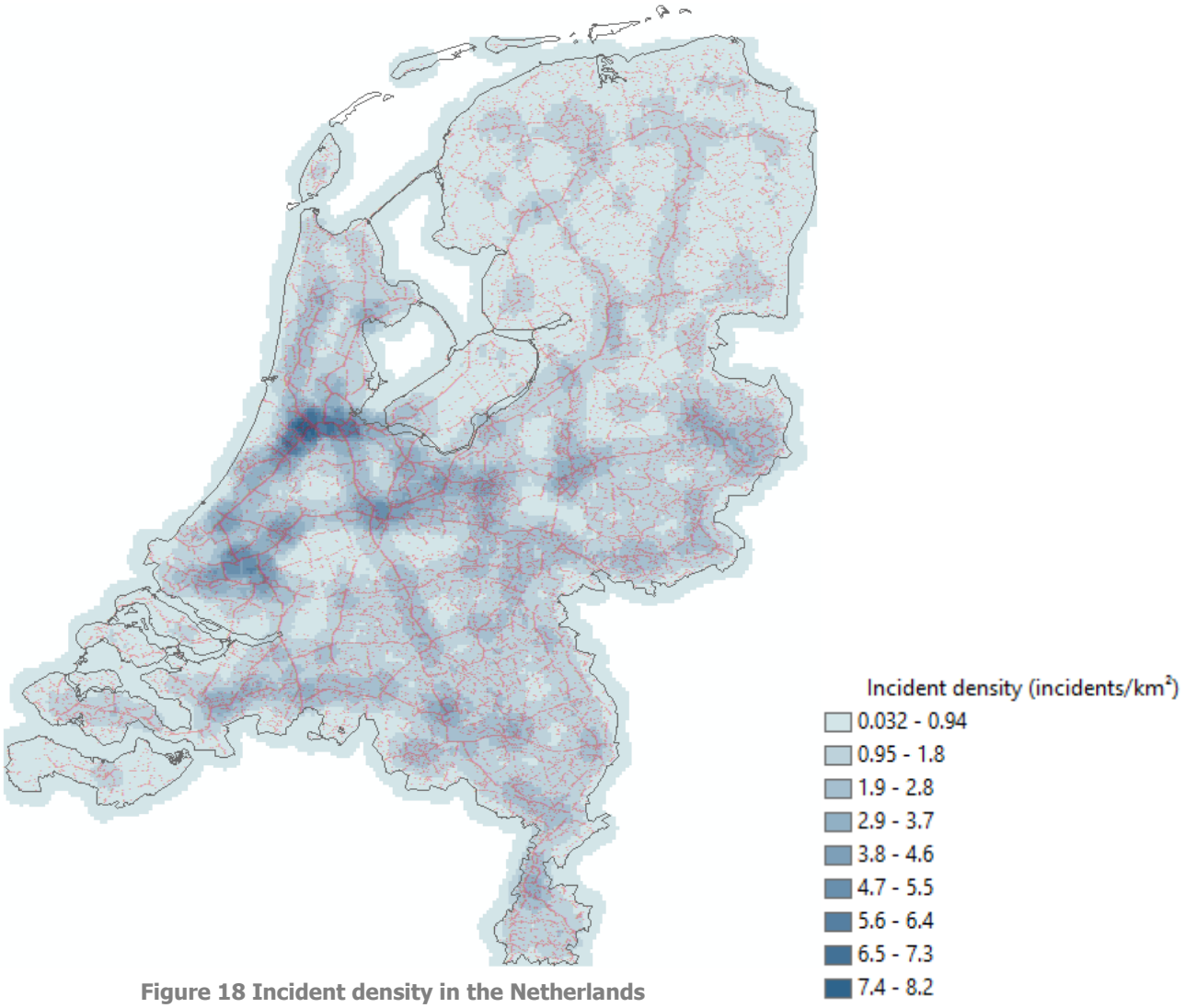


Figure 18 Incident density in the Netherlands (RWS, 2003-2011)

Study area definition and motivation

From the three different density analyses, it should be concluded that the area around Amsterdam is the most suitable study area for the research purposes of this thesis. Figure 16, Figure 17 and Figure 18 show that the highest density of state highways and secondary roads, geotagged tweets and incidents can be found in the area around Amsterdam.

Figure 19 shows the exact positioning of the defined study area. For this area Twitter data will be extracted and processed, as well as incident loggings will be gathered.

Figure 19 Geographic extent of the study area

4.4.Prequalification of temporal coverage of harvested Twitter data

Twitter data was collected in the year 2013 from the beginning of July until the end of December. Due to technical issues like connection failures, Twitter API issues or system hangs, the Twitter data could not be harvested for various smaller periods of time. It is important to avoid these time-out periods when performing analyses, in order to decrease the chances on biased results.

Before analyses took place on the Twitter data, a prequalification of the Twitter data was executed first. The main goal of this prequalification was to check the Twitter databases on missing data for any period of time, or in other words, finding 'gaps' in the data. A prequalification of the Twitter data was executed by creating different types of time trend curves, representing the amounts of data that were collected over time.

Temporal trends have been plotted for the following datasets:

- The total number of georeferenced tweets
- The total number of tweets that were filtered on traffic-related keywords

From the temporal trends in Figure 20 gaps in the data that last one or more days can be spotted easily by a steep decline in tweet count. Less visible in these curves are gaps of less than one day. Smaller dips in the curves can be observed, however these dips do not necessarily have to indicate time-outs in the harvesting process.

In order to get a more thorough insight into smaller gaps, all tweets were classified based on date and hour. All tweets that belong to the same hour of a particular date were grouped together. For example, all tweets that were sent on July 24, between 06:00 and 06:59:59 were grouped and classified as "2013072406". After classifying all tweets in groups of hours, it was checked for all dates if any classes were missing. If a date missed an hour-group, then a time-out of more than one hour took place on this date.

To avoid studying data that contain gaps of more than an hour, a table was created of all days for which the data contained gaps of more than one hour (Table 5). Hereafter in this thesis, this table is used to determine suitable time periods for which the harvested Twitter data can be analyzed.

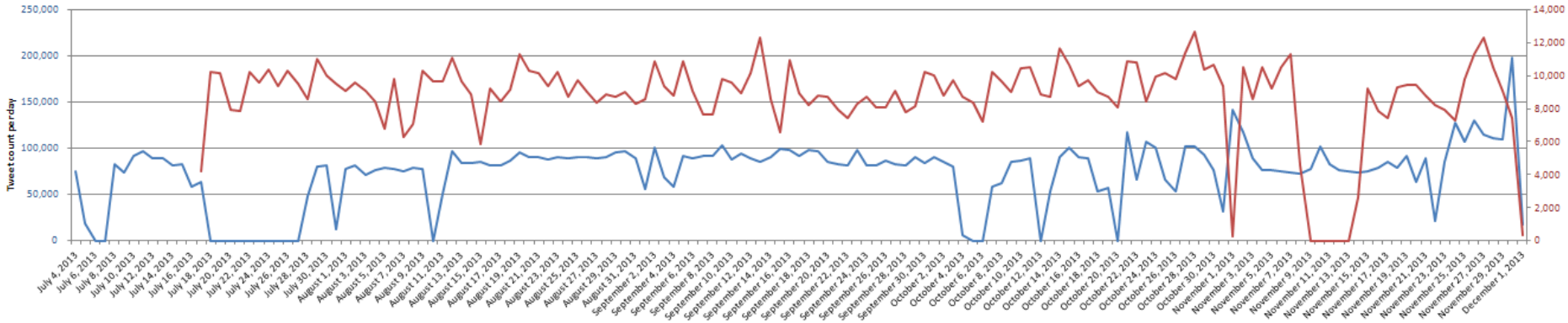


Figure 20 Temporal trends of harvested tweets. The red line represents the number of georeferenced tweets per day that are harvested. The blue line represents the number of non-georeferenced tweets per day that are harvested.

	July																												
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
A																															
B																															
	August																														
A																															
B																															
	September																														
A																															
B																															
	October																														
A																															
B																															
	November																														
A																															
B																															

Table 5 Overview of days for which data is harvested with or without time-outs. Green represents days for which no time-outs took place during harvesting. Red represents days for which time-outs did took place. 'A' represents georeferenced tweets, 'B' represents non georeferenced tweets.

4.5.Finding relations between datasets using correlation statistics

As has been discussed earlier, one of the main goals described in this methodology is to identify potentially valuable tweets. If tweets give information about an incident, or in other words if they increase the situational awareness, a tweet could be potentially valuable. Before the information that a tweet contains can be used however, it should be clear that a tweet is related to an incident or other event on the road. A very basic criterion that defines a potential relation between a tweet and an incident is the co-occurrence of an incident and a tweet in each other's proximity, both in space and time. Hence, it is expected that tweets that tell something about an incident's situation, are sent near the incident location and within a relatively short time after the incident happened.

In a first attempt to get some insight into the geographic Twitter data, a generic relation between the geographic Twitter data and the incident loggings was investigated. In upcoming sections, it is tried to identify incident-related events from the geographic Twitter data happening on or near the highways in the study area. This identification could bring more positive results, which means that a lot of incident-related events could be identified from the data if a strong relation would exist between the incidents happening on the roads and patterns in the Twitter data around these incident locations and times.

A first exploratory attempt to identify a relation between the geographic Twitter data and the incident management data was done using correlation calculations. A first assumption is made that a stronger relationship between incident loggings and Twitter data could be found if only those tweets are taken into account which were sent either from or very close to the highway. The closer tweets' locations are from a highway the more likely these tweets are sent by traffic participants. Moreover, it is assumed that tweets sent by traffic participants are more likely to have a relation with the traffic situation and thus with incident data. It is therefore, in order to increase chance on finding a relationship between twitter data and incident data, that only tweets are taken into account that were sent within a short distance from highways or secondary roads.

A hypothesis is formulated assuming that if tweets are sent as a reaction to incident events on the road, the Twitter data will contain more tweets at times directly after an incident has happened. In this view, two variables that could be related to each other are the time distance from an incident happening and the frequency of tweets that are sent near roads. More specifically, it could be expected that more tweets are sent at points in time closer to an incident happening. If this would be the case, then a negative correlation would exist between the variables tweet frequency and passed time (in minutes) after an incident occurrence.

In order to test this hypothesis, correlation calculations were made for different sets of Twitter data. As a reference measure of the variable of time after an incident happening, incident loggings from Rijkswaterstaat were used. A table was created that listed for every single minute in the test period the frequency of tweets and whether or not an incident happened in that minute. In a third column each single minute of the test period was translated to a relative time variable. This relative time variable, the time distance from a past incident, stands for the number of minutes that passed after the most recent incident took place. The table attributes 'time distance' and 'tweet frequency' that are listed for each single minute could be used as linear variables to do a correlation estimation.

In the form of a small sensitivity analysis, various different correlation estimations were run with different selections of geographic tweets regarding their distance from roads (see Figure 21). This sensitivity analysis was set up because it was questionable in which correlation estimation setup chances were highest on finding a relation between Twitter data and the incident data. It was taken into consideration that a selection of tweets very close to roads would increase the chance that these tweets were sent by traffic participants as a reaction on the traffic conditions, but on the other hand would deliver a small selection of tweets and the chance that relevant but 'misplaced' tweets were overlooked. Taking a broader selection of tweets into account that is selected based on a greater distance from the road could decrease the chance on overlooking

relevant tweets, however increases the chance on distorting the correlation because more non-incident-related tweets sent by non-traffic participants, for example citizens of Amsterdam living close to the highway, are taken into account in the correlation calculations.

The following selections of geographic tweets were used in the sensitivity analysis:

- All tweets either on or within 10 meters from highways in the study area
- All tweets either on or within 50 meters from highways in the study area
- All tweets either on or within 100 meters from highways in the study area

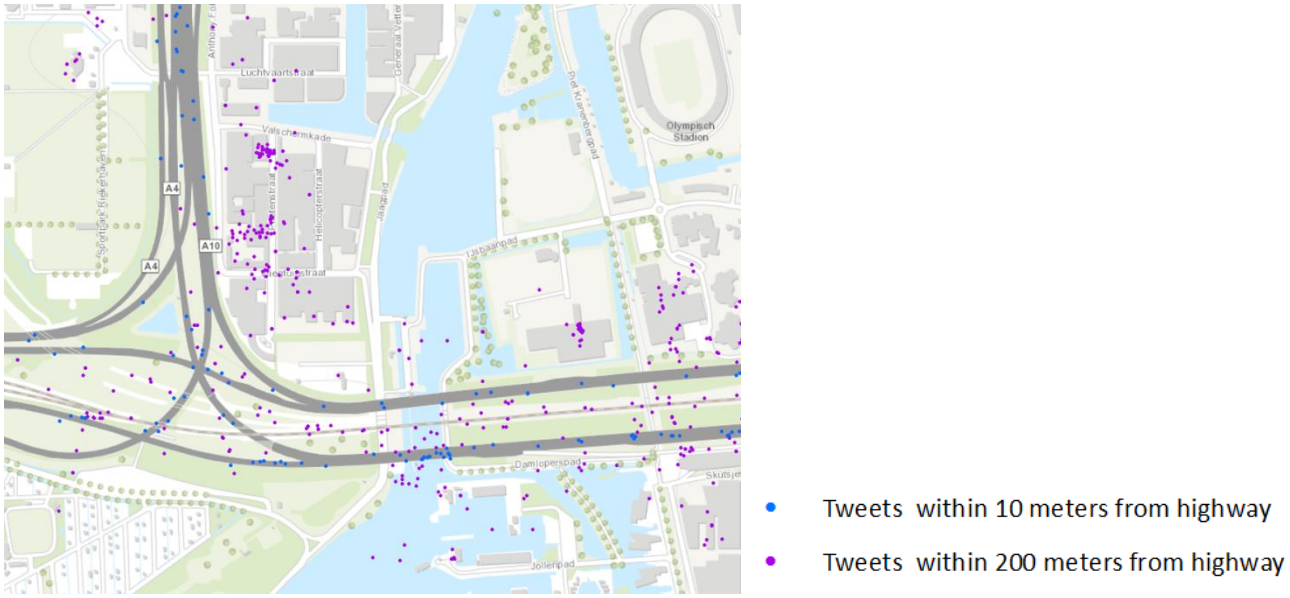


Figure 21 Example selection of tweets within different buffer distances from highways

Next to the varying selections of tweets, a second variable is altered for each run in the sensitivity analysis. Four different calculations of time distance from an incident happening were applied and compared in the sensitivity analysis. It was taken into consideration that the timestamps in the incident data always represent the moment in time when an incident is registered into the database and do not represent the actual time an incident took place. It is therefore expected that the incidents took place a few minutes earlier than the registered timestamps in the incident data. For this reason, four different types of time distance were calculated and applied in the sensitivity analysis:

v1: The time distance measured from the last occurred incident timestamp, registered in the incident loggings

v2: The time distance measured from the last occurred incident timestamp, registered in the incident loggings, minus 5 minutes delay between an incident happening and its registration.

V3: The time distance measured from the last occurred incident timestamp, registered in the incident loggings, minus 10 minutes delay between an incident happening and registration.

V4: The time distance measured from the nearest incident happening (both passed and upcoming incidents in time).

A calculation example of the different types of distances (v1 – v4) can be seen in Table 6.

Finally, the sensitivity analysis was carried out by running different correlation analyses on the variables of tweets frequency and time distance. Table 7 shows an overview of the different correlation analyses setups and outcome. In all cases, Pearson's correlations (Field, 2009) were calculated using SPSS software. A significant correlation between tweet frequency and distance from an incident happening could not be found for any of the different situations of the sensitivity analysis.

The absence of a relation between the tested datasets does not necessarily mean that the Twitter data does not contain incident-related Tweets. On the other hand it is probably a first indication that the identification of incident-related tweets asks for an extensive approach.

date	time	tweets	incident	distance (v1)	distance (v2)	distance (v3)	distance (v4)
16-8-2013	14:56:00	1	yes	0	5	3	0
16-8-2013	14:57:00	0		1	6	4	1
16-8-2013	14:58:00	0		2	0	5	2
16-8-2013	14:59:00	0		3	1	6	3
16-8-2013	15:00:00	0		4	2	7	4
16-8-2013	15:01:00	2		5	3	8	5
16-8-2013	15:02:00	3		6	4	9	6
16-8-2013	15:03:00	0	yes	0	5	10	0
16-8-2013	15:04:00	0		1	6	11	1
16-8-2013	15:05:00	0		2	7	12	2
16-8-2013	15:06:00	0		3	8	0	3
16-8-2013	15:07:00	3		4	9	1	4
16-8-2013	15:08:00	0		5	10	2	5
16-8-2013	15:09:00	0		6	11	3	6
16-8-2013	15:10:00	0		7	12	4	6
16-8-2013	15:11:00	0		8	0	5	5
16-8-2013	15:12:00	1		9	1	6	4
16-8-2013	15:13:00	1		10	2	7	3
16-8-2013	15:14:00	0		11	3	8	2
16-8-2013	15:15:00	0		12	4	9	1
16-8-2013	15:16:00	2	yes	0	5	10	0
16-8-2013	15:17:00	0		1	6	11	1
16-8-2013	15:18:00	0		2	7	12	2
16-8-2013	15:19:00	0		3	8	13	3
16-8-2013	15:20:00	0		4	9	14	4
16-8-2013	15:21:00	1		5	10	15	5
16-8-2013	15:22:00	0		6	11	16	6
16-8-2013	15:23:00	0		7	12	17	7
16-8-2013	15:24:00	5		8	13	18	8
16-8-2013	15:25:00	0		9	14	19	9
16-8-2013	15:26:00	0		10	15	20	10
16-8-2013	15:27:00	1		11	16	21	11

Table 6 Example of different calculations of time distances (v1 – v4) as input variables in the sensitivity analysis

	Tweets within 10m	Tweets within 50m	Tweets within 100m
Distance v1	0.000 (0.991)	0.001 (0.947)	0.009 (0.384)
Distance v2	-0.002 (0.929)	-0.002 (0,903)	0.009 (0.384)
Distance v3	0.004 (0.844)	-0.002 (0.890)	0.009 (0.395)
Distance v4	0.011 (0.591)	-0.002 (0.869)	0.008 (0.407)

Table 7 Results of the sensitivity analysis: Pearson's correlation estimations and their significant values between brackets.

