Centre for Geo-Information

Thesis Report GIRS-2015-09

Revealing spatial and temporal patterns from Flickr photography

A case study with tourists in Amsterdam

Sander van der Drift

April 2015

WAGENINGEN UNIVERSITY
WAGENINGEN UR

Revealing spatial and temporal patterns from Flickr photography

- A CASE STUDY WITH TOURISTS IN AMSTERDAM-

Sander van der Drift

(850120 199 110)

Supervisors

Dr. Ir. Arend Ligtenberg[1]

Dr. Ir. Ron van Lammeren[1]

[1] Laboratory of Geo-Information Science and Remote Sensing

Wageningen, The Netherlands

A thesis submitted in partial fulfilment of the degree of Master of Science

at Wageningen University and Research Centre

The Netherlands

April 2, 2015

Wageningen, The Netherlands

Thesis code number: GRS-80436

Thesis report: GIRS-2015-09

Wageningen University and Research Centre

Laboratory of Geo-Information Science and Remote Sensing

Abstract

This thesis presents an exploratory visual analytics approach to identify *temporal distributions*, *spatial clusters* and *popular routes* of tourists in Amsterdam by making use of geotagged photos from social media platform Flickr. Our methods combine the analytical strength of humans with the data processing power of computers, using geovisualisations and charts to explore data, find patterns, and draw conclusions from its outcomes. For this research, the metadata of 2,849,261 geotagged photos was harvested from Flickr and stored in a spatial database. From this dataset, 393,828 photos were located in the municipality of Amsterdam. Our semi-automatic classification method classified 39,1% of the users as tourist with a very high precision and recall. The temporal distribution of tourists and locals is compared for different temporal granularities. A method is presented to assess photo timestamps by making use of photos that contain a real clock. We implemented and improved an existing grid-based clustering method to explore Amsterdam's spatial distribution of tourists in Google Earth. The major tourist hotspots are detected using the density-based clustering algorithm DBSCAN. Finally, we estimated the most probable routes of tourists between subsequent photo locations and aggregated all route calculations into a route density map. A qualitative approach was used to validate the study outcomes by interviewing eight tourism experts of the municipality of Amsterdam. We found that their knowledge about the city bears a good resemblance with the detected spatial clusters and route density map of tourists. Despite several imperfections of geosocial data, we conclude that our methods provide meaningful insight into the spatial and temporal patterns of tourists in urban spaces and are a valuable addition to traditional tourism surveys.

Keywords: Flickr, Tourism, Geosocial Data, Temporal Distribution, Spatial Clustering, Touristic Routes

Acknowledgements

First and foremost, I would like to express my gratitude to my supervisors Arend Ligtenberg and Ron van Lammeren for being very interested in my research and giving me valuable tips and comments. It was great to have supervisors who shared my enthusiasm about the topic. Every meeting gave me a lot of motivation and new ideas for future work. I am confident that this research project wouldn't have reached the level it has without your contributions.

I would also like to use this opportunity to thank all other staff members of the laboratory of Geo-Information Science and Remote Sensing. The many inspiring courses and personal contact made me enjoy every single moment of my Master at Wageningen University.

Furthermore, I would like to express my gratitude to all tourism experts of the municipality of Amsterdam for their valuable comments on the results of my research project. Your expert judgement is a great contribution to this work.

I would like to thank my fellow students for the many great moments and new friendships. A special appreciation goes out to Vera van Zoest for proofreading my report. If you find a mistake in this work, I most likely added it after she was done.

Last but not least, I would like to thank my family and friends who are always there to support me.

Sander van der Drift
April 2015

Table of contents

List of figures

List of tables

List of code snippets

List of abbreviations

AGI	Ambient Geographical Information
API	Application Programing Interface
EPSG	EPSG Geodetic Parameter Set
GIS	Geographic Information System
HTTP	Hypertext Transfer Protocol
JSON	JavaScript Object Notation
KDD	Knowledge Discovery in Databases
KML	Keyhole Markup Language
MGI	Master of Geo-Information Science
OGC	Open Geospatial Consortium
OSM	OpenStreetMap
REST	Representational State Transfer
SQL	Structured Query Language
SRID	Spatial Reference Identifier
VGI	Volunteered Geographical Information
WKT	Well-Known Text
XML	Extensible Markup Language

