COMBINING AUTHORITATIVE AND VOLUNTEERED GEO-INFORMATION TO ANALYZE THE DISTRIBUTION OF TICK BITES

BERIHU ALEMAYEHU GIDEY

March, 2015

SUPERVISORS:

Dr. F.O. Ostermann
Dr. R. Zurita-Milla
Ms. I.Garcia Marti MSc (AOI)

COMBINING AUTHORITATIVE AND VOLUNTEERED GEO-INFORMATION TO ANALYZE THE DISTRIBUTION OF TICK BITES

BERIHU ALEMAYEHU GIDEY

Enschede, The Netherlands, March, 2015

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.
Specialization: Geo-informatics

SUPERVISORS:

Dr. F.O. Ostermann
Dr. R. Zurita-Milla
Ms. I. Garcia Marti MSc (AOI)

THESIS ASSESMENT BOARD:

Prof.dr. M.J. Kraak (Chair)
Ms A. Hofhuis MSc

ABSTRACT

The combination of cloud computing and volunteered geographic information presents a great opportunity to collect, store, process, and disseminate geo-information to understand societal problems related to tick borne diseases in the Netherlands. Wealth of literature showed that with the capability of Web 2.0 technology to handle user generated content, the World Wide Web has become data production environment. It has enabled volunteers to share geographic information that can be collected and used to solve real world problems. However, the quality of this information continues to remains debatable. In addition, cloud computing has been studied by many and found to be an alternative solution to the IT challenges in the field of geo-information science.

This research aims to combine volunteered geo-information and authoritative data in order to improve our understanding of tick bite distribution and tick bite risk and systematically evaluate the capabilities of cloud computing platforms by running the analysis where possible. The goal of combining the volunteered geo-information and authoritative data here is to address the VGI data quality issue whereas the evaluation of cloud platforms is to find out if clouds computing can really be an alternative solution to the IT challenge.

A combination of volunteered tick observations, geolocated *Flicker* photos, and land cover data was mainly used to perform a spatio-temporal analysis to understand the distribution and risks of tick bites. These datasets were first preprocessed, partially cleaned, and prepared for analysis.

The spatial and temporal distribution of the VGI datasets was performed using combination spatial analysis methods such KDE and Getis-ord GI* to identify the locations of hot spots for tick bites as well as supporting the visual analysis of finding relationships between the tick bite VGI data and geolocated photos. The analyses done using the above methods was supported by geovisual analysis in the cloud and spearman's rank correlation method to evaluate the temporal relationships of the two VGI datasets. As a result, we have identified that areas in the west-coastal, central, north eastern, and to a small extent the southern regions of the country to be the locations high incidents of tick bites. Indeed we found out that the areas of the identified hot spots to be strongly related to high vegetation cover which are mostly recreational areas. Also, we found out that the temporal region of high incidents to be in the months of June and July for the years 2011-2013.

The SaasS (CartoDB) was selected using a method called Analytic Hierarchy Process, used throughout the project to perform geovisual analysis to understand the distribution of tick bite incidents, and evaluated using SaaS quality model to establish an understanding on the maturity of the geospatial cloud platforms to support geo-information processing workflows. It was finally found to be a powerful solution for building intuitively understandable, easily sharable, dynamic, and interactive geovisualization products.

Keywords

Volunteered geographic information, authoritative data, cloud computing, SaaS, spatio-temporal analysis, geovisualization, geolocated photos, tick bite hot spots

ACKNOWLEDGEMENTS

First and foremost, I must acknowledge and thank my employer the FDRE, Information Network Security Agency for giving me this paid scholarship to peruse my study. It might have been an easy decision to fund the scholarship for the organization, but for me, it was one of the best opportunities I got in life.

I would like to express my special appreciation and thanks to my first supervisor Dr. F.O. Ostermann for his useful guidance, comments, motivation, and encouragements throughout the thesis work. I will always remember the moments you told me to "think simple" as I came to learn that "simple solutions are always beautiful". Furthermore, I would like to thank my second supervisor Dr. R. Zurita-Milla for his support and guidance.

I would like to thank Ms. I. Garcia Marti for her valuable comments, guidance, and support throughout the thesis work. I would also like to express my heartfelt gratitude for being so generous to give me the *"Natural language processing script"* she developed to use it in my work.

I must admit, I feel privileged to have met Dr. Tagel Gebrehiwot in my stay in the Netherlands. You were like an elder brother who is always there for his younger brother no matter what. What can I say Tagel? I know thank you is not enough for what you did for me, I wish I knew another word to express what I feel, but thank you. Thank you very much for everything.

Last but not least, I would like to thank my friends Tsegay Gebremedhin, Tesfay Kidanemariam, J.F. Beauprè, Jovani Yifru, and Adugna Girma for the wonderful time we had together. I will not forget the moments I had with you guys and the lessons I learned from all of you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Acronyms and definitions

Acronyms

API	Application Programming Interface
ESA	European Space Agency
AHP	Analytic Hierarchy Process
ESRI	Environmental Systems Research Institute
ISO	International Organization for Standardization
CSS	Cascading Style Sheet
CPD	Contextual Photo Density
JSON	JavaScript Object Natation
JSONP	JavaScript Object Natation with Padding
TBD	Tick Bite Density
CRS	Complete Randomness Hypothesis
KDE	Kernel Density Estimate
SaaS	Software as a Service
PaaS	Platform as a Service
IaaS	Infrastructure as a Service
SQL	Structured Query Language
REST	REpresentational State Transfer
XML	Extensible Markup Language
VGI	Volunteered Geographic Information

1. INTRODUCTION

1.1. Motivation and problem statement

The combination of cloud computing and volunteered geographic information presents a great opportunity for the collection, storage, processing, and dissemination of geo-information to solve societal problems. While cloud computing can provide the required resources when there is a limited geo-data infrastructure and the resources to maintain and expand them are scarce, volunteered geographic information on the other hand, can provide an up-to-date geo-information when there is a limited geospatial data to solve societal problems. The focus of this research is then in the line of using combination of volunteered geo-information and cloud computing to understand a health problem linked to tick bites in the Netherlands.

The consultations to general practitioners for tick bites and Lyme disease has increased three times in the years 2004 to 2009 in the Netherlands (Sprong et al., 2012). If infected by Borrelia bacterium, ticks can transmit Lyme disease, which is one of the infectious disease in both humans and animals in Europe (Vinh et al., 2014). Although several efforts were made to prevent human exposure to tick bites and promote timely removal, the Lyme disease infection incidents have continued (Sprong et al., 2012) . One of the efforts was involving the public to gather the information regarding the incidents using web 2.0 based application which will be discussed later in this chapter.

The development of the Web 2.0 framework has revolutionized the World Wide Web taking web application design and implementation a long way to rich internet application development (O'Reilly, 2007). Dynamic web sites developed in this framework became applications that serve as data entry and retrieval platforms. Indeed, these applications have now widely become platforms for handling user generated content (De Longueville, 2010).

The capability of the web 2.0 environment fueled crowdsourcing which is defined by Brabham (2008) as "*online, distributed problem-solving and production model*", one branch of which is the volunteered geospatial information (VGI). Volunteered geospatial information according to Goodchild (2007) is a profound transformation in how geographic data, information, and knowledge are produced and circulated. In this regard, geospatial data content can be obtained from geotagged social media content and used in different application domains. One major challenge in the area, however, is the data quality and the credibility of the data collected by volunteers. The observations can be biased and incomplete resulting in substantial information gap in the theme.

Geo-information scientists and analysts are facing information technology challenges in a massive scale due to the data volume, processing power, and spatio-temporal nature of geospatial information (Yang et al., 2011). The development and maintenance of spatial data infrastructures for storage, processing, analysis and dissemination to deal with this technology challenges is both expensive and difficult. As a solution to this challenge, it is preferable to outsource all or some part of the infrastructure component to organizations that specialized in building and maintaining cloud computing services. This allows cloud-based solutions to be developed using one of the cloud computing service levels such as PaaS (platform as a service) for developing and deploying solutions (Google developers academy ,2012) and SaaS (software as a service) for performing business functions.

One solution that leverages the "*distributed problem solving*" and "*online production model*" of crowdsourcing is the *tekenradar*[1] application. It was launched in 2012 and is being used as VGI data collection and dissemination platform as part of the research aimed at preventing the Lyme disease (Wageningen University, de Natuurkalender, & RIVM, 2012) in the Netherlands. The *tekenradar* application enables volunteers to report tick bites, and the authorities to disseminate information to the public. As explained earlier, the data collected from volunteers is assumed to be biased, in line with the data quality and credibility challenge of VGI.

To provide an effective solution to the societal problem, the two challenges outlined in the previous section must be addressed. Firstly, the bias in the volunteered observations should be assessed and minimized to possible minimum level. That is, the information gap that could arise as a result of the bias in the volunteered observations of tick bites should be filled before drawing conclusions and making decisions. Secondly, the maturity of geospatial cloud platforms in addressing the information technology challenge should be understood so that organizations can alternatively move to the cloud. This research is then aimed at addressing these challenges by combining data from different sources to improve our understanding of the tick bite distribution and evaluating cloud solutions throughout the process.

1.2. Research objectives

1.2.1. Overall objective

The main objective of this project is to combine social media content, tick bites observations collected by volunteers, and authoritative data in order to improve our understanding of tick bite distribution and tick bite risk. A secondary objective is to systematically evaluate the capabilities of cloud computing platforms to support this analysis and implement it where possible.

1.2.2. Specific objectives

1. Analyze the spatio-temporal distributions of reported tick bites and related social media data.
2. Investigate possible relationships between volunteered tick bite observations, contextual social media data and authoritative land cover data to improve understanding of tick bite risk.
3. Systematically identify and evaluate cloud platforms for supporting implementation of geospatial analyses to achieve objectives 1 and 2.

1.3. Research questions

1. How does the spatiotemporal distribution of the tick bite observations look like?
2. What are the relationships between the reported tick bites and the land cover?
3. How are observations of tick bites and related social activities represented in social media?
4. What are the relationships between reported tick bites and activities, as reported in social media?
5. How can the triangulation of data sources help in improving our understanding of tick bite risk by discovering hidden patterns?
6. Which functionalities do cloud computing platforms need to offer for this research?
7. To what extent did the use of cloud computing platforms improve the feasibility of the main tasks of this research?

[1] The *tekenradar* application can be accesses using the URL (http://www.tekenradar.nl/). Use Google Chrome browser and its automatic translation functionality to be able to read the content in English.

1.4. Innovation aimed at

This research aspires at using geolocated social media content to improve our understanding of the spatio-temporal distribution of tick bite incidents. Indeed, it aims at understanding and identifying the land cover and social activities that are related to high risk of tick bite and finally tries to evaluate the capability of geospatial cloud based solutions for implementing similar geospatial workflows.

1.5. Methodology adopted

The methodology that is adopted to conduct this research project is depicted in Figure 1. The main tasks, sequence of activates, input data, and intermediate outputs are described in the section that follows.

1. **Literature review**: At this stage, literature related to VGI, VGI data quality issues and cloud computing platforms for geo-information and available solutions were reviewed. In addition, social media platforms and the possible ways to collect data from these platforms was studied in detail. As a result of this task, the areas were well understood and the social media platforms were identified and data collection method defined.

2. **Social media data collection**: To collect the social media content, social media platforms that were studied in the previous stage such as *Twitter*[2] for geotagged textual data, *Flickr*[3] for geotagged photos were considered. The platforms have Application Programming Interfaces (API's) to programmatically collect and manipulate data. These API's (especially the GeoAPI's) provided by the platforms were also studied in parallel. In addition to this platforms, the *tekenradar* application was also studied to support the development of search vocabulary. The search vocabulary for harvesting the data represents the environments as well as outdoor activities that are potentially related to tick bites was developed. Finally, social media data harvest script to collect data reported from locations within the minimum bounding box of the Netherlands was developed and data collected from the selected source (*Flickr*) by using the *Flickr* photo search API's and Python script.

3. **Data preprocessing**: The first task executed in the data preprocessing stage was that both observations (tick bites and geolocated *Flickr* photos) that are located within the administrative boundary of the Netherlands were extracted. Both the datasets were first explored to find out if they can be used to identify patterns. In the process, the *Flickr* photos collected were observed to have a lot of noise and was then partially cleaned to minimize the noise.

4. **Analysis method and platform selection:** At this stage, the analysis methods to understand the datasets, analyze spatio-temporal distribution of tick bites, and comparing the tick bite observations and social media extract are studied and selected. The Kernel Density Estimation (KDE) method (Gatrell, Bailey, Diggle, Rowlingson, & Rowlingsont, 1996) was selected as a first order point pattern analysis method for analyzing the distribution of the datasets to identify hotspots. Another method selected to analyze the statistical significance of the hot spots identified in the first order point pattern analysis, was the Getis-Ord GI* (Ord & Getis, 2010) method. To understand the relationships of the each dataset to the land cover an overlay analysis in ESRI's ArcMap[4] to extract the actual land cover information was used, the result of which was an intermediate data for further analyses. Spearmann's rank correlation (Prion & Haerling, 2014) was also used to evaluate the association between the two VGI datasets, tick bite and photo data. Furthermore, the cloud solution for implementing the selected

[2] *https://twitter.com/*
[3] *https://www.flickr.com/*
[4] *http://www.esri.com/software/arcgis/arcgis-for-desktop*

processes there by performing the systematic evaluation was selected using Analytic Hierarchy Precess (AHP) (Godse & Mulik, 2009) for selecting software as a service as guiding principle

5. **Data analysis:** The tick bites dataset and the social media extract are analyzed for determining the spatio-temporal distribution of each. Furthermore, the datasets (social media extract and tick observations) were integrated with information from land cover and were analyzed separately and with respect to each other. Here, the method(s) selected in the previous stage were applied to perform the analyses. The anticipated result of the analyses is that it will be possible to improve our understanding of the tick bite distributions using both land cover and social media content there by identifying the missing information. The process of combining information from the three datasets (tick bite observations, contextual photos and land cover) is what we call in this project the "data triangulation" process. A sub methodology that I call data triangulation process is applied at the analysis stage of the main methodology (Figure 1). First, actual land cover of the tick bite observations and the contextual photos is extracted from the land cover data by intersecting each with the land cover. The individual environmental data is then extracted for each dataset using their land cover value. In addition to the environmental data, the potential outdoor activity data from the contextual photos plus the actual land cover information was extracted in the same way.

The data per each VGI dataset per each environment as well as the outdoor activity data of the contextual photos were aggregated using the same aggregating units for further analyses

The underlying VGI data that is used as an aggregate was used for correlation analysis to evaluate temporal relationships of the two datasets. Each of the two potentially related data are again aggregated using the same temporal resolution, plotted against each other and evaluated using the selected correlation analysis method.

6. **Evaluate the cloud platform(s):** The evaluation of cloud platforms starts in the selection process and continues through the analysis to the implementation of a prototype in the selected platform (SaaS in this case). Several geospatial-cloud platforms were identified and systematically evaluated to find the one the suits the requirement of the project. Finally, part of the analyses and the information dissemination or geovisualization prat of the process was implemented in the selected cloud platform. As a result of the process, the maturity of the geospatial cloud environment was studied and understood on "to what extent it can support such geospatial workflows".

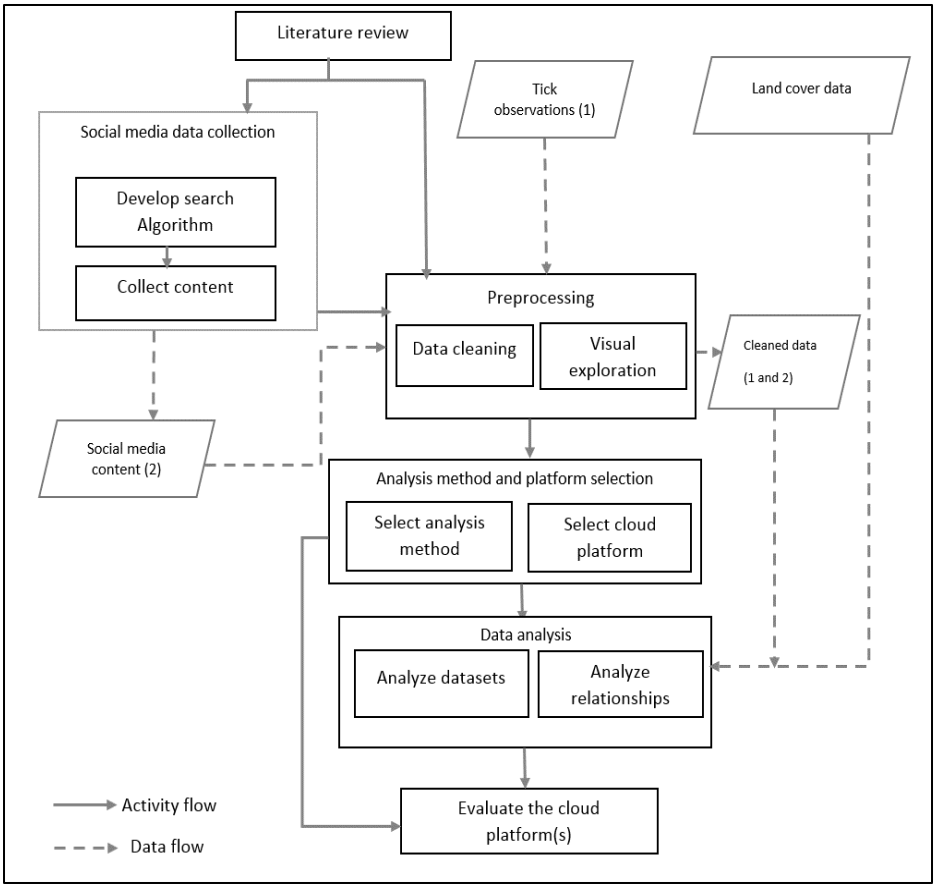


Figure 1. Adopted Methodology

1.6. Structure of the thesis

This thesis is composed of seven chapters. Chapter one describes the motivation and problem statement, research objectives, research questions, the innovation aimed at, and the methodology applied. Chapter two gives a brief overview of the related work to the research. Chapter three explains the data used in the research and the data preparation process. Chapter four describe the methods applied to answer the research questions and tools used to generate results. Chapter five presents the results obtained in the research. Chapter six discusses and explains the meaning and importance of the findings. The final chapter, chapter seven, presents the conclusions and recommendations.

2. LITERATURE REVIEW

2.1. Volunteered geographic information (VGI)

Geographic data and information have been produced and used by small groups of specialized professionals for specialized purposes for a long time (Brown et al., 2013). Over the years when geographic data and geographic information (GI) have been especially used for military and government consumption, it had been available for experts who are capable of using specialized GI tools (Brown et al., 2013). The advancement in web technologies and production of consumer focused GI tools made it possible for GI to be shared over the internet and widens the GI user community (Brown et al., 2013).The web continued to advance to the concept of "web 2.0" which turned the web into a platform which enables users to create their own content (O'Reilly, 2009). This unprecedented transformation of the web in to information sharing platform has indeed fueled the emergence of VGI. VGI according to Goodchild (2007) is special case of user generated content in that the data contributed in this case has the geographic location of the theme for which the data is produced .

According to (Sui, Elwood, & Goodchild, 2012)," *the phenomenon of volunteered geographic information is part of a profound transformation in how geographic data, information, and knowledge are produced and circulated*". There are different motivations for the production and forms of production of VGI. The motivation of individual contributors can be both positive and negative (Coleman, Georgiadou, Labonte, Observation, & Canada, 2009). Positively motivated contributors share information because of motivating factors such as helping others, professional interest, social reward and pride of a place (Coleman et al., 2009). Although very limited and less important, there are also negatively motivated contributors that do so for mischief, and criminal intent.

For positively motivated contributors several platforms were developed over the years. To produce geographic information about verifiable facts on the ground for example, OpenStreetmap[5] (which aims to create a free digital map of the world) and is implemented through the engagement of participants in a mode similar to software development in Open Source projects (Haklay, 2010)) and Google Map Maker[6] are enabling volunteers to produce geographic data and information. Another form of VGI platforms are social media platforms such as Flickr and Twitter in combination with the advancement of location enabled smart devices that accompany people's lives (Caverlee, Cheng, Sui, & Kamath, 2013). VGI from this platforms is available as "geo-social footprints" (Caverlee et al., 2013) of the people using these social media platforms. The data collected from these volunteers can be used for different applications as they represent social and spatial contexts.

Several studies have been conducted in using VGI to solve societal problems and solutions have been developed as well. These studies mainly focus on technology solutions for the collection, storage, and dissemination of VGI and such solutions are being developed and used widely. There are many applications that use VGI or crowdsourcing. One notable group of such applications cover the area of disaster management such as early warning systems (Sweta, 2014), pervasive heath computing solutions (Mooney, Corcoran, & Ciepluch, 2012), urban evacuation system for risk minimization (Oxendine & Waters, 2014), and agent-based indoor evacuation simulation (Goetz & Zipf, 2012). Other applications aimed at helping individuals that use user generated content include travel route recommendation using geotagged photos (Kurashima, Iwata, Irie, & Fujimura, 2010), characterization of urban landscape using geolocated tweets

[5] *http://www.openstreetmap.org/#map=5/82.569/-4.834&layers=T*
[6] *http://www.google.com/mapmaker*

(Frias-Martinez, Soto, Hohwald, & Frias-Martinez, 2012), mining tourist information from geolocated photos (Kurashima et al., 2010).

Although several researches on using VGI for solving societal problems are being conducted and solutions continue to be developed, the quality and credibility of VGI data remains debatable (Heipke, 2010). This is true especially when the data is obtained from individual volunteers and geotagged social media content. A recent study on mapping of the data shadows of hurricane sandy (Shelton, Poorthuis, Graham, & Zook, 2014) showed that relying only on the location content of geotagged social media data does not give a complete understanding of real world incidents.

Different researchers (Haklay, 2010; Hauff, 2013; Vandenbroucke, Bucher, & Crompvoets, 2013; Zielstra & Hochmair, 2013) tried to show the positional accuracy of user generated content which can introduce positional bias in to our understanding of geospatial phenomena. Several methods have been proposed to assess the quality of VGI. These methods include conceptual workflow for automatic assessment of VGI (Ostermann, 2011), photogrammetric approach to assess the quality of VGI (Canavosio-Zuzelski, Agouris, & Doucette, 2013) and automated matching procedure for assessing data completeness (Koukoletsos, Haklay, & Ellul, 2012).

Researchers proposed methods to enhance the data quality of VGI and improve the information content such as data cleaning (Xinlin Qian et al., 2009).Other approaches include crowd-sourcing, social and geographic approaches (Goodchild & Li, 2012) .

The combination of VGI data that is produced by volunteers participating as a social endeavor and geolocated social media footprints could give a better understanding than they can provide when analyzed independently. This is possible only if the two are related to each other both in space and time. To understand the distribution of tick bites, a growing social problem associated to tick-borne disease (in concrete, Lyme disease) in the Netherlands (Sprong et al., 2012), the two types of VG, social endeavor and social media, are used in this research. Also, using the concepts of VGI data cleaning to reduce the noise in both datasets is applicable.

2.2. Cloud computing

Cloud computing is a new computing model which refers to software applications delivered as services over the internet and the hardware devices and system software that host these application in remotely located data centers. The technological advancement and the in the processing and storage devices and the success of the internet are the driving force for this model (Q. Zhang, Cheng, & Boutaba, 2010). According to the National Institute for Standard and Technology (NIST[7]) (NIST, 2011) "*Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources*". These configurable computing resources are networks, servers, storage, applications and services.

Cloud computing is not as such a new technology in the computing industry. Most of the technologies used in cloud computing such as virtualization and utility computing had existed before (Q. Zhang et al., 2010). The innovation that cloud computing brought about is rather a new operational model that brings together existing technologies to run business differently. The technologies that made a cloud computing a reality are (Q. Zhang et al., 2010) *Grid computing*, *Virtualization*, *Utility computing*, and *Automatic computing*.

The essential characteristics of cloud computing as outlined by NIST (NIST, 2011) such as on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured services are inherited from the enabling technologies one way or the other. The on-demand self-service and rapid elasticity are basically

[7] *National Institute of Standard and Technology (http://www.nist.gov/)*

related to the concept behind utility computing and automatic computing respectively. Resource pooling in cloud computing is related to the distributed computing paradigm that coordinates network resources to achieve a specific goal. The computing in the cloud model is highly associated with virtualization in that it encapsulates the details of the hardware and networks and enables users to use software as utility.

Cloud computing services can be consumed by information technology (IT) consumers in either Software as a Service (SaaS), Platform as a Service (PaaS), or Infrastructure as a Service (IaaS) of the cloud computing service delivery models (NIST, 2011). SaaS also sometimes referred as Application as a Service refers to the multi-tenant platform where by common computing resources and a single instance of the application code and the underlying database are used by multiple customers simultaneously. The delivery mode PaaS refers to a platform that includes the systems and environments that enable developers to develop, test, deploy and host web based applications in the cloud. The computing resources provided under the IaaS delivery mode include the underlying hardware and network resources in the data centers on which cloud computing consumers can deploy and run any arbitrary software including operating systems and applications (NIST, 2011) .

These service models can be deployed as private, community, public, and hybrid deployment models (NIST, 2011). These different deployment models give an opportunity to consumers to choose one that that suits their needs. If a consumer for example, is concerned about the security and ownership of the data that is used in the cloud platform they have the alternative to deploy the cloud platform on their private network behind their own firewalls.

With the added advantages of cloud computing, there are also associated security and privacy challenges with this paradigm (Takabi, Joshi, & Ahn, 2010). The security challenge posed here are mainly the result of the unique architecture. The multi domain nature of cloud computing requires different security and privacy policies. As a result there is a noble frustration in consumers. That is, as the data and resources are located in remote locations (Takabi et al., 2010), there is no guarantee that service provider do not use the data for their own purposes in a way customers would not allow them to.

2.3. Cloud computing for Geosciences

Cloud computing for geospatial information has received a special attention in the research arena as the information technology infrastructure has continued to be a big challenge in the geospatial landscape. Researches such as Yang and collegues (2011) suggested that geospatial sciences have the capacity to shape the cloud computing because of the inherent characteristics of spatial data. The inherent characteristics that pose information technology challenge according to (Yang et al., 2011) are data intensity, computing intensity, concurrent access intensity and spatiotemporal intensity. In this regard, research studies (Huang, Yang, Nebert, Liu, & Wu, 2010; Yang, Raskin, Goodchild, & Gahegan, 2010; J. Zhang, 2010) suggested that one or more of cloud computing services in one of the deployment models can help solve the IT challenge.

Furthermore, researches have been conducted on using the cloud to implement geospatial workflows (Ji, Chen, Huang, Sui, & Fang, 2012) to manage geospatial processes for spatial analysis and decision support.

Even though cloud computing is a highly promising computing resources acquisition model for solving the IT constraint in Geosciences, there are also challenges (Dillon, Wu, & Chang, 2010) linked with it. The major challenges are security, costing model, charging model, service level agreement, migration (what to migrate), and interoperability issues (Dillon et al., 2010).

For public and hybrid clouds, cloud computing consumers store their data and run their applications using cloud computing service provider's resources in the service provider's premises. This makes it difficult for them to have little or no control over their data and applications. This may compromise the integrity, confidentiality and privacy of the data and services (Sinanc & Sagiroglu, 2013). One important security risk

that could happen in this regard is data leakage (Sinanc & Sagiroglu, 2013) that can have grave consequences. Solutions to overcome the challenges of cloud computing (Boampong & Wahsheh, 2012; Sinanc & Sagiroglu, 2013) security have been proposed for general purpose computing workflows.

The ubiquitous availability of PaaS solutions has proven that conventional geoprocessing functions can be migrated into the cloud (Yue, Zhou, Gong, & Hu, 2013). Developer can use the platforms to implement the geoprocessing algorithms in the proprietary GIS systems. This suggests that SaaS platforms that could be used to solve geospatial problems can be developed and made available.

There are several geospatial SaaS platforms serving different purposes. This applications include MangoMap[8], CartoDB[9], Geocommons[10], eSpatial[11], and MapCentia GC2[12] to name some. These powerful web mapping platforms provide their services to many customers. However, to my knowledge, there is no established knowledge on how much these geospatial SaaS applications are capable of supporting geospatial workflows from functionality, security, and availability, cost, and response time stand points. In this research we will systematically select geospatial SaaS platforms that fit the requirements for spatio-temporal analysis of VGI data related to tick bites in the Netherlands. Furthermore, the analyses processes and results of the research will be implemented in the selected cloud platform for evaluating the capabilities to support establish the knowledge in this area.