4.6.Geographic irregularity pre-analysis of tweet intensity patterns

4.6.1. Theoretical background of irregularity analysis

The previous chapter section explained that no clear relation could be found between the incident loggings and the pattern of Tweet frequency that was sent from or close to highways. Although this relation could not be found, certainty cannot be given that the Twitter database does not contain tweets that hold relevant information for incident management. In this section the identification of relevant tweets will be extended by a different approach: geographic irregularity analysis.

Geographic or "zonal" irregularity analysis is a detection method to find geo-social events, which is successfully applied in different studies by Lee et al (2011; 2013). , Wakamiya et al. (2013) and Fujisaka (2010). In order to detect unusual Twitter traffic in certain predefined geographic zones, the tweet intensity during specific periods of time is compared with the statistical regular Twitter pattern that can be found for these zones. If for a specific zone and time period the tweet intensity differs significantly from earlier found patterns, then this zone could be identified as being either unusually crowded, or an event takes place in this zone which induces many reactions on Twitter. In the above named studies, zonal irregularity is successfully applied to discover various big geo-social events in Japan. Both expected as well as unexpected events could be detected using the zonal irregularity analysis.

In literature the zonal irregularity analysis of tweet intensity patterns is only applied to small scaled datasets and study areas. For the analysis of Lee et al. (2011), geotagged tweets are used which were sent during a period of one and a half month and within the geographic extent of Japan. Moreover, the irregularity analysis of Lee et al. (2011) was only applied to identify crowded places. The goal of this thesis is not to identify crowded places but to identify incident-related events. On the other hand, crowded places and smaller scale events like road incidents could be discovered by a similar reaction in the tweet-intensity pattern: both will cause an unusual increase of tweets sent around a location and within a short period of time. It is therefore that the zonal irregularity analysis could be a solution for reaching the research objectives.

In regard of this thesis' objectives, it is interesting to test the zonal irregularity method for the identification of smaller scaled datasets and study areas. Not only would this bring more insight into the applicability of the zonal irregularity method, but it could also be very supportive in reaching the objectives of this thesis to identify incident-related tweets. It is for these reasons that a test case is set up to apply the zonal irregularity method on the available Twitter data in order to identify traffic- and incident-related tweets in the study area. Another important reason to use zonal irregularity analyses for reaching the research objectives is the geographic focus of this method. The zonal irregularity method doesn't require language processing of the tweet texts to detect events, which is a complex task and a field of study outside the scope of this thesis. Event detection will be based mainly on the geographical component of the available tweets.

In order to test the applicability of the zonal irregularity method for achieving the thesis' objectives in a structured way, the test case is set up in the form of a sensitivity analysis. A sensitivity analysis enables testing the zonal irregularity method in different setups. For the purpose of identification of traffic and incident-related tweets, it is favorable to apply the zonal irregularity analysis in different setups. Many uncertainties exist about the way regular tweet intensity patterns and events should be defined in order to detect traffic and incident-related events. It is therefore that a trial-and-error method like a sensitivity analysis could give more insight into these uncertainties by testing different analysis setups.

4.6.2. Execution of the geographic irregularity analysis

A zonal irregularity analysis, as it is applied by Lee et al. (2011), consists of different steps that need to be taken:

1 *Geographically located tweets should be collected.* This can best be done by developing a tweet gathering system, which is developed for this thesis as well. The Twitter data should be made ready for use in a GIS (Figure 22).

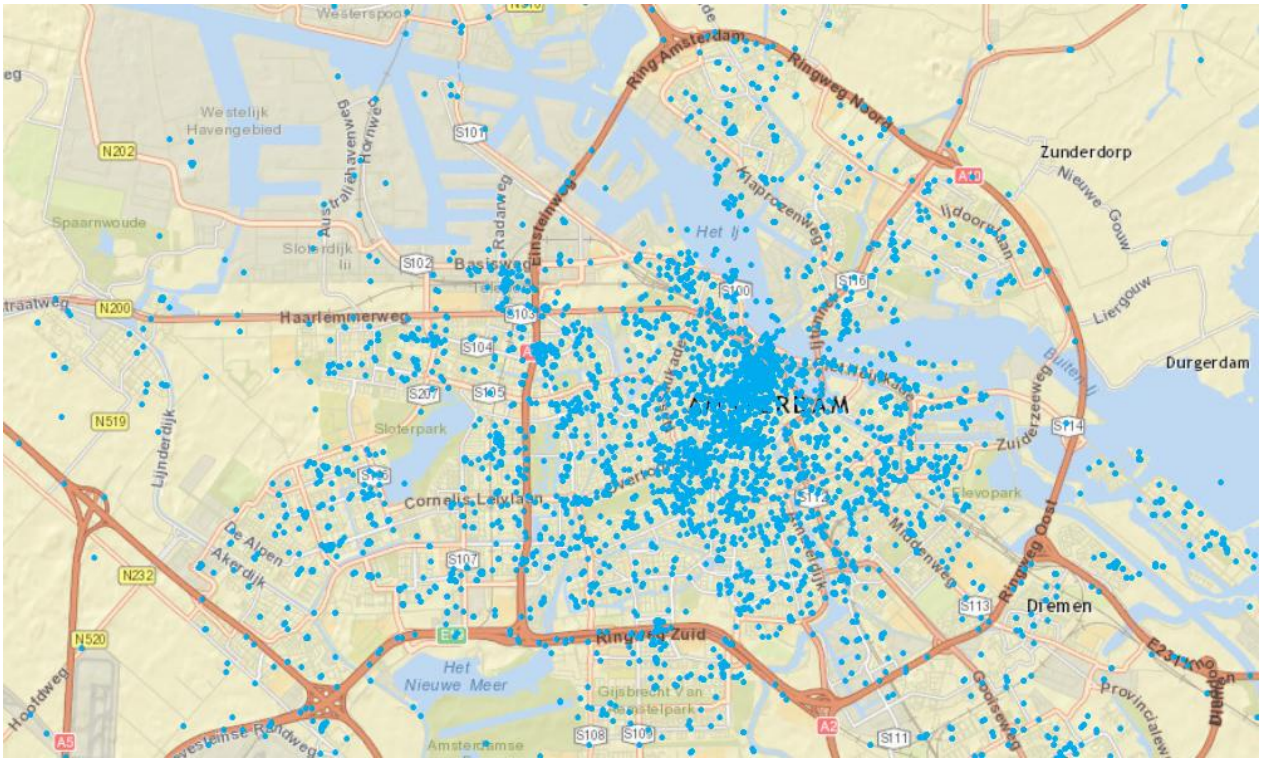


Figure 22 Example of geographically located tweets in the study area, visualized in a GIS

2 *Establishing zones for which a statistical regular tweet intensity pattern will be estimated.* There are different ways in which zones could be defined. Zones could be defined by administrational borders. For the study area the different districts of Amsterdam could be used for example (Figure 23). Another option is to use equally sized zones, for example the cells of a raster draped over the study area (Figure 24). In the study of Lee et al. (2011), a K-means clustering method was used to divide their study area into zones based on the geographical occurrences of their dataset. The centers of the clusters were used to define a Voronoi diagram that divided their study area into "socio-geographic boundaries".

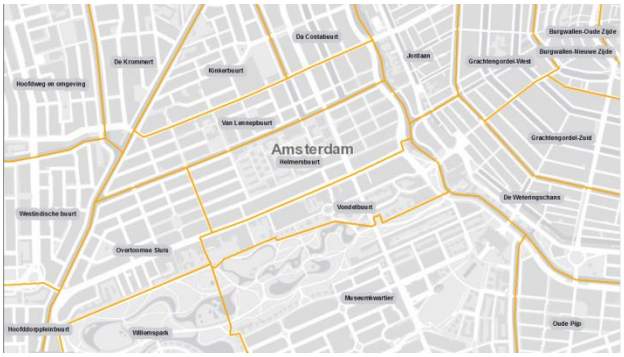


Figure 23 Zones defined by district borders

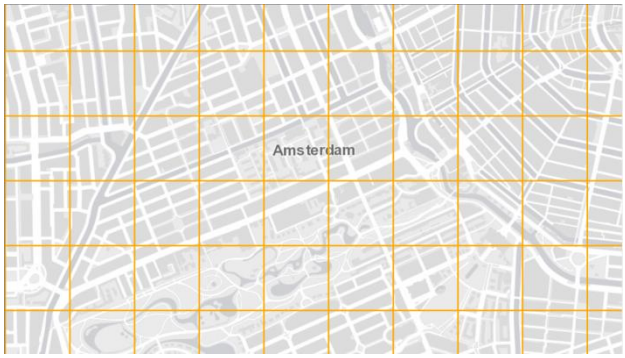


Figure 24 Zones defined by a raster

3 *Estimating statistical regularities for defined zones.* From a long-term harvested Twitter dataset the regular pattern of tweet intensity should be estimated for each zone in a specific time period, for example days or hours of the day. This regular pattern or geographic regularity will be used as threshold value to identify unexpected high tweet occurrences. A boxplot can be used to deal with the indicators of a high tweet intensity in a simple statistical way (Lee et al., 2011).

A boxplot (see Figure 25) gives insight into the distributions of tweet occurrences by representing a couple of sample statistics (Field, 2009):

- the lowest and highest score
- the bottom quartile: this is the range between which the lowest 25% of scores fall
- the median which is the middle score
- the top quartile: this is the range between which the highest 25% of scores fall.

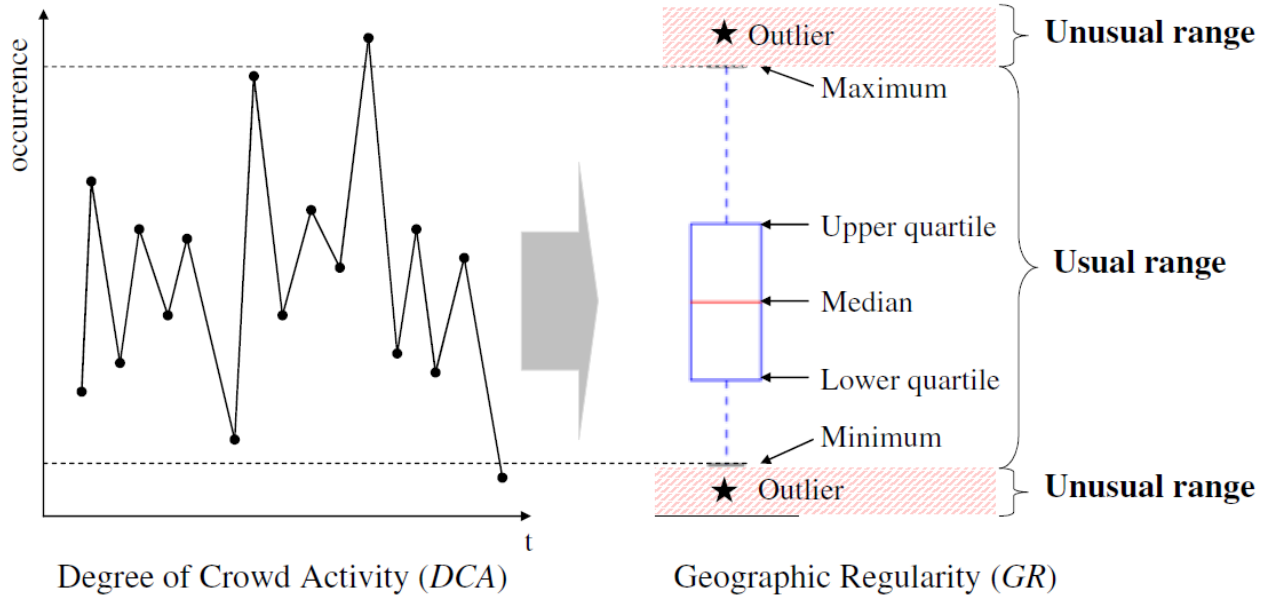


Figure 25 Boxplot-based geographical regularity construction (Lee et al., 2011)

The boxplot can be used to establish geographic regularities for all zones. A geographic regularity pattern is defined by a usual range and an unusual range of tweet intensity. In the study of Lee et al. (2011), the limits between usual and unusual range are defined by the lowest and highest occurrence scores. The definition of an outlier is not consistent in literature (Hodge & Austin, 2004), and therefore determining what data samples are outliers is a subjective task.

The definition of the usual range and the unusual range is very important for the results of an event detection analysis. If the unusual range is defined too narrow, there is chance that relevant events cannot be detected. On the other hand, if the unusual range is set too broad, then there is a chance that too many false positive events are detected. Because there is no uniform definition of how to define outliers, and therefore how to define the usual and unusual range, the definition of outlier is a variable in the analysis that is very suitable as input variable in a sensitivity analysis. This will be more elaborated on in the sensitivity analysis set-up.

4 *Detect irregularities for time series of tweets in established zones.* When the tweet-intensity of a certain zone for a specific period of time is higher than the usual range of the geographic regularity, then this could be defined as a unexpected high tweet intensity which could indicate the occurrence of an event.

4.6.3. Irregularity analysis applied on a broad scale to identify massive events

For the study area around the city of Amsterdam, an explorative zonal irregularity analysis was executed in order to check the functioning of the irregularity method for the available data that was collected earlier in this thesis. This explorative zonal irregularity analysis is a first check on the potential of the irregularity analysis to identify event-related tweets in the Twitter data.

Because a relatively low percentage of all tweets is geographically located, it is expected that bigger events which provoke many tweets to be sent, are easier to detect than smaller events which provoke only a few tweets. It is therefore that first an explorative geographic irregularity analysis is performed on the data in order to be able to discuss the potential of the geographic irregularity analysis. When event occurrences can be identified clearly from the Twitter data, a more extensive event detection can be performed on a larger scale in the form of a sensitivity analysis. This sensitivity analysis could result is the identification of smaller events, and hopefully incident-related events as well.

In the first explorative zonal analysis, the variance in tweet intensity per day for the whole study area is examined. Tweet intensity trends were plotted for all the available geographical Twitter data in the study area (Figure 19). Next to plotting tweet frequency trends for the study area, tweet frequency trends for all the available data from the complete data harvest area (Figure 14) were plotted. It is expected that if some events that can be detected from Twitter data are bound to a geographic location, differences can be found in the pattern of geographic tweet frequencies for different areas. For example, it is expected that some events are bound to the study area of Amsterdam, meaning that these events cannot be detected in the frequency pattern of other areas like the whole harvest area.

In order to validate this hypothesis, the geographic tweet frequency trend of the study area is compared to the tweet frequency trend of all available data from the harvest area. Outliers in both frequency trends should be identified. If outliers can be found in the frequency trends of the study area, and no outliers can be found in the data of the harvest area for the same moment in time, then this would be an indication that an event is happening that is only related to the study area of Amsterdam and not to the complete harvest area.

All geotagged tweets for both areas are first grouped per day. Then, for each day the frequency of tweets is counted and plotted into a frequency trend. In order to define outliers in the frequency pattern, the mean frequency per day and standard deviation from this mean was calculated. Outliers were defined as days that had a higher frequency than the standard deviation from the mean frequency per day for all data in a zone. The frequency trends for both areas can be found in Appendix 4.

From both frequency trends, outliers could be identified. Where possible, the probable causing event for these outliers was identified from the tweet texts:

Outliers in tweet frequency trends for the study area of Amsterdam:

Day	**Probable event responsible for outlier**
22th August	No event identified from tweets' texts.
25th August	No event identified from tweets' texts.
28th October	Heavy storm over the Netherlands
10th November	MTV European Music Awards (held in Amsterdam)
26th November	Soccer game: Ajax – Barcelona (held in Amsterdam)

Outliers in tweet frequency trends for the whole harvest area:

Day	**Probable event(s) responsible for outlier**
22th October	No event identified from tweets' texts.
27th October	No event identified from tweets' texts.
28th October	Heavy storm.
24th November	No event identified from tweets' texts.
26th November	Various Champions League soccer games.
27th November	No event identified from tweets' texts.
28th November	No event identified from tweets' texts.
30th November	No event identified from tweets' texts.

Remarkable in the outliers that were found is that the peaks in the frequency trends of the harvest area could not be found in the frequency trends of the study area of Amsterdam except for two dates: 28th October and 26th November. These 2 dates were defined as outliers in the frequency trends for the study area of Amsterdam as well as for the harvest area. It may not be coincidental that the events that were identified as being the probable event responsible for the outliers are events with a large geographical impact. For example the heavy storm of 28th October is an event that didn't only have impact on Amsterdam, but had impact over all of the harvest area as well.

Secondly, an outlier could be found in the study area of Amsterdam on the 26th of November because of an important Champions League soccer game that took place in Amsterdam: Ajax vs. Barcelona. Although this event took place in the study area of Amsterdam, the impact of this event could be found all over the harvest area as well. Most likely this event had impact on the harvest area as well because it is a popular subject among many soccer fans distributed over the harvest area.

Also interesting is the outlier that was found for the study area of Amsterdam on the 10th of November. A similar peak in the tweet frequency, much higher than the regular tweet intensity, was not found for the whole harvest area. This could be an indication that this particular event has more impact on the tweet frequency of the study area than on the total harvest area. The event that is found to be probably responsible for the outlier in the tweet frequency trend is the MTV European Music Awards event that was held in a concert hall in Amsterdam. Because many world famous artists were present at the MTV awards, this event attracted many visitors that were actively using Twitter to share their experiences. It is therefore not surprising that this event has a very local impact, and could only be detected in the frequency trend of the study area of Amsterdam. For the study area, at least 10.2% of all geotagged tweets mentioned the MTV awards. For the harvest area this percentage was less: at least 5.2%. Thus, this particular event had not a big enough impact on the total harvest area in order to be identifiable.