1 Introduction

1.1 Tourism in Amsterdam

Enjoying an evening picnic in the scenic Vondelpark, relaxing after a day of strolling along the canals and marvelling at paintings from Van Gogh and Rembrandt. Amsterdam is a major destination for tourists from both the Netherlands itself and abroad. The research and statistics department of the city of Amsterdam (O+S 2014) recorded a total number of 11.3 million overnight stays in 2013 and tourism is expected to grow by 29 per cent until 2025 (NBTC 2013). Many people and organisations benefit from this large amount of visitors. Their expenditure generates jobs for the local community and tax revenues for the municipality. However, the city struggles with the amount of tourists and day visitors it receives. Tourism, along with all its benefits, is causing an increasing pressure on Amsterdam's city centre and more and more people believe that it disturbs the way the city functions. The biggest problems experienced by residents of Amsterdam's city centre are overcrowding, high noise levels, litter and dangerous traffic situations (Berge and Jakobs 2013). To cope with these problems, planners and policy-makers need detailed information about the whereabouts of a tourist in a city (Edwards, Dickson et al. 2010).

Locations where people have been at specific moments in time are traditionally studied by expensive and time-consuming fieldwork methods like questionnaires, people counts and travel diaries. Gathering knowledge with these techniques provides limited coverage in both space and time (Wood, Guerry et al. 2013). Wolf, Guensler et al. (2001) assessed the feasibility to completely replace the traditional methods by means of advanced GPS data loggers. The researchers obtained very promising results and concluded that GPS tracking outperforms traditional fieldwork methods in terms of spatio-temporal accuracy. Girardin, Calabrese et al. (2008) used aggregated mobile phone data to analyse the spatial and temporal presence of tourists in Rome. This method is suitable to study the behaviour of groups from different nationalities in time, but only provides spatial information at neighbourhood level. Today's *geospatial web* (Haklay, Singleton et al. 2008) and the rise of *location-based social networks* (Roick and Heuser 2013) open new ways of exploring tourism dynamics in a city.

1.2 Geosocial data

Social media has gained a widespread popularity in the era of *Web 2.0*. Every day, users around the world generate a vast amount of content on social media platforms and this amount is expected to grow significantly in the years to come. Microblogging network Twitter recently reported that a stunning amount of 500 million Tweets are being sent every day (Twitter 2014). The popular photo sharing platform Flickr reported the 6 billionth upload on their platform in August 2011 (Flickr 2011). Kaplan and Haenlein (2010) defined social media as "a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of *User Generated Content*" (p.61).

Nowadays, a substantial part of the available social media data has a geographical component. As Sui and Goodchild (2011) described it, "social media moved from cyberspace to real place" (p.1740). Locations are stored in the form of a *geotag* or as a textual description. A geotag is metadata of an object, specifying a location in space and time (Janowicz, Raubal et al. 2009). In this report we refer to social media data with a spatial component as *geosocial data* (Croitoru, Crooks et al. 2013).

Geosocial data is a form of *ambient geographical information* (AGI), a term coined by Stefanidis, Crooks et al. (2013). AGI differs from the well-known *volunteered geographical information* (VGI) which was defined by Goodchild (2007). The main purpose of a creator of AGI is to share a message or media item while the geo-information is a by-product. This differs from OpenStreetMap, a successful example of VGI where members of a large community deliberately share geo-information to create a free-to-use map of the world. Stefanidis, Crooks et al. (2013) advocate that VGI is a form of *crowdsourcing* while AGI is a form of *crowd-harvesting*.

1.3 Visual analytics and knowledge discovery

Croitoru, Crooks et al. (2013) have identified the abundance of publicly available geosocial data as a valuable source of information and "the new big data challenge for the computational and geospatial communities" (p.2404). Recent studies have proven that geosocial data can indeed be used for knowledge discovery and decision-making. An example is a research project by Wood, Guerry et al. (2013) who successfully used Flickr images to investigate where people recreate. Another example is a method developed by Hollenstein and Purves (2010) to geographically identify city centres using metadata from Flickr photos. Estima and Painho (2013) used publicly available Flickr data for the quality control of land cover maps. These projects illustrate that geosocial data can be successfully utilized for purposes that differentiate from the purpose of the creator. However, collecting, analysing and interpreting this type of data requires highly skilled data scientists.