[8] *https://mangomap.com/*
[9] *http://cartodb.com/*
[10] *http://geocommons.com/*
[11] *https://www.espatial.com/*
[12] *http://www.mapcentia.com/en/geocloud/*

3. DATA

3.1. Overview

To take measures and develop awareness programmes in preventing tick-borne diseases such as Lyme, it is important to have a complete understanding of the environmental and social factors that are associated to tick bites. That is, having an actionable information on the high risk environments and social activities for which tick bites are linked to is vital in taking actions in protecting the incidents. Actionable information in this sense is that information which can give both inhabitants and authorities the possibility of taking informed decisions. To achieve this goal, combining data from different sources to understand the spatio-temporal distribution of tick bite incidents, social activities linked to tick bites and the number of people that are vulnerable to tick bites is essential. It is equally important to involve the community in collecting the data voluntarily and make the analyzed information easily reachable and intuitively understandable by the general public. That is what this research is aiming to achieve as stated in its objective. Following the primary objective of this research, three main datasets are considered. This datasets are:

1. The tick bite observations, from 2006-2014
2. Geolocated photos extracted from Flickr[13], from 2011- October,2014
3. Land cover data extracted from GlobCover (Bontemps et al., 2011) produced for the year 2009 obtained from European Space Agency[14] (ESA)

The primary dataset here is the tick bite observations dataset. The other two datasets are auxiliary datasets used to improve and understand the primary dataset. Another data that is used in this research for looking into the number of people that are susceptible to tick bite per each municipality is the official population data. The remainder of this section describes the details of the main three datasets separately.

3.2. Tick bite observations (tekenradar) dataset

The tick bite observations are collected by volunteers using the *tekenradar* application. The application provides volunteers with a step-by-step wizard to report tick bite incident. Volunteers report the environment in which they got the tick bite, the outdoor activity they were involved in, the date of the incident, the location of the incident, and other additional information.

The total number of observations collected by volunteers since the year 2006 including those collected using *takenradar* starting from 2012 is 33838. Only a subset of the data is used in this research. That is, observation for the years 2011 *to* June, 2014 that account for 67.84 % are considered in this research. From this particular dataset (the data for the years 2011- June, 2014), data within the administrative boundary of the Netherlands are used. The number of observations obtained as a result of the filtering process stated above since the year 2011 is 18788. The distribution per year of both the tick bite observations are presented in Table 1.

It is clearly visible from Table 1 that the number of tick bite reports is increasing year after year. The increase in the number of reports can be linked to a growing number of incidents or to growing number of participants in the reporting due to public awareness programmes. At this point, there is no explanation to why the number of tick bite reports is growing. However, it could be linked a growing risk as consultations to general practitioners for tick bites and Lyme disease in the last decade in the Netherlands (Sprong et al.,

[13] *https://www.flickr.com/*
[14] *http://www.esa.int/ESA*

2012) continued to increase. Either way, the phenomenon is worth understanding as it is still a societal problem

Table 1. Tick bite observations located in the Netherlands 2011- June, 2014

Year	Number of observations
2011	1210
2012	6356
2013	7695
2014*	3528

Note (2014): The data used for the year 2014 is six months observation since the rest was not available when this project started*

To understand the tick bite distribution and the risk, the spatio-temporal component in the location and date of incident as well as the context in the environment, outdoor activity, and the description provided by the volunteers are the central focus of this research. That is, the spatio-temporal distribution and the context should be understood. This can be achieved using both the spatio-temporal content and the context in each observation. However, there is a considerable missing and biased (mixed) information in the environment and activity components. Therefore, the tick bite observations discussed in this section cannot provide a complete understanding of the phenomenon.

Table 2. Tick bite observations per environment summary

Environment-Dutch	Environment_English	Observations	Per Environment (%)
bos	forest	6417	34.15
tuin	garden	4413	23.49
unknown*	unknown	1645	8.76
tuin-bos	garden - forest	1188	6.32
bos-heide	forest - heath	889	4.73
duinen	dunes	785	4.18
bos-weiland	forest - meadow	514	2.74
weiland	meadow	500	2.66
stadspark	city park	360	1.92
bos-duinen	forest - dunes	291	1.55
heide	heath	226	1.20
tuin-bos-weiland	garden - forest - meadow	179	0.95
tuin-weiland	garden - meadow	165	0.88
tuin-stadspark	garden - park	115	0.61
tuin-bos-heide	garden - forest - heath	113	0.60
moerasgebied	wetland	102	0.54
bos-heide-weiland	forest - heath - pasture	101	0.54
tuin-duinen	garden - dunes	92	0.49
bos-stadspark	forest park	76	0.40

Note (Table 2): 1.The data in (Table 2) only a subset of the whole data.
2. (unknown*) is a combination of *no value, weetniet, weet niet, and anders*

As can be observed from Table 2 and Table 3, the tick bite observation data that we have also suffers from incompleteness and noise in the environment and associated outdoor activity at the time of the incident. Out of the available data 8.76% occurred in unknown environment and 22.18% are associated with unknown outdoor activity. This will obviously lead to incomplete understanding of the spatio-temporal tick bite incidents under study.

Table 3. Tick observations per activity summary

Activity-Dutch	Activity - English	Observations	Per Activity (%)
wandelen	to hike/walk	4991	26.56
unknown	unknown	4167	22.18
tuinieren	gardening	3503	18.64
spelen	to play	2541	13.52
wandelen-spelen	walk - play	587	3.12
honduitladen	honduitladen	587	3.12
groenbeheer	green management	431	2.29
wandelen-tuinieren	walk - gardening	399	2.12
picknicken	picnic	307	1.63
wandelen-honduitladen	walk - honduitladen	281	1.50
wandelen-picknicken	walk - picnic	176	0.94
tuinieren-groenbeheer	gardening - green management	107	0.57
honduitladen-tuinieren	honduitladen gardening	96	0.51
hond uitlaten	dog walkers	95	0.51
wandelen-honduitladen-tuinieren	walk - honduitladen gardening	77	0.41
wandelen-picknicken-spelen	walk - picnic - play	48	0.26
wandelen	to hike	4991	26.56

Note: 1.The data in (Table 3) only a subset of the whole data.
2. (unknown*) is a combination of *no value, weetniet, weet niet, and anders*

3.3. Geolocated social media

Due to the missing and biased environmental and activity values observed in the data, the tick bite observations discussed in the previous section cannot provide a complete understanding as the context can be biased as a result of the incomplete environmental and activity information and ambiguous comments in the description provided by the volunteers. That is the missing data components in general are source of incomplete comprehension of the context. It follows then that using alternative sources to fill the missing values is indispensable. For this reason, geotagged social media content is collected and used as one of the auxiliary datasets to improve the tick bite observations there by improving our understanding. To collect the social media content, two social media platforms were considered in this project. These platforms are twitter for geotagged textual data and Flickr for geolocated photos. Both platforms have API's (application programming interface), to search for data. The API's provided by these source were first studied in parallel.

Twitter[15] was considered as a primary data source at the beginning assuming that historical public tweets can be retrieved. The data that can be obtained from twitter is in two ways, by streaming (which is not relevant for this project) and searching the archive by consuming the twitter developer API[16].

Due to their "terms and conditions", twitter provides search result for only the recent 9 or 10 days. To solve this problem, possible commercial alternatives such as Gnip[17] and Datasift[18] were considered for collecting the geotagged tweets. However, it was not possible to get data from these sources since the financial resources were not available. As a result, the primary social media source (Twitter) had to be replaced by *Flickr* and photos ad to be used instead of tweets.

Flickr was found a promising social media source as public photos can be obtained freely and for a longer duration. In addition to the availability of the media in this particular platform, the search functionality is flexible in such a way that media contents can be searched in many ways. It gives the programmer a freedom of choosing what portion of the content to retrieve.

The individual media element, photos in this case, has a rich information associated with it. In addition to the spatio-temporal information (longitude, latitude, and time), it contains the tags (comma delimited list of words associated with the photo), accuracy which is the numeric representation of the recorded accuracy level of the location information given as (World level is 1, Country is ~3, Region is ~6, City is ~11, and Street is ~16), numeric representation of the geo-context(not defined = 0, indoors =1 , outdoors =2), and description (free text written by the photo owner) among other things.

After understanding the platform API, data content of each media element and the required data for this research, social media data harvest process was conducted. This process of collecting geolocated photos is discussed in the following sections.

i. Developing the search vocabulary

To collect the contextual geolocated photos search vocabulary of the environment and the outdoor activity that can potentially be related to a tick bite incident was created from two sources. The first source of information was the "reporting forms" on the *tekenradar* application on which volunteers report the tick bites. The second source of this information used for this task is the tick bite dataset. This second source was used for obtaining additional search terms and filtering the already collected ones.

From the first source (*tekenradar* "reporting forms"), the *Dutch* key words representing the potential environment and outdoor activity related to tick bites were taken. In addition to that, the *English* translation of these terms was taken in to account which give the sum of the keywords and their translations as potential search terms.

Extracting key words from the available tick bite VGI dataset (*tekenradar* data), was performed with the help of Ms. Irene Garcia Marti. The whole dataset had to be programmatically processed using natural language processing (Bird, Klein, & Loper, 2009) script she developed in Python. The result of the natural language processing gave list of nouns and verbs, among others, and their frequency of occurrence in the whole dataset. From the results words for selected categories with a frequency of occurrence larger than 10 except for nouns in the whole dataset were considered as initial candidates. The resulting set of terms from this process and the previous process were manually analyzed to finally develop the search vocabulary.

Combination of the list of terms in Table 4 and selected terms from Table 5 were examined using the Flickr photo search API Explorer to search for results in order to develop the final search vocabulary.

[15] *https://twitter.com/*
[16] *https://dev.twitter.com/overview/api*
[17] *http://gnip.com/*
[18] *http://datasift.com/*

Table 4. List of key words from *tekenradar*

Environment	English translation	Outdoor activity	English translation
tuin	garden	wandelen	walk
bos	forest	Hond uitlaten	Dog walkers
heide	heath	tuinieren	gardening
weiland	meadow	picknicken	picnic
stadspark	City park	groenbeheer	green management
duinen	dunes	spelen	play
moergebied	wetlands		

Table 5. This table shows part of the results of the natural language analysis script

Nouns	Quantity Nouns	Verbs N	Quantity VBN	Verbs D	Quantity VBD	Verbs G	Quantity VBG	Verbs Z	Quantity VBZ
het	1431	een	276	was	346	camping	115	is	483
op	1400	en	144	had	219	wandeling	92	huis	49
van	1182	teken	100	met	52	kring	61	dus	34
een	894	met	72	gehad	41	omgeving	47	tijdens	24
en	872	been	56	gebied	35	ring	37	thuis	24
teek	824	opgelopen	51	bloed	27	kleding	27	pas	10
de	778	bossen	37	goed	25	besmetting	20	poes	7
ik	767	zitten	37	landgoed	23	behandeling	13	morgens	6
met	662	spelen	33	bed	18	ging	12	ws	5
bij	596	tekenbeten	31	natuurgebied	16	terschelling	12	zes	4
mijn	582	geen	28	fietstocht	14	boswandeling	11	s	4
door	369	buiten	23	huid	12	melding	11	ziekenhuis	2
heb	364	hebben	23	vermoed	10	verwijdering	10	inziens	2
teken	356	het	21	bosgebied	6	ontdekking	7	vakantiehuis	2
tuin	356	tussen	21	tot	5	stichting	5	kennis	2
opgelopen	319	fietsen	20	struikgewas	5	scouting	5	pannenkoekenhuis	2
er	316	gebeten	19	verspreid	4	leiding	4	paleis	2
niet	301	weken	18	geleid	4	ontsteking	4	circus	2
aan	270	bed	17	niet	4	verkleuring	4	cursus	2
na	247	wandelen	17	verwijderd	4	vereniging	4	steeds	2
naar	226	goed	16	onkruid	4	waarneming	3	scoutingkamptijdens	1
te	224	besmet	16	heeft	4	boscamping	3	walibikamperencursus	1
nog	214	heeft	15	ooglid	4	begroeing	3	afwachtenis	1
zijn	211	toen	13	speelt	3	richting	3	ligttijdens	1
ook	210	plukken	13	tocht	3	schatting	3	opnames	1
rode	204	binnen	13	gewandeld	3	bovenkleding	2	gewerkttijdens	1
jaar	190	laten	13	buitengebied	3	rekning	2	teekws	1
tekenbeet	188	zoeken	12	duingebied	3	kamping	2	nemenis	1
voor	182	gezeten	12	oid	3	bestrijding	2	hiercampingthuis	1
dat	182	struiken	11	begroeid	3	bedekking	2	mens	1
heeft	178	dat	11	dekbed	3	ontdektbesmetting	2	plukkentijdens	1
verwijderd	176	gaten	11	deed	3	mening	2	deels	1
dag	175	bosbessen	11	gemaaid	3	vakantiewoning	2	lymes	1
dagen	160	bloed	10	gemiddeld	3	afrastering	2	autofietsenws	1
veel	158	beten	10	huisartsenpost	2	beplanting	2	gelopenis	1
lyme	143	katten	9	mogelijkheid	2	tuintraining	2	volgezogenschoolreis	1
beet	141	plassen	9	wed	2	huiscamping	2	dubieus	1
maar	140	ben	9	gespeeld	2	pieterpadwandeling	2	bovenbeenpas	1
dit	140	de	9	ingelicht	2	overnachting	2	saptijdens	1
zat	138	mensen	8	hoeveelheid	2	schutting	2	bosws	1
deze	132	gelegen	8	gehadwas	2	plukkencamping	2	komterras	1
wel	131	mijn	8	behandeld	2	zwelling	2	brievenbus	1
kleine	131	had	8	ook	2	fotograferencamping	2	hoenderlootijdens	1
onder	129	kinderen	8	recreatiegebied	2	aanleiding	2	visseneveneens	1
onze	128	kunnen	7	terecht	2	zittencamping	2	arrows	1
al	120	personen	7	dicht	2	kringvorming	2	zittenbosjes	1

source: Irene Garcia-Marti

From Table 5, the words under the coluns labeled as "Nouns", "Verbs N", and "Verbs G" which represent nouns, past participle and gerends (Bird, Klein, & Loper, 2009, pp.183) were used to support the development of the search vocabulary. The rationale here is that the nouns can potentially represent the environment where as the verbs can represent the outdoor activites associated with the tick bite incident. As a result of limiting the frequency of ocuurrence for candidaite search term, there is a possibility for potentially helpful key words to be left out.

Figure 2. Flickr photo search API explorer (API Explorer: Flickr.photos.search)

Flickr has a platform where developers can showcase the applications they have created and where one can find new ways to explore Flickr media content called the App Garden[19]. Most of the functionalities provides by the Flickr API have been implemented as method explorers and can be used for free from this platform. In this research, the photo search [20] functionality of the API implemented as a photo explorer in the Apps Garden is used for filtering the potential search terms by looking the amount of data that can be obtained for each search.

By providing the function arguments and selecting the data format such as XML, JSONP or JSON, the API explorer for the photo search functionality provides information regarding the response and part of public photo that meet the search criteria. It was understood that the response containing the data is organized in pages of records. The response gives vital information to the user. It shows that, while using the API, one should be aware of going through all the pages to harvest the data provided for the search term. Indeed, it give how many records can be obtained on each page of the response. This helps in deciding on requesting the optimal number of records per page to minimize the social media data harvest time.

The potential search term collected both from *tekenrarar* and the natural language processing result of the collected information were used to search for photos and metadata of the result to filter them out. For the potential environmental and outdoor activity contexts, terms with 50 photos were taken in to consideration. Only search terms such as (teek, teken, and tick/s) were taken without testing. Based on the assessment, the final set of search vocabulary in Table 6 was developed.

[19] *https://www.flickr.com/services/apps/about/*
[20] *https://www.flickr.com/services/api/flickr.photos.search.html*

Table 6. Search terms used for harvesting geolocated photos

Environment and nouns	English translation	Outdoor activity and verbs	English translation
tuin	garden	Wandelen	walk
bos	forest	wandeling	walk
bossen	forests	tuinieren	gardening
bosbessen	blueberries	boswandeling	Forest walk
teek	tick	kamperen	camping
teken	ticks	spelen	playing
heide	heath		
weiland	meadow		
---	city park		

ii. Harvesting geolocated Flickr photos

Flickr has an open API that enable the platform users to write their own program to extract data from or present public Flickr content (like photos, video, tags, profiles or groups) in new and different way. The API can be consumed directly as Representational State Transfer (RESTful[21]) API or using free "API Kits" developed by other developers.

Most of the free "API Kits" that are available for use are developed for a specific use by their developers. There is no complete kit that can fit all purposes. In most cases there is little or no documentation to help other users. Some of the kits that was used in the process did not provide all the required attributes of the photos as required by this project. However, the RESTFul API was found to give similar results like that of the API explorer used for evaluating the search terms. Therefore, the RESTful API was consumed in the python script that was used to harvest the photo extracts.

The geolocated Flickr photos that satisfy the search query were obtained on a search term by search term basis in JavaScript Object Notation (JSON). The obtained JSON response was automatically converted in to comma separated (.CSV) files using the Python script for which the pseudo code is given in Figure 3. The first part (GetNumberOfPages function) in the script pseudo code uses a query that requests the first page of the response and extracts the number of pages available for the particular search term and returns the number of pages. The second function that follows which is defined *GetPhotoExtracts* requests all the required data from the public photos, coverts each record in to a comma separated content, and writes it into comma separated file. The search query is a URI[22] query designed to obtain data from a specific location in a specified time frame for a specified search term. Indeed the extra information and data format of the

[21] *http://www.restapitutorial.com/*

[22] *In the World Wide Web, a* ***query string*** *is the part of a uniform resource locator (URL) containing data that does not fit conveniently into a hierarchical path structure. The query string commonly includes fields added to a base URI by a Web browser or other client application, for example as part of an HTML form*

Example URI query string:
'https://api.flickr.com/services/rest/?method=flickr.photos.search&api_key=36ad3a871cb369669a14fd4a372c5ec3&text=garden&min_taken_date=2011-01-01&max_taken_date=2014-10-14&bbox=3.362556%2C50.753918%2C7.227944%2C53.51219&has_geo=1&extras=date_taken%2Cgeo%2Ctags&format=json&nojsoncallback=1&per_page=500'

requested data is included in the query. The data for all the search terms was then collected by manually providing the search term in the query and is summarized in *Table 7* and *Table 8*.

```
START Script
    var queryString= uriQueryString

    Function GetNumberOfPages(queryString)
            request = request.get(queryString)
            JSON_data = JSON_load(request.text)
            numberOfPages = Get_number_of_pages(JSON_data)
            return numberOfPages
    END Function

    Function GetPhotoExtracts(numberOfPages)
            CSVFile=openFile( fileName)
            For Page =1 step 1 To numberOfPages
                searchQuery = uriQueryString + '& page=' +page
                photo_request = request.get(searchQuery)
                JSON_Photo_data = JSON_load(photo_request.text)
                For line = 1 step 1 To numberOfPhotos(JSON_Photo_data)
                    CSVFile.Write(line)
                END For
            END For

            close CSVFile
    END Function
END Script
```

Figure 3. Pseudo code for the Python script used to harvest the geolocated Flickr photos

Table 7. Harvested information for geolocated Flickr photos per search term

Search term	Number of photos
bos	6283
bossen	589
boswandeling	507
forest	2809
bosbessen	63
camping	2049
kamperen	225
gardening	11752
garden	9583
tuin	3137
tuinieren	63
wandelen	5311
wandeling	6908
walk	9653
teek	6
teken	1632
heide	1655
heath	527
weiland	756
meadow	1403
Citypark	1945
spelen	1241
playing	6439

The data in given in Table 7 and Table 8 represent those that are located within the administrative boundary of the Netherlands. The observations that are collected from neighbouring countries as a result of the rectangular nature of the minimum bounding box were discarded by clipping using the administrative boundary.

Table 8. Harvested information for geolocated Flickr photos per search term per year

	Number of photos per year			
Search term	***2011***	***2012***	***2013***	***2014***
bos	1676	1610	1774	1223
bossen	84	247	193	65
boswandeling	43	176	251	37
forest	939	765	712	393
bosbessen	25	15	15	8
camping	665	350	756	278
kamperen	94	40	61	30
gardening	3515	3362	2601	2274
garden	31	110	7483	1959
tuin	827	848	881	581
tuinieren	10	41	8	4
wandelen	557	1388	1573	1793
wandeling	607	1763	2307	2231
walk	29	225	7410	1989
teek	2	1	2	1
teken	545	595	478	14
heide	468	456	585	146
heath	191	165	142	29
weiland	230	261	217	48
meadow	368	401	353	281
Citypark	840	380	350	375
spelen	435	374	307	125
playing	249	2356	2647	1187
Total	**12430**	**15929**	**31106**	**15071**

The initial exploration of the geolocated photos showed that the data is too noisy to be used for the actual analysis. The data had to be partially cleaned for the noise that could be introduce as a result of taking too many photos from the same location.

Before running the data cleaning the data was categorised in to two groups, the environment data and outdoor activity data. They were further classified by combining the Dutch search terms and their English translations in each category. It was after this process the data cleaning script was used.

The partially cleaned contextual photos are summarized in Table 9.

Table 9. Partially deaned photo extracts

Category	Keyword	Number of photos
Environmental	Forest	7152
	City park	1476
	Garden	4171
	Heath	1596
	Meadow	1958
Activity	Camping	1390
	Gardening	7747
	Playing	3136
	Walk	11872
	Forest walk	411

3.4. Land cover data

The tick bite observations as discussed in earlier sections has thematic bias as result of the missing environmental data. This data gap can lead to different understanding of the relationship to the environment on which the tick bite incidents occur. The bias in this thematic information has to be reduced using authoritative land cover data.

The geolocated *Flickr* photo extracts that are related to outdoor activities do no specifically contain information about the environment on which they were taken. Hence, if these extracts are to be used to solve the problem caused by the missing outdoor activity in the primary data, the authoritative data should again be used to extract their actual land cover information.

To improve the environmental information in the tick bite observations, land outdoor activity information in the *Flickr* photo extracts, official land cover data had to be used. The land cove data used in this project is an extract from the global land cover map produced by the GlobCover project. This data was obtained from the GlobCover portal[23] of the ESA[24] free of charge. According to product description and validation report (Bontemps et al., 2011),the original land cove map has:

- 22 land cover classes as described in its legend description
- A resolution of 300m and,
- Overall accuracy of the classification 67.0%.

[23] *The GlobCover Portal (http://due.esrin.esa.int/globcover/)provides access to the results of the GlobCover project. GlobCover is an ESA initiative which began in 2005 in partnership with JRC, EEA, FAO, UNEP, GOFC-GOLD and IGBP. The aim of the project was to develop a service capable of delivering global composites and land cover maps using as input observations from the 300m MERIS sensor on board the ENVISAT satellite mission. ESA makes available the land cover maps, which cover 2 periods: December 2004 - June 2006 and January - December 2009*

[24] *http://www.esa.int/ESA*

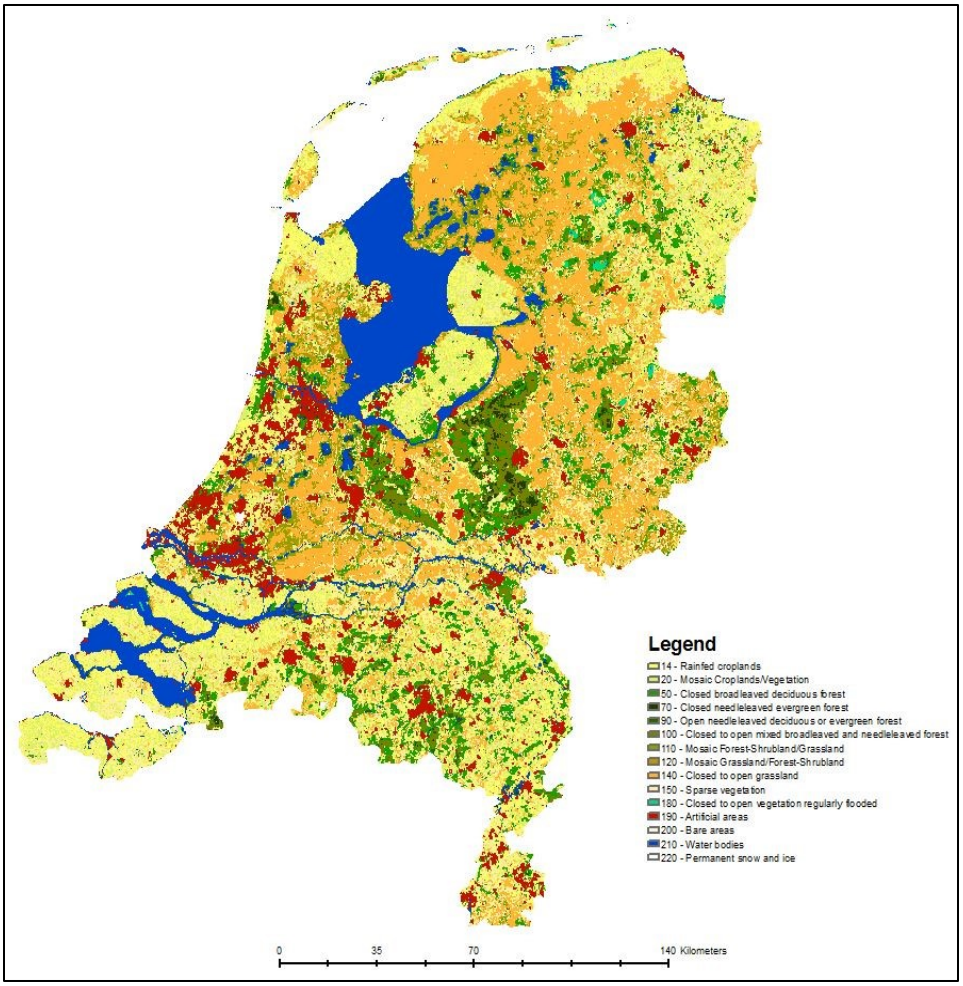


Figure 4. Globcover map extract.

This map is taken from the GlobCover project global land cover map. The details of the legend can be referred in the product description and validation report (Bontemps et al., 2011)

The original land cover data obtained from ESA cannot be used as is. Forest types, croplands, and other land cover classes are represented in multiple types. As far as this project is concerned, it is not necessary to define multiple forest types or other land cover classes since we are interested only on the environment. For example, the type of trees in the forest are not of interest to understand tick bite distribution. A generalized land cover forest is sufficient for this case. It was then found very important to reclassify the original land image to satisfy the needs of this project. So, all land cover types with multiple representations were classified as one class. That is all types of forest, all types of grass lands, and all types of croplands were each classified as one. The whole land cover dataset is then reclassified as in the following map which results in land cover classes with half of the original number classes.

Land cover classes represented on the map in Figure 5 are the land cover classes used in this project. Any discussion that refers to land cover classes is associated to this map.

Throughout the analysis process, the land cover polygon feature class extracted from the image is used. A feature class in ArcGIS "*is a collection of geographic features with the same geometry type (such as point, line, or polygon) the same attributes, and the same spatial reference*". A feature, again in the same technology, is "*representation of a real-world object on a map*". Any use of the "feature class" or "feature" in this thesis is to means these definitions.

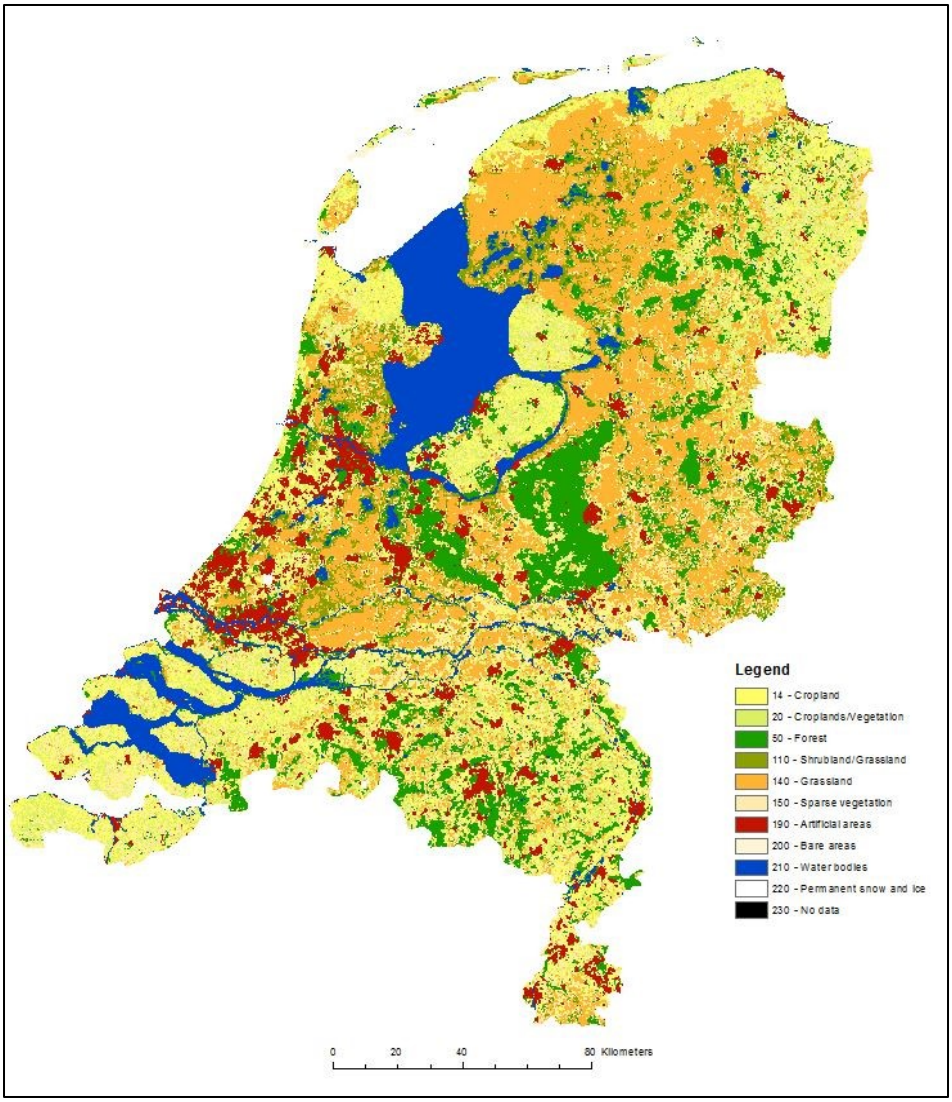


Figure 5. Reclassified land cover map.