The first attempt to identify geographical events from geotagged tweets by using irregularity analysis, has shown that the method is suitable for finding local geo-socio events from the available Twitter data. Because zonal regularity was estimated for very broad areas, only relatively massive events that provoke hundreds of tweets could be identified. Interesting for further investigation is that events were found with a different geographical impact. Both events were found that could be identified from tweet frequency trends in both areas, as well as events that could only be found in one of the study area's frequency trends.

4.6.4. Explorative zonal irregularity analysis setup for the study area around Amsterdam

From the preceding section it became clear that massive events can be detected from geographic irregularity indications in the Twitter data. A next question is whether smaller events can be detected using geographic irregularity indicators as well. This is an important question for the thesis' objectives as it is assumed that traffic incident-related events have much smaller dimensions than the events that are found in section 4.6.3. In the irregularity analysis of the preceding section geographic irregularities were only identified by comparing the daily tweet frequencies for the complete study area. In order to detect smaller scaled events, geographic irregularities should be identified in smaller geographic zones or time periods. In an attempt to find smaller scaled events in the study area, geographic regularity patterns are estimated based on the daily tweet frequencies in smaller zones.

A zonal regularity analysis was performed, by taking the following steps in a GIS:

1. Tweets were grouped by day of the year. Hence, a dataset containing all geotagged tweets was split into multiple point datasets that contained all geotagged tweets of only one day.
2. Zones are defined by overlaying the study area with a grid, dividing the study area in equal grid cells of 5000 by 5000 meters. Each grid cell represents a zone for the zonal regularity analysis.
3. By spatially joining the point data in each dataset to the overlapping grid cells, the tweet frequency was calculated for each grid cell and each day of the year. The tweet frequencies were stored in a frequency table where each row represents a grid cell, and each column the day of the year.
4. Statistical regularities were calculated for each grid cell based on the clusters of tweets in these grid cells. The statistical regularities were expressed as the mean plus the standard deviation of the daily tweet frequencies.
5. For all frequency values of each tweet cluster in the frequency table, the deviation of this frequency value from the regularity value of the according grid cells was calculated.
6. High deviations from the regularity values were identified for specific tweet clusters. In case a tweet cluster showed a high deviation from the regularity value, it was checked whether the tweets in this cluster reveal the occurrence of an event or not.

An important observation was done when checking the clusters of tweets that were identified by the zonal regularity analysis. It was found that many of the identified high geographic irregularities could not be explained by the occurrence of an event, but could be explained as the result of a single Twitter user that was sending tweets very frequently. In this context it is explained by Walther and Kaisser (2013) that the number of tweets is not a good indicator for the occurrence of an event, as they often found series of tweets from the same person, issued at the same location. According to Walther and Kaisser, these so called monologues do not describe real-world events. In the Twitter data collected for the study area around Amsterdam, these found monologues seldom describe a real-world event as well. Moreover because monologues are very disturbing for the event detection process, they can be seen as noise in the data.

In order to improve performance of the geographic irregularity analysis, it is tried to avoid monologues in the regularity analyses. Monologues are removed from the data by deleting all tweets that are issued by the same person within the same day and grid cell, hence in the same tweet cluster. After monologues are deleted from the data, the geographic irregularity analysis was performed again. Clusters of tweets that showed deviations from the mean plus standard deviations of 150% or higher were checked whether they described real-world events or not.

As a result of the geographic irregularity analysis, 25 events were identified (Table 8). Hence, from the analysis explained in this section it becomes clear that irregularity method is working, and different events can be detected from the spatial Twitter data in the study area.

Date	Identified Event	Description
13-07-2013	A Day at the Park	Dance music festival
16-07-2013 / 17-07-2013	Meeting of Jongerenorganisatie Vrijheid en Democratie (JOVD)	Meeting of political party
24-08-2013	Mysterland and Dekmantel festival	Dance music festivals
5-09-2013	HISWA te Water	Boat show
27-09-2013	UNSEEN	Photography festival
01-10-2013	Ajax – Milan	UEFA Champions Leage soccer match
15-10-2013	The Case - Amsterdam	Marketing event
16-10-2013 /17-10-2013	Amsterdam Dance Event (ADE)	Dance music festival held at various venues in Amsterdam
28-10-2013	Games for Health Europe	Conference
28-10-2013	Stormy weather	During the storm event in Amsterdam, higher tweet activity could be found around train stations, for example train station Sloterdijk.
29-10-2013	Jay-Z concert	Concert in a large music venue
29-10-2013	Ajax – ASWH	Soccer match
6-11-2013	Flora Holland Trade Fair	Flower fair
6-11-2013	Ajax – Celtic	UEFA Champions Leage soccer match
6-11-2013	SAP TechEd	Technical conference
7-11-2013	The National concert	Concert in a large music venue
8-11-2013	Adobe Digital Marketing Journey	Technical conference
10-11-2013	MTV European Music Awards (EMA)	Music event in large music venue
19-11-2013	The Netherlands – Colombia	UEFA World Championship qualification soccer match
25-11-2013	DroidConNL	Technical conference
27-11-2013	Accountantsdag / Week van de Ondernemer	Accounting conference
28-11-2013	FD Gazellenuitreiking	Journalistic prize-giving ceremony
29-11-2013	Kodaline concert	Concert
29-11-2013	De Nieuwe Wibaut-day	Meeting of residents
29-11-2013	DWDD University	TV broadcast

Table 8 Events identified in the area of Amsterdam using the geographic regularity analysis

4.7.Geographic regularity analysis around highways in the study area

4.7.1. Motivation for sensitivity analysis

The preceding section described that the geographic regularity analysis is a successful method to identify large geo-socio events from the available geotagged Twitter data around Amsterdam. It became clear that tweets that are sent near an event can be identified solely based on their spatio-temporal component. An important question remains unanswered however, which is whether this method is suitable for identification of small-scaled events as well. With respect to the research objectives of this thesis, it is necessary to identify tweets that are related to incidents or other events on or near roads. The dimensions of these events are most probably much smaller than the events that are detected in section 4.6.4, however many uncertainties exist about the impact of road incidents on Twitter data. It is unclear whether road incidents and other road events provoke clusters of tweets, and secondly, what the dimensions of these clusters are in space and time.

In order to investigate the data on the presence of tweet clusters that are provoked by traffic-related events, geographic regularity analysis is applied to identify tweet clusters near roads. Because uncertainty exists about the spatio-temporal and quantitative dimensions of any possible traffic-related tweet cluster the geographic regularity analysis should be run various times with different variables. In order to compare the outputs of the different runs in a structured way, a sensitivity analysis is set up.

4.7.2. Sensitivity analysis workflow design

A sensitivity analysis is often used in GIS-based research to validate a model's outputs and investigate the impact of the different input variables on the model's output. In Saltelli (2008) a possible definition of sensitivity analysis is given: "The study of how uncertainty in the output of a model (numerical or otherwise) can be apportioned to different sources of uncertainty in the model input." The sensitivity analysis that will be carried out in this thesis, has the purpose to increase the chance on finding traffic-related tweet clusters of any possible spatial, temporal or quantitative dimension. By means of a sensitivity analysis, the impact will get visible of the individual input variables on the output events that are found by the analysis. If more insight is gained into the impact of the input variables, then a possible next step is trying to calibrate the model and its input variables on available incident registrations of Rijkswaterstaat. In this way, a relation can be identified between the intensity of tweets and events on or near roads.

The sensitivity analysis is carried out by running multiple runs of a workflow for the geographic regularity analysis on Twitter data in the study area. For each run of the geographic regularity analysis, a single input variable is changed consistently in order to measure its impact on the analysis' output results. On the complete setup of the sensitivity analysis setup will be elaborated more on in section 4.7.4. The current section will elaborate more on the design of the geographic regularity analysis workflow.

In Figure 26 a flow chart overview of the geographic regularity analysis workflow can be seen. The flow chart items can be distinguished in three types: input variables, tools (python scripts or ArcGIS ModelBuilder models) and outputs. In order to refer to the flow chart items in the text below, each flow chart item is marked with a reference number/letter.

The first step in the sensitivity analysis workflow is running the *CreateZones* model (I) and *ClusterTweets* script (II):

Create zones

The *CreateZones* model (Appendix 5) will create the zones for which geographic regularities will be estimated eventually. The zones are created based on the light-version of the Dutch National Road File (Figure 27). In this light-version of the Dutch National Road File, all Dutch primary and secondary highways are simplified and represented by a single line only in order to facilitate data linking between data and the Dutch National Road File (Rijkswaterstaat, 2014). Because roads are represented by a single line, this light-version of the Dutch National Road File was especially useful for creating zones for the geographic regularity analysis.

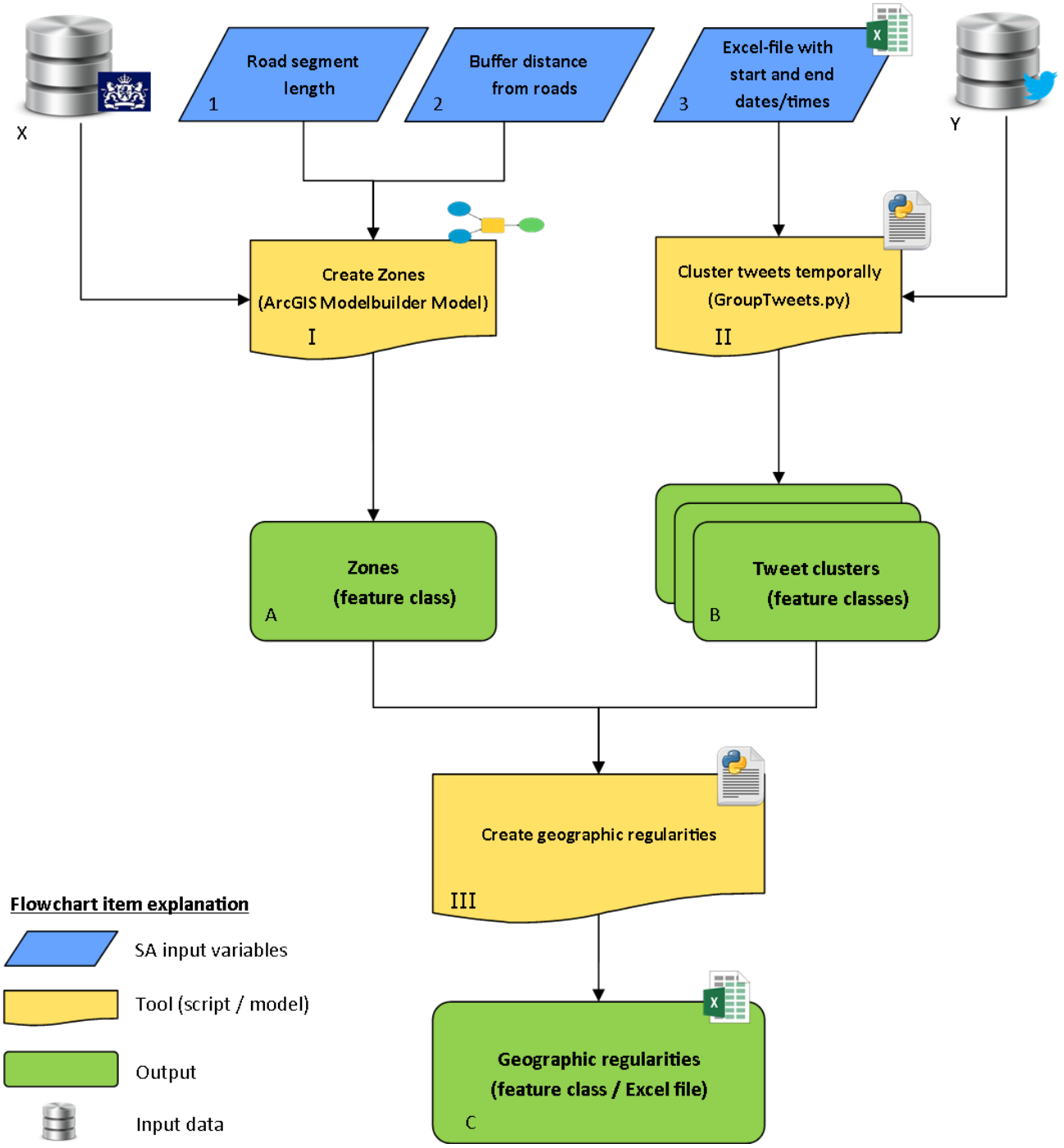


Figure 26 Flow chart overview of the geographic regularity analysis workflow

The working of the *CreateZones* model can be explained by two steps. In the first step, all lines in the input road file are split into segments of a specific length. This *road segment length* (1) in which the lines are split, is one of the input variables of the sensitivity analysis and must be specified before running the model. In a second step the *CreateZones* model creates buffers around all the split line segments. This *buffer distance from roads* (2) is the second input variable of the sensitivity analysis and must be specified before running the model as well.

The output of the *CreateZones* model is an ArcGIS feature class with polygon features for each zone (Figure 28). Each zone is a buffer that was created around all single split line segments of the Dutch National Road File in the final step of the *createZones* model.

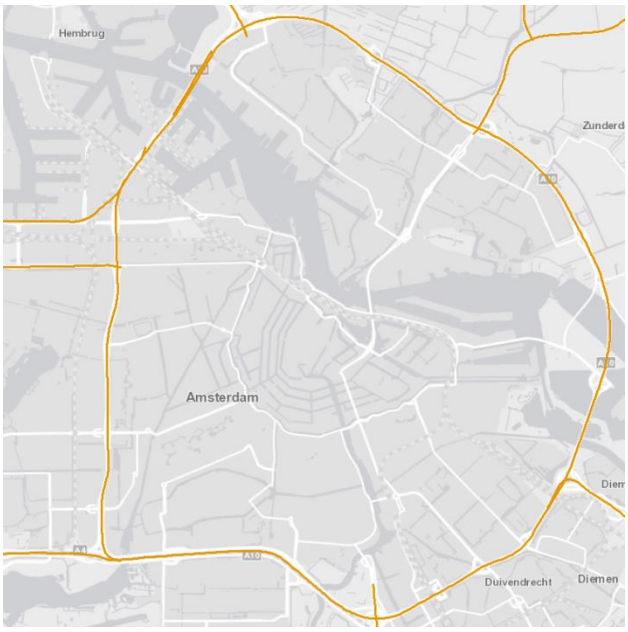


Figure 27 Original light-version of the Dutch National Road File, in the area of Amsterdam

Figure 28 Zones created by using the *CreateZones* model, using 2000 m as variable for *road segment length* and 150 m as variable for *buffer distance from roads*.

Cluster tweets

The second type of input data that is required for running the *CreateGeographicRegularities (*III*)* script are clusters of tweets (B). These clusters are created by the *ClusterTweets* script (Appendix 6). The *ClusterTweets* script makes queries on the Twitter database, based on start and end dates/times that are read from an Excel file (3). The Excel file should be created manually before running the script. Selections on the Twitter database, based on the Excel values are subsequently exported to separate feature classes. The output of the script is a database with separated feature classes which contain temporal clusters of tweets. If for example Tweets are clustered per day, then each feature class will contain al tweets that are sent during one day.

Create geographic regularities

The third step of the sensitivity analysis workflow is running the script *CreateGeographicRegularities* (III) (Appendix 7) which counts for each zone of the zones input feature class (A) the number of tweets of each Tweet cluster (B) that overlap the zone. In other words, the script counts the number of tweets that were sent within a specific zone, within a specific time frame. An example scenario of the working of the CreateGeographicRegularities script can be seen in Figure 29. For each zone, the number of tweets counted per time period are written to a table (C) in which each row represents a geographic zone, and each column represents a time period. The cells in the table thus represent the number of tweets that are counted for a specific zone for a specific time period.

	Time period 1	Time period 2	Time Period 3
Zone 1	1	1	1
Zone 2	2	3	1
Zone 3	2	1	1
Zone 4	1	2	1

Figure 29 Example scenario of the working of the CreateGeographicRegularities script. In the image left, a situation can be seen where tweets of different time periods are overlapping different zones. The table on the right is the result of the CreateGeographicRegularities script when doing calculations for the situation in the image on the left.

Identify irregularities in table with geographic regularities

After the *CreateGeographicRegularities* (III) finished running, the model's output table can be used to define and find geographic irregularities. First the geographic regularity threshold value should be defined. This threshold value defines whether clusters of tweets should be detected as being a potential event or not. If the number of tweets in a specific cluster exceeds the threshold value, then this cluster is identified as being a potential event.

Geographic regularity threshold values (GRTV) are calculated for each zone, based on the interquartile range in the distribution of tweet occurrences (Figure 25):

(Formula 1) $\boldsymbol{GRTV = Q_3 + k(Q_3 - Q_1)}$

where:
Q_1 = first quartile
Q_3 = third quartile
k = constant

When the GRTVs are calculated for each zone, irregular tweet clusters can be identified in the table.

Compare found irregularities with incident management loggings

In order to value the performance of the different runs of the sensitivity analysis, the results are valued based on incident management loggings from Rijkswaterstaat. In a similar fashion as tweets are counted for each zone and time period, the occurrence of road incidents are counted for each zone and time period as well. For each detected event in the Twitter data, it is checked if a road event happened in the same zone and time period according to the incident management loggings. These checks give the following details for each sensitivity analysis run, which can be used as indicators for their performance:

- The number and percentage of irregularities in the Twitter data, that match a road event
- The number and percentage of irregularities in the Twitter data, that do not match a road event

If more detected irregularities match road events occurrences in space and time, then this indicates a better performance of a sensitivity analysis run.

4.7.3. Differences and similarities compared with regularity analysis in literature

The regularity analysis as it is applied in this thesis is based on the research of Lee et al. (2011). The main principle of detecting events from Tweets, based on their spatio-temporal characteristics, is adopted from the work of Lee et al. (2011) and other studies that were carried out following this work (Lee et al., 2013; Wakamiya et al., 2013).