One of the difficulties, according to Thatcher (2014), is that creators of geosocial data often have other intentions than a researcher. Unlike data produced by official authorities or with VGI initiatives, raw geosocial data tends to be unstructured and comes in a variety of different forms. Furthermore, users of social media continuously add new data, resulting in a high velocity and big volume. The field of geo-information science has moved from a data-poor to a data-rich era, aptly described as an avalanche of data by Miller (2010). The characteristics of large databases with geosocial data correspond with the characteristics of big data; volume, velocity and variety (Croitoru, Crooks et al. 2014). These large volumes of data have to be processed before it reveals its spatial and temporal patterns. Traditional GIS tools are often not suitable for analysing this type and amount of data. Sui and Goodchild (2011) emphasize that the development of new tools and methods to study the spatial dynamics of large datasets is a challenge for the years to come.

A scientific domain that deals with gaining understanding and discovering patterns in today's complex big spatial data sets is geovisual analytics (Andrienko, Andrienko et al. 2011). As explained by Keim, Andrienko et al. (2008), visual analytics combines the analytical strength of humans with the data processing power of computers. An important characteristic of geovisual analytics are geovisualisations, which are essential in understanding large spatio-temporal datasets (Wood, Dykes et al. 2007). They help researchers to explore data, find patterns, and draw conclusions from its outcomes.

Geovisual analytics is often used in combination with *Knowledge Discovery in Databases* (KDD). KDD deals with theories, methods and practices to extract meaning from large and complex databases (Mennis and Guo 2009). The KDD process consists of different steps, including data selection, data pre-processing, data transformation, data mining and interpretation (Fayyad, Piatetsky-Shapiro et al. 1996). An overview of the steps is given in figure 1-1. The process does not necessarily have to be executed in this order. It is an iterative method where steps can be skipped or repeated, depending on the requirements and expertise of the researchers. Miller and Han (2009) emphasize that the exact steps taken in the KDD process are difficult to specify before executing the work. Promising methods to derive knowledge from complex social media datasets have been developed over the last couple of years but Croitoru, Crooks et al. (2013) emphasize that new knowledge discovery methods are required that can cope with the dynamic and unstructured nature of geosocial data.

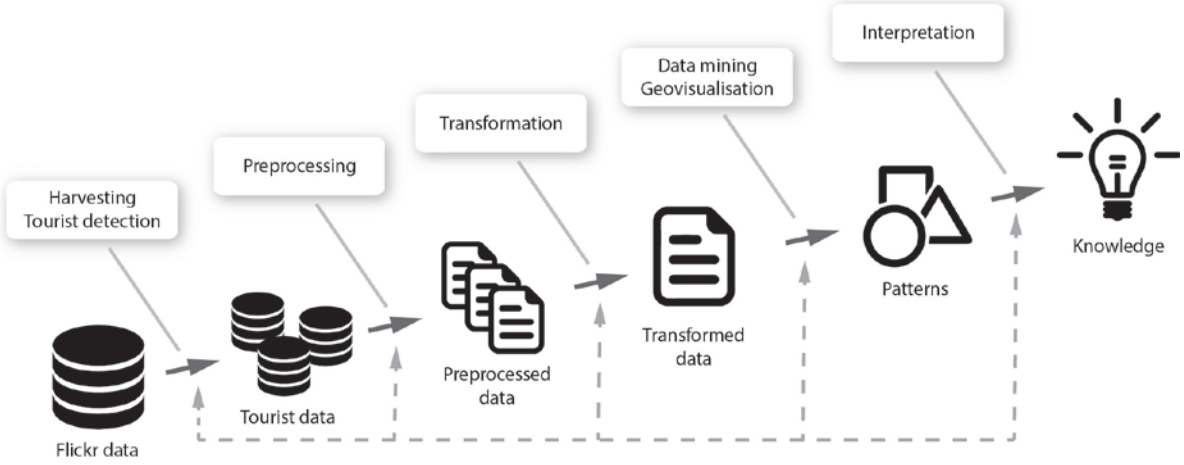


Figure 1-1: Knowledge Discovery in Databases, adapted from Fayyad (1996)