4. METHODS AND TOOLS

4.1. Overview

In this chapter, the methods and tools used to achieve the research objective and answer the research questions are discussed. The methods section explains the methods used to prepare the datasets for analysis, select evaluate the cloud platform (SaaS) that was used to partially implement the tasks in this project. It continues with the discussion of the spatial, temporal, and statistical analysis methods used to understand the distribution of tick bites. These methods were indeed used to find out if there is a relationship between the two VGI datasets so that we can use the Geolocated photos to improve our understanding of the tick bite distributions and risks by inference. Finally, the software tools that are used to generate the outputs of this research discussed in the tools section.

4.2. Methods

4.2.1. Partial data cleaning and preparation methods

The initial exploration of the geolocated photos showed that the data a lot of noise. The noise is believed to be introduced as a result of taking too many photos from the same location by the same person. This action is a normal observable behaviour. So, taking this behaviour into consideration and the results of the preliminary exploratory analysis an algorithms to clean the data from this noise was developed and used.

Before running the data cleaning the data was categorised in to two groups, the environment data and outdoor activity data by combining the resulting data from Dutch search terms and their English translations in each category. From each, category, photos that are taken by the same person, from the same location, on the same date, and have the same title were identified. If there are multiple photos that satisfy the condition, only one of them was taken and the rest discarded. The geolocated photos obtained as a result of this process are used in this project. The pseudo code for the script used to clean the photo extracts is given in Figure 6.

```
START Script

    Function ReadFile(filePath)
            fileContent_List = Read_CSV_File(filePath)
            return fileContent_List
    END Function

    Function GetPhotoExtracts(fileContentList)
            CSVFile = openFile(fileName)
            list1 = fileContentList[1:]
            list2 =fileContentList[2:]
            cleanDataList = []
            For countList1 = 1 step 1 To length(list1) # Loop1
                For countList2 = 1 step 1 To length(list2) # Loop2
                    IF
                     list1[photo_owner] == list2[photo_owner] AND
                     list1[taken_date] == list2[taken_date] AND
                     list1[photo_location] == list2[photo_location] AND
                     list1[photo_title] == list2[photo_title] AND
                     Continue   Loop2
                Continue Loop1
                    ELSE
                     cleanDataList.upend(list1[countList1])

                    END IF
                END For
            END For
            For countList3 = 1 step 1 To length(cleanDataList) # Loop3
                CSVFile.Write(cleanDataList[countList3])

            END For

            close CSVFile
    END Function
END Script
```

Figure 6. Pseudo code for photo data cleaning script

To perform the analyses using the selected methods the datasets had to be prepared. First, to understand the distribution per municipalities of the both tick bites and photos, the two point data representing both phenomena are aggregated using the municipal boundaries of the Netherlands. The density of the tick bites and photos per municipality scaled by 10 to avoid rounding to zero was calculated as:

$$TBD = 10 * \frac{NT_M}{A_M} \quad (1)$$

$Where$:

$TBD\ is\ tick\ bite\ density$

$NT_M\ \ is\ number\ of\ tick\ bites\ per\ municipality\ and$

$A_M\ \ is\ area\ of\ municipality\ in\ square\ kilometrs$

$$CPD = 10 * \frac{NP_M}{A_M} \quad (2)$$

$Where$:

$CPD\ is\ conetxtual\ photo\ density$

$NP_M\ \ is\ number\ of\ photos\ per\ municipality\ and$

$A_M\ \ is\ area\ of\ municipality\ in\ square\ kilometrs\ and$

The resulting dataset was used to evaluate the relationships between the actual numbers of the individual phenomenon and the density per municipality of each. An exploratory analysis of the actual number per municipality of each and their respective densities showed that the number and density are not highly associated. Hence, another bias can be introduced because of the area of the administrative units. That is, large number per municipality does not mean high risk if the municipality also covers very large area and vie-versa. Therefore, the values TBD and CPD were chosen to be used to perform the analyses at this aggregate level.

To investigate the relationships between the VGI datasets and the land cover the VGI datasets were intersected with the land cover to extract the land cover type. The unknown and mixed land cover information in the tick bites and as well as the land cover information for outdoor activity related photos were obtained using this process.

For the analysis of relationships between the two VGI datasets, the points representing tick bites and photos in each context were first aggregated using a victor grid cell of size 1000. This vector grid was created using the Create fishnet tool in ESRI'S ArcMap which can be referred in the ArcGIS Resources[25] page.

Another equally important analysis that was done in this project is understanding the actual risk of tick bite to the inhabitants. To address this societal problem, it is crucial to know how many people are vulnerable to tick-borne diseases and where. It is then very important to use a combination of the tick bite observations and the population to evaluate the risk to the residents in each municipality. The risk of tick bite for the years 2012 and 2013 is evaluated for each municipality. To create the risk map depicting the risk of tick bites, the aggregated observations over the municipalities were divided to 1000 inhabitants per municipality. The resulting data was used to identify the municipalities with high risk of tick bite per 1000 residents in each.

[25] *http://resources.arcgis.com/en/help/main/10.2/index.html#//00170000002q000000*

4.2.2. Cloud platform selection and evaluation methods

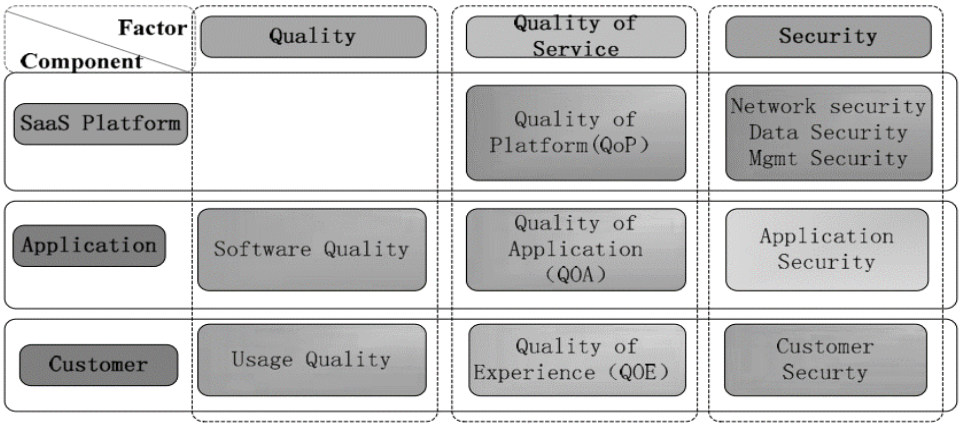
The secondary objective of this project focuses on evaluating geospatial cloud platforms for implementing geo-information work flows. It also highlights that the evaluation is done by implementing the main tasks of the primary objective in this project where applicable. To achieve the goal, two types of evaluation shall be performed. The first type is software evaluations for section while the second type is evaluating the selected software for its ability to support the requirements of this project. This in turn leads to evaluating the maturity of the selected platform to support similar workflows.

The first stage of the evaluation, which is the evaluation for selection was done using the AHP (Saaty, 1990) as a guiding principle. According to (Godse & Mulik, 2009), selecting a software as service is a multi-criteria decision making problem (MCDP) and needs a hierarchical method of selection. In their research they applied the AHP to make decisions on selecting software as a service. For selecting the SaaS for this project the basic principles discussed in their approach were used. However, a binary values 0 or 1 were given for the attributes to be evaluated instead of using the "local" and "global" weights discussed in their approach.

The core idea of using the AHP in SaaS selection is minimizing the bias that is introduced as a result of personal judgement. However, creating the weights for each criteria in this project where only a single person is involved in the process does not make any difference in reducing the bias. In addition, after the initial filtering of the solutions, there were only two final candidates to select from. So, taking the basic principles to create the hierarchy and selecting the final product using a binary value was found appropriate.

The second stage of the cloud platform evaluation is finding out whether the promised functionalities are available and fit the requirements of the project. To do so, an evaluation checklist of requirements to evaluate the SaaS was developed. This checklist is related to functionalities for performing geovisual analysis of the datasets and implement geovisualization solution for information sharing in SaaS platform.

The evaluation of the selected geospatial cloud computing platform was done based on the quality model for SaaS (Wen & Dong, 2013). This model defines three quality factors for SaaS namely security, quality of service, and software quality. The model also decomposes the quality factors in to the three roles namely customer, platform, and application. The model is summarized in Figure 7.



This Figure is taken from (Wen & Dong, 2013).

Figure 7. Quality model for SaaS

The evaluation of the selected platform (SaaS) was performed from the "Customer" role perspective. That is to say, it is evaluated for "Usage quality", "Quality of experience (QOE)", and "Customer Security" factors described in this model. For the reasons of time and the scope of evaluation, only one component of the software was used in this project.

4.2.3. Spatial Analysis methods

To understand the two VGI datasets representing potentially related phenomena, spatial analysis methods are employed to understand the distribution of each. These geospatial phenomena as far as this project is concerned are the tick bites incidents and the outdoor activities represented by contextual social media data (geolocated *Flicker* photos). The spatial analysis methods used here are mainly to understand how both phenomena are distributed in space and identify the location of hot spots. Indeed, they are used to visually investigate the possible relationships between the two in order to use the geolocated photos to improve our understanding of the tick bite distribution and risks.

The missing outdoor activity information in the tick bite reports can be inferred from the contextual photos provided that the tick bite reports and the photos representing outdoor activities are related in space and time. If the tick bites and the photos are distributed with the similar patterns in space and time, then there is a high probability that the two phenomenon are related.

To infer the social activities using the photos, in case they are highly related, the analyses were performed at multiple levels. That is, first all the tick bite observations were analyzed for relationships with all the activity related photos. This process continued to individual environments and activities such as tick bite happening in the forest versus activity photos located in the forest, tick bites located in built-up areas versus activity photos located in built-up areas, tick bite happening in the croplands versus activity photos located in the croplands, and tick bites located in grassland versus activity photos located in grassland. This process continued with all of the land cover classes and activity photos that are located in these areas.

To perform the spatial analyses described above to understand the spatial distribution of tick bites as well as the contextual photos and investigate the relationships, two analysis methods were used. These methods are Kernel Density Estimation (KDE) and Geti-Ord GI*.

Kernel Density Estimation (KDE) was used to identify hot spots in the distribution of point processes represented by both VGI datasets. KDE for point features calculates smooth density surface for point events in a two dimensional geographic space (Xie & Yan, 2008). This method is implemented in several spatial data analysis tools. For this particular project, ***ESRI's ArcMap*** was used.

To evaluate the statistical significance of the identified hot spots, Getis-Ord GI* for cluster analysis was used. The Getis-Ord GI*(Ord & Getis, 2010) cluster analysis method is used to identify the locations of statistically significant hot spots and cold spots of tick bites and contextual photos. The method is particularly useful when action is needed based on the location of one or more clusters. Since we are trying to find out the risky areas, the identified locations with high density of tick bites should be tested for their spatial autocorrelation.

The outputs from the Getis-Ord GI* statistics method are used to test the complete spatial randomness (CSR) hypothesis which states that point events occur within a given study area in a completely random fashion.

4.2.4. Temporal analysis method

A separate methodology that is depicted in Figure 8, was used to perform the temporal analysis. The methods applied to investigate the temporal distribution of both VGI datasets are discussed in this section.

The temporal analysis methods used here are mainly to understand how both phenomena represented by the VGI datasets are distributed in time and examine if they follow similar patterns. The use of this temporal analyses is twofold. First, it helps to identify the yearly temporal windows of high incidents of tick bites. Second, it assists in finding out whether we can use the contextual social media to improve our understanding of the tick bite distribution and tick bite risks.

The temporal analysis method used in this project is a combination of temporal plotting for visual analysis and investigating the monthly distribution. Both datasets are aggregated using the same temporal resolution and plotted to investigate their individual distribution. The resulting plots are also used to understand the similarity of the datasets by looking in to the temporal regions of absolute maximum and absolute minimum of each data set. In addition the underlying data is analyzed using the monthly distribution to identify the actual number and proportion of high incidents during month of the year.

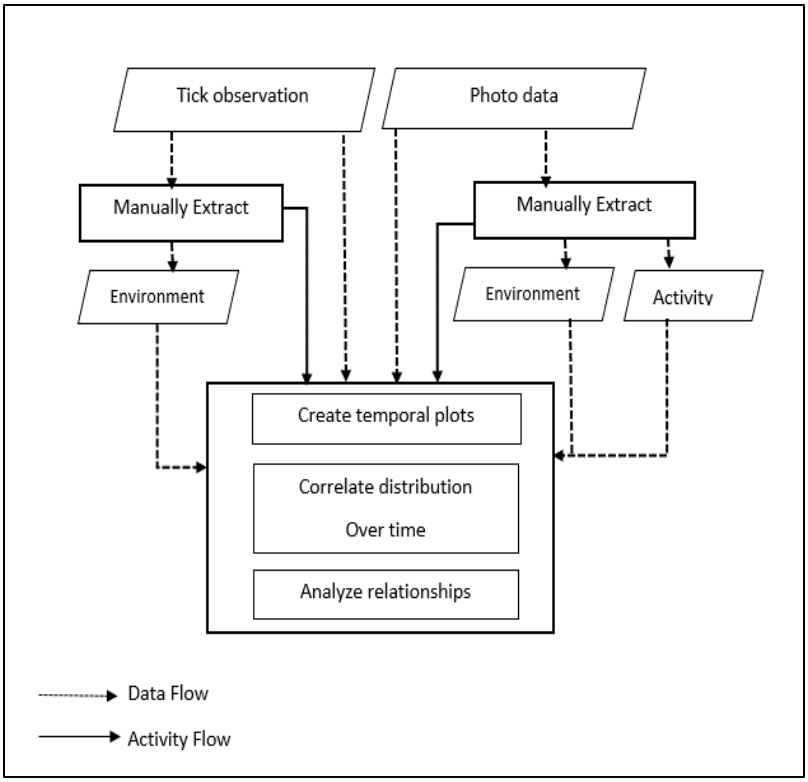


Figure 8: Temporal analysis methodology

4.2.5. Statistical analysis method

The tick bite observations and the contextual *Flickr* photos are assumed to be related in space and time. One of the main tasks of these research then is finding out if there is a similarity between two main datasets and using the *Flickr* photos to improve the understanding of tick bite distribution by exposing hidden patterns in the tick bite observations.

To analyse the relationships between the two datasets especially in time, the ***Spearman's correlation*** is used. ***Spearman's correlation*** method is used with ordinal or nonparametric(non-normally distributed) interval or ratio data (Prion & Haerling, 2014) to measure the strength and direction of the relationship between two datasets. This method was used to mainly evaluate the temporal relationships between the two VGI datasets in order to infer the pattern from the photo extracts to understand hidden patterns of tick bite distribution and risk, in case there is a strong relationship between them. It is also applied in analysing the relationship between the two datasets as aggregated per municipality and land cover features for the same purpose.

The Spearman correlation coefficient returns a value between -1 and -1, with 0 denoting no relationship at all. The higher the absolute value of the number, the stronger the relationship between the two variables. A positive correlation means that both variables move in the same direction and negative correlation means that the variables move in opposing directions.

The "rule of thumb" (Mukaka, 2012) for interpreting Spearman correlation coefficient (ρ) results are as follows:

- $0 \leq \rho < 0.30$ is negligible
- $0.30 \leq \rho < 0.50$ is weak
- $0.50 \leq \rho < 0.70$ is moderate
- $0.70 \leq \rho < 0.90$ is strong, and
- $0.90 \leq \rho \leq 1.00$ is considered very strong.

This method does not have pre assumptions of the data to be tested for association and was found to be appropriate for this research. Therefore, the method was applied to calculate the correlation coefficient and probability (p-value) to test the correlation between the datasets.

4.3. Tools

Different tools were used to perform the social media harvest, analysis and dissemination of the results of the project. The analysis tasks, especially the spatial distribution of both VGI datasets are performed both in the local machine and the cloud platform. This is mainly because it is difficult to produce analysis results for a paper based work in the cloud platform. The following section discusses the spatial data analysis (both spatial analysis and spatial statistics) tools, temporal analysis tools and geovisual analytics tools.

4.3.1. Spatial data analysis tools

The ArcGIS for desktop[26] family software was used to organize the data in a file geodatabase, analyze the spatial distribution of both tick bite observations and photos as well as the identifying hot spots of tick bite risks. The ArcGIS Spatial analyst tools[27] and Spatial statistics tools [28] were used to identify hotspots and calculate local spatial statistics respectively.

4.3.2. Temporal analysis tools

To perform the temporal analysis of the datasets, Anaconda[29], which is Continuum Analytics' data analysis environment, was used. The software package, Anaconda, is a free collection of powerful packages for Python that enables large-scale data management, analysis, and visualization for Scientific Analysis, Engineering, Machine Learning, and many more.

4.3.3. Geovisualization tools

CartDB Editor, which is a SaaS, was used to run part of the analysis and implement the final geovisualization prototype. For the spatial analysis, it was used to identify the areas of high density for both tick bite incidents and contextual photos. It was also used to implement the geovisualization prototype to showcase how the results of the analysis can be easily shared in an effective and intuitive way to help authorities and the public in making informed decisions. The out puts of the project can be shared easily and almost real time using the capability of this platform to create animated maps to show the spatio temporal distribution of tick bite incidents so that interested parties can have both the spatial and temporal understanding of the phenomena. This platform was not only used to create the maps but also to evaluate the maturity of the geospatial cloud computing landscape to support geospatial workflows.

[26] *http://www.esri.com/software/arcgis/arcgis-for-desktop*
[27] *http://www.esri.com/software/arcgis/extensions/spatialanalyst*
[28] *http://blogs.esri.com/esri/arcgis/2010/07/13/spatial-statistics-resources/*
[29] *http://docs.continuum.io/anaconda/index.html*

5. RESULTS

5.1. Overview

In this chapter the analysis and SaaS evaluation results are presented. The sections and sub sections are structured in such a way that they can address the research questions. The first and last sections 5.2 and 5.6 is linked to the cloud platform selection and evaluation results. Sections 5.2 and 5.6 related with research questions 6 and 7 respectively. Section 5.3 and its subsections related to the distribution of the tick bites. The research question 1 and partly research question 2 are addressed in this. The section that follows Section 5.4 and its subsections are related to the results of the analysis of the contextual Flickr photos. The results are used to answer research question 3. Finally, the results of relationships among the datasets that we call in this research the "data triangulation" are presented in Section 5.5. The results are used to address research question 4 and partly research question 2.

5.2. Selection of cloud computing platforms

One part of the evaluation of the geospatial cloud computing platforms that was done in this project is the systematic selection of SaaS platform for performing the analysis tasks when applicable. In this phase, out of the five online solutions one was selected after going through the evaluation process. This first stage selection is summarized in Table 10.

Table 10. First stage evaluation results

Functionalities	MangoMap[30]	Geocommons[31]	eSpatial[32]	CartoDB[33]	NCVA-GVA[34]*
Identify clusters/ hot spots	no	no	yes	yes	no
Import data /multi-data support	yes	yes	yes	yes	yes
Create density maps/choropleths	yes	yes	yes	yes	yes
Create temporal representation	no	yes	yes	yes	yes
Delivered as SaaS	yes	yes	yes	yes	no

(*) *The NCVA geovisual analytics platform is a web application which is available for educational purposes only. Although it has powerful visual impression, it cannot be branded as a SaaS. This solution can be an alternative if the purpose is only visualization.*

As a result of the filtering two online mapping software solutions (eSpatial and CartoDB) were found to have all the functionalities required. These two cloud based solutions were further evaluated to select one. The final evaluation based on the AHP model for selecting SaaS solutions is summarized in Table 11.

[30] *https://mangomap.com/*
[31] *http://geocommons.com/*
[32] *https://www.espatial.com/*
[33] *http://cartodb.com/*
[34] *http://ncva.itn.liu.se/ncva?l=en*

Table 11. AHP based doud platform selection result

Goal	Factors	Attributes	eSpatial	CartoDB
SaaS selection	Functionality	Identify clusters	1	1
		Create density maps	1	1
		Create multi-layer maps	1	1
		Create temporal representation	1	1
	Usability	Learnability	1	1
		Efficiency	1	1
		Memorability	1	1
		Satisfactory documentation	1	1
	Architecture	Integration (has API?)	0	1
		Reliability	1	1
		ESRI shape file support	0	1
		Security	0	1*
	Pricing	Pay as you go (monthly option)	0	1
		Free option	1**	1***
	Result		**10/14**	**14/14**

Note:

(*) the enterprise version of the system is provided as a private cloud and so data can be deployed in the customer's own premises.

(**) limited number of records (10000) and limited number of maps

(***) limited size of data (50MB) and unlimited number of maps

As can be seen from Table 11 CartoDB which is a cloud based open-source mapping platform that can be used from simple visualizations to complex but highly scalable geospatial applications and analytics selected. The products that can be creating in the software are simple, choropleths, category, density, intensity, and animated maps to name some. The animated maps that can be created are especially applicable for a spatio-temporal representation of datasets like those we have in this project.

5.3. Spatio-temporal analysis of tick bite observations

5.3.1. Tick bite density per municipality

To analyze the spatial distribution of tick bites, TBD which was calculated from the number of tick bites aggregated using the municipalities as an aggregation units and the area of each municipality was used to calculate the density. The density maps in Figure 9 was then created using the resulting data to find out the municipalities with high and low density of tick bite incidents.

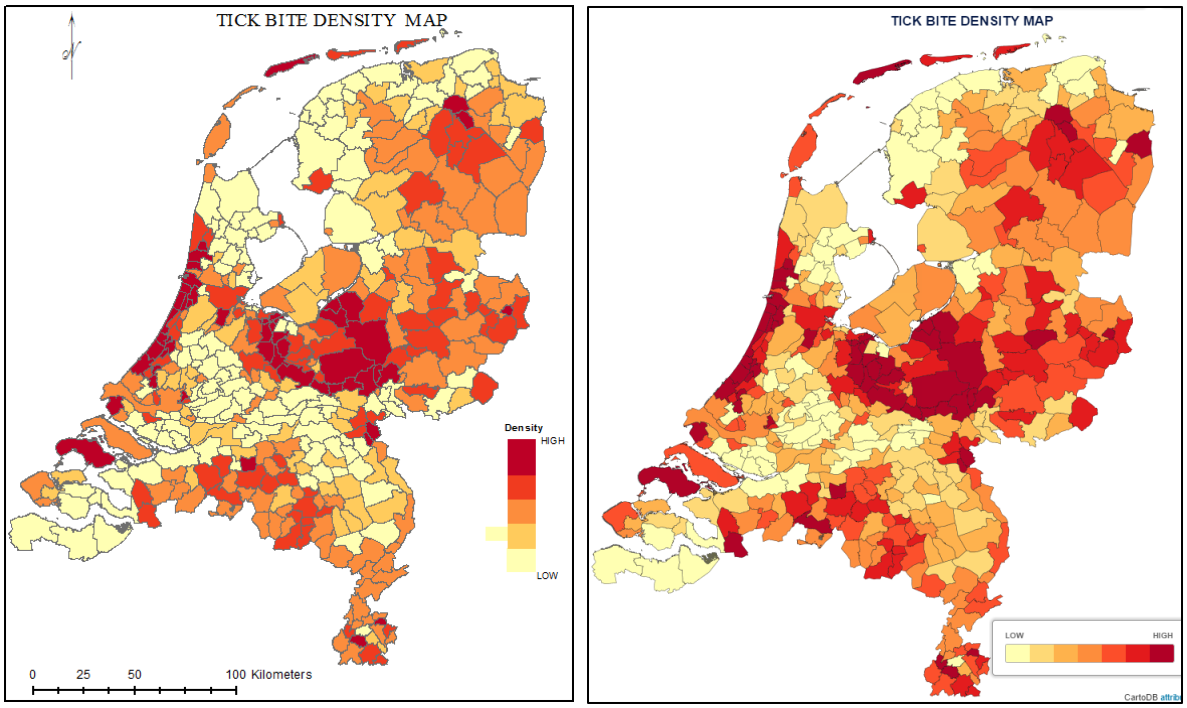


Figure 9: Tick bite density maps, ArcMap (left) and CartoDB (right)

The result in Figure 9 shows concentration of tick bites in municipalities from low (yellow) to high (dark red). The map to the left represents a density map classified as (value = TBD, Classes = 7, Classification = Geometrical Interval) created in ArcMap and the map to the right shows density map with (value=TBD, Classes = 7, Quantification = Quantile) created using CartoDB Editor[35]. The choice of different classification methods here is because, CartoDB has not implemented 'Geometrical Interval' options. The small difference that can be observed in the two maps is the result of the classification methods.

As can be observed from the maps in Figure 9, there are contiguous municipalities with high and low density of tick bites throughout the west coast, the central, north eastern and southern part of the country. The municipalities at the center are those that accumulate more forest surfaces. Municipalities located throughout the west coast are visited by many people for recreation.

5.3.2. Tick bite hot spots

We have seen from the previous result that there are neighboring administrative units with high and low densities. It is also important to know whether the underlying data (tick bite VGI) exhibits similar pattern or not. To investigate the underlying pattern, the KDE method was run in ArcMap using the tick bite observations. The same data was used to create a T*orque heat* [36]map in CartoDB to investigate the spatio-temporal distribution of tick bites. The resulting KDE map and the static version of the *heat* map are presented in Figure 10.

[35] *The account that is used for this project is based on the free license plan. CartoDB can put down free accounts at any time. The author cannot guarantee the availability of the mapping product that is created in the platform. The published map Figure 9 (right) can be accessed using this the link* https://berihu.cartodb.com/viz/0c76cf42-b237-11e4-9fad-0e0c41326911/public_map

[36] *Torque Heat maps leverage the combination of heat maps and Torque to investigate the spatio-temporal location of hot spots of point process data* http://blog.cartodb.com/introducing-heatmaps/

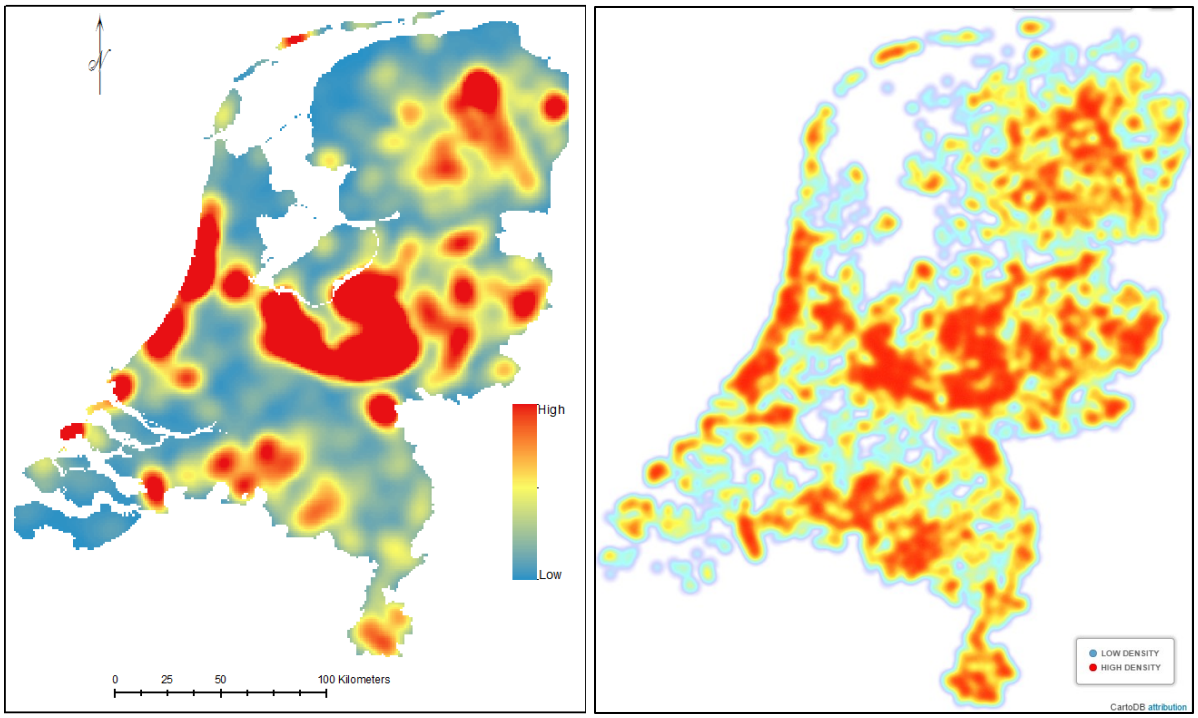


Figure 10: tick bite observations' kernel density (left) and torque heat (right) map

The KDE map (left) in Figure 10 was created in ArcMap and represents (kernel density for 1000 raster cell size, symbolized with 1 standard deviation stretch) whereas, the *heat* map (right) was created with the same notion (identifying actual hot spots) and the same input, but with different method in CartoDB. The *heat map* was created with (Marker size=15, Threshold=0.4, Resolution=4) and depicts the tick bites hot spots. The animated version of this map can be accessed from the free license[37] account with which it was created.