The studies of Lee and Wakamiya and this thesis, share a common research goal: detecting unusual regional social activities from Twitter (or different microblogs). Although the geographic regularity analysis is applied in this thesis following the work of Lee et al. (2011), there are some main difference in the way the geographic regularity analysis is implemented in this thesis and the study of Lee et al. (2011):

- In the study of Lee et al. (2011), three different types of indicators for crowd activity are used:
  - Degree of crowd activity based on Tweets, which is the number of tweets sent within a zone, within a specific period of time.
  - Degree of crowd activity based on crowd, which is the number of individual Twitter users found within a zone, within a specific period of time.
  - Degree of crowd activity based on moving crowd, which is the number of moving users related to a zone's boundary within a specific period of time. For each time period and zone, the number of users that came into the zone or stayed within the zone were calculated.

  The use of above mentioned different crowd activity indicator types is especially useful if Twitter activity in relatively crowded areas are monitored. For the research goals of this thesis, it is considered sufficient to calculate only the degree of crowd activity based on tweets for each zone and period of time. Degree of crowd activity based on crowd, and based on moving crowd are not taken into consideration in the thesis' analysis, because:
  - The zones that are monitored show a relative low Twitter activity compared to other area's in the center of Amsterdam. Because of this lower Twitter activity, it is expected that the number of tweets and the number of Twitter users are very close to each other.
  - Because of the high number of monologues it is chosen to dissolve tweets within a zone based on individual Twitter users. This will improve detection performance of traffic-related events as it is assumed to be unlikely that a Twitter user will write more than 1 message about a traffic situation.
  - Degree of crowd based on moving crowd is not calculated because it is assumed that relevant tweets will be most often from traffic participants which are always moving Twitter users.

- In order to determine an adequate zone size Lee et al. (2011) used a K-means clustering method which divided their study area into zones based on the geographical occurrences of their dataset. The center of each cluster was used to define a Voronoi diagram that divided the study area into "socio-geographic boundaries".

  The use of a K-means clustering method is not applied in the geographic regularity analysis of this thesis, because the focus lies on analyzing tweets that are sent near highways of secondary roads in the study area. This means that it is probably irrelevant to take the geographic distribution of all tweets into account when defining zones for the analysis. Moreover, it is considered unnecessary to cover the whole area with zones. Only zones that cover highways or secondary roads are meaningful to the analysis. Hence, in order to focus on the distribution of the tweets near highways and secondary road instead, zones are based on the geographic position of roads.

  Because it is aimed to find clusters of tweets near or on the roads, it isn't very useful to cover the study areas with rasters in order to define zones (as described in section 4.6.4). In order to increase

chances on finding clusters on or near roads, it is presumed that it is better to base the zones on the geometry of the roads, instead of the geometry of rasters randomly positioned over the study area. Moreover it is expected that different tweets about the same road event could be sent with a relatively long distance (along the roads) from each other. For example it could be that two traffic participants are tweeting about the same traffic jam, however they are kilometers away from each other because one of the traffic participant is at the head of a traffic jam and the other traffic participant is at the traffic jam's tail.

Another reason to search tweet clusters in elongated shaped zones, is to decrease the presence of irrelevant tweets in the zones which disturb the geographic regularity estimation. The clusters that are aimed at to find, are expected to consist of only a few tweets. The more geographic regularity zones are deviating from roads, the more chance many tweets are taken into account during the geographic regularity estimation, lowering the chance that small cluster of tweets near the road can be identified.

- In the study of Lee et al. (2011), geographic regularities are estimated based on time periods of six hours: morning (6AM – 12 AM), afternoon (12AM – 6 PM), evening (6PM – 12PM), night (12PM – 6AM).

  For the geographic regularity analysis, the same time periods are applied as in the study of Lee et al. (2011). Next to these time periods of six hours, time periods of 12 hours and 24 hours are used as well. Because Twitter intensity is much lower for the geographic areas that are analyzed in this thesis compared to the areas that are analyzed by Lee et al. (2011), it is assumed that more adequate geographic regularities could be estimated for longer time periods.

4.7.4. Sensitivity analysis setup

The geographic regularity analysis is performed in different runs, with different model variables set for each run. In this way the sensitivity analysis is shaped. The following model variables are manipulated each run:

1. **Area of zones**

   A. Buffer distance from roads

   The buffer distance from roads, on which the width of the zones is based, has impact on the number of tweets that are used to estimate geographic regularities. It is expected that if geographic regularities are estimated for zones with a wider area around the roads, more tweets are taken into account that are not relevant to the road incidents. It is expected that the more irrelevant tweets are taken into account during geographic regularity estimation, the more difficult it becomes to detect relevant road events. So the model's performance to detect road incidents is likely to decrease with buffer distance. In order to test this assumption, different buffer distances are used to estimate geographic regularities: 50, 100, 150 and 200 meters.

   B. Segment length of roads

   The segment length of roads, on which the length of the zones is based, has impact on the number of tweets that are potentially related to a road incident. For example, it could be that a tweet about a road incident was sent after a car passenger traveled a few hundreds of meters away from the road incident. For this reason it could be that relevant tweets are missed when zones with a short length are used to estimate geographic regularities for. On the other hand it could be as well that noise increases due to the occurrence of irrelevant tweets. In order to test the impact of the segment length variable on the model's performance to detect incident-related events, different segment lengths are used to estimate geographic regularities: 500, 1000, 1500, 2000 meters.

2. **Time interval of the tweet clusters**

   The time interval for which geographic regularities are estimated has impact on the number of tweets that will be available for this estimation. Peaks of incident occurrences and tweet frequencies near roads in the study area take place between 8AM-10AM and 4PM-6PM. It would therefore be expected that it would be easier to detect road events based on geographic regularities estimated for these peak time periods instead of whole days. When estimating geographic regularities for whole days, it is likely that more irrelevant tweets are taken into account which will have a negative impact on the model's performance. On the other hand however, if short time periods are chosen to estimate geographic regularities for, it could be the case that too few tweets are available to base geographic regularities on.

   As a baseline, geographic regularities are first estimated based on tweet frequencies for whole days. Thereafter, geographic regularities are estimated for time periods of 12 hours, and six hours.

3. **The geographic regularity threshold values definition**

   Geographic regularity threshold values (GRTV) are calculated for each zone, based on the interquartile range in the distribution of tweet occurrences (see formula 1). The constant k has impact on the sensitivity of the geographic regularity threshold. If a low value is chosen for k, then there's a higher chance on detecting road event occurrences, however the chance that these occurrences are relevant are lower than when a higher value for k is chosen. In order to test the impact of the constant k on the model's performance to detect road events, different k values are used to estimate geographic regularities: K=1.5 K=2 and K=3.

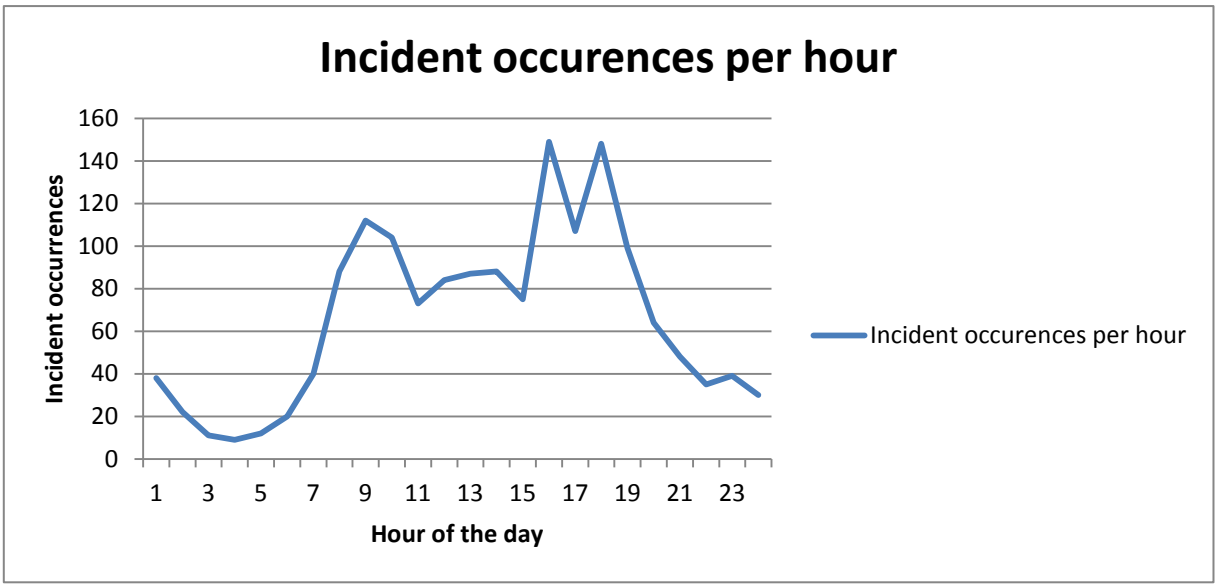


Figure 31 Incident occurrences per hour on highways and secondary roads in the study area, counted on all available dates. Source: Rijkswaterstaat

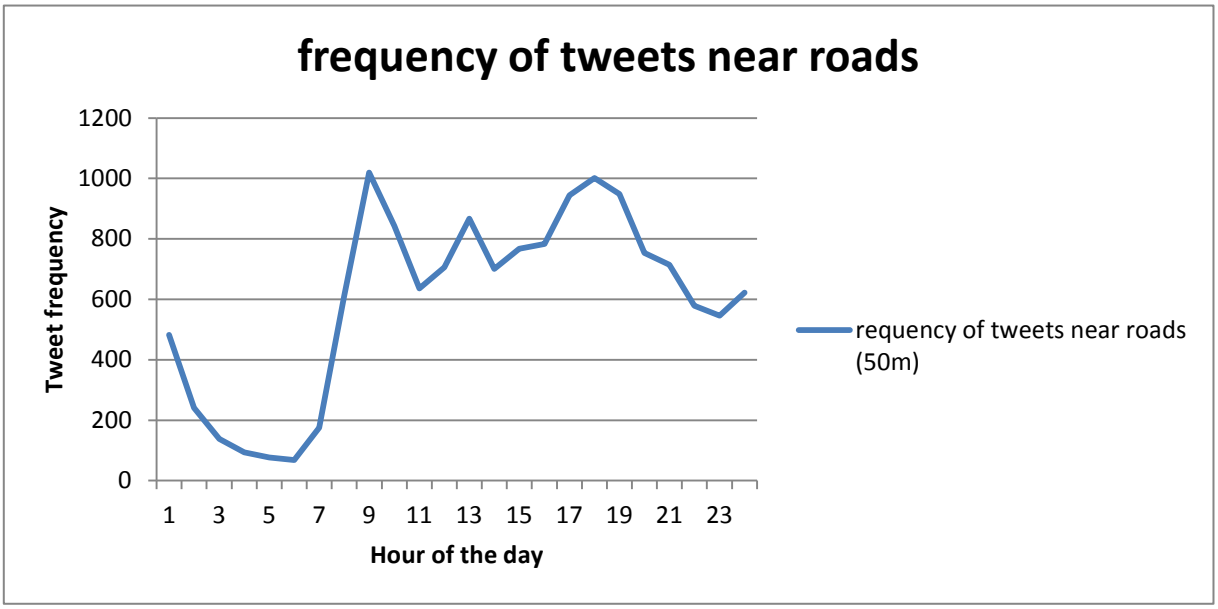


Figure 30 Tweet frequencies counted within 50 meters of highways and secondary roads in the study area counted on all available dates.

5. Results

5.1. Geographic regularity analysis results

In this section the results of the different model runs of the entire sensitivity analysis are listed. In total the model is run 112 times, each run with different input variables specified (see section 4.7.4). The results of the model runs are expressed in two values: the number of events (in other words "irregularities") that are detected in the run, and the percentage of detected events that co-occur with registered road events in the same spatio-temporal zone.

The percentage that is estimated for each model run can be used as an indicator of the performance of the model run to detect road events. A percentage of 0 would mean that there's no indication that any detected irregularity is related to any registered road event. A percentage of 100% would mean that there's a very strong indication that detected irregularities are related to registered road events. In the next section (section 5.2), the model runs that returned the highest percentages are evaluated. For these runs it is verified if the indication for a relation between detected and registered events is correct.

For the first set of model runs, tweets grouped per day were used for estimating geographic regularities. In order to define the geographic regularity threshold value, a constant 'k' of 1.5 was used. Different buffer distances and road segment lengths were entered as input variables for each model run. The results of this first set of model runs can be found in Table 9.

A second and third set of model runs were carried out in order to verify the impact of the constant 'k'. All runs of the first set are repeated in the second and third set. Equal variables are used, except for the constant 'k' a value of 2 and 3 is used respectively in the second and third set of model runs. The results of the second and third set of model runs can be found in Table 10 and Table 11.

Moreover, new sets of series of analysis runs were carried out in order to get insight into the impact of the temporal factor, the time period for which geographic regularities are estimated, on the model's performance. In these subsequent series of runs, the geographic regularities are estimated based on tweets clustered per six hours. Tweets issued on separate days are further split into groups of tweets issued between 12AM – 6 AM, 6AM – 12 PM, 12PM – 6PM and 6PM – 12PM and used as input-data in the model. The results of this four series of model runs can be found in Table 12, Table 13, Table 14 and Table 15.

From the tables that show results of model runs using tweets grouped per day, the following most important observations can be summarized:

- Performance decreases with increasing "buffer distance from roads."
- Performance increases with increasing "road segment length."
- Performance decreases with increasing "constant k."

From the tables that show results of model runs using tweets grouped per six hours, the following most important observations can be summarized:

- Performance of runs that are based on mornings and afternoons is much better compared to runs that are based on nights and evenings.
- The average performance of runs based on mornings and afternoons is comparable with runs based on days.

k=1.5	Buffer distance from roads			
Road segment length	50 m	100 m	150 m	200 m
500 m	161 (8.29%)	134 (6.49%)	121 (5.99%)	124 (6.15%)
1000 m	97 (9.73%)	73 (8.18%)	63 (7.60%)	61 (7.32%)
1500 m	78 (12.85%)	75 (13.00%)	62 (11.03%)	45 (10.34%)
2000 m	55 (12.25%)	59 (12.94%)	45 (10.84%)	45 (12.50%)

Table 9 - Results for the 1st set of model runs, using tweets grouped per day and a constant 'k' of 1.5. The numbers represent the number of events that are found in occurrence with a road incident for the same zone and time period. Between brackets the percentages are given of the total number of found events that co-occure with a road incident.

k=2	Buffer distance from roads			
Road segment length	50 m	100 m	150 m	200 m
500 m	158 (8.27%)	127 (6.40%)	107 (5.69%)	107 (5.72%)
1000 m	87 (9.34%)	59 (7.41%)	49 (6.92%)	43 (6.20%)
1500 m	57 (10.71%)	40 (8.81%)	39 (8.71%)	23 (7.01%)
2000 m	37 (9.92%)	39 (10.24%)	28 (8.24%)	29 (10.25%)

Table 10 - Results for the 2nd set of model runs, using tweets grouped per day and a constant 'k' of 2. The numbers represent the number of events that are found in occurrence with a road incident for the same zone and time period. Between brackets the percentages are given of the total number of found events that co-occure with a road incident.

k=3	Buffer distance from roads			
Road segment length	50 m	100 m	150 m	200 m
500 m	135 (7.71%)	118 (6.38%)	103 (5.95%)	94 (5.67%)
1000 m	77 (9.20%)	43 (6.22%)	33 (5.51%)	33 (5.92%)
1500 m	45 (10.39%)	32 (7.69%)	29 (7,67%)	9 (3.42%)
2000 m	26 (8.78%)	22 (6.69%)	21 (7.17%)	21 (10.05%)

Table 11 - Results for the 3rd set of model runs, using tweets grouped per day and a constant 'k' of 3. The numbers represent the number of events that are found in occurrence with a road incident for the same zone and time period. Between brackets the percentages are given of the total number of found events that co-occure with a road incident.

Night (12AM-6AM)	Buffer distance from roads			
Road segment length	50 m	100 m	150 m	200 m
500 m	7 (2.4%)	11 (2.34%)	12 (1.87%)	15 (1,87%)
1000 m	9 (3.2%)	15 (3.60%)	17 (3.09%)	18 (2.94%)
1500 m	11 (2.64%)	22 (3.93%)	20 (3.82%)	23 (4.05%)
2000 m	15 (5.66%)	20 (5.70%)	23 (5.6%)	22 (4.7%)

Table 12 - Results for the 4th set of model runs, using tweets grouped per six hours and a constant 'k' of 2. The numbers represent the number of events that are found in occurrence with a road incident for the same zone and time period. Between brackets the percentages are given of the total number of found events that co-occure with a road incident.

Morning (6AM-12PM)	Buffer distance from roads			
Road segment length	50 m	100 m	150 m	200 m
500 m	101 (10.5%)	120 (8,94%)	128 (8.3%)	133 (8.26%)
1000 m	53 (6.16%)	49 (4.82%)	41 (4.31%)	34 (3.70%)
1500 m	60 (8.34%)	59 (7.63%)	50 (6.67%)	37 (5.57%)
2000 m	50 (9.01%)	52 (9.56%)	37 (6.95%)	23 (5.15%)

Table 13 - Results for the 5th set of model runs, using tweets grouped per six hours and a constant 'k' of 2. The numbers represent the number of events that are found in occurrence with a road incident for the same zone and time period. Between brackets the percentages are given of the total number of found events that co-occure with a road incident.

Afternoon (12PM-6PM)	Buffer distance from roads			
Road segment length	50 m	100 m	150 m	200 m
500 m	46 (3.79%)	55 (3.69%)	64 (3.81%)	58 (3.38%)
1000 m	62 (6.60%)	59 (5.71%)	56 (5.66%)	54 (5.79%)
1500 m	14 (15.91%)	12 (10.62%)	10 (7.41%)	43 (7.14%)
2000 m	59 (10.97%)	53 (10.58%)	47 (10.59%)	41 (9.49%)

Table 14 - Results for the 6th set of model runs, using tweets grouped per six hours and a constant 'k' of 2. The numbers represent the number of events that are found in occurrence with a road incident for the same zone and time period. Between brackets the percentages are given of the total number of found events that co-occure with a road incident.

Evening (6PM- 12AM)	Buffer distance from roads			
Road segment length	50 m	100 m	150 m	200 m
500 m	21 (2,46%)	26 (2.24%)	35 (2.63%)	31 (2.25%)
1000 m	39 (5.32%)	37 (4.31%)	39 (4.59%)	33 (4.15%)
1500 m	40 (6.91%)	36 (5.40%)	28 (4.90%)	24 (4.15%)
2000 m	37 (7.39%)	32 (6.10%)	24 (6.06%)	19 (5.31%)

Table 15 - Results for the 7th set of model runs, using tweets grouped per six hours and a constant 'k' of 2. The numbers represent the number of events that are found in occurrence with a road incident for the same zone and time period. Between brackets the percentages are given of the total number of found events that co-occure with a road incident.

5.2.Evaluation of model runs with highest performance

For each individual model run in the seven series of model runs listed in section 5.1, model performance is estimated by a percentage. This percentage represents the part of the total number of detected events that co-occur with a road incident in the same time period and geographic zone. It is for those runs with the highest percentage expected to have the highest chance on finding a relation between the detected events and the events that are registered by Rijkswaterstaat.

In order to validate the performance of the most successful model runs, correlation statistics are used. In a similar way as done in section 4.5, the relation between issued tweets and past minutes after occurrences of incidents are valued. For model runs with a higher performance, it should be expected that a stronger relation exists between the tweet clusters that are detected in these runs as irregularities and the incidents that co-occur in the same spatio-temporal zones. Although some runs have better performance values, this can always be the result of coincidental co-occurrence of many irregularities and road incidents. The high performance of a model run would therefore be more justified if a significant correlation could be found between co-occurring tweets and incidents which resulted in this high performance of the model.

In order to make correlation calculations between the tweet clusters that are identified as irregularities, and the road incidents that co-occur in the same spatio-temporal zone, the following steps are taken:

- Extract all tweets from the Twitter database within spatio-temporal zones that are marked as irregularities.
- Extract all incidents from Rijkswaterstaats inicident loggings within spatio-temporal zones that are marked as irregularities.
- Create tables that list for every minute of the studied period the tweet frequency and the time distance in minutes from an incident occurrence (see Table 6 on page 43).
- Pearson's correlations (Field, 2009) are calculated for the table attributes *tweet frequency* and *time distance*.

For all 18 model runs listed in Table 16, Pearson's correlations (Field, 2009) were calculated using SPSS software. For all model runs very weak negative correlations were found, of which all but two are significant at the 0.01 level. Compared to the results that were found for correlation calculations in section 4.5, the results in Table 16 are relatively better.

Before results are interpreted, it should be considered that correlation statistics cannot give an indication about the direction of causality (Field, 2009). This means for this analysis that it cannot be concluded that (a small part of) road incidents cause more tweets to be issued around the location and time of these incident occurrences. In order to know if there's a genuine relation between detected tweet clusters and road incident, tweets' texts should be read and checked if they contain information related to the incidents. Because this is a time consuming task, it is decided to check only the tweets of detected irregularities in the model run with the highest correlation efficient.

In Table 16 it can be seen that the highest correlation coefficient was found for the model run with rank 11. Because this model run has the strongest correlation between tweets in irregularity clusters and road incidents, it is assumed that the highest chance on finding relevant tweets is for irregularity clusters in this model run. The 4665 tweets that were part of irregularity clusters in this model run are manually sorted on relevancy. Each tweet's text is checked if it described a road incident or the road condition. If the text was traffic-related, and a photo was attached to the tweet, then this photo was checked on any traffic or incident-related information. Only 48 tweets of the total 4665 tweets could be related to the road traffic (Appendix 8). Photographs were sent as attachment with 10 of the 48 tweets.

rank	performance of run (%)	k	road segment length (m)	buffer distance (m)	time-period	correlation
1	15.91	2.0	1500	50	afternoon	-.027 (0.01)**
2	13.00	1.5	1500	100	day	-.052 (0.01)**
3	12.94	1.5	2000	100	day	-.049 (0.01)**
4	12.85	1.5	1500	150	day	-.057 (0.01)**
5	12.50	1.5	2000	200	day	-.050 (0.01)**
6	12.25	1.5	2000	50	day	-.001 (.941)
7	11.03	1.5	1500	150	day	-.057 (0.01)**
8	10.97	2.0	2000	50	afternoon	-.034 (0.01)**
9	10.84	1.5	2000	150	day	-.059 (0.01)**
10	10.71	2.0	1500	50	day	-.041 (0.01)**
11	10.59	2.0	2000	150	afternoon	-.102 (0.01)**
12	10.58	2.0	2000	100	afternoon	-.049 (0,01)**
13	10.50	2.0	500	50	morning	-.006 (.116)
14	10.39	3.0	1500	50	day	-.041 (0.01)**
15	10.34	1.5	1500	200	day	-.065 (0.01)**
16	10.25	2.0	2000	200	day	-.050 (0,01)**
17	10.24	2.0	2000	100	day	-.049 (0,01)**
18	10.05	3.0	2000	200	day	-.050 (0,01)**

Table 16 Rank list of 18 of the 112 model runs that scored best performances. The columns "k", "road segment length (m)", "buffer distance (m)" and "time period" list the input variables of each model run. The column "correlation" lists the correlation coefficient and its significance level in brackets that were calculated for each run (see page 62 for explanation).

**** correlation is significant at the 0.01 level (2-tailed)**

6. Discussion

In this chapter the results of analyses that are carried out in the methodology are evaluated and reflection is given upon the methodology. First the results are evaluated against the data quality criteria that are defined in section 4.1. In this way the potential value of Twitter as information source for incident management is discussed. Thereafter, the methodology of the research is evaluated. It is discussed how the design of the geographic regularity analysis, and the preprocessing of the input data could have affected the results and how the research could have been improved. Finally, there are some questions that remain unanswered in this thesis. It is explained why these questions are important to answer in further research.

6.1. Evaluation of results against the data quality criteria

Based on preliminary investigations and literature research a couple of use cases are defined in chapter 3 for which it was expected that Twitter as a potential source of (geo-)information could support different practices in the incident management process:

- Incident report verification
- Incident detection
- Incident communication to road users

In order to be a valuable information source for incident management, different criteria on the Twitter data quality are defined in section 4.1. It is now discussed whether the Twitter data could meet these criteria:

(criterion 1) A tweet should hold sufficient indications or evidence about the location from which it was sent.

It is discovered that a coordinate pair is probably the only evidence about the location of a tweet that is really useful for incident management. Information about incidents must be able to locate an incident scene with precision. If no coordinate pair would be available then geolocalizing tweets with enough precision would become a seemingly impossible task. In Appendix 8 it can be seen that twitter users most of the time don't provide details about their location, and if they do, location details are too inaccurate to be useful. Sometimes Twitter users use the highway number and the driving direction in their tweets. In rare occasions Twitter users note the hectometer from the road signs at which they are at. Sometimes, pictures give an indication of the location of an incident but it would be impossible to use images in automation processes for localizing tweets.

(criterion 2) Tweets should be sent within relevant time limits after an incident happened.

Since there hasn't been found tweets that could clearly be linked to incident occurrences it remains unanswered whether incident-related tweets are sent within relevant time limits after an incident happened or not. In section 4.5 a possible relation between incident occurrences and Twitter activity around roads is investigated using correlation calculations. From these calculations it should be concluded that no relation could be found between the time after an incident happened, and the tweets that are sent after this happening.

(criterion 3/3B) Tweets should bring detailed information updates about the incident situation or road traffic conditions

In section 5.2 all tweets of the best performing model run were investigated. Only 48 tweets of the total 4665 tweets could be related to the road traffic (Appendix 8). The majority of these 48 tweets didn't tell anything more than that a traffic jam was present. Regarding the tweets' texts, only two tweets (from users with car troubles) gave some detailed information about a road event.

Photographs were sent as attachment with 10 of the 48 tweets. When investigating these photographs (Appendix 9) it becomes clear that most photos do not bring detailed information about an incident situation or about road conditions. Most photographs only show cars in a traffic jams which doesn't bring any useful information. The only photograph that might be useful for incident management is a photo of a trailer with a flat tire (Appendix 9, image 2).

(criterion 4) A spatio-temporal relation should exist between road traffic conditions and patterns in Twitter data traffic.

Correlation calculations are done for the 18 best performing model runs of the sensitivity analysis. From the results of this calculations in Table 5 it was discovered that very weak negative correlations exists between co-occurring tweets from irregularities and incident events. The fact that nearly all correlation calculations are significant is an indication of a possible relation between the issued tweets around road incident events. Unfortunately the fact that all correlation coefficients are very small (near zero) is an indication that a relation between issued tweets around incident occurrences can only be found in a small part of the sample size. This would mean that based on the spatio-temporal characteristics of tweets, there's a small chance that tweets can be found that are related to road incidents. Because these chances are so small however, it should be concluded that a useful spatio-temporal relation between Twitter data and registered incidents does exist. Twitter data does not meet criterion 4.

6.2.Evaluation of impact of individual model input parameters

From the tables of model runs' results in chapter 5, it can be observed that each model input parameter has a different impact on the models outcome.

The shape of geographic zones (road segment length and buffer distance)

A particular trend can be observed for changes in the model's input variable "road segment length". For the majority of model runs, performance increases when road segment length is increased and other input parameters are held equal. This trend could have different explanations. One explanation could be that relevant tweet clusters around incidents can be found over long distances, however this explanation is very hard to demonstrate since so few relevant tweet clusters are found. Maybe a more obvious explanation is that performance increases because irregularities are calculated for bigger zones when longer road segment lengths are used. In bigger zones, there's a higher chance on finding an irregularity cluster that co-occurs with a road incident, hence there's a higher chance on getting a better performance of the model run.

The results showed that input parameter "road segment length" has a relatively high impact on the model's performance. It is doubtful however whether it is justified to conclude that using longer road segment lengths as input parameter really improves the performance of the model in the sense that more irregularities are found that are relevant to road incidents. Moreover it should certainly be taken into consideration that events detected in long zones provide less detailed information about the location of a possible event than when this same event was detected in a small zone. It is therefore that a better performance of a model run's result not necessarily means that this is a more useful result.

For the variable "buffer distance" a less clear and consistent trend can be observed compared to the variable "road segment length", though it can be seen that there's a general trend that an increasing buffer distance has a negative impact on the model's performance. This trend is expected because it is assumed that tweets issued further away from roads are less relevant to road events. In order to reduce noise in the input data and to increase chance on finding relevant tweet clusters, it is better to keep the zones that are used for irregularity calculations as small as possible around roads.

There's no doubt that the geometry of the input zones has impact on the success of the regularity analysis. This is for example also the reason that in other studies an ideal geometry of the zones is determined by

using an analytical approach. In Lee et al. (2011) for example, a K-means clustering method is applied to estimate zones more convenient for analysis. This method is not applied in this thesis, as it became clear that it wouldn't be useful to take all tweets into account for analysis. Only tweets issued near roads are analyzed by manually creating zones around roads.

Time period of tweet clusters
The choice of time period for which geographic regularities are estimated seems to have a relatively high impact on the model's performance. Interesting differences can be observed when comparing the mean of the performance values of the model runs that are done using tweets grouped by six hours (Tables on page 61) with the runs that are done with the same input variables but using tweets grouped by days (Table 10):

Input variable "Time period of tweet clusters"	Average performance value (%)
Days	8.12
Night	3.59
Morning	7.12
Afternoon	7.56
Evening	4.64

The difference between the average performance values of model runs for morning, afternoon and day periods is relatively low, whereas the difference between the average performance values of model runs for night, evening and day periods is relatively high. This observation can be expected since most incidents take place during mornings and afternoons (Figure 31), which increases the potential performance of the model runs that are done for these periods.

An important observation is that it seems that there's not much performance to gain by running models that estimate geographic regularities for shorter time periods. From the total number of 112 model runs, 18 runs resulted in a percentage of 10% or higher. It is remarkable that although more runs were based on time periods of six hours, a majority (72%) of the 18 best scoring events are based on time periods of 24 hours (Table 16). This is not what was expected on forehand. It was assumed that more relevant tweet clusters would be found when doing model runs for shorter time periods. For example, when doing irregularity estimations for whole days, tweets in found irregularities can be issued both in the morning as in the evening. If one tweet is issued in the morning and another tweet is issued in the evening, they both cannot be about the same road event although they may be in the same irregularity cluster. By doing irregularity estimations for shorter time periods like six hours, there´s less chance that noisy data (like irrelevant tweets sent a few hours after an incident) are taken into account during event detection. Maybe time ranges of six hours are still too wide to see clear improvements in the model's performance.

Although it was expected that estimating geographic regularities on sorter time periods would bring better performance results, it was no option to do analysis runs for time periods shorter than six hours. When estimating geographic regularities for time periods shorter than six hours, too few tweets are available for each run to base proper irregularity estimations on. For the majority of model runs that are done for time period of six hours the geographic regularity threshold values of the geographic zones (GRTV, see page 54) was '0'. If the GRTV is 0, then each issued tweet in this zone is automatically detected as irregularity. Of course, chances are very low that in those situations relevant events are found.

Constant K
For the constant 'k' a trend can be recognized that performance of runs decreases when K is higher. The impact of K is relatively low. For instance, an increase of K from 1.5 to 2 (33%) results in a mean decrease of the model's performance with 14.7%. An increase of K from 2 to 3 (50%) results in a mean performance decrease of 12%. The constant K is only useful to tweak the model's performance if next to relevant events

too many irrelevant events are found. Since, there are so few relevant events found, it seems premature to use the constant K for performance improvement in the analyses in this thesis.

6.3. Discussion about methodology

It may be hard to conclude on the usefulness of Twitter as a valuable information source for incident management if only a very small part of the Twitter source is used. The analyses that are performed in this thesis only make use of tweets that contain a coordinate pair in their metadata. As this is only a very small percentage of the total number of tweets (approximately 1% according to Schulz et al., 2013), conclusions in this thesis are only well-grounded for this small part of the data.

There is a possibility that the analyses in this thesis could have been improved if, next to the 'geographic tweets', tweets were also included that do not contain geographic coordinates. However, this would ask for complex analyses. According to Schulz et al. (2013), geolocalizing tweets is a difficult task which requires natural language processing to find spatial indicators for the tweets in tweets' texts. Aside from the complexity it is doubtful whether geolocalizing tweets would bring additional geotagged tweets with enough spatial detail. For instance, Schulz et al. (2013) were able to correctly localize 92% of the data within a 30 kilometer radius, a scale much too large to be of any use for incident management purposes. In Daly et al. (2013) however, much better geotagging results were found. In their attempt to geotag tweets in the city of Dublin, 100% of the successfully geotagged tweets were right about the location within an error range of 2 kilometers. At least 50% of the successfully geotagged tweets were right about the location within an error range of 500 meters. It would be interesting to know if similar error ranges could be achieved when traffic-related tweets in the study area of Amsterdam are geotagged.

To the best of our knowledge, detection of small-scale events using zonal regularity analysis hasn't been attempted prior to the present study. From the results of this thesis, it appears that the quantity of available data is the biggest shortcoming for detecting relevant small-scale events from the Twitter data using the zonal regularity analysis. As discussed earlier, running the sensitivity analysis for relatively small spatio-temporal zones often resulted in a GRTV of '0' because most of the times no tweets are counted for these small zones. In these cases where the GRTV for a zone is '0' because of a lack of sufficient available tweets, each single tweet that is issued in this zone is automatically detected as an event. It would perhaps be more efficient to use the point clustering technique, which is used in Sugitani et al. (2013) and Walther and Kaisser (2013), for these situations where there is so little Twitter activity. By using a point clustering technique, single tweets can never be identified as relevant events. An important drawback of this technique however, is that clusters are detected within a certain radius from each other. From the results of this thesis, it has been observed that clusters can better be detected within long zones around roads, instead of detecting clusters within a certain radius in order to reduce noise.

The sensitivity analysis that was carried out in this study was applied on the study area of Amsterdam. It was expected that in this area the highest chance existed that incident-related tweets would be found, because of the high Twitter activity and incident occurrences in this area. Within the scope of this thesis, the zonal regularity analysis is not applied on a different study area with different characteristics for Twitter activity and incident occurrences. It would have been interesting to apply the zonal regularity analysis on highways outside urban areas. Maybe around roads in a rural area there's less Twitter activity and fewer road incidents take place, though it could be that model runs for highways in rural areas would perform better because for example due to a lower impact of data noise. The question how useful Twitter could be as an information source for incident management on highways and secondary roads outside urban areas remains unanswered.

Another point of reflection is about the way performance is measured for all model runs in the sensitivity analysis. In order to measure performance of the model run, incident registrations of Rijkswaterstaat are used

to validate the potential relevancy of a found irregularity. If an irregularity was found in co-occurrence with a registered incident, then this irregularity contributed to the performance of the model run. If a irregularity did not co-occur with a registered incident it did not contribute to the model run performance. It would be interesting to use different data sources than the incident registrations to measure performance. For example, inductive loop detector loggings from Rijkswaterstaat could be used as well to find a relation between Twitter activity and detected events on highways. To give another example, it would also be interesting to investigate a possible relation between identified irregularities in Twitter and registrations in the P2000 network (e.g. in Stronkman (2011), P2000 registrations and tweets are used to detect events). In P2000 registrations incidents are also registered if they are not located on highways or secondary roads. Using P2000 registrations for validation of model runs, it would also be possible to measure performance of geographic regularity analyses that are carried out for zones that cover streets instead of highways.