Both maps in the result of Figure 10 show density estimates from low (blue/cyan) to high (red). As can be seen from the maps, the tick bite observations are observed to be clustered in the west-coast, the central, north eastern and southern part of the country. As explained in the results in section 5.3.1 the location of the hot spots, mainly the central and the west- coastal areas, are highly visited for recreation.

Having observed clustering of the point locations of the tick bite incidents in the analyses done so far in both platforms, it is important to look into the statistical significance of the identified hot spots. That is if there is statistically significant spatial relationship among the areas of high incidence of tick bites.

5.3.3. Spatial Statistical analysis

The results in Figure 9 and Figure 10 showed that there are clusters of tick bite hot spots. As the maps are representation of the same reality, testing one of them for statistical significance suffices. Hence, to find out whether the distribution per municipality of the tick bites is based on complete randomness or there is a spatial correlation among the municipalities, the complete spatial randomness hypothesis was tested. The complete spatial randomness hypothesis is stated as follows:

$\boldsymbol{H_0}$: The clustering of municipalities with similar density of tick bites is completely random

$\boldsymbol{H_a}$: The clustering of municipalities with similar density of tick bites is not random

[37] *CartoDB can put down free accounts at any time. The author cannot guarantee the availability of the mapping product that is created in the platform. The published tick bite intensity map can be accessed using this the link* https://beribu.cartodb.com/viz/eaf82e80-ad57-11e4-b5b7-0e9d821ea90d/map

To test this hypothesis, local spatial statistical analysis method for hot spot analysis was used. The method used is Getis_ Ord Gi* which creates the resulting hot spot map showed in Figure 11 which was created using Hot Spot Analysis (Getis_ Ord Gi*) tool in ArcMap with (value = TBD, Threshold distance = 25 Kilometers). This result shows statistically significant hot spots (Red), cold spots (Blue) and not significant (Yellow). As can be seen from the map, there are evidences for spatial relationship.

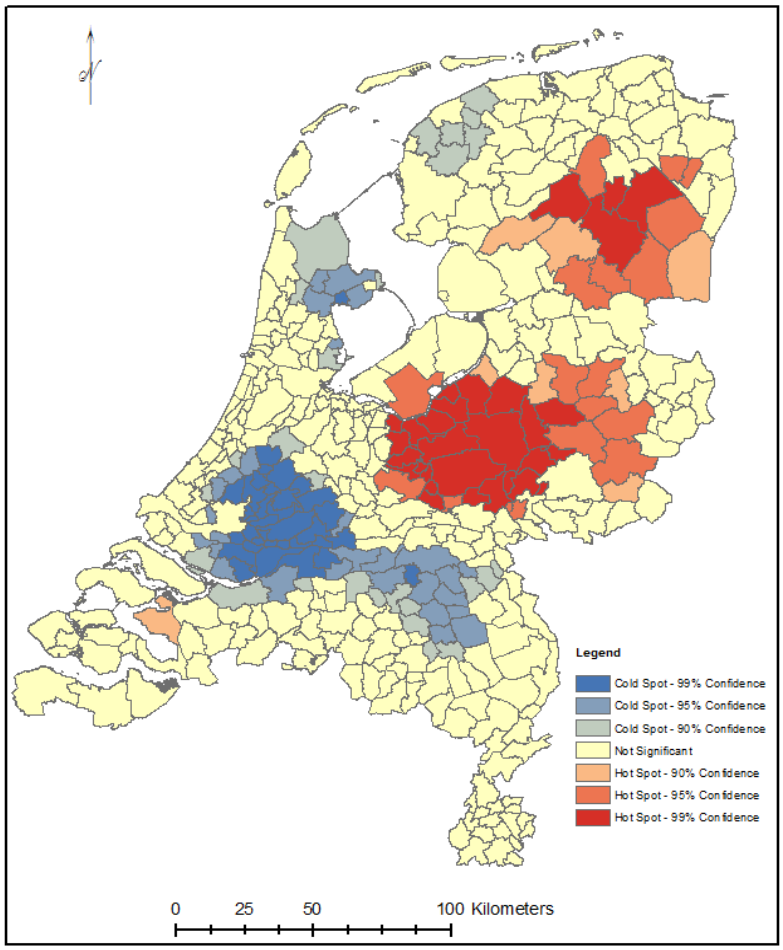


Figure 11: Tick bite observations' hotspot analysis result 2011-2014

The term "confidence" in the legend of the map indicates that it can be said with the given percent of certainty that the density values are spatially associated. That is, areas of high density and areas of low density are related to each other at the indicated level of certainty.

To collect more statistical evidence whether to reject or not to reject the null hypothesis, the data is further analyzed on yearly basis. The method described above for the whole dataset with the same parameters was applied. The analysis results for each year are given in Figure 12 below.

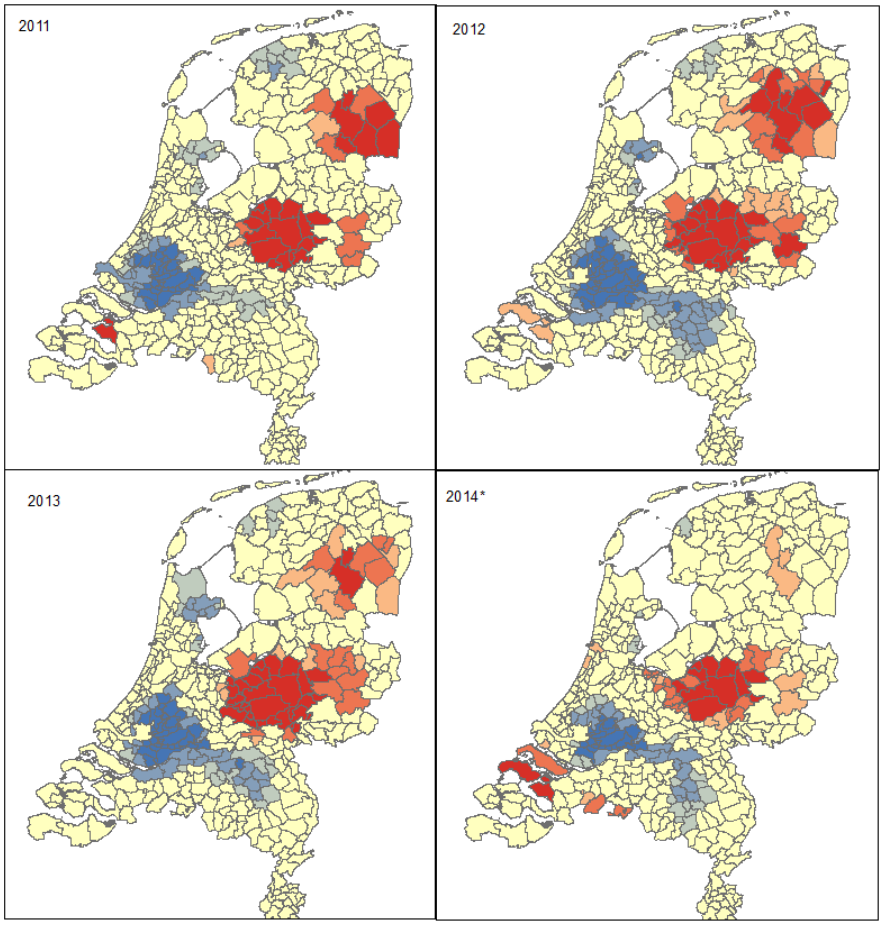


Figure 12: Yearly tick bite observations' hotspot analysis result

The result in Figure 12 again shows statistically significant hot spots (Red), cold spots (Blue) and not significant (Yellow) for all years in similar locations like what was observed for the whole dataset all together. Even for the six months data of (2014*), the results alike the others.

5.3.4. Temporal distribution of tick bites

Understanding the spatial distribution of tick bites in space alone is halfway to understand the phenomenon as the phenomenon happens in a certain space at a certain time. The analysis and understanding of the temporal characteristics of such phenomenon is essential to have a complete picture.

Individuals reporting a tick bite may do the reporting immediately or in a later time. There is also a possibility that they do not remember the exact date they get the tick bite. To get a representative temporal analysis result one should take in to account the temporal scale within which the data is aggregated. The possible temporal scale for understanding the seasonality of the tick bite is day, week or month. For this project the data is aggregated by week and the temporal distribution of the tick bites is analyzed accordingly.

It is also possible that they do not follow the same pattern year after year. So, a null hypothesis for the correlation between 2012 and 2013 temporal pattern is stated as:

$$\boldsymbol{H_0}: \rho = 0 \quad (3)$$

$$\boldsymbol{H_a}: \rho > 0 \quad (4)$$

where ρ is the spearman's correlation coefficient

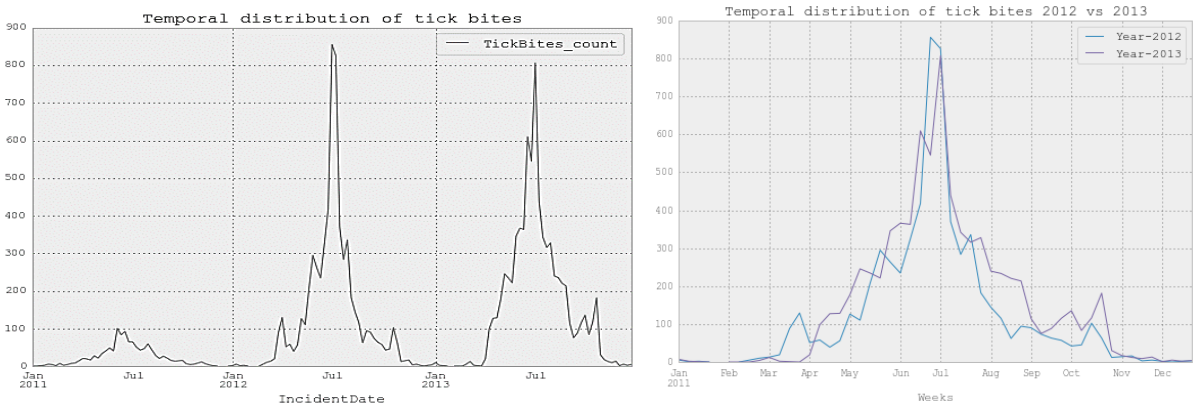


Figure 13: Tick bite observations' temporal distribution plot 2011-2013 (left), 2012 vs 2013 (right)

For the years 2011-2013 Figure 13 (left), the maximum tick bite incidents are observed during the summer season. It is evident from the temporal distribution of the tick bites that the large number of tick bite incidents occurred in the years 2012 and 2013.

The regions of maximum incidents for the years 2012 and 2013 Figure 13 (right), look to be strongly related. A spearman's correlation method was run to test the null hypothesis stated in equations 3 and 4 for the two years and strong positive correlation ($\rho = 0.77, p<0.001$) was observed.

Table 12. Monthly distribution of tick bites for 2011-2013

Month	Yearly distribution of tick bites					
	2011	2011(%)	2012	2012(%)	2013	2013(%)
January	13	1.07	12	0.19	15	0.19
February	18	1.49	9	0.14	6	0.08
March	45	3.72	251	3.96	22	0.29
April	94	7.77	243	3.84	415	5.39
May	189	15.62	886	13.99	1052	13.67
June	366	30.25	1682	26.55	2023	26.29
July	225	18.60	2107	33.26	2044	26.56
August	121	10.00	513	8.10	995	12.93
September	75	6.20	329	5.19	510	6.63
October	36	2.98	263	4.15	525	6.82
November	24	1.98	48	0.76	66	0.86
December	4	0.33	12	0.19	22	0.29
Total	**1210**	**100.0**	**6355**	**100.00**	**7695**	**100.00**

As can be observed from the Table 12 the months June and July are with the highest proportion of tick bite observations for all the years. The number of tick bite observations in 2011 are observed to be by far smaller than the other two years.

5.4. Spatio-temporal analysis of related *Flickr* photos

5.4.1. Representation of tick bite in social media

Data representing actual tick bite or Lyme disease which have the required attributes to perform spatio-temporal analysis were not found in selected social media platform. Although large number of photos were found, the main component that is the geographic location was missing in these photos. The geolocated Flickr photos representing tick bites (that is searched for the search terms representing ticks) turned out to be effectively zero (only 6 photos were found for the years 2011-2014).

5.4.2. Contextual environmental and outdoor activity data

To understand if there is a possible link between tick bites and the contextual *Flick* photos, the distribution of the *Flickr* photos should first be analyzed. It is important to find out how they are distributed over the country, where the high concentration is, and whether there is any spatial correlation. First, the CPD (contextual photo density) was calculated to get the photos per unit area of each municipality. Then the analysis is made in a similar fashion like that of the tick observations.

A density maps Figure 14 was then created using the resulting data to find out the municipalities with high and low density of photos.

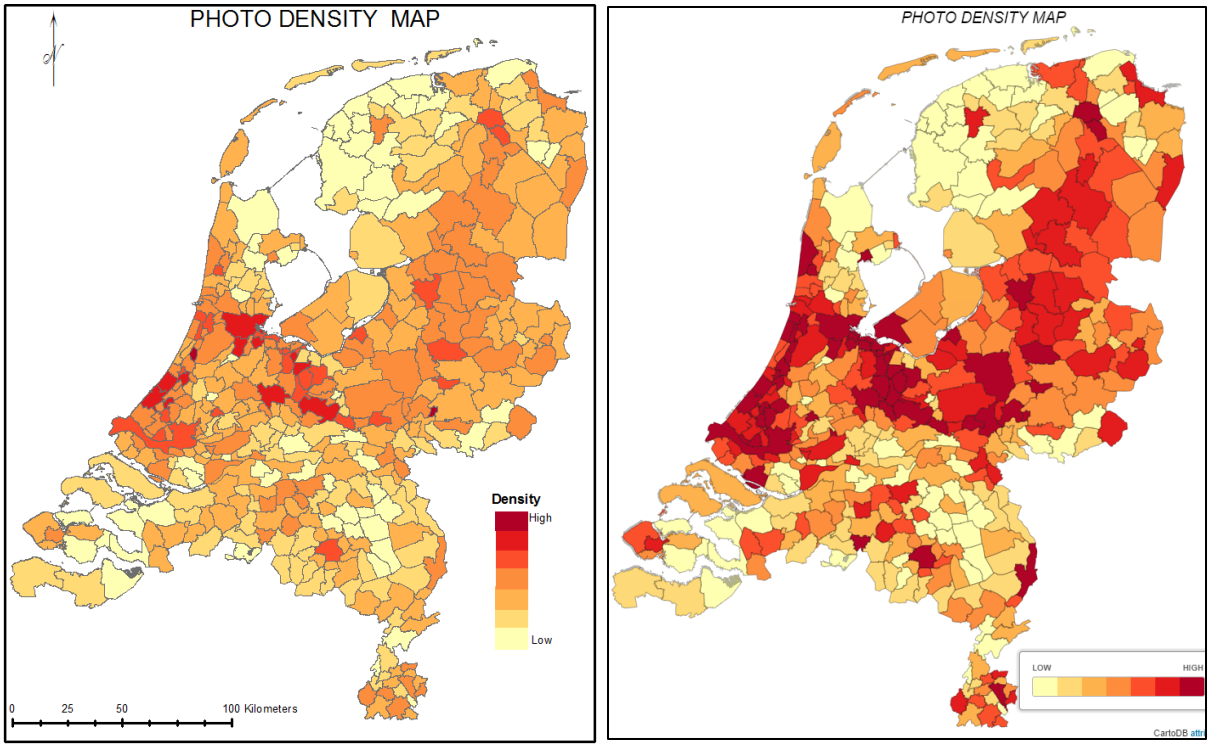


Figure 14: Flickr photo density maps, ArcMap (left) and CartoDB (right)

The result given in Figure 14 shows concentration of photos in municipalities from low (yellow) to high (dark red). The map to the left represents a density map classified as (value = CPD, Classes = 7, Classification = Geometrical Interval) created in ArcMap and the map to the right shows density map with

(value=CPD, Classes = 7, Quantification = Quantile) created using CartoDB Editor[38]. The choice of different classification methods here is because, CartoDB has not implemented 'Geometrical Interval' options. The difference that can be observed in the two maps is the result of the classification methods.

As can be observed from the maps in Figure 14, there are contiguous municipalities with high density of photos in the middle west-coast, the central, and north eastern part of the country. The municipalities at the center are those that accumulate more forest surfaces. Municipalities located in the central and central west-coast are visited by many people for recreation and tourism.

5.4.3. Photo hot spots

Taking the actual point features representing the location of photos as an input to investigate the underlying pattern, the KDE method was run in ArcMap using the contextual photos. The same data was used to create a T*orque heat* map in CartoDB to investigate the spatio-temporal distribution of the contextual photos. The resulting KDE map and the static version of the *heat* map are presented in Figure 15.

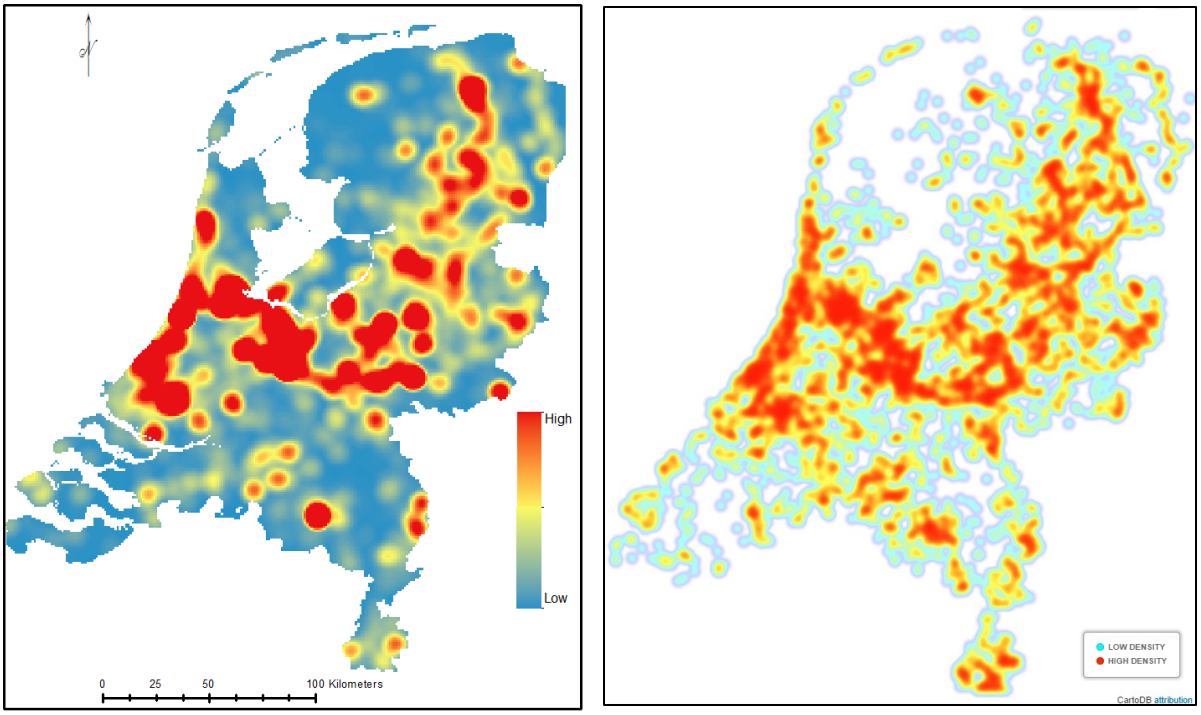


Figure 15: Flickr photos kernel density (left) and heat (right) map

The KDE map (left) in Figure 15 is created using Kernel density tool in ArcMap and represents (kernel density of 1000 raster cell size, symbolized with 1 standard deviation stretch) whereas, the *heat* map (right) was created with similar notion (identifying actual hotspots) using the same input, but with different method in CartoDB. The *heat* map was created with (Marker size=15, Threshold=0.4, Resolution=4) and depicts

[38] *The account that is used for this project is based on the free license plan. CartoDB can put down free accounts at any time. The author cannot guarantee the availability of the mapping product that is created in the platform. The published map can be accessed using this the link* http://beribu.cartodb.com/viz/e849ff1c-ade1-11e4-907a-0e018d66dc29/public_map

the photo hot spots. The animated version of this map can be accessed from the free license[39] account on which it was created.

Both maps in the result in Figure 15 show density estimates from low (blue) to high (red). As can be seen from the maps, the photos are observed to be clustered around the middle west-coast, the central, and north eastern part of the country. As discussed in the results in section 5.4.2 the location of the hotspots mainly the central and the middle west- coastal areas are highly visited for recreation and tourism.

Having observed clustering of the point locations of the photos in the analyses done so far in both platforms, it is important to look into the statistical significance of the identified hot spots. That is if there is statistically significant spatial relationship among the areas with large number of contextual photos.

5.4.4. Spatial statistical analysis

To find out whether the distribution per municipality of the photos is based on complete randomness or there is a spatial correlation among the municipalities, Getis_ Ord Gi* used to test the complete spatial randomness hypothesis. The complete spatial randomness hypothesis is stated as follows:

$\boldsymbol{H_0}$: The clustering of municipalities with similar density of photos is completely random

$\boldsymbol{H_a}$: The clustering of municipalities with similar density of photos is not random

The resulting hot spot map is showed in Figure 16. This map is created using Hot Spot Analysis (Getis_ Ord Gi*) tool in ArcMap and represents (Significant hot spots and cold spots for photo density values per municipality at 25 Kilometers distance threshold).The map shows statistically significant hot spots (Red), cold spots (Blue) and not significant (Yellow). As can be seen from the map, there are evidences for spatial relationship.

[39] *CartoDB can put down free accounts at any time. The author cannot guarantee the availability of the mapping product that is created in the platform. The published photo hot spot map can be accessed using this the link* https://berihu.cartodb.com/viz/b671dfd8-b2b8-11e4-870f-0e4fddd5de28/map

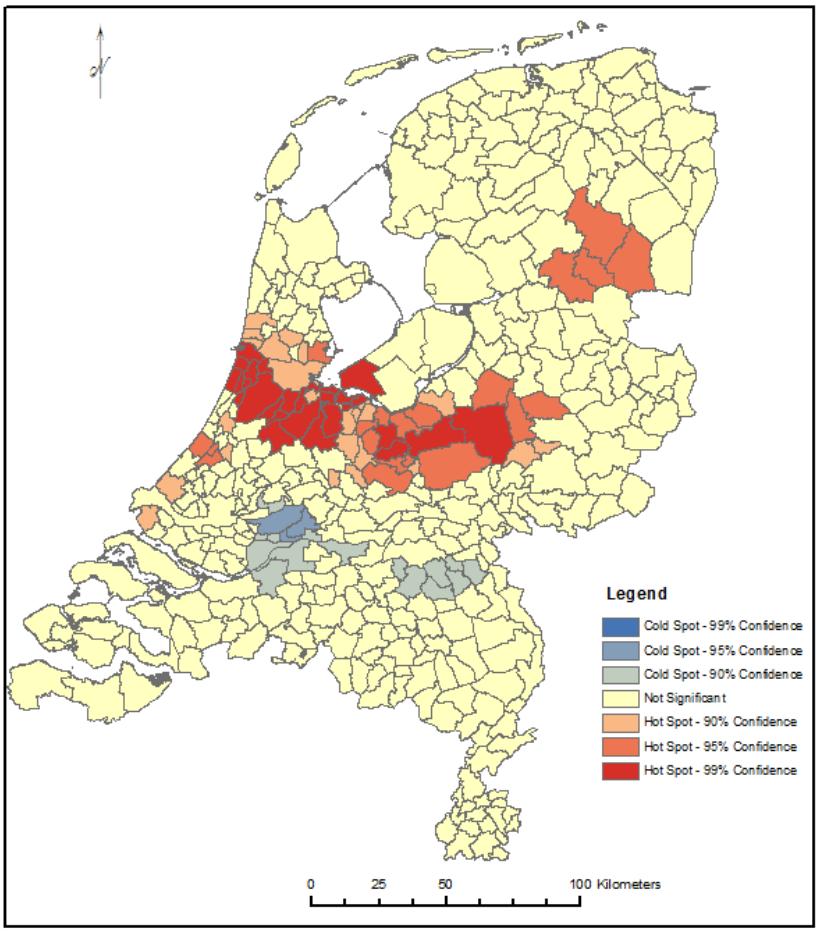


Figure 16: Flickr photos hotspot analysis result 2011-2014

To collect the statistical evidence whether to reject or not to reject the null hypothesis, the data is further analyzed on yearly basis. The method described above for the whole dataset with the same parameters was applied. The analysis results for each year are given in Figure 17.

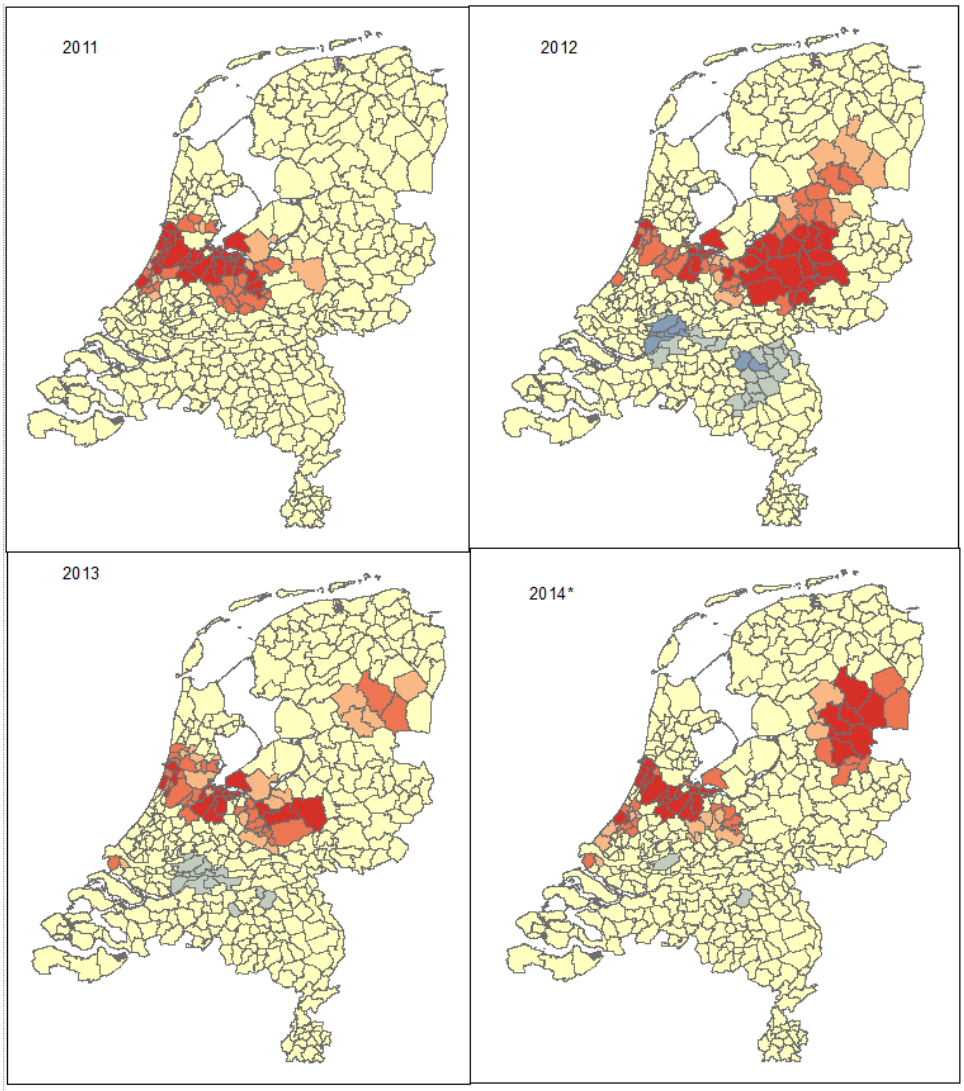


Figure 17: Yearly Flickr photos hotspot analysis result

Note: (2014*) represents the data for six months (January-June)

The result in Figure 17 shows statistically significant hot spots (Red), cold spots (Blue) and not significant (Yellow). Significant hot spots and cold spots of photos are observed for all the years. However, distribution of the photos over time is not similarly located over the country.

5.4.5. Temporal distribution of photos

Here again understanding the spatial distribution of photos in space alone is not sufficient to understand the phenomenon as it happens in space and time. Special attention shall be given to the temporal analysis of the photos as the primary purpose of studying them is to find out whether the information in these geolocated photos can be used to improve our understanding of the tick bite distribution by inference.

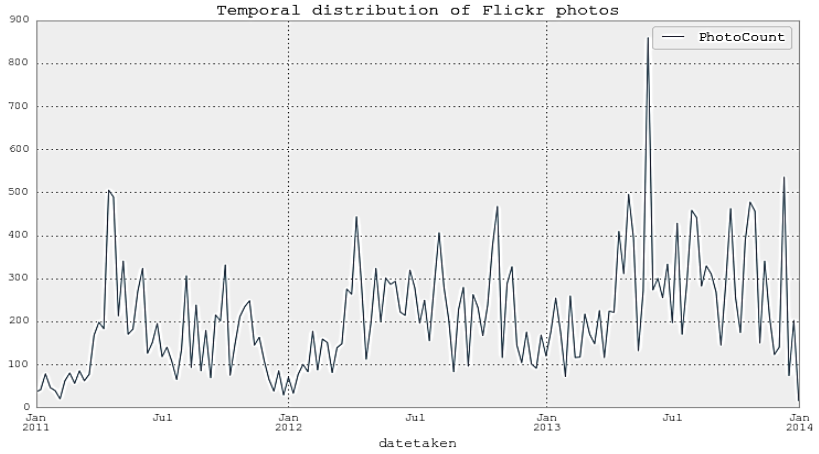


Figure 18: Temporal plot for Flickr photos distribution

As can be seen from the temporal plot, the temporal distribution of the photo is highly fluctuating which makes it difficult to clearly identify the location of absolute maximum and absolute minimum.

The above fact can be observed from the Table 13 as there is high variability in the monthly distribution of the proportion of photos for all the years. The months with large number of photos (larger than 10%) are for 2011 (April, May, and October), for 2012 (April, May, June, and October), and for 2013(April, June, August, and October). In general the number of photos per year tend to be increasing.

Table 13. Monthly Distribution of photos 2011-2013

Month	Yearly distribution of tick bites					
	2011	2011(%)	2012	2012(%)	2013	2013(%)
January	241	2.98	301	2.74	743	5.21
February	219	2.71	584	5.32	571	4.01
March	458	5.66	841	7.67	849	5.96
April	1470	18.16	1173	10.69	1358	9.53
May	1091	13.48	1181	10.76	1313	9.21
June	785	9.70	1135	10.35	1806	12.67
July	496	6.13	1065	9.71	1138	7.99
August	754	9.31	954	8.70	1698	11.91
September	771	9.52	972	8.86	1253	8.79
October	986	12.18	1296	11.81	1581	11.09
November	575	7.10	892	8.13	932	6.54
December	250	3.09	577	5.26	1011	7.09
Total	**8096**	**100.0**	**10971**	**100.00**	**14253**	**100.00**

Due to the high variability observed in Figure 18 and Table 13, it was found very important to have a closer look into the distribution. That is the photos are examined on a yearly basis with the same temporal scale separately.

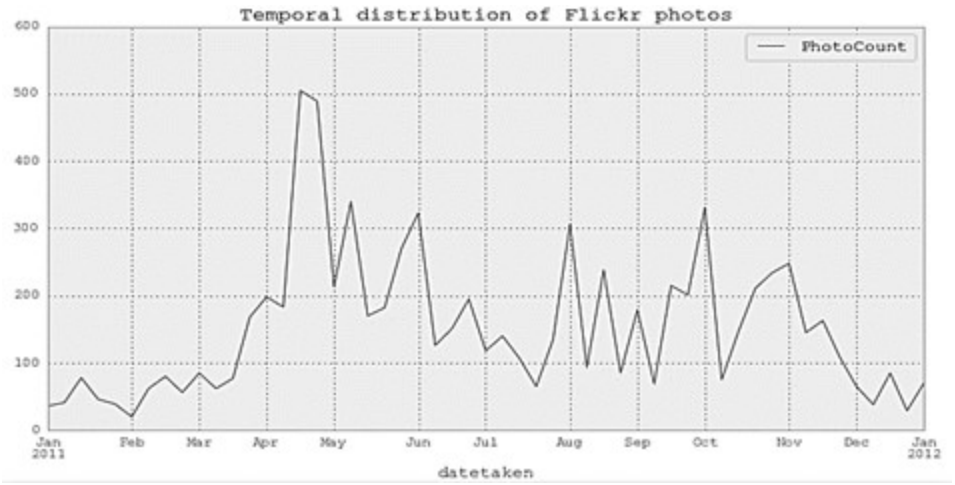


Figure 19: Temporal plot of Flickr photos for 2011

Although there is still high variability in the distribution of photos the maximum can be easily identified from Figure 19. It was observed that the maximum number of photos is in the days 17-31 April, of the year. Another large number for consecutive weeks is observed from 18, September to 2, October of the year. The weeks of the first peak were in the weeks of flower parade and Queen's Day in that year in the Netherlands. The second peak again is the weeks of flower parade.

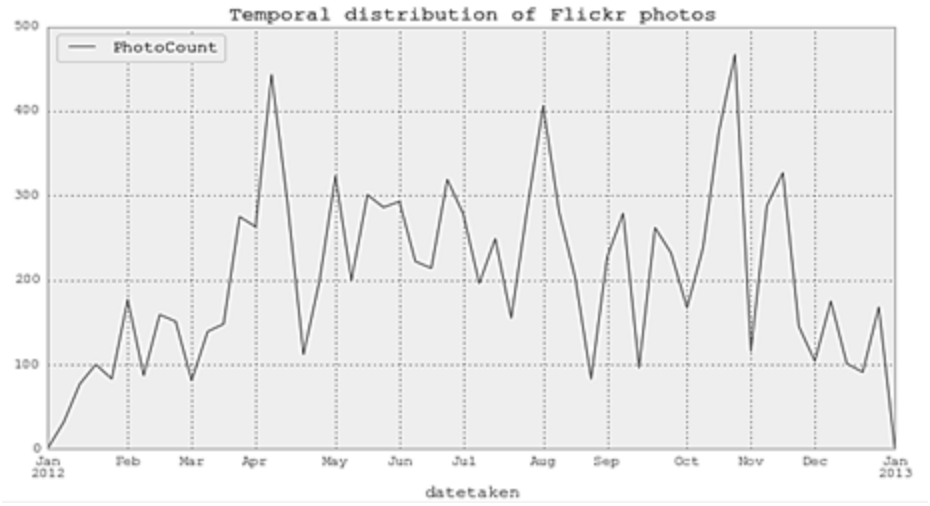


Figure 20: temporal plot of Flickr photos 2012

Here again, irrespective of still high variability in the distribution of photos three maximum can be easily identified from Figure 20. The first maximum is from 1 -15, April, in the particular year. Another large number for consecutive weeks is observed from 29, July to 25, August, and a third peak is in the days 14-28, October. News archives were searched to explain the first and third peaks, however there were no major events in the time frame. The weeks of the second peak are the weeks of many summer festivals.

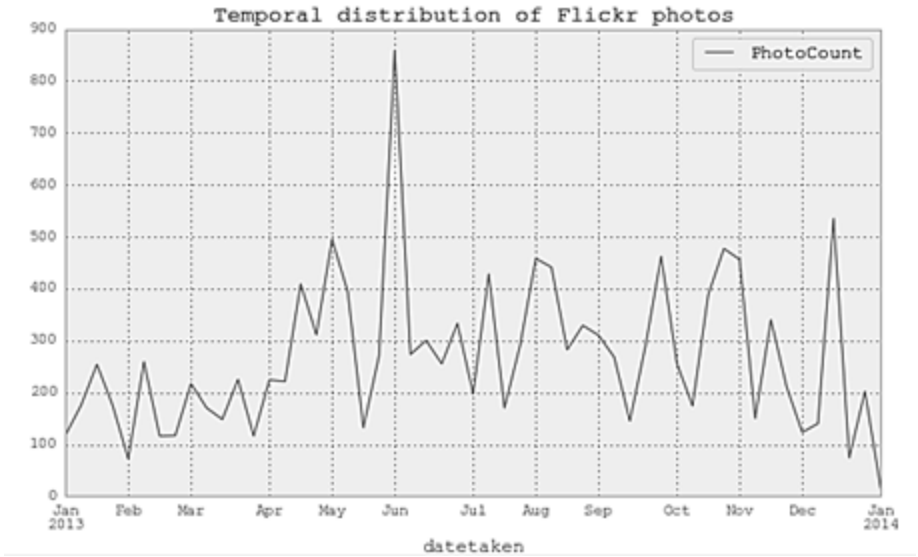


Figure 21: Temporal plot of Flickr photos for 2013

There is only one high peak that can be easily identified from Figure 21. The maximum is from 2 -8, June in the year 2013. News archives were searched to explain the peak, however there was no information about major events in the time frame. The variability in the other weeks is very high in the year.

5.5. Analysis of relationships among datasets

As discussed in earlier chapters, there is bias due to the missing and mixed information in the attributes of the tick bite information (environment and activity) that should be minimized using other data sources. The auxiliary datasets (land cover data and the social media extract) are considered as the sources to address this. The unknown environmental information as well as the mixed environmental information as a result of multiple land cover values reported for a single tick bite incident can be obtained from the land cover data. Indeed, provided that the tick bite observations and photo extracts are strongly related in space and time, the missing or biased information related to the activity can also be minimized using the activity related photos and their description by implication. In this section, the relationship between tick bite observations and land cover as well as the relationship between the tick bite observations and photo extracts are analyzed.

5.5.1. Relationship between tick bites and land cover

The environment information associated to tick bites that suffers from mixed land cover values amounts to more than 20% of the total. In addition to mixed land cover information 8.6 % of the data does not have environmental information. To fill the information gap that is evident to be introduced as a result of the missing values and the bias in the mixed types, the environmental information is extracted from the land cover data. To do so, the point data representing tick bite observations was spatially intersected with the land cover data resulting in a tick bite observation data with actual land cover on which the incident occurred.

The summary of the tick bite observations created from the intersection between the tick bites as reported and the land cover data Table 14 is given as follows.

Table 14. Summary of the distribution of tick bites per land cover

Land cover	Observed tick bites	Percentage (%)
Forest	8001	42.59
Build-up areas	2721	14.48
Sparse vegetation	1974	10.51
Rain-fed Croplands	1893	10.08
Grassland	1438	7.65
Croplands and Vegetation	1427	7.60
Shrub-land and grassland	1060	5.64
Water bodies	235	1.25
Bare areas	16	0.1

It can clearly be seen form Table 14 that more than 57% of the tick bites occur either in a forest or in build-up areas. It can also be observed that 77.66% of the incidents happened in areas covered by only four of the land cover classes.

5.5.2. Relationship between photos and land cover

As discussed in the social media data harvest section of Chapter 3, the contextual social media data (*Flicker* photos) are collected using search terms developed form the tick bite observations and the report forms used in the application (*tekenradar.nl*) used to report the incidents. The actual land cover information in these photos especially those collected for the outdoor activities is not available. So, the actual land cove on which the photos are taken is extracted by spatially intersecting the point dataset representing the photo extracts and the land cover dataset. The result is summarized in Table 15.

Table 15. Summary of the distribution of photos per land cover

Land cover	Number of photos	Percentage (%)
Forest	13115	32.11
Build-up areas	11089	27.15
Sparse vegetation	3508	8.59
Rain-fed Croplands	3856	9.44
Grassland	2860	7.00
Croplands and Vegetation	2935	7.19
Shrub-land and grassland	2669	6.53
Water bodies	763	1.87
Bare areas	46	0.11

It can clearly be seen form Table 15 that more than 59.26% of the photos are taken either in a forest or in build-up areas.

5.5.3. Relationship between photos and tick bites

To find out whether the photos can be used to improve or confirm the understanding gained form the tick bite dataset, the spatio-temporal relationship between the two should be strong. That is the distribution in space and time of the two datasets should be strong enough indicating that they are linked to one and the same event so that information from one can be used to improve understanding the other. A strong relationship in space in this case is meant there are large number of photos in places with large number of tick bites and small number of photos in areas of small number of tick bite incidents. A strong relationship in time at the same time means there are large numbers of photos in time slots with large number of tick bites.

The tick bite observations and photos are aggregated using the same aggregation units of both space and time. First, both VGI datasets are aggregated using the land cover. That is the polygons representing individual land cover features that are extracted from the land cover image. On the one hand, the datasets summarized as in Figure 22 using the land cover classes to have a general idea of the relationship. On the other hand, the analysis is performed using the individual features in the dataset and their associated attribute values representing the number of tick bite and number photos per each feature to find out if they are related locally. The values of two attributes (number tick bites and number of photos) of the individual land cover features were used as an input for the scatterplot Figure 23 to explore the relationship and spearman's correlation method to perform hypothesis testing. The null and alternative hypotheses for this case are stated as:

$$\boldsymbol{H_0}: \rho = 0 \quad (5)$$

$$\boldsymbol{H_a}: \rho \gg 0 \quad (6)$$

where ρ is the spearman's rank correlation coefficient

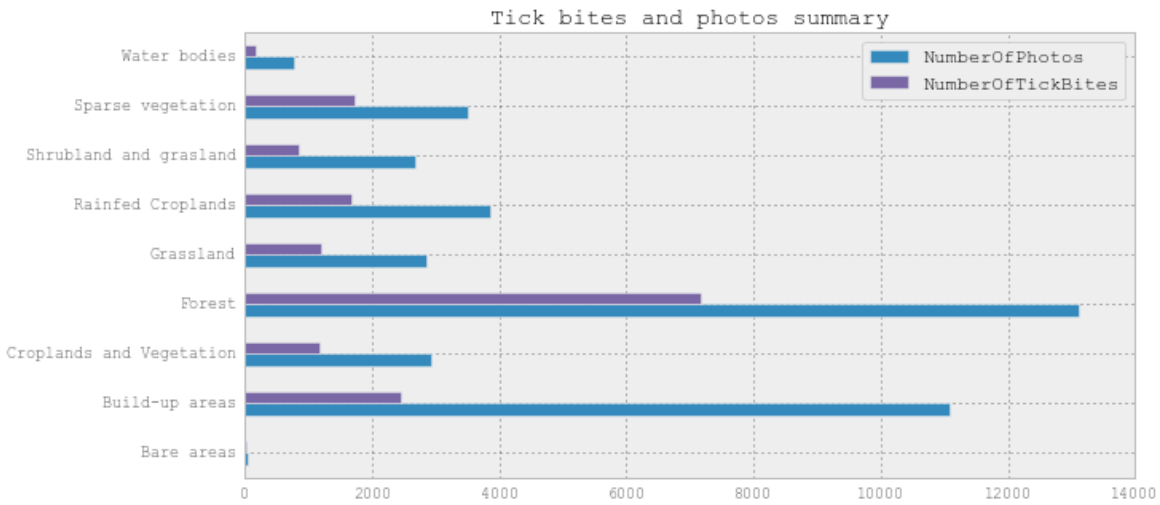


Figure 22: photos versus tick bite summary per land cover

The generalized summary in Figure 22 above shows that large number of tick bite per land cover are associated with large number of photos per land cover. It is worth investigating these relationship using the individual features to find out if they are indeed related locally.

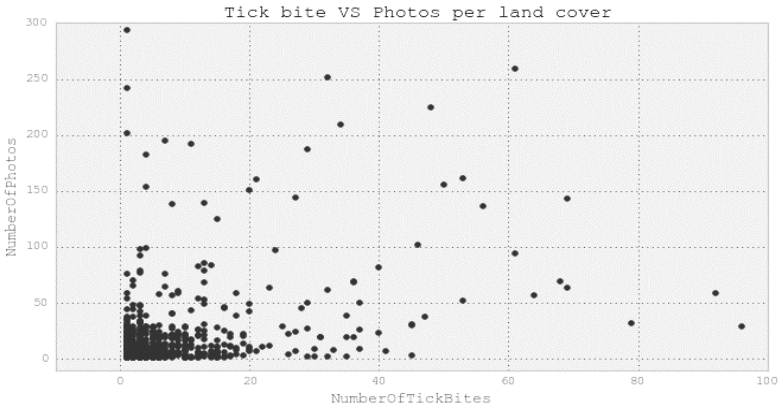


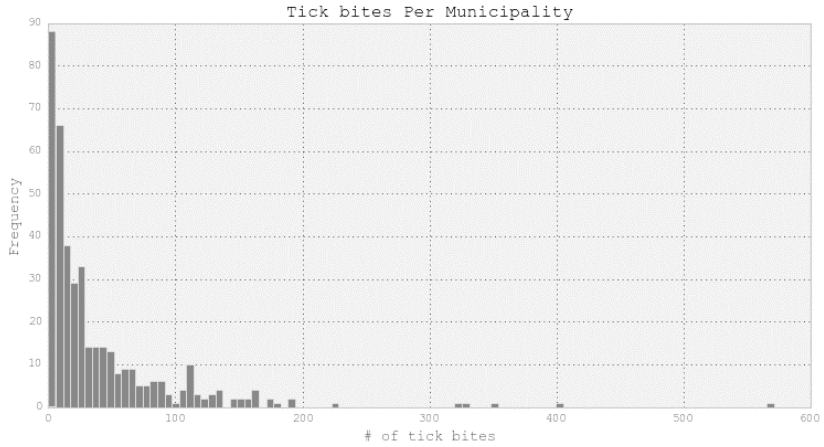
Figure 23: photos versus tick bite scatter plot

In this analysis, land cover feature classes for which both the number of tick bite and number of photos are zero were excluded. A spearman's correlation method was then run to determine the relationship between the two attributes. A positive correlation between the two ($\rho = 0.40, p<0.001$) was observed.
To collect more statistical evidences, the data used to identify the hot spots of municipalities with high densities of tick bites and photos in previous sections were used to perform bivariate analysis. The relationship between the numbers and the densities of both tick bites and photos were analyzed to investigate the relationship in space for the same null hypothesis stated above.

Table 16. Summary of the data extracted for the municipality aggregate sorted by number of photos

Municipality	# of Tick bites	# of photos	TBD	CPD
Amsterdam	193	3263	0.98	16.60
Rotterdam	107	1491	0.39	5.41
Lisse	15	1482	0.93	92.31
Utrechtse Heuvelrug	350	1434	2.61	10.69
Utrecht	83	1250	0.84	12.60
's-Gravenhage	118	1076	1.39	12.70
Apeldoorn	571	988	1.67	2.90
Baarn	111	761	3.36	23.06
Amstelveen	58	706	1.32	16.02
Midden-Drenthe	115	660	0.33	1.91
Eindhoven	47	596	0.53	6.71
Wassenaar	124	577	2.35	10.95
Groningen	111	571	1.33	6.82
Ede	403	549	1.26	1.72
Doesburg	4	480	0.31	37.04
Deventer	95	477	0.71	3.55
Westerveld	226	446	0.80	1.58
Zwolle	57	414	0.48	3.47

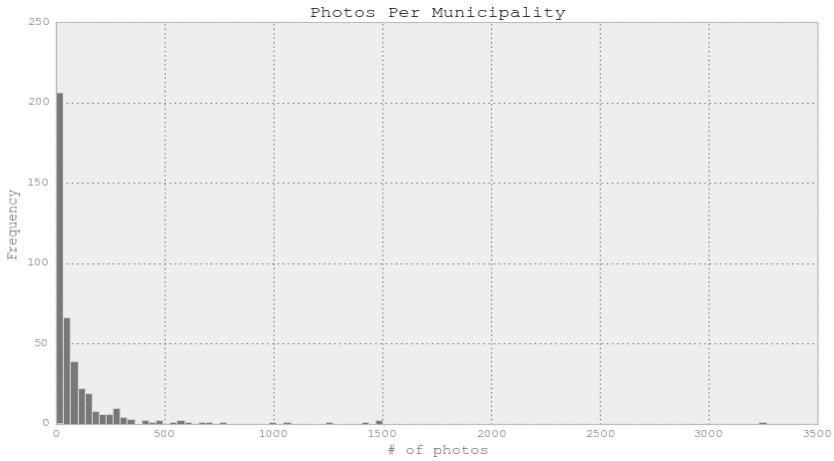
Note that: Table 16 shows part of the municipalities. The complete list is provided in the Appendix B. of this document



Summary statistics:

N	408
Mean	40.12
std	58.64
Min	0.00
25%	7.00
50%	19.00
75%	50.00
Max	571

Figure 24: Number of tick bite per municipality



Summary statistics:

N	408
Mean	99.84
std	243.32
Min	0.00
25%	11.00
50%	31.00
75%	91.50
Max	3263

Figure 25: Number of photos per municipality

As can be observed from Figure 24 and Figure 25, the histograms are skewed right indicating that there are few number of municipalities with an extreme values (outliers) in both datasets. The standard deviation (std) in both datasets indicates that the values are highly dispersed. However, the histogram and measure of standard deviation representing the photos shows more dispersion than the one representing the tick bites.

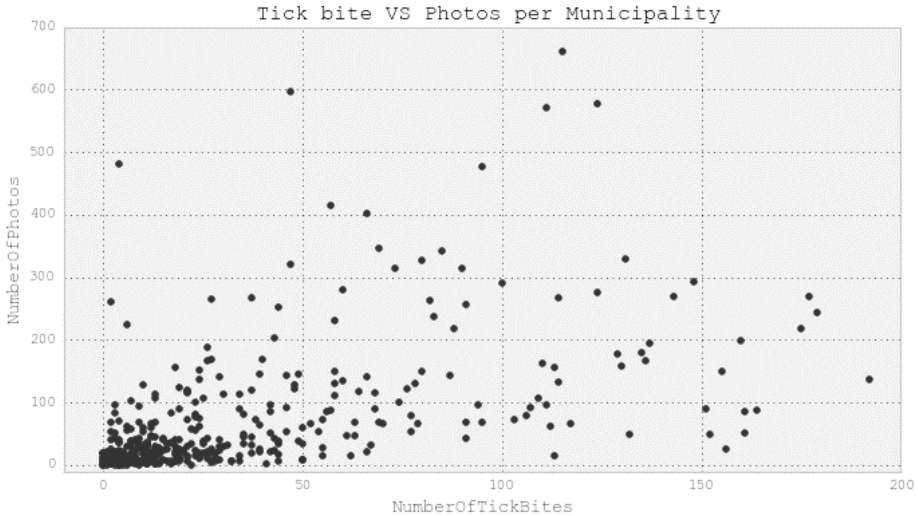


Figure 26: tick bite vs photos per municipality scatterplot

A spearman's correlation method was run to determine the relationship between the two datasets as distributed in the municipalities. A positive correlation between the two ($\rho = 0.48, p<0.001$) was observed.

Having observed the relationship over the municipalities, it is essential to look at the relationship in time of the two datasets. The relationship over time is evaluated for the full 3 years as follows.

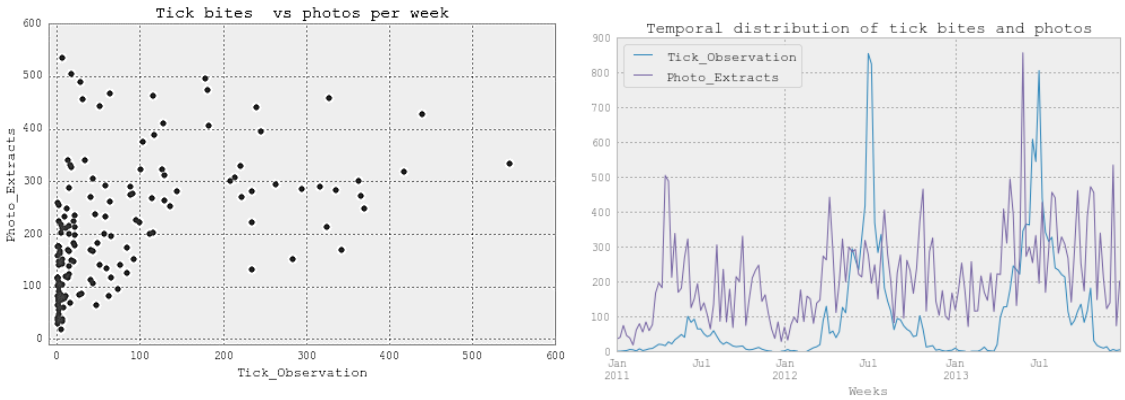


Figure 27: Scatterplot (left) and temporal plot (right) for tick bite and photo data per week for the period 2011-2013

A spearman's correlation method was run to determine the temporal relationship between the two datasets as distributed in the weeks over the years 2011-2013. A positive correlation between the two ($\rho = 0.51$, $p<0.001$) was observed.

At this point we have all environmental information about the tick bites and we know what the tick bite prone areas are. And also we know that municipalities with large number of tick bites have large number of photos. Besides, we know the two datasets are related in time to a certain extent. Therefore, it crucial to identify the social activities that are associated with the tick bite incidents. Here is where the contextual social media data comes to play.

The results of the analyses for relationships between the two VGI datasets to improve our understanding of the social activities related to tick bite incidents is presented as follows. From the land cover by land cover analyses, only the results for forest and built-up area are included in this thesis. This is because, the results indicating the association between the two VGI datasets of observations at the level for other land cover classes are similar or less than what is already observed in the context of forest and built-up areas.

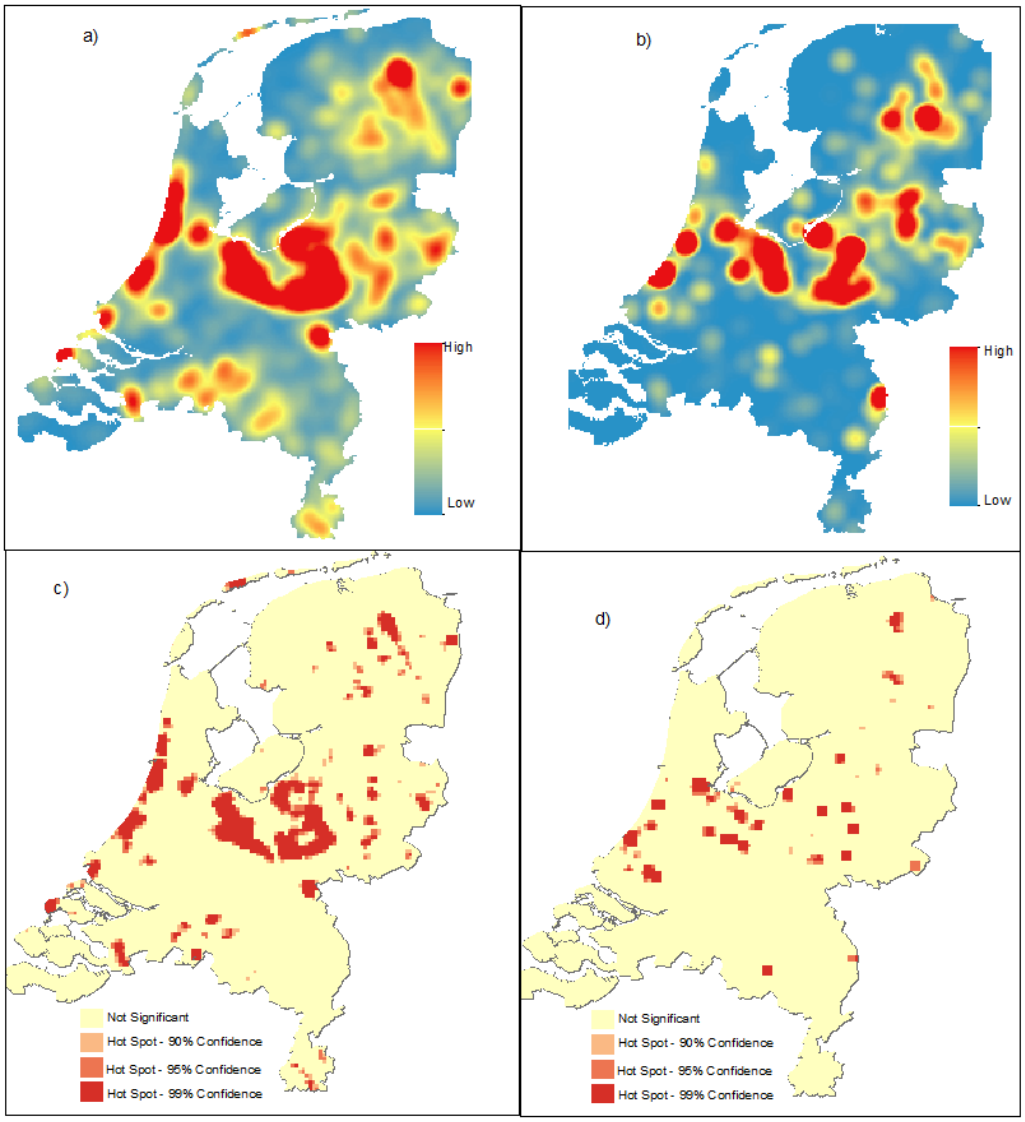


Figure 28: kernel density estimate of tick bites (a), kernel density estimate of all activity photos (b), tick bite hotspots for the same data(c) and photo hotspots for the same data (d)

In Figure 28 a) and b) are created using Kernel density tool in ArcMap and represent (kernel density with 1000 raster cell size, symbolized with 1 standard deviation stretch).These results show density estimate from low (blue) to high (red). As can be seen from the map, the tick bites and the activity photo are observed to be clustered in similar areas of the country.