A final important point to reflect upon regarding the methodology that is used, is the way noise elimination is applied in the geographic irregularity analysis. Noise elimination is an important part of the methodology in related studies (Abel et al., 2012a; Schulz et al., 2012; Sugitani et al., 2013). By filtering out tweets from the input data that are irrelevant in any case, performance of event detection can often be improved. For example, tweets are removed that are posted by 'bots' or that are retweeted. The only noise elimination measure that is taken in the geographic irregularity analysis of this thesis is the removal of tweets with identical user ids that are found in the same zone for the same time period (24 or 6 hours). It is assumed that Twitter users generally do not issue more than one tweet about an incident situation. Whether this measure really had a positive effect on the model's performance is a question that is not answered in this thesis. It could have been interesting as well to apply more noise elimination measures and investigate what the impact is on the model's performance. In this exploratory study however, harsh noise elimination measures were avoided because it was no so clear which type of tweets could be removed without negative consequences for the model's performance.

7. Conclusion and recommendations

The main research objective of this thesis was to investigate how useful Twitter data is as a source of spatio-temporal information for incident management in the Netherlands. It was expected that Twitter could be a valuable source of information for three types of incident management practices: incident report verification, incident detection and incident communication to road users. It should be concluded from the evaluation of results and the discussion however, that geotagged tweets in Twitter data cannot contribute as a useful information source for any of these three incident management practices in the area around Amsterdam.

The most important shortcoming of Twitter data for use in geographical analysis in general is that only a very limited number of tweets can be geolocated by a coordinate pair that is available in the metadata of the tweets. As a consequence it is difficult to detect small-scale events in the large amounts of data that are generated through Twitter every day. Real-world events that are successfully identified in this thesis using spatio-temporal analysis are detected from massive and temporary crowds at festivals, fairs, concerts or sport events. Without sophisticated noise filtering, machine learning or language processing, it seems challenging to detect small-scale events with analysis that is solely based on the spatio-temporal characteristics of tweets.

The geographic regularity analysis designed by Lee et al. (2011) is a practical design of spatio-temporal analysis that can be used for event detection in Twitter data, although it seems that this analysis can only be applied on large geographic zones. When the regularity analysis is applied on small geographic zones around highways, as done in this thesis, no useful results are found. Many tweet clusters are detected as irregular Twitter traffic, but when taking a closer look at these tweets they do not seem coherent and they cannot be linked to real-world events most of the times. This is also the case when searching for traffic-related events. Only a small part of the tweets that are sent on or near highways and secondary roads are actually about traffic conditions or incident events. Moreover, only a small part of the traffic-related tweets hold detailed information in the form of descriptive texts or images.

In order to be able to contribute to the incident management practices, information in tweets should meet hard criteria on information quality. It is shown that only a negligible minority of the traffic-related tweets meet these criteria in order to be valuable for incident management. Next to these findings, correlation calculations showed that there's no valuable relation between incident occurrences and Twitter activity near roads. Altogether, chances are thus low that events on highways and secondary roads will trigger tweets to be issued which can be geolocated accurately and provide relevant information about these events. For this reason relevant tweets are hard to detect by using spatio-temporal analysis.

Since this thesis only focused on events on highways and secondary roads, it is interesting to continue the search for applications that can benefit from information in geotagged Twitter data. This suggestion for further research is confirmed by many examples of useful applications of Twitter data described in the literature background of this thesis. For further research on small-scale event detection using spatio-temporal analytical approaches, it would be most challenging to deal with the small number of geotagged tweets that are available in the total Twitter stream.

An unanswered question remains how much potentially useful geographic information is hidden in the non-geotagged tweets. For the use cases of incident management, but certainly for other applications as well, information in tweets is useless if these tweets cannot be positioned accurately on a map. From literature it can be concluded that geotagging of tweets without coordinate-pairs is still a complex task and the accuracy of the estimated tweet locations is often limited. Especially for detecting small-scale events like traffic incidents the low availability of geotagged tweets is an obstacle.

8. References

The Basics | JSON - JavaScript Object Notation [online]. Retrieved December 3, 2013. Available on the world wide web: <http://www.json.com/>.

JSON - Wikipedia, the free encyclopedia [online]. (2013). Retrieved December 3, 2013. Available on the world wide web: <http://en.wikipedia.org/wiki/JSON>.

Abel, F., C. Hauff, G. Houben et al. (2012a). Semantics filtering search= twitcident. exploring information in social web streams. Paper presented at the Proceedings of the 23rd ACM Conference on Hypertext and Social Media, pp.285-294.

Abel, F., C. Hauff, G. Houben et al. (2012b). Twitcident: Fighting fire with information from social web streams. Paper presented at the Proceedings of the 21st International Conference Companion on World Wide Web, pp.305-308.

Ammerlaan D., M. Bijl, E. Klem et al. (2013). Jaarmeting 2012 Incident Management No. BB3979)Rijkswaterstaat Dienst Verkeer en Scheepvaart.

Barbera, P. (2013). Access to Twitter Streaming API via R [online]. Retrieved 07/04, 2013. Available on the world wide web: <http://cran.r-project.org/web/packages/streamR/streamR.pdf>.

Bontcheva, K. & D. Rout. (2012), Making sense of social media streams through semantics: a survey. Semantic Web

Cantino, A. (2013). twitter_to_csv [online]. Retrieved 07/04, 2013. Available on the world wide web: <https://rubygems.org/gems/twitter_to_csv>.

Verkeersinformatie - Beleid in uitvoering: de stand van zaken: Verkeersinformatie - Beleid in uitvoering: de stand van zaken: Verkeersinformatie - Beleid in uitvoering: de stand van zaken: (2006).

Daly, E.M., F. Lecue & V. Bicer. (2013). Westland row why so slow?: Fusing social media and linked data sources for understanding real-time traffic conditions. Paper presented at the Proceedings of the 2013 International Conference on Intelligent User Interfaces, pp.203-212.

Dawson, R. (2012). Which countries have the most Twitter users per capita?[online]. Retrieved April 16, 2013. Available on the world wide web: <http://rossdawsonblog.com/weblog/archives/2012/02/which-countries-have-the-most-twitter-users-per-capita.html>.

Drolenga J. (2011). Jaarmeting Incident Management No. GM-0062217). De Bilt: Grontmij.

Eurlings, C. M. P. S. (2010, Richtlijn eerste veiligheidsmaatregelen bij incidenten met eenzijdig aanrijdgevaar. *Staatscourant,* pp. 1-2.

Fernandes, J., P. Oliveira, C. Silva et al. (2012), Route Social Network. Procedia Technology 5, pp.547-555.

Field, A. (2009). Discovering statistics using SPSSSage publications.

Fujisaka, T., R. Lee & K. Sumiya. (2010). Discovery of user behavior patterns from geo-tagged micro-blogs. Paper presented at the Proceedings of the 4th International Conference on Uniquitous Information Management and Communication, pp.36.

Graser, A. (2012). Tweets to QGIS [online]. Retrieved 07/04, 2013. Available on the world wide web: <http://anitagraser.com/2011/10/01/tweets-to-qgis/>.

Haslam, N. (2012). How I Got Twitter Data Onto SQL Server [online]. Retrieved 07/04, 2013. Available on the world wide web: <http://architects.dzone.com/articles/how-i-got-twitter-data-sql>.

Hecht, B., L. Hong, B. Suh et al. (2011), Tweets from Justin Bieber's Heart: The Dynamics of the "Location" Field in User Profiles.

Hodge, V.J. & J. Austin. (2004), A survey of outlier detection methodologies. Artificial Intelligence Review 22(2), pp.85-126.

Immers, L. (2007), Guide to professional Incident Management.

Knibbe W. J. (2007). Inventarisatie beleidseffecten incidentmanagement. Rotterdam: Rijkswaterstaat Adviesdienst Verkeer en Vervoer.

Krishnamurthy, B., P. Gill & M. Arlitt. (2008). A few chirps about twitter. Paper presented at the Proceedings of the First Workshop on Online Social Networks, pp.19-24.

Lee, R., S. Wakamiya & K. Sumiya. (2011), Discovery of unusual regional social activities using geo-tagged microblogs. World Wide Web 14(4), pp.321-349.

Lee, R., S. Wakamiya & K. Sumiya. (2013), Urban area characterization based on crowd behavioral lifelogs over Twitter. Personal and Ubiquitous Computing 17(4), pp.605-620.

Lunden, I. (2012). Twitter May Have 500M+ Users But Only 170M Are Active, 75% On Twitter's Own Clients [online]. Retrieved April 16, 2013. Available on the world wide web: <http://techcrunch.com/2012/07/31/twitter-may-have-500m-users-but-only-170m-are-active-75-on-twitters-own-clients/>.

Mai, E. & R. Hranac. (2013). Twitter interactions as a data source for transportation incidents. Paper presented at the Transportation Research Board 92nd Annual Meeting, (13-1636)

McCarroll, N. (2012). A commandline tool for reading tweets from the twitter streaming API [online]. Retrieved 07/04, 2013. Available on the world wide web: <http://www.mccarroll.net/snippets/twitstreamer/index.html>.

Megally, M. (2012), Information extraction from social media for route planning.

Morris, M.R., S. Counts, A. Roseway et al. (2012). Tweeting is believing?: Understanding microblog credibility perceptions. Paper presented at the Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, pp.441-450.

Newcom Research & Consultancy B.V. (2013). Verrijken in plaats van bereiken No. 2013.11049.2732). Enschede / Amsterdam:

Rijkswaterstaat. (2011). IM guide to work processes (GuideRijkswwaterstaat.

Rijkswaterstaat. (2014). Nationaal Wegen Bestand Wegen Light [online]. Retrieved June 22, 2014. Available on the world wide web: <https://data.overheid.nl/data/dataset/nationaal-wegen-bestand-wegen-light>.

Rios M. (2013). The geography of Tweets, Twitter Inc. <https://blog.twitter.com/2013/geography-tweets-3>

Sakaki, T., M. Okazaki & Y. Matsuo. (2010). Earthquake shakes twitter users: Real-time event detection by social sensors. Paper presented at the Proceedings of the 19th International Conference on World Wide Web, pp.851-860.

Sakaki, T., M. Okazaki & Y. Matsuo. (2012), Tweet Analysis for Real-Time Event Detection and Earthquake Reporting System Development.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D. et al. (2008). Global sensitivity analysis: the primerJohn Wiley & Sons.

Schulz, A., A. Hadjakos, H. Paulheim et al. (2013), A Multi-Indicator Approach for Geolocalization of Tweets. Proceedings of the Seventh Internation Conference on Weblogs and Social Media (ICWSM)

Schulz, A., P. Ristoski & H. Paulheim. (2012), I See a Car Crash: Real-time Detection of Small Scale Incidents in Microblogs.

Semiocast. (2012). Twitter reaches half a billion accounts - More than 140 millions in the U.S.[online]. Retrieved April 16, 2013. Available on the world wide web: <http://semiocast.com/en/publications/2012_07_30_Twitter_reaches_half_a_billion_accounts_140m_in_the_US>.

Steenbruggen, J., P. Nijkamp & M. van der Vlist. (2013), Urban traffic incident management in a digital society. An actor-network approach in information technology use in urban Europe.

Stronkman, R. (2011), Exploiting Twitter to fulfill information needs during incidents. Delft: TU, pp.24-29.

Sugitani, T., M. Shirakawa, T. Hara et al. (2013). Detecting local events by analyzing spatiotemporal locality of tweets. Paper presented at the Advanced Information Networking and Applications Workshops (WAINA), 2013 27th International Conference On, pp.191-196.

Terpstra, T., A. de Vries, R. Stronkman et al. (2012). Towards a realtime twitter analysis during crises for operational crisis management. Paper presented at the Proceedings of the 9th International ISCRAM Conference. Vancouver, Canada,

The Netherlands Traffic Management Centre (VCNL). (2005). The Roles of the Emergency Services in Incident Management in the Netherlands [Het Rood - Blauwe boekje] (First English edition ed.). Utrecht: Verkeerscentrum Nederland.

Tumasjan, A., T.O. Sprenger, P.G. Sandner et al. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. Paper presented at the Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, pp.178-185.

Twitcident. (2013a). More about Twitcident / Twitcident [online]. Retrieved 07/08, 2013. Available on the world wide web: <http://twitcident.com/#about>.

Twitcident. (2013b). Projecten / Twitcident [online]. Retrieved 06-04, 2013. Available on the world wide web: <http://twitcident.com/#projects>.

Twitter. (2011). Visualization: Europe (Archive), Twitter Inc. <http://www.flickr.com/photos/twitteroffice/5331000428/in/set-72157633647745984>

Twitter. (2012a). Places [online]. Retrieved October 15, 2013. Available on the world wide web: <https://dev.twitter.com/docs/platform-objects/places>.

Twitter. (2012b). The Streaming APIs [online]. Retrieved 07/02, 2013. Available on the world wide web: <https://dev.twitter.com/docs/streaming-apis>.

Twitter. (2012c). Twitter Blog: Twitter turns six [online]. Retrieved April 16, 2013. Available on the world wide web: <http://blog.twitter.com/2012/03/twitter-turns-six.html>.

Twitter. (2013a). Frequently Asked Questions [online]. Retrieved 07/03, 2013. Available on the world wide web: <https://dev.twitter.com/docs/faq#6861>.

Twitter. (2013b). OAuth [online]. Retrieved 07/04, 2013. Available on the world wide web: <https://dev.twitter.com/docs/auth/oauth#user-context>.

Twitter. (2013c). REST API v1.1 Limits per window by resource [online]. Retrieved 07/03, 2013. Available on the world wide web: <https://dev.twitter.com/docs/rate-limiting/1.1/limits>.

Twitter. (2013d). Streaming API request parameters [online]. Retrieved October 2, 2013. Available on the world wide web: <https://dev.twitter.com/docs/streaming-apis/parameters#locations>.

Twitter. (2013e). Streaming API request parameters [online]. Retrieved October 2, 2013. Available on the world wide web: <https://dev.twitter.com/docs/streaming-apis/parameters#track>.

Twitter. (2013f). Tweets [online]. Retrieved 07/03, 2013. Available on the world wide web: <https://dev.twitter.com/docs/platform-objects/tweets>.

Twitter. (2013g). Users [online]. Retrieved October 15, 2013. Available on the world wide web: <https://dev.twitter.com/docs/platform-objects/users>.

Wakamiya, S., R. Lee & K. Sumiya. (2013). Social-urban neighborhood search based on crowd footprints network. Social Informaticspp. 429-442Springer.

Walther, M. & M. Kaisser. (2013). Geo-spatial event detection in the twitter stream. Advances in Information Retrievalpp. 356-367Springer.

Wanichayapong, N., W. Pruthipunyaskul, W. Pattara-Atikom et al. (2011). Social-based traffic information extraction and classification. Paper presented at the ITS Telecommunications (ITST), 2011 11th International Conference On, pp.107-112.

Zanten W. v. & J.d. Veth. (2011). Economische Wegwijzer - Economische schade door files blijven aanpakken (Factsheet No. 15531.1111107)TNO.

Zhao, S., L. Zhong, J. Wickramasuriya et al. (2011), Human as real-time sensors of social and physical events: A case study of twitter and sports games. arXiv preprint arXiv:1106.4300

9. Appendices

Appendix 1: List of abbreviations

ANWB	Algemene Nederlandsche Wielrijders-Bond
API	Application programming Interface
DRIP	Dynamisch Route Informatie Paneel
GIMA	Geographical Information Management and Application
GIS	Geographic Information System
GPS	Global Positioning System
JMA	Japan Meteorological Agency
JSON	JavaScript Object Notation
NNCI	NNCI
REST	Representational state transfer
RWS	Rijkswaterstaat
SMART	Specific, Measurable, Assignable, Realistic, Tme -related
TSNC	Truck Salvage Notofication Centre
VCNL	Verkeers Centrale Nederland
VID	Verkeersinformatiedienst
WGS	World Geodetic System

Appendix 2: tweet responded by the streaming API in JSON format


```
{
"retweeted": false,
"entities": {
        "hashtags": [{
                "text": "trafficjam",
                "indices": [90, 101]}],
        "urls": [],
        "symbols": [],
        "user_mentions": []},
"favorited": false,
"source": "<a href=\"http://twitter.com/download/iphone\" rel=\"nofollow\">Twitter for iPhone</a>",
"retweet_count": 0,
"coordinates": {
        "coordinates": [5.57270923, 51.30295885],
        "type": "Point"},
"created_at": "Fri Jun 21 12:24:29 +0000 2013",
"in_reply_to_status_id": null,
"in_reply_to_user_id_str": null,
"id": 348053808376594433,
"text": "Dorpskern afgesloten ivm kermis, schrikbarend wat een verkeer er overdag door zurrik komt
#trafficjam",
"favorite_count": 0,
"lang": "nl",
"in_reply_to_user_id": null,
"filter_level": "medium",
"geo": {
        "coordinates": [51.30295885, 5.57270923],
        "type": "Point"},
"user": {
        "follow_request_sent": null,
        "notifications": null,
        "profile_use_background_image": true,
        "default_profile": true,
        "description": null,
        "favourites_count": 0,
        "contributors_enabled": false,
        "is_translator": false,
        "name": "Rob van Hooff",
        "verified": false,
        "created_at": "Fri Oct 15 08:41:32 +0000 2010",
        "protected": false,
        "profile_link_color": "0084B4",
        "profile_background_color": "C0DEED",
        "id": 202993807,
        "statuses_count": 646,
        "profile_background_image_url": "http://a0.twimg.com/images/themes/theme1/bg.png",
        "friends_count": 136,
        "default_profile_image": false,
        "followers_count": 93,
        "profile_sidebar_border_color": "C0DEED",
        "location": "Soerendonk, Netherlands",
        "profile_background_image_url_https": "https://si0.twimg.com/images/themes/theme1/bg.png",
        "profile_text_color": "333333",
        "profile_image_url_https":
"https://si0.twimg.com/profile_images/2926590221/872e67deb87171037c26a9a0800847e1_normal.jpeg",
        "url": null,
        "lang": "en",
        "profile_image_url":
"http://a0.twimg.com/profile_images/2926590221/872e67deb87171037c26a9a0800847e1_normal.jpeg",
        "time_zone": "Amsterdam",
        "following": null,
        "profile_background_tile": false,
        "screen_name": "RobvanHooff",
        "id_str": "202993807",
```

```
        "profile_sidebar_fill_color": "DDEEF6",
        "listed_count": 1,
        "utc_offset": 3600,
        "geo_enabled": true},
"in_reply_to_status_id_str": null,
"place": {
        "attributes": {},
        "full_name": "Cranendonck, North Brabant",
        "place_type": "city",
        "id": "71458c401c6d4b4a",
        "bounding_box": {
                "coordinates": [[[5.5156653, 51.2209132], [5.5156653, 51.3542543], [5.6722413,
                51.3542543], [5.6722413, 51.2209132]]],
                "type": "Polygon"},
        "country": "The Netherlands",
        "url": "http://api.twitter.com/1/geo/id/71458c401c6d4b4a.json",
        "name": "Cranendonck",
        "country_code": "NL"},
"id_str": "348053808376594433",
"in_reply_to_screen_name": null,
"truncated": false,
"contributors": null
}
```