It is then important to evaluate the statistical significance of each for using activity photos to understand the social activities better. Figure 28 c) and d) are created using Hot Spot Analysis (Getis_ Ord Gi*) tool in ArcMap.

Map (c) represents significant hot spots and cold spots for (value = *number of tick bites per grid cell*, Threshold distance = 3.4 kilometres). Map (d) represents significant hot spots and cold spots for (value = *number of photos per grid cell*, Threshold distance = 3.4 kilometres) for tick bites and activity related photos. In both cases, *Red* indicates statistically significant hot spots and *Yellow* represents not significant.

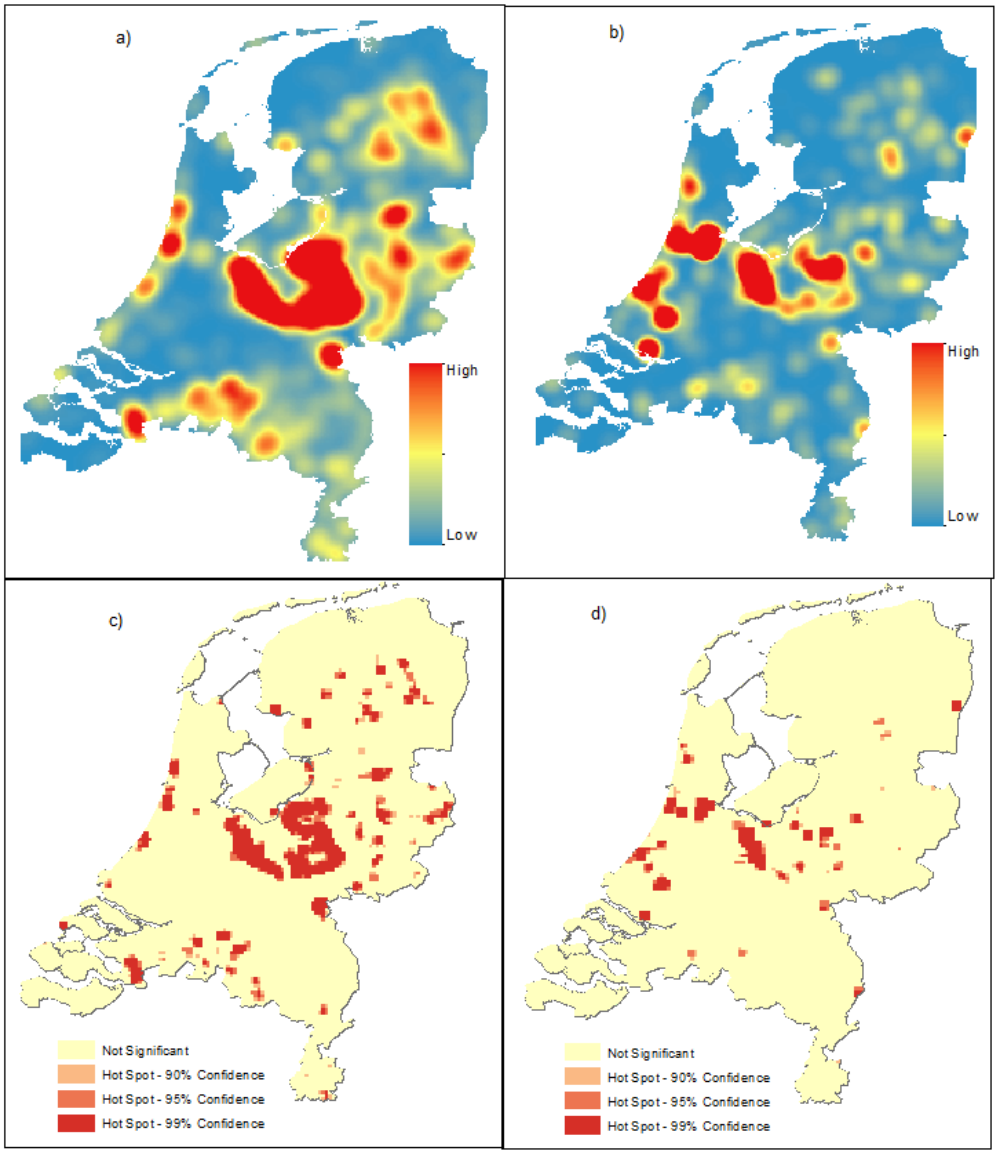


Figure 29: kernel density estimate of tick bites (a), kernel density estimate of activity photos (b), tick bite hotspots for the same data(c) and photo hotspots for the same data (d) located in the forest

In Figure 29 a) and b) are created using Kernel density tool in ArcMap and represent (kernel density with 1000 raster cell size, symbolized with 1 standard deviation stretch).The results a) and b) show density estimate from low (blue) to high (red). As can be seen from the map, the tick bites located in forest and the activity photo located in forest are observed to be clustered in different locations.

It is then important to evaluate the statistical significance of each to confirm the difference. Figure 29 c) and d) are created using Hot Spot Analysis (Getis_ Ord Gi*) tool in ArcMap.
Map (c) represents significant hot spots and cold spots for (value = *number of tick bites per grid cell*, Threshold distance = 3.4 kilometres). Map (d) represents significant hot spots and cold spots for (value = *number of photos per grid cell*, Threshold distance = 3.4 kilometres) for forest. In both cases, *Red* indicates statistically significant hot spots and *Yellow* represents not significant.

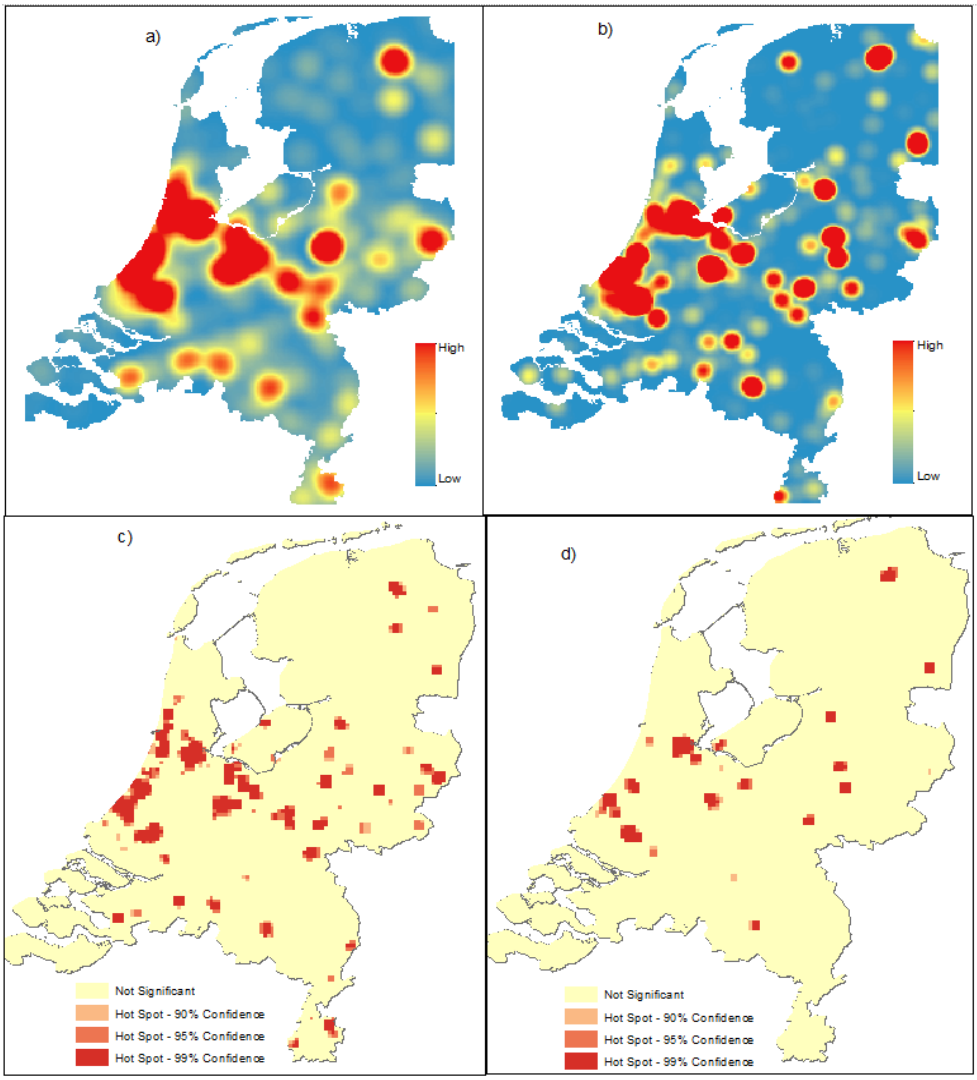


Figure 30: kernel density estimate of tick bites (a), kernel density estimate of activity photos (b), tick bite hotspots for the same data(c) and photo hotspots for the same data (d) locate in built-up areas

In Figure 30 a) and b) are created using Kernel density tool in Spatial Analyst Tools tool box of ArcMap and represent (kernel density with 1000 raster cell size, symbolized with 1 standard deviation stretch).The results a) and b) show density estimate from low (blue) to high (red). As can be seen from the map, the tick bites located in Built-up areas and the activity photo located in built-up areas are observed to be clustered in similar areas of the country.

It is then important to evaluate the statistical significance of each to confirm the similarity. Figure 30 c) and d) are created using Hot Spot Analysis (Getis_ Ord Gi*) tool in ArcMap.
Map (c) represents significant hot spots and cold spots for (value = *number of tick bites per grid cell*, Threshold distance = 3.4 kilometres). Map (d) represents significant hot spots and cold spots for (value = *number of photos per grid cell*, Threshold distance = 3.4 kilometres) for built-up areas resulting in different maps. In both cases, *Red* indicates statistically significant hot spots and *Yellow* represents not significant.

The two datasets in each context are temporally analyzed to investigate their association in time for which the scatterplots, temporal plots and correlation are given as follows. The data are aggregated at a temporal scale of 1 week.

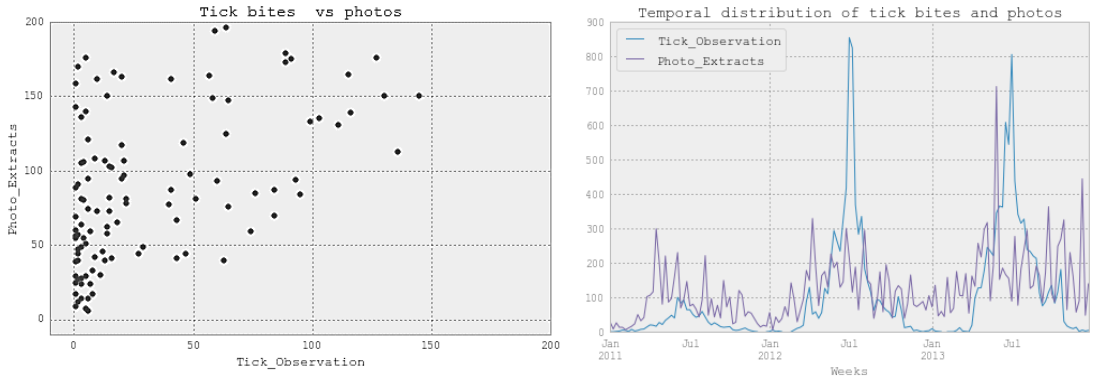


Figure 31: scatter plot (left) and temporal plot (d) of tick bite versus outdoor activity photo extracts

From the scatter plot and the temporal graph of Figure 31, it can be seen that the two datasets are weakly related. In addition a spearman's correlation was run to determine their association at this temporal scale. A positive correlation between the two (ρ= 0.50, p<0.0001) was observed.

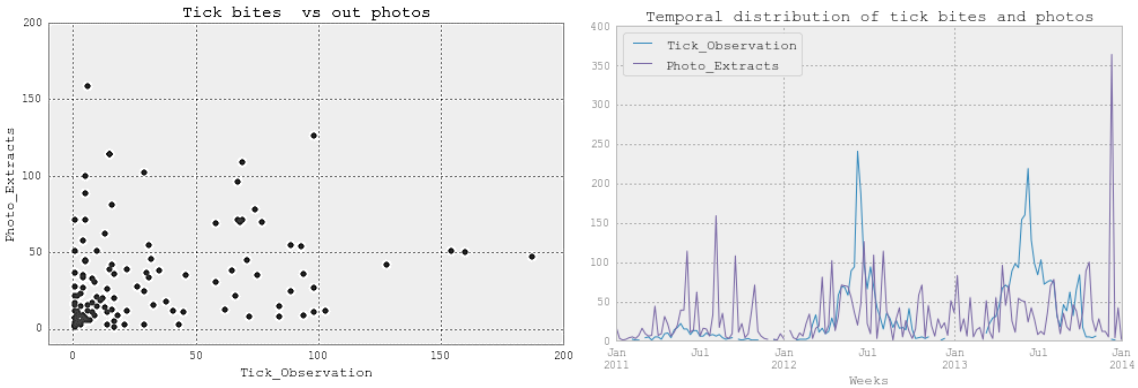


Figure 32: scatter plot (left) temporal plot (right) for tick bite and photos located in Forest

From the scatter plot and the temporal graph of Figure 32, it can be seen that the two datasets are weakly related. In addition a spearman's correlation was run to determine their association and a negligible, positive correlation (ρ= 0.12, p>0.13) was observed.

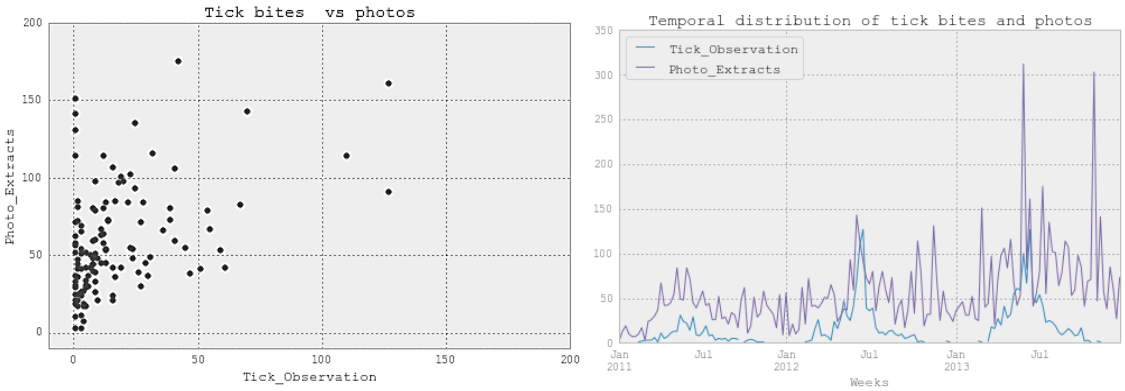


Figure 33: scatter plot (left) temporal plot (right) for tick bite and activity photos located in built-up areas

From the scatter plot and the temporal graph of Figure 33, it can be seen that the relationship is weak. In addition a spearman's correlation was run to determine their association and a negligible, positive correlation between them (ρ= 0.04, p=0.59) was observed.

5.5.4. Relationship between tick bites and population

The calculated number of tick bite per municipality divided per 1000 inhabitants was used to identify the number of people that could be at risk of getting a tick bite. The results are given under.

Table 17. Tick bite to person ratio for highest 10 in 2012 and 2013

Municipality	Tick bite to person 2012	Tick bite to person 2013
Rozendaal	20:1000	39:1000
Schiermonnikoog	17:1000	14:1000
Terschelling	19:1000	12:1000
Vlieland	3:1000	11:1000
Ameland	10:1000	7:1000
Alphen-Chaam	5:1000	7:1000
Bloemendaal	5:1000	6:1000
Westvoorne	5:1000	5:1000
Westerveld	7:1000	4:1000
Haren	6:1000	4:1000

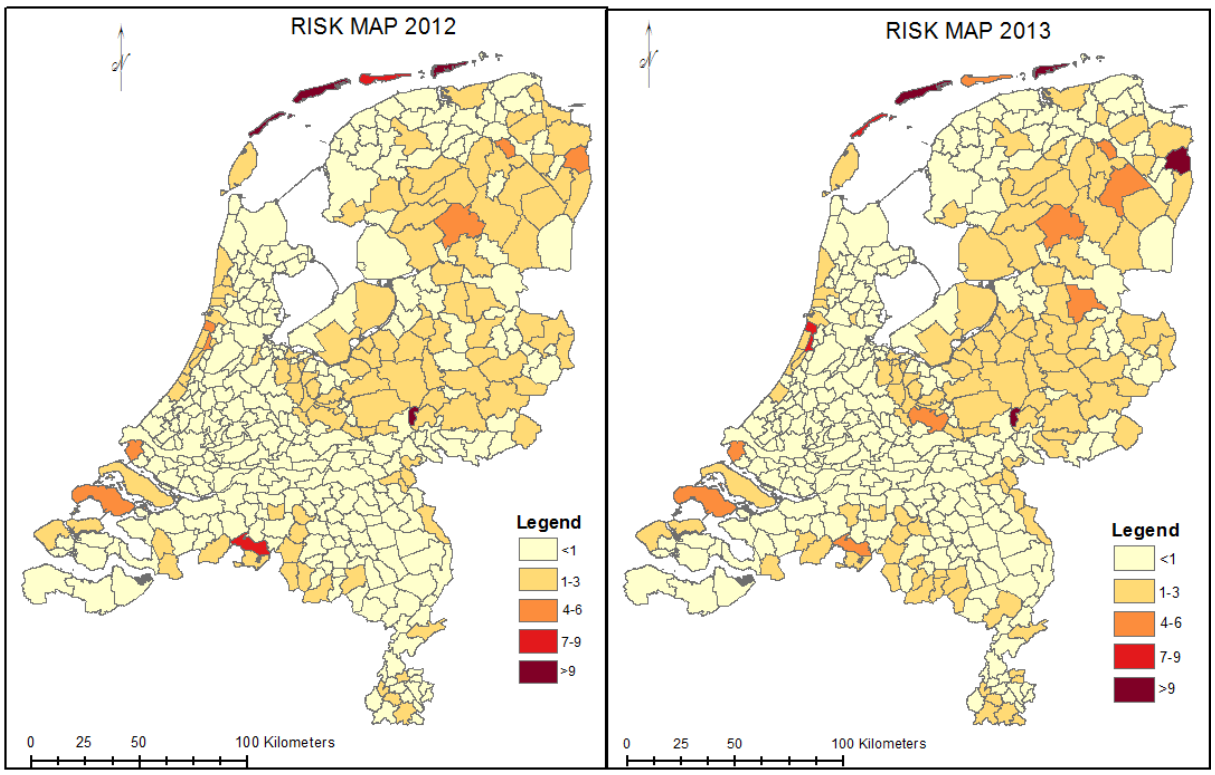


Figure 34: Tick bite risk maps calculated per 1000 persons per municipality 2012 &2013

As can be observed from Table 17, the maximum number of tick bites per municipality per 1000 inhabitants in 2012 and 2013 are 20 and 39 respectively. The ratio for 274 municipalities, which accounts for roughly 67% was found to be less than 1:1000.
The map in Figure 34 shows the risk of tick bites per municipality per 1000 residents. In both years the **four** of municipalities with high risk of tick bite ratio are located in the northern islands of the country.

5.6. Evaluation of cloud computing platform

Out of the whole package of CartoDB, *CartDB Editor*[40] was used to execute some of the analysis tasks performed in previous sections mainly the geovisual analysis and implement the multi-layer, dynamic and interactive geovisualization prototype that shows the spatio-temporal distribution of tick bites as part of this project.

CartDB was evaluated for its *CartDB Editor* for the security, software quality and quality of service based on the **Customer** Role. The results of implemented prototype that is publicly available in the platform and the evaluation result for the qualities of the platform from the customer perspective are presented in this section.

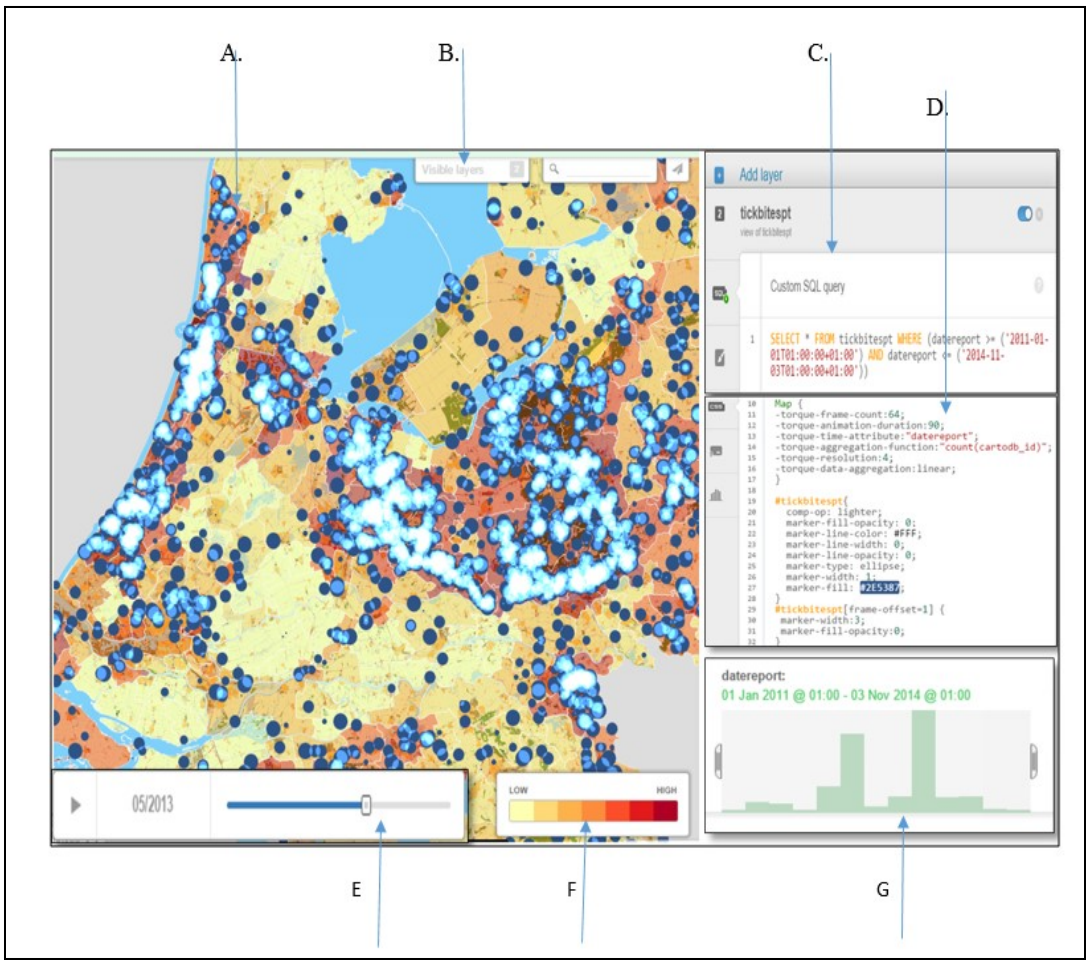


Figure 35: Screenshot of the multi-layer Geovisualization prototype in CartoDB

[40] *CartoDB Editor is the online data management and geovisual analytics component of the CartoDB platform. Interested readers are referred to its official documentation on* http://docs.cartodb.com/cartodb-editor.html

Key Figure 35:

A. Map area showing multi-layer map created from land cover (WMS), the tick bite density and the animated map representing spatio-temporal distribution of tick bites
B. Layer selector to switch between layers.
C. Postgre+postGIS styled SQL query builder.
D. CSS environment for managing the display style of the map
E. Animated map controller
F. Map legend
G. Graphical representation of the underlying data over time

The evaluation result for the security, software quality, and quality of service considering the Customer role is evaluated using the evaluating model for SaaS developed by Wen & Dong (2013) are summarized as follows.

Table 18. Quality measure for security of level 2 (standard SaaS)

Quality Metrics (customer security)	**Comply**
Secure data transfer	✓
Service level agreement	✓
Risk management (for enterprise license only)	✓
End point security	✓

Table 19. Quality measure for Usage quality of level 2 (standard SaaS)

Quality Metrics (Usage quality)	**Comply**
Multi-tenant	✓
Data isolation	✓
Multi-user (for enterprise license only)	✓
Interoperability	✓
Fault tolerance	✓
Configuration (for enterprise license only)	✓

Table 20. Quality measure for Quality of Experience of level 2 (standard SaaS)

Quality Metrics (Quality of Experience)	**Comply**
Service availability	✓
Response time*	✓
Usability*	✓
User documentation	✓
User support	✓

Note:

(*) for the response time and usability testing, no quantitative measures were applied. It was only evaluated by reviewing success stories, testimonials, and by using the geovisualization on computers, tablets and smartphones.

The software is deployed in amazon web services (AWS) cloud infrastructure which is ISO 27001[41] certified infrastructure for its security. Hence, the evaluation results for customer security can be taken at that level.

[41] *http://www.iso.org/iso/home/standards/management-standards/iso27001.htm*

6. DISCUSSIONS

6.1. Overview

In this chapter the analysis and SaaS evaluation results are discussed. The sections and sub sections are structured in such a way that they can address the research questions. The discussion on the analysis of the individual datasets (Sections 6.2 and 6.3) answers the research questions related to the distribution of the tick bite observations and the distribution of contextual social media. Research questions 1 and 3 are answered in sections 6.2 and 6.3 respectively. The section that follows (Section 6.4) is related to the results of the analysis of the relationships among the datasets. Research questions 2, 4 and 5 are answered in sections 6.4. The last section, Section 6.5, is related to the results on the cloud platform selection and evaluation performed in this research and answers research questions 6 and 7.

6.2. Spatio-temporal analysis of tick bite observations

As can be observed form the density map created from the aggregated data with the municipalities as aggregation units (Section 5.3.1, Figure 9), there are contiguous municipalities with high density of tick bites along the west-coastal, the central, the southern, and the north-eastern regions of the country. The identified actual location of hotspots presented in the KDE map and heat map of Section 5.3.2, Figure 10 are also in the regions of the country where there are municipalities with high density of tick bites. This suggests that there is a spatial relationship among the municipalities which are affected by tick bites. The result at this point cannot be taken certainly to conclude that the pattern in the municipalities is not a random with certainty. Hence, the clusters of high and low density were evaluated for their spatial relationships to further investigate the environmental and social conditions that lead to such a clustering.

The complete spatial randomness hypothesis testing done in this regard showed that there are indeed statistically significant hot spots and cold spots of municipalities with high and low density of tick bites respectively. The statistical evidence was collected from the two levels of analysis, one performed for all the years and the other performed for each year. On the first level analysis, the result of Getis-Ord GI* statistics for all the tick bites (2011-2014, Section 5.3.3, Figure 11) showed the areas with high density of tick bites are not based on complete randomness. Therefore, the complete spatial randomness (CRS) null hypothesis at this level was rejected in favour of the alternative hypothesis. That is the municipalities with high and low density of tick bites are spatially auto-correlated. However, the tick bites can also be random when investigated at a finer temporal resolution (on a yearly basis in this case). So, a second level local spatial statistical testing for the CSR null hypothesis was applied to the observations in each year. The results of the analysis (Section 5.3.3, Figure 12) showed that there are statistically significant hotspots for each year in similar locations like those observed for all the data. As a result, the CSR null hypothesis at this level was also rejected in favour of the alternative hypothesis again. So, we can certainly say that the hot spots observed both in the density maps and hot spot maps are the result of spatial associations of the tick bite observations.

The regions observed to have high incidents of tick bites, the west-coast, the country are areas frequently used by people for recreation. According to information from *Holland.cm*[42], the west-coast of the country from northern tip of North-Holland to the most southern stretch of beach in South-Holland there are excellent walking paths and are chosen by tourists and locals for recreation.

The areas of tick bite hot spots around the central part of the country are area with high vegetation cover which are also used by people for recreation. These areas are also home to the national parks like *Utrechtse Heuvelrug, De Hoge Veluwe* and *Veluwezoom.*

[42] *http://www.holland.com/uk/tourism/article/dutch-coast-6.htm*

The hot spots in the northern region are located around the parks such as *Drentsche Aa, Drents-Friese Wold* and *Dwingelderveld*.

The temporal distribution of the tick bites (Section 5.3.4, Figure 13 and Table 12) showed that the incidents are seasonal. Although there are differences in the number of tick bite reports among the years, high incidents are associated with the summer seasons for each year. The region of absolute maximum of the number of tick bite for the years 2011-2013 is in June and July. For all the years, the number of incidents starts to significantly increase in May and decrease in August. Winter months in general, showed very low number of incidents of tick bites. The pattern was also found to be strongly related (ρ =0.77, p<0.0001) as can be observed from the correlation results. We can say from the results that the incidents follow strong similarity throughout the year with a 99% confidence.

The observed seasonality of tick bite incidents was confirmed by geovisual analysis performed in the in CartoDB. The animated "*Torque heat"* map showed that the significantly large clusters of tick bite incidents are pronounced in June and July of each year.

The year 2011 has by far small number of tick bite observations when compared with the other years. This could be related to the development of the *tekenradar* application in 2012 which enabled volunteers to contribute in collecting the data. If that is the case, it strongly shows the VGI project has contributed a lot towards the understanding of the increasing social problem linked to tick bites.

6.3. Spatio-temporal analysis of related *Flickr* photos

The aim of the analysis of the spatio-temporal distribution of the *Flickr* photos in this research is to find out if the data can be used to identify hidden patterns, if any, in the tick bite observations and improve the our understanding of the distribution and risk. That is why the geolocated photos data that is used in the project is collected within the context of tick bites. Any analysis that is done on the data is therefore, in the context of tick bite incidents. So, the procedure, methods, analysis parameters such as spatial and temporal resolution as well as the depth of the analysis is the same with that of the tick bite observations discussed in the previous section.

As can be observed form the density maps of Section 5.4.2, Figure 14 geolocated photos per municipalities, there are contiguous municipalities with relatively high density of photos in middle west-coast, central and north eastern part of the country. The hot spot maps in Section 5.4.3, Figure 15 which were created to identify the location of clusters of photos, again confirmed that the actual hotspots are located in the areas where there are municipalities with high density of photos. This gives the indication that that there is a spatial relationship among the photos and indeed among municipalities with high concentration of contextual geolocated photos. To ascertain the availability of spatial relationship among the observed clusters and make sure that is not a result of random process, the clusters are evaluated for statistical significance.

The statistical evidence collected at two levels to test the complete spatial randomness hypothesis showed that there are indeed statistically significant hotspots and cold spots of municipalities with high and low concentration of geolocated photos. On the first level analysis, the result of Getis-Ord GI* statistics for all the photos (2011-2014, Section 5.4.4, Figure 16) representing environment and outdoor activities showed the areas with high density of photos are not based random process. Therefore, the CSR null hypothesis at this level was rejected in favour of the alternative hypothesis. However, the spatio-temporal process represented by these photos can also be random when investigated on a yearly basis. So, a second level local spatial statistical analyses for each year was run. The results of the analyses also showed that there are statistically significant hot spots for each year as well. However, the location of the statistically significant hot spots for every year are located in different locations of the country as shown in section 5.3.4, Figure 17 indicating spatial randomness over the years. Indeed, the location of the hot spots for the whole dataset

and for the yearly datasets showed obvious differences. As a result, we failed to reject the CSR null hypothesis. Therefore, the observed hot spots could be the result of spatially random processes. That is to say that over the years, the location of hot spots for distribution of the photos are not consistently located. They rather tend to be random.

The general temporal distribution of the geolocated photos Section 5.4 showed that there is a low seasonality in the distribution of the photos. Even though the context for which the photos are collected is related to the tick bites, their distribution does not show pronounced seasonality as such like that of the tick bites. The absolute maxima for some of the time slots (weeks in this case) seem to be associated with mass events (like the flower parade) that happens regularly each year.

The geovisual analysis performed in the in CartoDB also confirmed that the availability of significantly large clusters of photo is less dependent on time.

6.4. Analysis of relationships among datasets

An intersection between point processes representing tick bite incidents and land cover data enabled to reduce the bias in the tick bite observations in multiple ways. Firstly, the missing environmental information in the dataset is obtained from the land cover data. This gives the 8.6% observations which were reported with not known, other, or empty value in their environment attribute were given the proper land cover value. Secondly, the observations with multiple environmental information were assigned with an official land cover type which gives a better insight in to the whole tick bite observations dataset. In total the thematic bias associated to environmental information is reduced by 28.6 % which led to knowing the proper land cover. It was finally understood that the four types of land cover classes account for 77.66% of tick bites and only two types, forest and build-up areas, account for 57% of the tick bites.

The intersection of the geolocated photos and the land cover data similarly improves the information in the photos in two ways. Firstly, the photos collected under the context of the environment were assigned the actual land cover on which they were taken. This gives the opportunity to make further analysis on the relationships between the tick bite incidents and the photos for reducing the bias in the information given the free text comment of the tick bites by either confirming or replacing it all. Secondly, the photos obtained for the outdoor activities were given the land cover information as a result of the intersection. In this case again, the resulting data was suitable to be used in trying to find out the relationship between the tick bites and outdoor activities associated with the incidents by comparing the two datasets that share the same land cover.

To have a better understanding of the tick bites, finding out a way and getting the missing information as a result of the unknown outdoor activity was one of the main tasks of the project. To do so, both VGI datasets (tick bite observations and photo) were used as inputs in the process of finding the relationships. To get the missing outdoor activity from the contextual photos so that they can be used to improve our understanding of the social activities linked to tick bite incidents, both datasets had to be investigated for their relationships in space and time.

The application of Kernel density estimates and local spatial statistics methods were found to show no significant relationship between the two datasets. That is the tick bite hot spots and activity photo hot spots were found to be significantly different for the whole datasets (Section 5.5.3, Figure 28). The datasets were analyzed further at a land cover level such as tick bites located in the forest and photos representing activities in the forest. Although tick bite incidents located in a land cover and photo representing activities in the same type of land cover showed hotspots in the land cover types, the location of those hot spots were found to be in different locations (see Section 5.5.3, Figure 30 for the built-up areas).

The visual analysis was support by the spearman's correlation method applied for each. The spearman's correlation analysis for both datasets aggregated on a week temporal resolution was found to show no significant relationship between the two datasets. The spearman's correlation coefficient for almost all the analyses was in the range of insignificant to weak according to the "rule of thumb" defined in this thesis. For example, for the results that included in section 5.5.3 the correlation results for corresponding VGI datasets in relation to forest and built-up areas were found to be (ρ= 0.12, p=0.14) and (ρ= 0.04, p=0.59) respectively. These statistical evidences are not strong enough to reject the stated null hypothesis which essentially says the relationship is weak.

Not all of the spatial relationship were not without similarities. Those that showed similarities of any kind were further taken in to consideration for temporal analysis as only the spatial relationship could not be an evidence for a strong association between the two data sets. Even the few that were found to be moderate (result not included here) undeniably showed that the probability that the similarity happens by chance for the relationships was not significantly small to say that they are related. The value was found to be p> 0.56 for all the other.

The results of the thorough analyses performed to find a strong relationship between the two datasets so that the geolocated photos can be used to improve our understanding of hidden patterns in the tick bite distribution and risks, if any, concluded otherwise. That is the geolocated social media (*Flickr* photos in this project) could not be used to improve our understanding of tick bite distribution and risks.

The pattern observed in the risk map of section 5.5, Figure 34 for the years 2012 and 2013 are similar to the tick bite density maps and hot spot maps of section 5.3, Figures 9 and 10 in sense that they are clustered throughout the west-coast, in the central and north eastern part of the country. The picks however are located in the northern islands of the country.

As can be observed from (Section 5.5.4, Table 17and Figure 34), the maximum number of tick bites per municipality per 1000 inhabitants in 2012 and 2013 are 20 and 39 respectively. In both years the **four** of municipalities with highest risk of tick bite ratio are located in the northern islands of the country. The ratio for 274 municipalities, which accounts for roughly 67% was found to be less than 1:1000. This in general could also be associated with the areas being used for recreational purposes. But, we do not really know why the pattern appears the way it is.

6.5. Evaluation of cloud computing platforms

The application of AHP based SaaS selection process (Section 5.2) has made it easy to find the appropriate SaaS to perform the geovisual analysis to understand the tick bite distributions and implement geovisualization prototype. The selection of suitable SaaS platform was approached as a multi-level decision making problem. Hence, two level selection process was applied. The first level was done by down selecting the platforms using the functionalities that are required by this project and the type of application they are that stems from the secondary objective of this project. The second level was done using the basic principles of AHP framework and using binary values (1 if comply and 0 otherwise) for the two final candidates in the category

For the first stage down-selection, the AHP method was helpful in formulating a structured thinking to approach the process. At this stage the usability, architecture, pricing and other factors were ignored. Taking only the functionality was sufficient to do the task.

Geospatial SaaS products that passed the first stage of evaluation were further evaluated using the AHP for SaaS selection method as a guiding principle. Although all the parameters were used in this stage, the method was not applied as is since there were no experts to rank the products. Taking the factors and the associated attributes related the requirements of this project, a binary value was assigned to each attribute (1 if it

complies and 0 if not) to evaluate whether the candidate product satisfies the requirement. Based on the selection process CartoDB was found to have the required functionalities for this project as presented in section 5.2, Table 11.

The SaaS select for the project was used to implement some of the tasks including creating the density maps and performing the hot spot analysis using the point processes VGI datasets. All the online maps both the static as well as the dynamic and interactive ones created in the platform are available for access online.

Finally, the SaaS platform was formally evaluated for its qualities from customer perspective using SaaS quality model developed by Wen & Dong (2013).The evaluation performed in this project must not be mistaken for a complete evaluation of the platform. It was applied only for the customer perspective of the evaluation model. This is because, since only one part of the solution (CartDB Explorer) was used in this project, the platform could not be evaluated for the application development and the platform as a whole. For the evaluation performed here the solution satisfies the requirements that one "standard level" (Wen & Dong, 2013) SaaS should comply. Therefore the CartoDB is at least a standard level geospatial SaaS according to the model used.

A summary of the license plans and the near-real time geovisualization functionality is given below to show how effectively the solution can be used in projects with real time geospatial data requirements.

CartoDB is provided as a service on demand with five license plans. These license plans are called "Free", "Magellan", "John Snow", "Mercator" and "Enterprise". The price ranges from 0 - $299 per month for the first four plans. The enterprise license plan which can also be deployed on customers' premises is only available at an annual price of $7999. With the enterprise license, customers can grow their database as much as they need and pay per extra GB of data. They also get a customized Service Level Agreement and are able to concurrently access their data.

Starting from the "John snow" plan, which is priced at $49 US dollars per month, customers can sync their data from Goggle drive and Dropbox which are another SaaS platforms. With this license plan and beyond, it is possible to produce near real-time geovisualization solutions. For instance, from using geosocial media data for performing a complete geo-information processing work-flow, a simple and effective solution can be developed in a short time by writing a data harvest script like the one used in this project for which the pseudo code is give in section 3.3, Figure 3. This can be simply achieved by collecting and automatically adding the data into comma separated (CSV) file on your local Dropbox folder and *synchronise your data using CartoDB Editor*[43]. The script can continue collecting and writing the data while CartoDB synchronizes the file with the displayed visualization at user defined intervals.

[43] *http://blog.cartodb.com/synced-tables-create-real-time-maps-from-data-anywhere/*

7. CONCLUSIONS AND RECOMMENDATIONS

7.1. Conclusions

The spatial analysis performed at different temporal scales (2011- June 2014, and yearly basis) on the tick bite observations in this research showed the entire west-coast, the central and north-eastern regions of the country to be the hot spots for tick bite incidents. The west-coastal and central regions are chosen by people for recreation and tourism. The hot spots are found to be located in regions with large vegetation cover. It was, therefore, clear from this that tick bite incidents are highly dependent on the vegetation cover.

The analysis of temporal distribution of the tick bites showed that the incidents are seasonal. Irrespective of the differences in the number of tick bite reports among years, high incidents are associated with the summer season. The temporal region for absolute maximum of the number of tick bite incidents for the years 2011-2013 is in the months of June and July. The pattern of increase and decrease was indeed found to be strongly related.

The year 2011 has by far smaller number of tick bite observations when compared with the other years. There is a high possibility that this is associated with the development of the *tekenradar* application in 2012, which enabled volunteers to contribute in collecting the data. Even though additional evidence is required to confirm, it strongly shows that the VGI project has contributed a lot towards the understanding of the increasing social problem linked to tick bites.

Using authoritative data to reduce the bias in the VGI information contributed in obtaining official value to 28.6% of the ***unknown*** and **mixed** land cover values for the "*environment"* attribute in the tick bite VGI dataset used in this project. As a result, it was found out that 42. 59% of tick bite incidents occurred in areas covered by forest and 14.48% in built-up areas both of which account for 57% percent of the total. Therefore, it is clear from this that using authoritative data to minimize the bias that could be introduced in volunteered geo-information will benefit VGI users in getting more complete data for their further analysis.

In this research, it was found out that tick bites are not well represented in social media. Even though there is a rich amount of information related to the contexts represented in tick bite observations, the spatio-temporal distribution of the contextual geolocated social media data is far from similar with that of the tick bite observations. This data was also found to be randomly distributed when compared on a year by year basis whereas the tick bite observations showed similar spatio-temporal distribution. Therefore, the contextual geolocated social media data that was used in this research (*Flickr* photos) could not be used to improve our understanding of tick bite distributions.

An investigation of the tick bite observation with respect to the population for the years 2012 and 2013 in each municipality in the Netherlands showed that **four** of the **ten** with higher ratio of tick bite incidents per 1000 inhabitants are located in the northern islands of the country. It was also found out that the municipalities with a ratio of more than 1:1000 were found to be 43%.

CartoDB which is an open source SaaS for geospatial data storage and visualization was selected and used for performing the analysis, especially geovisual analysis, in this project and was found to be mature enough to implement similar projects. It was found to comply at least the "standard level" as defined by the quality model for SaaS (Wen & Dong, 2013) from the "Customer Role" perspective defined in the same model. This SaaS helped in building intuitively understandable, easily sharable, dynamic, and interactive geovisualization prototypes that were created as part of this project.

7.2. Recommendations

The following recommendations are made based on the research outputs and findings in this thesis:

1. In this research we have only tried to understand the spatial and temporal distribution of tick bite incidents using VGI data and land cover data. The population data used to evaluate the risk is also the official population data of each municipality. Therefore, other data sets such as temperature, surface humidity as well as the daily population of the municipalities as a result of people's movement should be included to explain the distribution of the tick bites.
2. The contextual geolocated photos were only cleaned for obvious noise that is believed to be traduced as a result of multiple photos taken by the same person from the same location in the same day. So, it is better to develop a robust noise cleaning algorithm and clean such geolocated social media VGI datasets to use them in identifying hidden pattern in VGI dataset collected through public endeavor.
3. The geospatial SaaS evaluated in this project was only evaluated based on the requirements of this thesis. It was indeed evaluated from the customer perspective. It is then recommended to separately study the maturity of geospatial cloud computing for other geospatial workflows from all framework, application and customer perspectives.

LIST OF REFERENCES

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python. Text* (First Edit., Vol. 43, p. 479). O'Reilly Media. doi:10.1097/00004770-200204000-00018

Boampong, P. A., & Wahsheh, L. A. (2012). Different facets of security in the cloud, 5. Retrieved from http://dl.acm.org/citation.cfm?id=2331762.2331767

Bontemps, S., Defourny, P., Bogaert, E., Arino, O., Kalogirou, V., & Perez, J. (2011). GLOBCOVER 2009 - Products Description and Validation Report.

Brabham, D. C. (2008). Crowdsourcing as a Model for Problem Solving: An Introduction and Cases. *Convergence: The International Journal of Research into New Media Technologies*, *14*(1), 75–90. doi:10.1177/1354856507084420

Brown, M., Sharples, S., Harding, J., Parker, C. J., Bearman, N., Maguire, M., … Jackson, M. (2013). Usability of Geographic Information: Current challenges and future directions. *Applied Ergonomics*, *44*(6), 855–865. doi:10.1016/j.apergo.2012.10.013

Canavosio-Zuzelski, R., Agouris, P., & Doucette, P. (2013). A Photogrammetric Approach for Assessing Positional Accuracy of OpenStreetMap© Roads. *ISPRS International Journal of Geo-Information*, *2*(2), 276–301. doi:10.3390/ijgi2020276

Caverlee, J., Cheng, Z., Sui, D. Z., & Kamath, K. Y. (2013). Towards Geo-Social Intelligence : Mining , Analyzing , and Leveraging Geospatial Footprints in Social Media. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 33–41.

Coleman, D. J., Georgiadou, Y., Labonte, J., Observation, E., & Canada, N. R. (2009). Volunteered Geographic Information : The Nature and Motivation of Produsers ∗, *4*, 332–358. doi:10.2902/1725-0463.2009.04.art16

De Longueville, B. (2010). Community-based geoportals: The next generation? Concepts and methods for the geospatial Web 2.0. *Computers, Environment and Urban Systems*, *34*(4), 299–308. doi:10.1016/j.compenvurbsys.2010.04.004

Dillon, T., Wu, C., & Chang, E. (2010). Cloud Computing: Issues and Challenges. In *2010 24th IEEE International Conference on Advanced Information Networking and Applications* (pp. 27–33). IEEE. doi:10.1109/AINA.2010.187

Frias-Martinez, V., Soto, V., Hohwald, H., & Frias-Martinez, E. (2012). Characterizing Urban Landscapes Using Geolocated Tweets. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing* (pp. 239–248). IEEE. doi:10.1109/SocialCom-PASSAT.2012.19

Gatrell, A. C., Bailey, T. C., Diggle, P. J., Rowlingson, B. S., & Rowlingsont, B. S. (1996). Point Spatial application pattern analysis geographical epidemiology. *Transactions of the Institute of British Geographers*, *21*(1), 256–274.

Godse, M., & Mulik, S. (2009). An Approach for Selecting Software-as-a-Service (SaaS) Product. In *2009 IEEE International Conference on Cloud Computing* (pp. 155–158). IEEE. doi:10.1109/CLOUD.2009.74

Goetz, M., & Zipf, A. (2012). Using Crowdsourced Geodata for Agent-Based Indoor Evacuation Simulations. *ISPRS International Journal of Geo-Information*, *1*(3), 186–208. doi:10.3390/ijgi1020186

Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal*, *69*, 211–221. doi:10.1007/s10708-007-9111-y

Goodchild, M. F., & Li, L. (2012). Assuring the quality of volunteered geographic information. *Spatial Statistics*, *1*, 110–120. doi:10.1016/j.spasta.2012.03.002

Google developers academy. (2012). What is Cloud Computing? Retrieved June 02, 2014, from https://developers.google.com/appengine/training/intro/whatiscc

Haklay, M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design*, *37*(4), 682–703. doi:10.1068/b35097

Hauff, C. (2013). A study on the accuracy of Flickr's geotag data. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '13* (p. 1037). New York, New York, USA: ACM Press. doi:10.1145/2484028.2484154

Heipke, C. (2010). Crowdsourcing geospatial data. *ISPRS Journal of Photogrammetry and Remote Sensing*, *65*(6), 550–557. doi:10.1016/j.isprsjprs.2010.06.005

Huang, Q., Yang, C., Nebert, D., Liu, K., & Wu, H. (2010). Cloud computing for geosciences. In *Proceedings of the ACM SIGSPATIAL International Workshop on High Performance and Distributed Geographic Information Systems - HPDGIS '10* (pp. 35–38). New York, New York, USA: ACM Press. doi:10.1145/1869692.1869699

Ji, X., Chen, B., Huang, Z., Sui, Z., & Fang, Y. (2012). On the use of cloud computing for geospatial workflow applications. In *2012 20th International Conference on Geoinformatics* (pp. 1–6). IEEE. doi:10.1109/Geoinformatics.2012.6270263

Koukoletsos, T., Haklay, M., & Ellul, C. (2012). Assessing Data Completeness of VGI through an Automated Matching Procedure for Linear Data. *Transactions in GIS*, *16*(4), 477–498. doi:10.1111/j.1467-9671.2012.01304.x

Kurashima, T., Iwata, T., Irie, G., & Fujimura, K. (2010). Travel route recommendation using geotags in photo sharing sites. In *Proceedings of the 19th ACM international conference on Information and knowledge management - CIKM '10* (p. 579). New York, New York, USA: ACM Press. doi:10.1145/1871437.1871513

Mooney, P., Corcoran, P., & Ciepluch, B. (2012). The potential for using volunteered geographic information in pervasive health computing applications. *Journal of Ambient Intelligence and Humanized Computing*, *4*(6), 731–745. doi:10.1007/s12652-012-0149-4

Mukaka, M. M. (2012). Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal : The Journal of Medical Association of Malawi*, *24*(3), 69–71. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3576830&tool=pmcentrez&rendertype=abstract

NIST. (2011). The NIST Definition of Cloud Computing. Retrieved November 11, 2014, from http://faculty.winthrop.edu/domanm/csci411/Handouts/NIST.pdf

O'Reilly, T. (2007). What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. Retrieved June 03, 2014, from http://mpra.ub.uni-muenchen.de/4578/1/MPRA_paper_4578.pdf

O'Reilly, T. (2009). *What is Web 2.0 (Google eBook)* (p. 12). "O'Reilly Media, Inc." Retrieved from http://books.google.com/books?hl=en&lr=&id=NpEk_WFCMdIC&pgis=1

Ord, J. K., & Getis, A. (2010). Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. *Geographical Analysis*, *27*(4), 286–306. doi:10.1111/j.1538-4632.1995.tb00912.x

Ostermann, F. (2011). A Conceptual Workflow For Automatically Assessing The Quality Of Volunteered Geographic Information For Crisis Management. Retrieved August 24, 2014, from http://itcnt05.itc.nl/agile_old/Conference/2011-utrecht/contents/pdf/shortpapers/sp_122.pdf

Oxendine, C. E., & Waters, N. (2014). No-Notice Urban Evacuations: Using Crowdsourced Mobile Data to Minimize Risk. *Geography Compass*, *8*(1), 49–62. doi:10.1111/gec3.12104

Prion, S., & Haerling, K. A. (2014). Making Sense of Methods and Measurement: Spearman-Rho Ranked-Order Correlation Coefficient. *Clinical Simulation in Nursing*, *10*(10), 535–536. doi:10.1016/j.ecns.2014.07.005

Saaty, T. L. (1990). How to make a decision: The analytic hierarchy process. *European Journal of Operational Research*, *48*(1), 9–26. doi:10.1016/0377-2217(90)90057-I

Shelton, T., Poorthuis, A., Graham, M., & Zook, M. (2014). Mapping the data shadows of Hurricane Sandy: Uncovering the sociospatial dimensions of "big data." *Geoforum*, *52*, 167–179. doi:10.1016/j.geoforum.2014.01.006

Sinanc, D., & Sagiroglu, S. (2013). A review on cloud security. In *Proceedings of the 6th International Conference on Security of Information and Networks - SIN '13* (pp. 321–325). New York, New York, USA: ACM Press. doi:10.1145/2523514.2527013

Sprong, H., Hofhuis, A., Gassner, F., Takken, W., Jacobs, F., van Vliet, A. J. H., … Takumi, K. (2012). Circumstantial evidence for an increase in the total number and activity of Borrelia-infected Ixodes ricinus in the Netherlands. *Parasites & Vectors*, *5*(1), 294. doi:10.1186/1756-3305-5-294

Sui, D., Elwood, S., & Goodchild, M. (2012). *Crowdsourcing Geographic Knowledge : Volunteered Geographic Information (VGI) in Theory and Practice*. Dordrecht: Springer. doi:10.1007/978-94-700-4587-2_1

Sweta, L. O. (2014). Early Warning Systems and Disaster Management using Mobile Crowdsourcing - MDIwMTMxNDI5.pdf. *International Journal of Science and Research*. Retrieved August 13, 2014, from http://www.ijsr.net/archive/v3i4/MDIwMTMxNDI5.pdf

Takabi, H., Joshi, J. B. D., & Ahn, G. J. (2010). Security and privacy challenges in cloud computing environments. *IEEE Security and Privacy*, *8*(6), 24–31. doi:10.1109/MSP.2010.186

Vandenbroucke, D., Bucher, B., & Crompvoets, J. (Eds.). (2013). *Geographic Information Science at the Heart of Europe*. Cham: Springer International Publishing. doi:10.1007/978-3-319-00615-4

Vinh, V. H., Lionel, A., Cristina, S., Didier, R., Philippe, P., & Frédéric, P. (2014). Monitoring human tick-borne disease risk and tick bite exposure in Europe: Available tools and promising future methods. *Ticks and Tick-Borne Diseases*, *in press*(6), 47p. doi:10.1016/j.ttbdis.2014.07.022

Wageningen University, de Natuurkalender, & RIVM. (2012). Lyme disease. Retrieved June 02, 2014, from http://www.tekenradar.nl/

Wen, P. X., & Dong, L. (2013). Quality model for evaluating SaaS service. *Proceedings - 4th International Conference on Emerging Intelligent Data and Web Technologies, EIDWT 2013*, 83–87. doi:10.1109/EIDWT.2013.19

Xie, Z., & Yan, J. (2008). Kernel Density Estimation of traffic accidents in a network space. *Computers, Environment and Urban Systems*, *32*(5), 396–406. doi:10.1016/j.compenvurbsys.2008.05.001

Xinlin Qian, Liping Di, Deren Li, Pingxiang Li, Lite Shi, & Liefei Cai. (2009). Data cleaning approaches in Web2.0 VGI application. In *2009 17th International Conference on Geoinformatics* (pp. 1–4). IEEE. doi:10.1109/GEOINFORMATICS.2009.5293442

Yang, C., Goodchild, M., Huang, Q., Nebert, D., Raskin, R., Xu, Y., … Fay, D. (2011). Spatial cloud computing: how can the geospatial sciences use and help shape cloud computing? *International Journal of Digital Earth*, *4*(4), 305–329. doi:10.1080/17538947.2011.587547

Yang, C., Raskin, R., Goodchild, M., & Gahegan, M. (2010). Geospatial Cyberinfrastructure: Past, present and future. *Computers, Environment and Urban Systems*, *34*(4), 264–277. doi:10.1016/j.compenvurbsys.2010.04.001

Yue, P., Zhou, H., Gong, J., & Hu, L. (2013). Geoprocessing in Cloud Computing platforms – a comparative analysis. *International Journal of Digital Earth*, *6*(February 2015), 404–425. doi:10.1080/17538947.2012.748847

Zhang, J. (2010). Towards personal high-performance geospatial computing (HPC-G). In *Proceedings of the ACM SIGSPATIAL International Workshop on High Performance and Distributed Geographic Information Systems - HPDGIS '10* (pp. 3–10). New York, New York, USA: ACM Press. doi:10.1145/1869692.1869694

Zhang, Q., Cheng, L., & Boutaba, R. (2010). Cloud computing: State-of-the-art and research challenges. *Journal of Internet Services and Applications*, *1*(1), 7–18. doi:10.1007/s13174-010-0007-6

Zielstra, D., & Hochmair, H. H. (2013). Positional accuracy analysis of Flickr and Panoramio images for selected world regions. *Journal of Spatial Science*, *58*(2), 251–273. doi:10.1080/14498596.2013.801331