Appendix 3: Python script for gathering data from Twitter's streaming API

```
# twitstreamer.py
# Copyright (C) 2012 Niall McCarroll

from urllib.parse import quote
import time
import json
import select
import datetime
import time
import os
from os.path import exists
import sys
import argparse
import logging
import random
import hashlib
import hmac
from hashlib import sha1
import base64
import csv

from http.client import HTTPSConnection

# you must fill in the following values before you can run this script!
consumer_key = ""
consumer_secret = ""

access_token = ""
access_secret = ""

# formatter class for storing tweets as CSV rows
class csvformatter(object):

    csv.register_dialect('quotedcsv', delimiter=',', quoting=csv.QUOTE_ALL)

    def __init__(self,columns,write_header):
        self.columns = columns
        self.writer = None
        self.writer = csv.writer(sys.stdout,dialect="quotedcsv")
        if write_header:
            self.writer.writerow([col[0] for col in self.columns])

    def write(self,raw,obj):
        row = []
        for (col,decoder) in self.columns:
            if col in obj:
                row.append(obj[col])
            else:
                row.append("")
        self.writer.writerow(row)


# formatter class for storing tweets as JSON objects
class jsonformatter(object):

    def __init__(self):
        self.file = sys.stdout

    def write(self,raw,obj):
        s = json.dumps(obj)
        self.file.write(s+"\n")

# formatter class for storing tweets as raw JSON objects
class rawformatter(object):

    def __init__(self):
        self.file = sys.stdout

    def write(self,raw,obj):
        s = json.dumps(raw)
        self.file.write(s+"\n")


# utility class for streaming tweets from the twitter API
```

```
class twitstream(object):

    def __init__(self,options):
        self.options = options
        self.track = options.track
        self.locations = options.locations
        self.count = 0
        self.checkpoint_count = 0
        self.start_time = 0
        self.checkpoint_time = 0

        def date_decoder(s):
            return time.strftime('%Y-%m-%d %H:%M:%S', time.strptime(s,'%a %b %d %H:%M:%S +0000 %Y'))

        def geo_decoder(g,index):
            try:
                return str(g["geo"]["coordinates"][index])
            except:
                return ""

        # define which columns to create from each tweet object
        # as a list of column name, extractor-function pairs
        # an extractor function extracts the value of the column from
        # the tweet object
        # if extractor-function is set to None, the column name
        # is used as the lookup key in the tweet object
        self.columns = [("id",None),
                ("created_at",lambda x: date_decoder(x["created_at"])),
                ("geo_lat",lambda x: geo_decoder(x,0)),
                ("geo_lon",lambda x: geo_decoder(x,1)),
                ("from_user_name",lambda x:x["user"]["name"]),
                ("from_user_screen_name",lambda x:x["user"]["screen_name"]),
                ("iso_language_code",lambda x: x["user"]["lang"]),
                ("text",None)]

        self.formatter = self.createFormatter(self.columns)

    # create and return a formatter object
    def createFormatter(self,columns):
        if self.options.format == "json":
            return jsonformatter()
        elif self.options.format == "raw":
            return rawformatter()
        elif self.options.format == "csv":
            return csvformatter(columns,True)
        else:
            return csvformatter(columns,False)

    # generate a nonce used in the OAuth process
    def generate_nonce(self):
        random_number = ''.join(str(random.randint(0, 9)) for i in range(40))
        m = hashlib.md5((str(time.time()) + str(random_number)).encode())
        return m.hexdigest()

    # generate an OAuth Authorization header to add to each request
    # see https://dev.twitter.com/docs/auth/authorizing-request
    def generate_authorization_header(self,method,url,query_parameters):
        nonce = self.generate_nonce()
        s = ""
        params = {}
        for key in query_parameters.keys():
            params[key] = query_parameters[key]

        params["oauth_nonce"] = nonce
        params["oauth_consumer_key"] = consumer_key
        params["oauth_token"] = access_token
        params["oauth_signature_method"] = "HMAC-SHA1"
        params["oauth_version"] = "1.0"
        params["oauth_timestamp"] = str(int(time.time()))

        sortkeys = [k for k in params.keys()]
        sortkeys.sort()
        for k in sortkeys:
            if s != "":
                s += "&"
            s += quote(k,'')
            s += '='
```

```
            s += quote(params[k],'')

        base_string = quote(method,'')+"&"+quote(url,'')+"&"+quote(s,'')

        signing_key = consumer_secret+"&"+access_secret

        tok =
base64.standard_b64encode(hmac.new(signing_key.encode(),base_string.encode(),sha1).digest()).decode('
ascii')

        params["oauth_signature"] = tok

        auth_header = "OAuth "
        auth_keys = [k for k in params.keys()]
        auth_keys.sort()
        first = True
        for k in auth_keys:
            if k.startswith("oauth"):
                if not first:
                    auth_header += ", "
                auth_header += k
                auth_header += '="'
                auth_header += quote(params[k])
                auth_header += '"'
                first = False
        return auth_header

    def sample(self):
        url = "https://stream.twitter.com/1.1/statuses/sample.json"
        query = {}
        self.start_time = int(time.time())
        while True:
            try:
                auth_header = self.generate_authorization_header("GET",url,query)
                conn = HTTPSConnection("stream.twitter.com")
                logging.getLogger("twitstream").debug("calling:
https://stream.twitter.com/1.1/statuses/sample.json")
                conn.request("GET","/1.1/statuses/sample.json",None,{'User-
agent':'Mozilla/5.0','Authorization':auth_header})
                self.stream(conn)
            except Exception as ex:
                logging.getLogger("twitstream").error(str(ex))

    def filter(self):
        url = "https://stream.twitter.com/1.1/statuses/filter.json"
        query = {}
        querystring =""
        if self.track:
            query["track"] = self.track
            querystring += "track="+quote(self.track)

        if self.locations:
            query["locations"] = self.locations
            if querystring:
                querystring += "&"
            querystring += "locations="+quote(self.locations)

        self.start_time = int(time.time())
        running = True
        while running:
            try:
                auth_header = self.generate_authorization_header("POST",url,query)
                logging.getLogger("twitstream").debug("calling:
https://stream.twitter.com/1.1/statuses/filter.json?"+querystring)
                conn = HTTPSConnection("stream.twitter.com")
                conn.request("POST","/1.1/statuses/filter.json?"+querystring,"",{'User-
agent':'Mozilla/5.0','Authorization':auth_header})
                running = self.stream(conn)
            except Exception as ex:
                logging.getLogger("twitstream").error(str(ex))

    def stream(self,conn):
        resp = conn.getresponse()
        data = bytes()

        while True:
            ready = select.select([conn.sock],[],[],90.0)[0]
```

```
            if not ready:
                # twitter api is designed to send a dummy message every 30 seconds
                # but we have not recieved anything in 90 seconds, timeout reading and restart the
connection
                logging.getLogger("twitstream").error("timeout - retrying connection")
                return True
            newdata = resp.read(65536)
            data += newdata
            pos = data.find(b'\r\n')
            while pos > -1:
                line = data[:pos]
                data = data[pos+2:]
                if line:
                    try:
                        j = line.decode("utf-8")
                        status = json.loads(j)
                        if "text" not in status:
                            if "delete" not in status:
                                logging.getLogger("twitstream").info("not a
status?:"+json.dumps(status))
                        else:
                            self.write(status)
                        if options.maxtweets and self.count > options.maxtweets:
                            logging.getLogger("twitstream").info("collected
"+str(options.maxtweets)+",terminating")
                            return False
                    except Exception as ex:
                        logging.getLogger("twitstream").error(str(ex))
                pos = data.find(b'\r\n')


    # call the twitter search API to fetch tweets matching search term
    def start(self):
        if self.track or self.locations:
            self.filter()
        else:
            self.sample()


    def write(self,r):
        obj = {}
        for (col,decoder) in self.columns:
            try:
                if decoder:
                    obj[col] = decoder(r)
                elif col in r:
                    obj[col] = str(r[col])
            except:
                obj[col] = None
        self.formatter.write(r,obj)
        self.count += 1

        if self.options.interval:
            t = int(time.time())
            if self.checkpoint_time == 0:
                self.checkpoint_time = t

            lastinterval = (t - self.checkpoint_time)
            interval = (t - self.start_time)
            if lastinterval > self.options.interval:
                rate = self.count / interval
                recentcount = self.count - self.checkpoint_count
                lastrate = recentcount / lastinterval
                self.checkpoint_time =  t
                self.checkpoint_count = self.count
                logging.getLogger("twitstream").info("recent: %d records in %d secs (%.2f records per
second).  overall: %d records in %d secs (%.2f records per
second)."%(recentcount,lastinterval,lastrate,self.count,interval,rate))

if __name__ == '__main__':

    if consumer_key == "" or consumer_secret == "" or access_token == "" or access_secret == "":
        print("Error - please define the variables consumer_key,consumer_secret,access_token and
access_secret at the start of this program")
        sys.exit(-1)
```

```
    parser = argparse.ArgumentParser(description="stream tweets from the twitter streaming APIs",
usage="python3 twitstreamer.py")

    parser.add_argument('-t', "--track", dest='track', type=str, help='track option to filter tweets,
for example to recieve whiskey related treets, -t=whiskey for more information see
https://dev.twitter.com/docs/streaming-apis/parameters#track')
    parser.add_argument("-v","--verbose",dest="verbose",action="store_true",help="display verbose
messages")
    parser.add_argument("-i","--interval",dest="interval",type=int,default=300,help="define number of
seconds interval for reporting statistics")
    parser.add_argument("-l","--locations",dest="locations",type=str,help="supply location filter in
form of a bounding box lon_sw,lat_sw,lon_ne,lat_ne (example for London: -l=-
0.563000,51.280430,0.278970,51.683979 for more information see
https://dev.twitter.com/docs/streaming-apis/parameters#locations",default="")
    parser.add_argument("-f","--format",dest="format",type=str,help="supply format as json or
csv",choices=["csv","csv_noheader", "json", "raw"], default="csv")
    parser.add_argument("-m","--max",dest="maxtweets",type=int,help="limit the number of tweets
retrieved to the specified number")


    options = parser.parse_args()

    if options.verbose:
        logging.getLogger("twitstream").setLevel(level=logging.DEBUG)
    else:
        logging.getLogger("twitstream").setLevel(level=logging.INFO)

    handler = logging.StreamHandler(sys.stderr)
    handler.setFormatter(logging.Formatter("%(asctime)s - %(name)s - %(levelname)s - %(message)s"))
    logging.getLogger("twitstream").addHandler(handler)
    tw = twitstream(options)
    tw.start()
```

Appendix 4: Tweet frequency trends

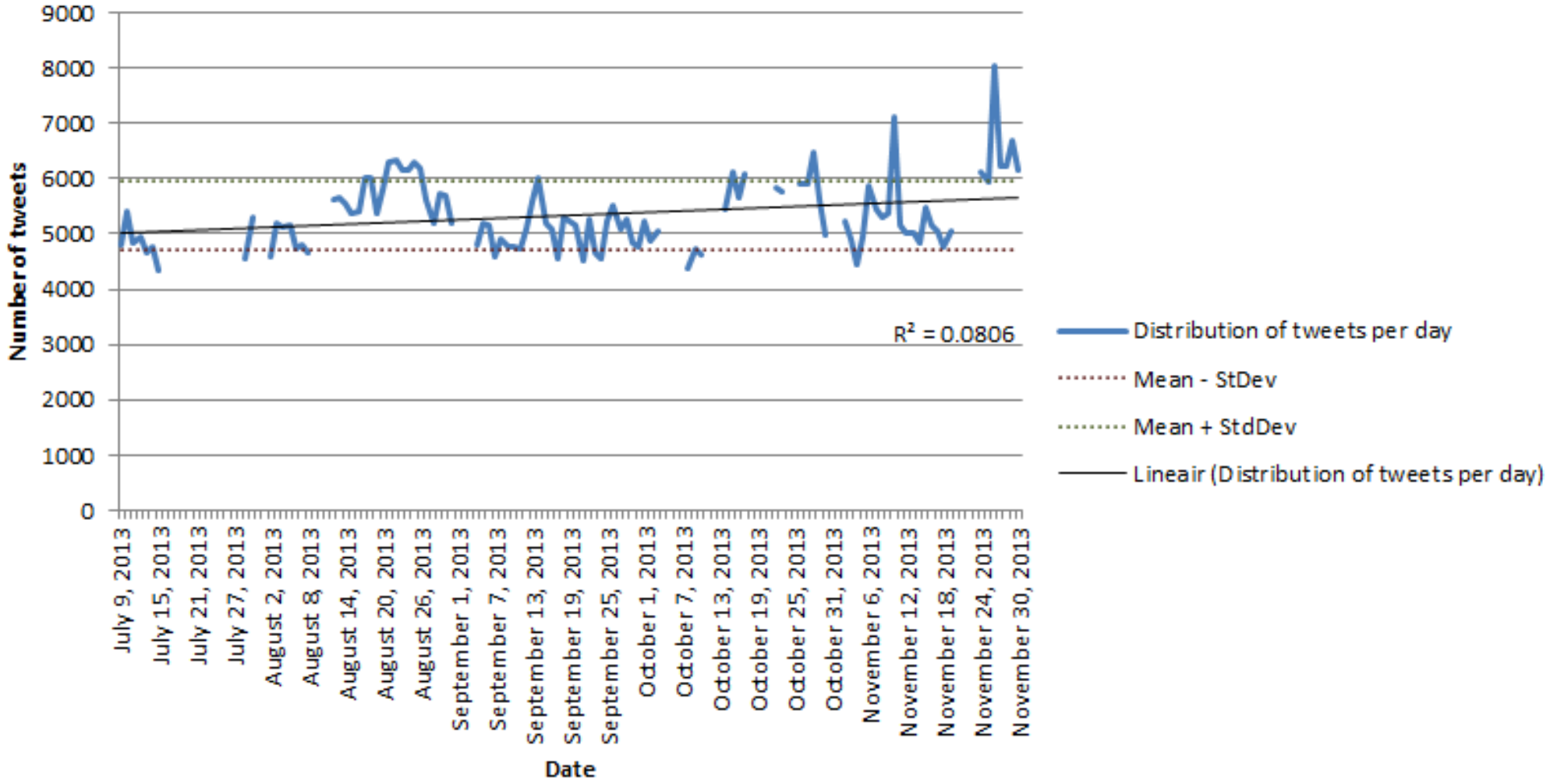


Figure 32 Tweet frequency (tweets per day) trend for the study area.

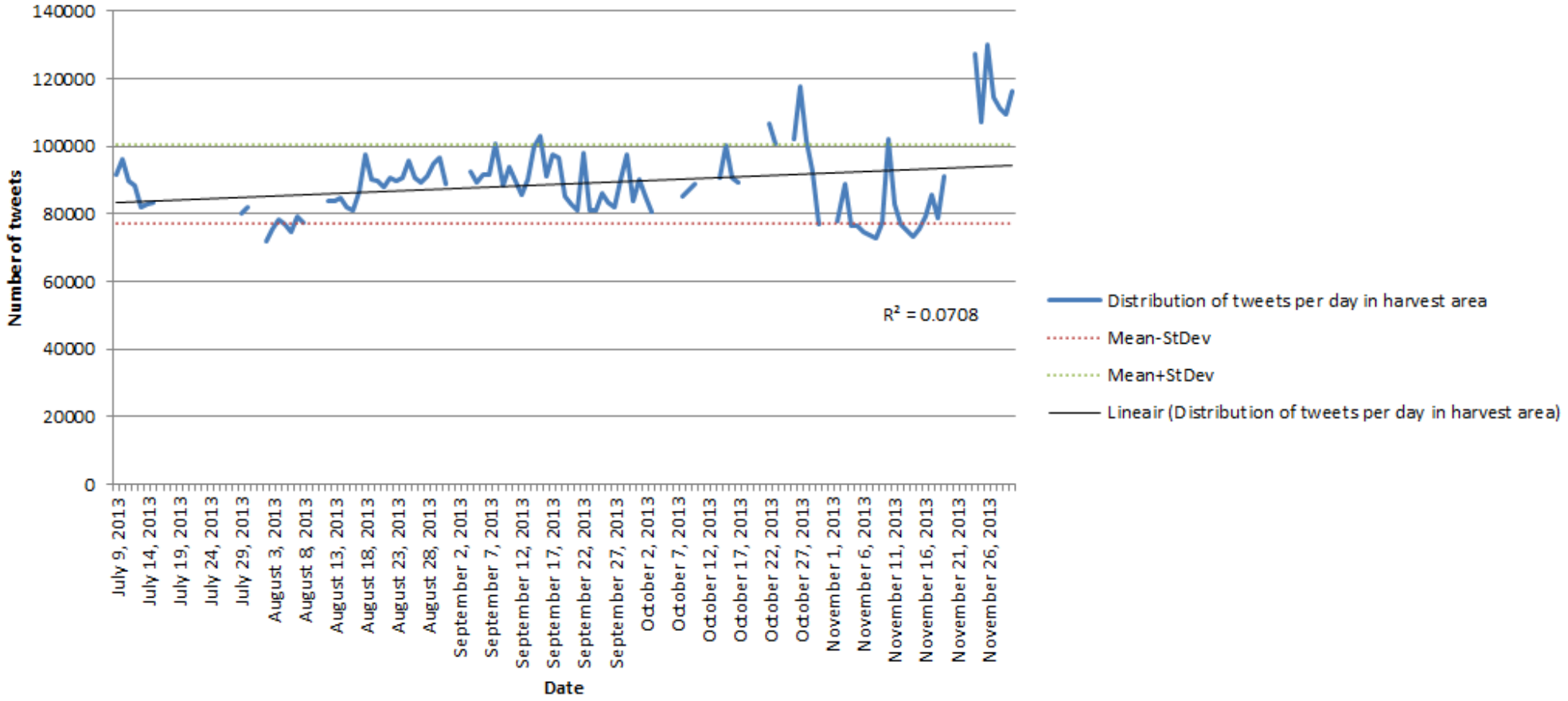
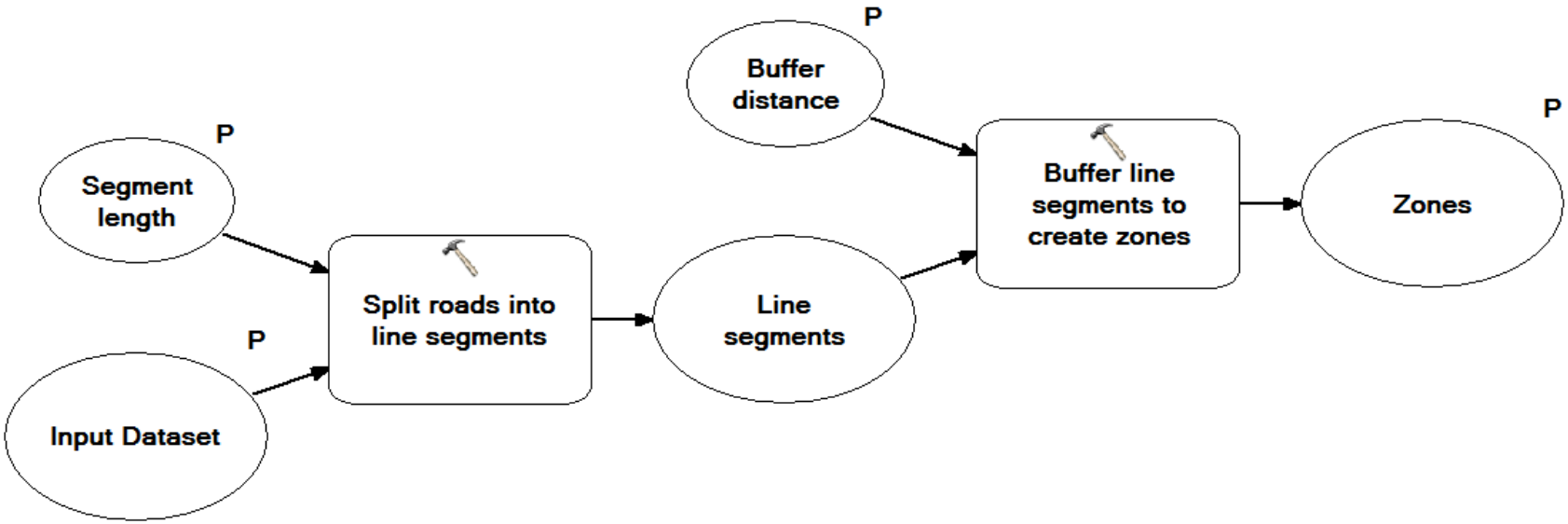


Figure 33 Tweet frequency (tweets per day) trend for the harvest area.

Appendix 5: CreateZones model



Appendix 6: ClusterTweets.py

```
#-------------------------------------------------------------------------------
# Name:        GroupTweetsTemporal
# Purpose:     Group tweets based on their date/time
#
# Author:      Roeland Steur
#
# Created:     17-5-2014
# Copyright:   (c) Roeland Steur 2014
#-------------------------------------------------------------------------------

import arcpy, os, xlrd
# Set workspace and local variables
# The following lines need to be adjusted manually:
#
ws = arcpy.env.workspace = r"D:\Twitter\gdb\ANALYSIS\19_GroupTweetsTemporal\tweet_fc.gdb"
xls = r"D:\Twitter\gdb\ANALYSIS\19_GroupTweetsTemporal\input_where_clause_3.xlsx"
input_Twitter_feature_class = "day_192"
output_database = r"D:\Twitter\gdb\ANALYSIS\19_GroupTweetsTemporal\output.gdb"
#
print "The script will group the tweets in feature class '{0}' into temporal groups based on the
inputs in file '{1}'.".format(input_Twitter_feature_class, xls)
x = raw_input("Check settings in the printed statement and press enter to continue.")
#
arcpy.env.overwriteOutput = True
# Create a list of all values that are in the Excel file
#
inplist = []
#
workbook = xlrd.open_workbook(xls)
worksheet = workbook.sheet_by_index(0)
num_rows = worksheet.nrows - 1
num_cells = worksheet.ncols - 1
curr_row = -1
while curr_row < num_rows:
        curr_row += 1
        row = worksheet.row(curr_row)
        inplist.append(row)
# Remove the first line from the list as it contains the column headers
#
inplist.pop(0)
# Set the values from the list as input values for the Select_analysis tool:
#
for row in inplist:
    where_clause_start_time = row[5].value
    where_clause_end_time = row[6].value
    output_name = row[4].value
    output_file = "{0}/{1}".format(output_database,output_name)
    where_clause = "TIMESTAMP_UTC0200 >= date '{0}' and TIMESTAMP_UTC0200 < date
'{1}'".format(where_clause_start_time, where_clause_end_time)
    # Select the tweets based on the input values from the list, and export them to the output
database
    arcpy.Select_analysis(input_Twitter_feature_class, output_file, where_clause)
    print "The feature class '{0}' was copied to the output database.".format(output_name)

print "done!"
```

Appendix 7: CreateGeoRegularities.py

```
#-------------------------------------------------------------------------------
# Name:        Geographic regularity maker
# Purpose:     Generate results for a sensitivity analysis
#
# Author:      Roeland Steur
#
# Created:     11-04-2014 - 16-5-2014
# Copyright:   (c) Roeland Steur 2014
#-------------------------------------------------------------------------------

import arcpy
from arcpy import env
import os
arcpy.env.overwriteOutput = True

# Set local variables for cleaning the zones
#
zone_workspace =
r"D:\Twitter\gdb\ANALYSIS\16_zonal_irregularity_sensitivity_analysis_development\zone_fc.gdb"
twitter_workspace =
r"D:\Twitter\gdb\ANALYSIS\16_zonal_irregularity_sensitivity_analysis_development\tweet_fc.gdb"
clean_twitter_workspace =
r"D:\Twitter\gdb\ANALYSIS\16_zonal_irregularity_sensitivity_analysis_development\clean_tweet_fc.gdb"
output_workspace =
r"D:\Twitter\gdb\ANALYSIS\16_zonal_irregularity_sensitivity_analysis_development\output.gdb"
excel_output_folder =
r"D:\Twitter\gdb\ANALYSIS\16_zonal_irregularity_sensitivity_analysis_development\output"
input_dataset = "zones3"
output_dataset= "geographic_regularities"
append_dataset = "append_table"
wild_card = ""
field_type = ""
#Set workspace
#
arcpy.env.workspace = zone_workspace
# Give some infromation to the user and let him/her check it.
#
print "For all zones in feature class '{0}' in database '{1}' geographic regularities will be
calculated, based on Twitter data in the database {2}.".format(input_dataset, zone_workspace,
twitter_workspace)
x = raw_input("Press enter to continue.")
print "Script is running..."
# Check if output dataset exist, and if so, delete it
#
if arcpy.Exists(append_dataset):
    arcpy.Delete_management(append_dataset)
model_output = "{0}/{1}".format(output_workspace, output_dataset)
if arcpy.Exists(model_output):
    arcpy.Delete_management(model_output)
# Backup the original input data
#
arcpy.FeatureClassToFeatureClass_conversion(input_dataset, zone_workspace, append_dataset)
# List all fields of the output dataset
#
fields =  arcpy.ListFields(append_dataset)
# Clean the zones feature class table by deleting all fields that are not required
#
print "The following fields will be deleted from the zones file '{0}': ".format(input_dataset)
for field in fields:
    if not field.required:
        print field.name
        arcpy.DeleteField_management(append_dataset, field.name)
# Add and calculate a new field called zone_id to the zones feature class
#
arcpy.AddField_management(append_dataset, "zone_id", "TEXT")
```

```
arcpy.CalculateField_management(append_dataset, "zone_id", "!OBJECTID!", "PYTHON_9.3")
# Change the workspace and define some new variables for the analysis
#
dropFields = ["created_at","text","retweeted","in_reply_to_user_id_str",
"in_reply_to_status_id_str","source","place_full_name","place_place_type",
"lang","id_str","user_name","user_id_str","user_location",
"user_followers_count","user_statuses_count","user_utc_offset","TIMESTAMP_UTC0200","TIMESTAMP_MONTH",
"TIMESTAMP_WEEK","TIMESTAMP_WEEKDAY","TIMESTAMP_YEARDAY", "TIMESTAMP_HOUR", "ET_ID"]
arcpy.env.workspace = twitter_workspace
arcpy.env.overwriteOutput = True
zones = "{0}/{1}".format(zone_workspace, append_dataset)
# List all the feature classes in the workspace
#
featureclassesoutput = arcpy.ListFeatureClasses()
for fc in featureclassesoutput:
    # Join the zone OIDs of the tweet intersecting zones to the tweets
    #
    arcpy.SpatialJoin_analysis(fc,"{0}/{1}".format(zone_workspace,
append_dataset),"clean_{0}".format(fc), "JOIN_ONE_TO_ONE","","", "INTERSECT")
# Dissolve the newly joined twitter feature classes and copy the output to a clean workspace
#
CleanTwitterFeaturesList = arcpy.ListFeatureClasses("clean*")
for fc in CleanTwitterFeaturesList:
    newname =  fc.replace("clean_", "")
    output_clean_tweet_fcs = "{0}/{1}".format(clean_twitter_workspace,newname)
    arcpy.Dissolve_management(fc, output_clean_tweet_fcs, ["zone_id", "user_id_str"])
    arcpy.Delete_management(fc)
# Change workspace
#
arcpy.env.workspace = clean_twitter_workspace
# Create the actual geographic regularities
#
CleanTwitterFeatures = arcpy.ListFeatureClasses()
for fc in CleanTwitterFeatures:
    print fc
    fields =  arcpy.ListFields(fc)
    # Delete all fields that are not editable
    #
    print ">The following fields will be deleted from tweet file '{0}':".format(fc)
    for field in fields:
        if not field.required:
            print field.name
            arcpy.DeleteField_management(fc, field.name)
    # Process: Add Field
    #
    arcpy.AddField_management(fc, fc, "LONG", "", "", "", "", "NULLABLE", "NON_REQUIRED", "")
    # Process: Calculate Field
    #
    arcpy.CalculateField_management(fc, fc, "1", "VB", "")
    # Create a new fieldmappings and add the two input feature classes.
    #
    fieldmappings = arcpy.FieldMappings()
    fieldmappings.addTable(zones)
    fieldmappings.addTable(fc)
    # First get the "fc" fieldmap. "fc" is a field in each twitter feature class.
    # The output will have the geographic zones with the count number of intersecting tweets per
zone.
    #Setting the field's merge rule to sum will aggregate the single tweets into a count number for
each zone.
    #
    FieldIndex = fieldmappings.findFieldMapIndex(fc)
    fieldmap = fieldmappings.getFieldMap(FieldIndex)
    # Get the output field's properties as a field object
    #
    field = fieldmap.outputField
    # Rename the field and pass the updated field object back into the field map
    #
```

```
    field.name = fc
    field.aliasName = fc
    fieldmap.outputField = field
    # Set the merge rule to sum and then replace the old fieldmap in the mappings object with the
updated one.
    #
    fieldmap.mergeRule = "sum"
    fieldmappings.replaceFieldMap(FieldIndex, fieldmap)
    #Run the Spatial Join tool, in order to count all tweets that intersect with a geographic zone.
    #
    arcpy.SpatialJoin_analysis(zones, fc, "regularities_{0}".format(fc), "#", "#", fieldmappings)
    # Add a field to the append_table in order to be able to append the count field to
    #
    # arcpy.AddField_management(append_dataset, fc, "LONG")
    #
    arcpy.JoinField_management(zones, "zone_id", "regularities_{0}".format(fc), "zone_id", fc)
# Copy output table to output workspace in order to safely clean the Twitter workspace
#
arcpy.FeatureClassToFeatureClass_conversion(zones, output_workspace, output_dataset)
print ">The output file '{0}' is copied to the output location
'{1}'.".format(output_dataset,output_workspace)
# Export geographic regularities to an Excel file
#
arcpy.TableToExcel_conversion("{0}/{1}".format(output_workspace,output_dataset),
"{0}/geographic_regularities.xls".format(excel_output_folder))
# Clean the Twitter workspace
#
FeaturestoDelete = arcpy.ListFeatureClasses()
for fc in FeaturestoDelete:
    arcpy.Delete_management(fc)
# Notify user that script completed
#
print ">The twitter workspace '{0}' is cleaned".format(clean_twitter_workspace)
print ">Script completed running!"
```

Appendix 8: Traffic-related tweets

The table below shows traffic-related tweets that are found in the irregularity clusters of the best performing model run of the sensitivity analysis. In left column, text of tweets are listed. In the right column it is listed if tweets meet the criteria for the use cases defined in section 4.1.

Tweets	Meet criteria for UC
File op de busbaan...	no
File bij A9 op de #A2 richting Amsterdam 1 baan open reden snap ik niet	no
Het is druk op de weg..	no
En toen stond ik stil #a10 #ongeluk http://t.co/GGEwxSWdrO	no
Viele, pfoeeee.	no
@#A10 stil voor de coentunnel hmp 26.5	no
We zijn weer gezellig bij amsterdam .. kanker files altijd bij die kankerstad	no
File (@ Brug) http://t.co/5A8b49gGU1	no
Pf ongeluk op snelweg ,	no
Pech klapband A4 #carecaverhuur bij Schiphol http://t.co/yiuzFudjVl	yes
A1 Rut A6 R- RYKSW 290,9 Ongeval/wegvervoer/letsel (beknelling personenauto) (personen in object: 1) (H.E.)	no
Lifeliner vertrokken vanaf vumc naar A6 r bij rutte.	no
Daar sta je dan met pech.. Zwaai even als je lang rijdt #22,6 Re (@ A9) [pic]: http://t.co/Hw73hKGCFM	no
Woehaaaaaa, sta compleet stil al n half uur op de A10	no
Interessante plek om in de file te staan. En ik wilde alleen maar even naar Landmarkt. http://t.co/6N73Sjx35M	no
Pfff 1tunnel dicht gelijk al.het verkeer lam. Zeeburgertunnel waarom.....	no
Jaja nou dat hielp niet want hij was schijnbaar niet de enige met dat idee #meerfile	no
Ben je lekker uit sta je in de file-_- #coentunnel	no
Lekker in de file (@ NH Schiphol Airport - @nh_hoteles) http://t.co/EcqQ4MwCNE	no
Getver file @ Coentunnel http://t.co/Dl3fdFAt4A	no
@A10Verkeersinfo ongeval nu http://t.co/PnWBsHO1KG	no
@A10Verkeersinfo ongeval file http://t.co/NCN2cUONEa	no
Wachten... (@ Brug Over De Ringvaart) http://t.co/qW4Wkzls1I	no
Sta langs de kant van de haarlemmerweg met kapotte ruitwissers vol in een regenbui! Niet echt handig	yes
Auto te water, niet door nieuwe rotonde #amstelveen http://t.co/Sg35SsGDzZ	no
Tering file door die klote coentunnel hoor	no
14:30 Wo 1 Auto te water Van Cleefkade Aalsmeer - Woasv Hvasv http://t.co/nXT6i8KL3N	no
Eerst foute tomtom en nu in de file door waarschijnlijk iets van een ongeluk ofzo	no
Ik haat auto's die langzaam rijden, files, werkzaamheden, en vooral de muziek die Skyradio draait 😩	no
File. Moet plassen. Stom.	no
File À 10 zondagmiddag niet te geloven	no
Shit man :s, file.	no
ruzie midden op de snelweg, auto van die persoon staat stil op de snelweg	no
In de file net onder de landingsbaan van Schiphol, lekker vliegtuigjes met Juma kijken dan maar	no
3 sleepauto's, brandweerauto en al 30 min. stil staan op de afrit A4.....	no
Even lekker uitwaaien @A1 http://t.co/Bcd3NvZPgA	no
Brug open op snelweg.. Blijf dit toch wel erg oldskool vinden, helemaal op nieuwe weg.. #welkeslimmerikbedenktdit http://t.co/TRRJKMeS6U	no
Waarom ik altijd? #A9brugstaataltijdopenalsikerlangswil http://t.co/Zyy1vtzKvE	no
Pff file!	no
File (@ Route Amsterdam - Groningen) http://t.co/O5AxZXtfZC	no
Alles staat open, behalve de carpool strook.... #frak #brugkarma http://t.co/dp3fMiR3UD	no
Dik 20 minuten vertraging om van ring a10 noord de tunnel in te komen, waarom hoor je dat niet bij @anwbverkeer? http://t.co/lBbaAzUw0U	no
@vid @meldkamervid als je uit Je raam kijkt staat het stil richting rottepolderplein...	no
Wat een filee jezakk	no
Filee 🚌	no
Hè, file voor de Coen. Hoe kan dat nou?	no
Lekker in de #file #A9	no
En natuurlijk in de file bij Amsterdam...	no

Appendix 9: Photographs added with traffic-related tweets of Appendix 8

Image 1

Image 2

Image 3

Image 4

Image 5

Image 6

Image 7

Image 8

Image 9

Image 10