APPENDIX A. PYTHON SCRIPTS USED IN THE PROJECT

A.1. Geolocated *Flickr* photo harvesting script

```
Created on Sun Nov 16 05:26:00 2014
@author: B.A.Gidey
"""
import requests
import json
import os
import unicodedata
os.chdir('D:\MSc\THESIS\DATA\RAW_PHOTOS')
uriQuery='https://api.flickr.com/services/rest/?method=flickr.photos.search&...'
'''
 The Uri query should contain all parameters
 Obtain the number of pages of the data returned by the http request for the search term
'''
def getNumberOfPages():
    r = requests.get(uriQuery)
    commit_data=r.text
    data=json.loads(commit_data)
    return data['photos']['pages']
'''
  Extract photo information from all available pages for the search term
  and write the resulting data in to a csv file on the local hard disk
'''
def getPhotoExtract(numberOfPage):
    csvFile=open('eggs.csv','a')
    for page in range( numberOfPage) :
        searcurl=uriQuery+'&page='+str(page+1)
        r = requests.get(searcurl)
        commit_data=r.text
        data=json.loads(commit_data)
        for i in range(len(data['photos']['photo'])):
            pid=data['photos']['photo'][i]['id']
            powner=data['photos']['photo'][i]['owner']
            ptitle=data['photos']['photo'][i]['title']
            ptitle=unicodedata.normalize('NFKD', ptitle).encode('ascii','ignore')
            dateTaken=data['photos']['photo'][i]['datetaken']
            tags=data['photos']['photo'][i]['tags']
            tags=unicodedata.normalize('NFKD', tags).encode('ascii','ignore')
            lat=data['photos']['photo'][i]['latitude']
            lon=data['photos']['photo'][i]['longitude']
            accuracy=data['photos']['photo'][i]['accuracy']
            str1=str(pid)+','+str(powner)+','+str(ptitle)+','+str(dateTaken)
            str2=str(tags)+','+str(lon)+','+str(lat)+','+str(accuracy)
            strinput=str1+str2
            csvFile.write('\n'+ strinput)
    csvFile.close()
getPhotoExtract(getNumberOfPages())
```

A.2. *Flickr* photo partial cleaning script

```
"""
Created on Thu Nov 27 16:06:23 2014
@author: BAGS
"""
import csv
import os

rawDataPath=r'D:\MScTHESIS\DATA\ANALYSIS_DATA\PhotoExtractNL_RAW'
cleanedDataPath=r'D:\MScTHESIS\DATA\ANALYSIS_DATA\PhotoExtractNL_CLEANED'


def ReadRawCSV(path):
    os.chdir(path)
    output = []
    dataFile = open( 'NLMeadow.csv','r') #open the file in read universal mod
    csvreader=csv.reader(dataFile,delimiter=',',quotechar='|')
    for row in csvreader:
        output.append(row)
    dataFile.close()
    return output
def CleanPhotoData(lst,path):
    os.chdir(path)
    with open('NLMeadow.csv', 'wb') as envFile:
        envWriter = csv.writer(envFile, delimiter=',',quotechar='|', quoting=csv.QUOTE_MINIMAL)
        envWriter.writerow(lst[0])
        l1=lst[1:]
        l2=lst[2:]
        for i in range(len(l2)):
            for j in range(len(l2)):
             if ((l1[i][1]==l2[j][1] and l1[i][2]==l2[j][2] and l1[i][3]==l2[j][3] and l1[i][4]==l2[j][4] )== False):
              envWriter.writerow(l1[i])
        envFile.close()

# do the data cleaning
CleanPhotoData(ReadRAWCSV(rawDataPath),cleanedDataPath)
```

A.2. Correlation *analysis* script

```
"""
Created on Mon Dec 08 15:55:00 2014

@author: BAGS
"""
import pandas as pd
import os
from scipy.stats import spearmanr
import numpy as np
import datetime
pd.options.display.mpl_style = 'default'

'''
 The function generates summary report for both photos and tick bites per land cover
'''
def GenerateSummaryReport():
    '''
    read photo data
    '''
    os.chdir('D:\MScTHESIS\DATA\ANALYSIS_RESULTS')
    tempPhoto_df=pd.read_csv('NLPELC10Interset.csv')
    photos_df= pd.DataFrame(tempPhoto_df)
    photos_data= zip( photos_df['COVETYPE'],photos_df['photoCnt'])
    newPhoto_df=pd.DataFrame(photos_data, columns=['Land_Cover','NumberOfPhotos'])
    photosDataFrame=pd.DataFrame(newPhoto_df.groupby('Land_Cover')['NumberOfPhotos'].sum())
    photos=photosDataFrame.sort('NumberOfPhotos',ascending=True)
    '''
    read tick bite data
    '''
    os.chdir('D:\MScTHESIS\DATA\ANALYSIS_RESULTS')
    tempTickBite_df=pd.read_csv('NLTOLC10Interset.csv')
    tickBites_df= pd.DataFrame(tempTickBite_df)
    tickBites_data= zip( tickBites_df['COVETYPE'],tickBites_df['tickCnt'])
    newTickBites_df=pd.DataFrame(tickBites_data, columns=['Land_Cover','NumberOfTickBites'])
    tickBitesDataFrame=pd.DataFrame(newTickBites_df.groupby('Land_Cover')['NumberOfTickBites'].sum())
    tickBites=tickBitesDataFrame.sort('NumberOfTickBites',ascending=True)
    '''
     Print data summary
    '''
    print (photos['NumberOfPhotos'],photos['NumberOfPhotos']/photos['NumberOfPhotos'].sum()  *100)
    print (tickBites['NumberOfTickBites'],tickBites['NumberOfTickBites']/tickBites['NumberOfTickBites'].sum() *100)

    '''
    Plot summary reports
    '''
    photos.plot(kind='bar',figsize =(10,5),title='Photos per land cover summary')
    tickBites.plot(kind='barh',figsize =(10,5),title='Tick bites per land cover summary')

'''

 The function calculates spearmans correlation for both datasets aggregated by land cover classes
'''
def EvaluateCorrelationByLandcover():
    os.chdir('D:\MScTHESIS\DATA\ANALYSIS_RESULTS')
    temp_df=pd.read_csv('NLTOPELANDCOVER10FINAL.csv')
    temp_dfNZ=pd.read_csv('NLTOPELANDCOVER10FINALNZ.csv')
    summ_df= pd.DataFrame(temp_df)
    corr_df=pd.DataFrame(temp_dfNZ)
    corr_data= zip( corr_df['FID'],corr_df['Sum_photoC'],corr_df['Sum_tickCn'])
    summ_data= zip( summ_df['COVETYPE'],summ_df['Sum_photoC'],summ_df['Sum_tickCn'])
    corrDataframe=pd.DataFrame(corr_data, columns=['FeatureID','NumberOfPhotos','NumberOfTickBites'])
    newDataframe=pd.DataFrame(summ_data, columns=['Land_Cover','NumberOfPhotos','NumberOfTickBites'])
    df1=pd.DataFrame(newDataframe.groupby('Land_Cover')['NumberOfPhotos','NumberOfTickBites'].sum())
    print   newDataframe.describe()
    print   df1.describe()
```

```
    print 'Spearmansr:',spearmanr(corrDataframe['NumberOfTickBites'],corrDataframe['NumberOfPhotos'])

    df1.plot(kind='barh',figsize =(10,5),title='Tick bites and photos summary')


    corrDataframe.plot(kind='Scatter',color='purple',xlim=(-10,100),ylim=(-10,300),x='NumberOfTickBites',
            y='NumberOfPhotos', figsize =(10,5),title='Tick bite VS Photos per land cover')

'''

 The function calculates spearmans correlation for both datasets aggregated by Municipality
'''
def EvaluateCorrelationByMunicipality():
    os.chdir('D:\MScTHESIS\DATA\ANALYSIS_RESULTS')
    temp_df=pd.read_csv('NLMunicipalitiesFinalAggregate.csv',index_col='FID')
    photos_df= pd.DataFrame(temp_df)
    corr_data= zip(photos_df['Sum_photoC'],photos_df['Sum_tickCn'])
    corrDataframe=pd.DataFrame(corr_data, columns=['NumberOfPhotos','NumberOfTickBites'])
    print 'Spearmansr:',spearmanr(corrDataframe['NumberOfTickBites'],corrDataframe['NumberOfPhotos'])
    corrDataframe.hist(bins=100, figsize =(15,5))

    corrDataframe.plot(kind='scatter',color='purple',xlim=(-10,200),ylim=(-10,700),x='NumberOfTickBites',
            y='NumberOfPhotos',figsize =(10,5), title='Tick bite VS Photos per Municipality')
'''
 The function calculates spearmans correlation for both datasets aggregated by Municipality
'''
def EvaluateTemporalCorrelation():
   os.chdir('D:\MScTHESIS\DATA\ANALYSIS_RESULTS')
   temp_df=pd.read_csv('NLTO_3YBuiltUP.csv',parse_dates=['datereported'],index_col='datereported')
   mydata=temp_df.resample('1W',how={'ticksCnt': np.sum})
   #os.chdir('D:\MScTHESIS\DATA\ANALYSIS_DATA\PhotoExtractNL_CLEANED')
   tempphoto_df=pd.read_csv('NLPE_3YBuiltUP.csv',parse_dates=['datetaken'],index_col='datetaken')
   myphotodata=tempphoto_df.resample('1W',how={'photosCnt': np.sum})

   smallerdata=min(len(mydata),len(myphotodata))-1

   ticksData=mydata['ticksCnt'][:]

   photodata=myphotodata['photosCnt'][:]
   time_df= pd.DataFrame(zip(pd.date_range('1/1/2011', '10/1/2014',freq='1W'),ticksData,photodata),
               columns=['Weeks','Tick_Observation','Photo_Extracts'])
   time_df.plot(kind='scatter',xlim=(-10,200),ylim=(-10,200),x='Tick_Observation',y='Photo_Extracts',figsize =(8,5),
           color='purple',title='Tick bites vs photos ')
   time_df.plot(figsize =(8,5), x=time_df['Weeks'], title='Temporal distribution of tick bites and photos')

   print 'Spearmansr:',spearmanr(mydata['ticksCnt'][:smallerdata],myphotodata['photosCnt'][:smallerdata])

'''
Call the functions
'''

GenerateSummaryReport()
EvaluateCorrelationByLandcover()
EvaluateCorrelationByMunicipality()
EvaluateTemporalCorrelation()
```

APPENDIX B. TICK BITES AND PHOTOS SUMMARY

Municipality	# of Tick bites	# of photos	TickBite Density	Photo Density
Amsterdam	193	3263	0.98	16.60
Rotterdam	107	1491	0.39	5.41
Lisse	15	1482	0.93	92.31
Utrechtse Heuvelrug	350	1434	2.61	10.69
Utrecht	83	1250	0.84	12.60
's-Gravenhage	118	1076	1.39	12.70
Apeldoorn	571	988	1.67	2.90
Baarn	111	761	3.36	23.06
Amstelveen	58	706	1.32	16.02
Midden-Drenthe	115	660	0.33	1.91
Eindhoven	47	596	0.53	6.71
Wassenaar	124	577	2.35	10.95
Groningen	111	571	1.33	6.82
Ede	403	549	1.26	1.72
Doesburg	4	480	0.31	37.04
Deventer	95	477	0.71	3.55
Westerveld	226	446	0.80	1.58
Zwolle	57	414	0.48	3.47
Venlo	66	401	0.51	3.11
Hardenberg	69	346	0.22	1.09
Soest	85	342	1.83	7.37
Arnhem	131	329	1.29	3.24
Harderwijk	80	326	2.07	8.42
Almere	47	321	0.34	2.30
Amersfoort	73	315	1.14	4.93
Coevorden	90	314	0.30	1.05
Enschede	148	292	1.04	2.05
Bergen (NH.)	100	290	1.02	2.96
De Wolden	60	279	0.27	1.23
Hilversum	124	276	2.68	5.95
Rheden	177	270	2.10	3.20
Ommen	143	270	0.79	1.48
Leiden	37	267	1.59	11.47
Emmen	114	266	0.33	0.77
Haarlemmermeer	27	265	0.15	1.43
Hellendoorn	82	262	0.59	1.88
Korendijk	2	260	0.03	3.26
Bodegraven-Reeuwijk	9	31	0.10	0.35
Roermond	30	30	0.42	0.42
Horst aan de Maas	29	30	0.15	0.16
Houten	13	30	0.22	0.51
Beuningen	11	30	0.23	0.64
Hillegom	7	30	0.52	2.23
Oirschot	21	29	0.20	0.28
Geldrop-Mierlo	18	29	0.57	0.92
Onderbanken	9	29	0.42	1.37
Gorinchem	6	29	0.27	1.32
Noordwijk	55	28	1.54	0.79
Bedum	10	28	0.22	0.62
Kerkrade	9	28	0.41	1.26
Oegstgeest	14	27	1.76	3.39
Rucphen	27	26	0.42	0.40
Uden	16	26	0.24	0.39
Beesel	10	26	0.34	0.89
Terschelling	156	25	1.80	0.29
Ubbergen	39	25	1.00	0.64
Vlieland	19	25	0.52	0.68
Leek	13	25	0.20	0.39
Moerdijk	11	25	0.07	0.15
Haarlemmerliede en Spaarnwoude	2	25	0.09	1.18
Smallingerland	43	24	0.34	0.19
Noordwijkerhout	23	24	0.98	1.02
Nuenen, Gerwen en Nederwetten	15	24	0.44	0.71
Waterland	8	24	0.14	0.43
Bergambacht	4	24	0.11	0.63
Montfoort	2	24	0.05	0.63
Sittard-Geleen	29	23	0.36	0.29
Vaals	20	23	0.84	0.96
Someren	14	23	0.17	0.28
Gaasterlkn-Sleat	66	22	0.61	0.20
Nijkerk	18	22	0.26	0.31
Appingedam	5	22	0.20	0.90
Maassluis	3	22	0.30	2.17
Zundert	39	21	0.32	0.17

Hof van Twente	91	257	0.42	1.19	Waalre	20	21	0.88	0.93
Raalte	44	251	0.26	1.46	Uitgeest	1	21	0.04	0.94
Aa en Hunze	179	244	0.64	0.87	Hollands Kroon	42	20	0.11	0.05
Winterswijk	83	238	0.60	1.71	Haaren	26	20	0.44	0.34
Vlagtwedde	58	230	0.34	1.35	Rijswijk	14	20	0.97	1.38
Molenwaard	6	224	0.05	1.77	Oudewater	2	20	0.05	0.50
Bloemendaal	323	222	7.99	5.49	Heemstede	37	19	3.84	1.97
Zeist	175	218	3.60	4.48	Huizen	28	19	1.75	1.18
Rhenen	88	217	2.01	4.96	Bergen (L_y)	27	19	0.25	0.18
Dalfsen	43	202	0.26	1.21	Mook en Middelaar	25	19	1.33	1.01
Haren	160	198	3.15	3.90	Schinnen	12	19	0.50	0.79
De Bilt	137	195	2.04	2.90	Lingewaard	10	19	0.14	0.27
Zutphen	26	187	0.61	4.36	Nieuwegein	8	19	0.31	0.74
Renkum	135	179	2.86	3.79	Purmerend	7	19	0.29	0.77
Barneveld	129	178	0.73	1.01	Heerhugowaard	5	19	0.13	0.48
Delft	40	168	1.66	6.98	Zevenaar	5	19	0.09	0.33
Delfzijl	27	168	0.20	1.23	Den Helder	24	18	0.51	0.38
Ermelo	136	166	1.59	1.94	Leiderdorp	5	18	0.41	1.47
Olst-Wijhe	26	166	0.22	1.40	Nederlek	2	18	0.06	0.58
Borger-Odoorn	110	162	0.40	0.58	Nuth	2	18	0.06	0.54
Bronckhorst	130	158	0.45	0.55	Schoonhoven	0	18	0.00	2.60
Tynaarlo	113	156	0.77	1.06	Alblasserdam	0	18	0.00	1.79
Leeuwarden	18	155	0.21	1.85	Ameland	44	17	0.74	0.29
Staphorst	24	151	0.18	1.11	Achtkarspelen	19	17	0.18	0.16
Noordenveld	155	150	0.75	0.73	Zuidhorn	18	17	0.14	0.13
Steenwijkerland	80	150	0.25	0.47	Best	14	17	0.40	0.48
Lelystad	58	149	0.23	0.58	Vlissingen	10	17	0.28	0.48
De Marne	39	146	0.23	0.86	Rijnwoude	2	17	0.03	0.29
Voorst	49	145	0.39	1.15	Nederweert	13	16	0.13	0.16
Tubbergen	46	143	0.31	0.97	Zwijndrecht	4	16	0.18	0.70
Ooststellingwerf	87	142	0.38	0.63	Duiven	3	16	0.09	0.45
Heerenveen	66	140	0.47	1.00	Barendrecht	1	16	0.05	0.74
Heiloo	29	140	1.53	7.36	Alphen-Chaam	113	15	1.21	0.16
Dordrecht	24	137	0.24	1.39	Bladel	21	15	0.28	0.20
Lochem	192	136	0.89	0.63	Aalsmeer	7	15	0.22	0.46
Zeewolde	60	134	0.24	0.53	Zederik	4	15	0.05	0.20
Berkelland	114	132	0.44	0.51	Bergeijk	62	14	0.61	0.14
Veere	58	131	0.43	0.97	Slochteren	55	14	0.35	0.09
Wijdemeren	78	130	1.02	1.70	Valkenswaard	34	14	0.60	0.25
Laren	48	128	3.87	10.31	Voorschoten	24	14	2.08	1.21
Kampen	10	128	0.07	0.85	Aalten	23	14	0.24	0.14
Heerlen	19	124	0.42	2.72	Zoeterwoude	3	14	0.14	0.64
Putten	76	121	0.89	1.42	Hulst	11	13	0.05	0.06
Haarlem	48	121	1.50	3.77	Oostzaan	8	13	0.50	0.81
's-Hertogenbosch	37	119	0.40	1.30	Lingewaal	6	13	0.11	0.24

Schagen	21	119	0.12	0.69	Waddinxveen	5	13	0.17	0.44
Nijmegen	64	118	1.11	2.05	Tiel	3	13	0.09	0.37
Westland	21	117	0.26	1.44	Druten	3	13	0.07	0.31
Oisterwijk	68	116	1.04	1.78	Neerijnen	3	13	0.04	0.18
Goirle	21	115	0.50	2.72	Loppersum	3	13	0.03	0.12
Vlaardingen	13	113	0.49	4.23	Beemster	0	13	0.00	0.18
Hoogeveen	34	112	0.26	0.87	Landgraaf	24	12	0.97	0.49
Zaanstad	30	112	0.36	1.35	Halderberge	24	12	0.32	0.16
Noordoostpolder	58	111	0.12	0.24	Veldhoven	21	12	0.66	0.38
Rijssen-Holten	109	107	1.15	1.13	Sint-Michielsgestel	13	12	0.22	0.20
Maastricht	25	106	0.42	1.76	Enkhuizen	8	12	0.61	0.92
Woerden	13	106	0.14	1.14	Landsmeer	6	12	0.23	0.45
Zwartewaterland	7	102	0.08	1.17	Ridderkerk	5	12	0.20	0.48
Assen	74	100	0.89	1.20	Langedijk	3	12	0.11	0.44
Overbetuwe	23	100	0.20	0.87	Schermer	3	12	0.05	0.19
Zandvoort	111	96	3.29	2.85	Vlist	2	12	0.04	0.21
Dronten	94	95	0.28	0.28	Boxmeer	28	11	0.25	0.10
Stichtse Vecht	42	95	0.39	0.89	Skarsterlen	26	11	0.12	0.05
Winsum	3	95	0.03	0.93	Gennep	24	11	0.48	0.22
Pijnacker-Nootdorp	9	93	0.23	2.41	Gilze en Rijen	19	11	0.29	0.17
Breda	107	91	0.83	0.71	Steenbergen	19	10	0.13	0.07
Leudal	46	91	0.28	0.55	Eemnes	4	10	0.13	0.32
Epe	151	90	0.96	0.57	Hilvarenbeek	50	9	0.52	0.09
Doetinchem	68	89	0.85	1.12	Montferland	50	9	0.47	0.08
Roosendaal	34	89	0.32	0.83	Son en Breugel	19	9	0.72	0.34
Zoetermeer	19	89	0.51	2.40	Urk	6	9	0.51	0.76
Velsen	164	88	3.34	1.79	Grave	3	9	0.11	0.32
Oldambt	57	88	0.24	0.37	Geertruidenberg	3	9	0.10	0.30
Nunspeet	161	86	1.25	0.67	Cranendonck	29	8	0.37	0.10
Heusden	42	86	0.52	1.06	Menterwolde	15	8	0.18	0.10
Dinkelland	56	85	0.32	0.48	Bernheze	12	8	0.13	0.09
Blaricum	17	84	1.53	7.55	Marum	11	8	0.17	0.12
Midden-Delfland	3	82	0.06	1.66	Vianen	6	8	0.14	0.19
Bunnik	35	81	0.93	2.16	Geldermalsen	6	8	0.06	0.08
De Ronde Venen	23	80	0.20	0.68	Giessenlanden	4	8	0.06	0.12
Bergen op Zoom	106	78	1.14	0.84	Hardinxveld-Giessendam	1	8	0.05	0.41
Gulpen-Wittem	77	78	1.05	1.06	Cromstrijen	0	8	0.00	0.15
Meppel	23	77	0.40	1.35	Weststellingwerf	44	7	0.19	0.03
Elburg	24	75	0.38	1.17	Schiermonnikoog	34	7	0.77	0.16
Heerde	38	73	0.47	0.91	Brunssum	32	7	1.85	0.40
Teylingen	21	73	0.63	2.18	Echt-Susteren	28	7	0.27	0.07
Opsterland	103	72	0.45	0.32	Eersel	27	7	0.32	0.08
Stadskanaal	55	72	0.46	0.60	Reusel-De Mierden	21	7	0.27	0.09

Diemen	4	70	0.31	5.38	Oost Gelre	10	7	0.09	0.06
Leusden	69	69	1.17	1.17	Cuijk	7	7	0.12	0.12
Losser	63	69	0.63	0.69	Krimpen aan den IJssel	5	7	0.56	0.78
Gouda	9	69	0.50	3.81	Deurne	18	6	0.15	0.05
Wageningen	95	68	2.94	2.10	Landerd	9	6	0.13	0.08
Groesbeek	91	68	2.06	1.54	Eemsmond	9	6	0.05	0.03
Oss	13	67	0.08	0.42	Rijnwaarden	4	6	0.08	0.12
Muiden	2	67	0.13	4.33	Bunschoten	2	6	0.06	0.19
Schouwen-Duiveland	326	66	1.38	0.28	Oud-Beijerland	1	6	0.05	0.31
Goeree-Overflakkee	117	66	0.44	0.25	Leeuwarderadeel	1	6	0.02	0.14
Heeze-Leende	52	66	0.50	0.63	Meerssen	32	5	1.16	0.18
Texel	79	65	0.48	0.40	Borne	19	5	0.73	0.19
Loon op Zand	70	65	1.38	1.28	Veenendaal	18	5	0.91	0.25
Zuidplas	12	65	0.19	1.01	Maasgouw	17	5	0.29	0.09
Hengelo	39	64	0.63	1.04	Noord-Beveland	16	5	0.18	0.06
Oldebroek	24	62	0.24	0.63	Oude IJsselstreek	12	5	0.09	0.04
Weesp	6	62	0.27	2.84	Leerdam	9	5	0.26	0.15
Woensdrecht	112	61	1.22	0.66	Borsele	9	5	0.06	0.04
Haaksbergen	50	59	0.47	0.56	Neder-Betuwe	4	5	0.06	0.07
Middelburg	12	59	0.24	1.19	Binnenmaas	4	5	0.05	0.07
Wijk bij Duurstede	10	59	0.20	1.17	Papendrecht	1	5	0.09	0.46
Naarden	22	57	0.91	2.35	Gemert-Bakel	27	4	0.22	0.03
Medemblik	7	57	0.06	0.45	Culemborg	14	4	0.45	0.13
Venray	23	56	0.14	0.34	Laarbeek	13	4	0.23	0.07
Kaag en Braassem	6	56	0.08	0.78	Beek	8	4	0.38	0.19
Tilburg	77	54	0.65	0.45	Voerendaal	7	4	0.22	0.13
Weert	46	54	0.44	0.51	Ten Boer	6	4	0.13	0.09
Lansingerland	10	54	0.18	0.96	Graft-De Rijp	4	4	0.18	0.18
West Maas en Waal	2	54	0.02	0.63	Zeevang	3	4	0.07	0.10
Brummen	54	53	0.64	0.62	Menameradiel	3	4	0.04	0.06
Westvoorne	161	51	2.76	0.87	Ouderkerk	0	4	0.00	0.14
Boxtel	42	51	0.65	0.79	Asten	13	3	0.18	0.04
Hoorn	3	50	0.14	2.40	Grootegast	11	3	0.13	0.03
Castricum	152	48	2.76	0.87	Boarnsterhim	10	3	0.06	0.02
Bellingwedde	132	48	1.20	0.44	Woudrichem	7	3	0.14	0.06
Bussum	35	48	4.29	5.89	Stede Broec	5	3	0.32	0.19
Sluis	12	47	0.04	0.17	Dongen	4	3	0.13	0.10
Katwijk	63	46	2.42	1.76	Renswoude	3	3	0.16	0.16
Oosterhout	61	46	0.83	0.63	Westervoort	2	3	0.26	0.38
Buren	13	45	0.09	0.31	Edam-Volendam	2	3	0.12	0.18
Leidschendam-Voorburg	37	44	1.04	1.24	Drechterland	1	3	0.02	0.05
Woudenberg	35	44	0.95	1.19	Wijchen	17	2	0.24	0.03
Etten-Leur	12	44	0.21	0.79	Dantumadiel	14	2	0.16	0.02

Wierden	26	43	0.27	0.45	Tholen	13	2	0.08	0.01
Twenterand	15	43	0.14	0.40	Veghel	11	2	0.14	0.03
Terneuzen	9	43	0.03	0.16	Sint Anthonis	11	2	0.11	0.02
Rozendaal	91	42	3.26	1.50	Koggenland	6	2	0.07	0.02
Heumen	43	42	1.04	1.01	Reimerswaal	5	2	0.04	0.02
Nieuwkoop	3	42	0.03	0.46	Dongeradeel	5	2	0.03	0.01
Hoogezand-Sappemeer	29	41	0.40	0.56	Harlingen	3	2	0.12	0.08
Helmond	15	41	0.27	0.75	Strijen	2	2	0.04	0.04
Alphen aan den Rijn	8	41	0.14	0.71	Opmeer	1	2	0.02	0.05
Lopik	4	41	0.05	0.52	Ferwerderadiel	1	2	0.01	0.02
Albrandswaard	2	41	0.08	1.73	Littenseradiel	1	2	0.01	0.02
Vught	19	40	0.55	1.16	Veendam	41	1	0.52	0.01
Werkendam	18	40	0.15	0.33	Maasdriel	13	1	0.17	0.01
Valkenburg aan de Geul	49	39	1.33	1.06	Uithoorn	9	1	0.46	0.05
Hattem	25	39	1.03	1.61	Lemsterland	9	1	0.10	0.01
Beverwijk	8	39	0.43	2.07	Sint-Oedenrode	8	1	0.12	0.02
Drimmelen	7	39	0.06	0.33	Stein	6	1	0.26	0.04
Almelo	44	38	0.63	0.55	Scherpenzeel	4	1	0.29	0.07
Brielle	7	38	0.22	1.22	Simpelveld	4	1	0.25	0.06
Alkmaar	17	36	0.54	1.15	Maasdonk	4	1	0.11	0.03
Schiedam	11	35	0.55	1.76	Hendrik-Ido-Ambacht	3	1	0.25	0.08
Schijndel	9	35	0.22	0.84	Franekeradeel	2	1	0.02	0.01
Bernisse	4	35	0.07	0.58	Sliedrecht	1	1	0.07	0.07
Zaltbommel	4	35	0.04	0.39	Boskoop	1	1	0.06	0.06
Roerdalen	44	34	0.50	0.38	Baarle-Nassau	22	0	0.29	0.00
IJsselstein	11	34	0.51	1.57	Kollumerland en Nieuwkruisland	9	0	0.08	0.00
Goes	11	34	0.12	0.36	Mill en Sint Hubert	6	0	0.11	0.00
Tytsjerksteradiel	50	33	0.31	0.20	Wormerland	5	0	0.11	0.00
Eijsden-Margraten	35	33	0.45	0.42	Kapelle	4	0	0.10	0.00
Capelle aan den IJssel	25	33	1.62	2.14	Pekela	4	0	0.08	0.00
Peel en Maas	22	33	0.14	0.20	Aalburg	3	0	0.06	0.00
Spijkenisse	16	33	0.53	1.09	Millingen aan de Rijn	1	0	0.10	0.00
Hellevoetsluis	16	33	0.49	1.01	het Bildt	0	0	0.00	0.00
Waalwijk	11	33	0.16	0.49	Boekel	0	0	0.00	0.00
Ouder-Amstel	11	32	0.43	1.24					
Heemskerk	67	31	2.43	1.12					
SMdwest-Fryslnn	37	31	0.08	0.06					
Oldenzaal	31	31	1.41	1.41					