1.4 Objective and research questions

Ashworth and Page (2011) mentioned that "it is curious that very little attention has been given to questions about how tourists actually use cities" (p.7). Most municipalities have very detailed statistics about hotels and controlled touristic sites such as attractions and museums. What lacks is detailed information about the spatial and temporal behaviour of a tourist in public urban spaces. The rise of geosocial media provides new state-of-the-art methods of gathering knowledge about tourists that can be used for decision-making and planning purposes.

The objective of this exploratory research project is to develop, implement and test methods that reveal the spatial and temporal patterns of tourists from a large dataset of geotagged Flickr photos. We aim to achieve the objective by answering the following research questions.

RQ-01: *What methods are available to detect spatial and temporal patterns from geosocial data?*

RQ-02: *What methods need to be to implemented to identify temporal distributions, spatial clusters and popular routes of tourists from the metadata of Flickr photos?*

RQ-03: *How well do the identified temporal distributions, spatial clusters and popular routes resemble the spatial and temporal behaviour of tourists?*

1.5 Study area

With this research project we contribute to the project *Beautiful Noise* of the recently founded Amsterdam Institute For Advanced Metropolitan Solutions (AMS). The study area of Beautiful Noise is the municipality of Amsterdam, highlighted with a light blue transparent fill in the right map of figure 1-2. All temporal analyses that are presented in this work are based on data in the whole municipality. The spatial analyses are only executed for the touristic city centre. This area is highlighted with a dark blue transparent fill and includes all neighbourhoods encircled by the city centres ring road and the areas around the major museums and the *Vondelpark*.

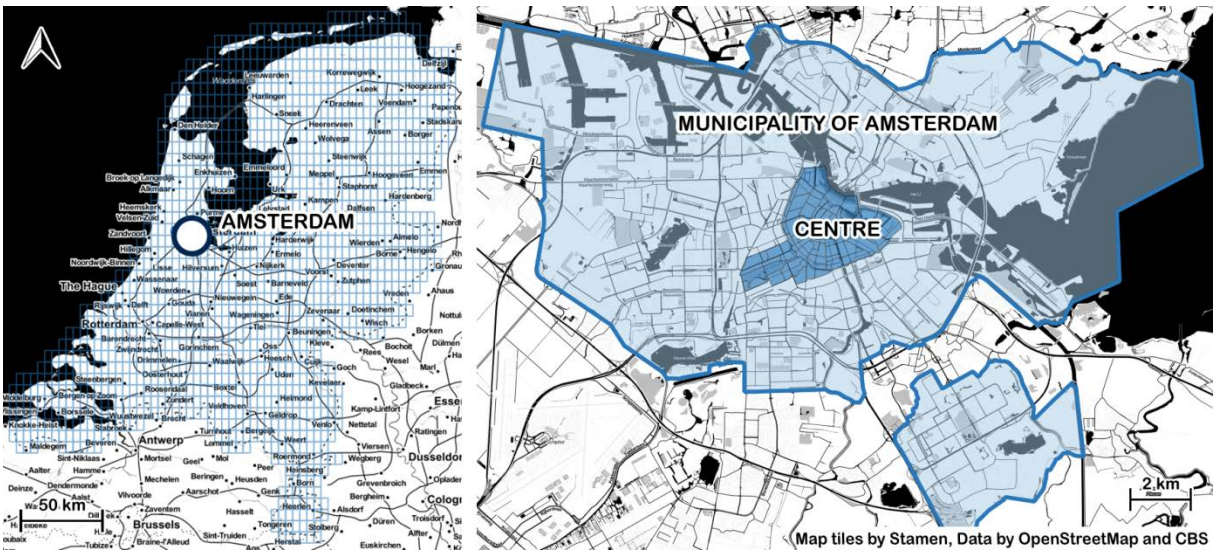


Figure 1-2: Study area

1.6 Reading guide

Recent studies in the field of geo-information science proved that the abundance of publicly available *geosocial data* is a new source for decision-making and planning purposes. Chapter 2 describes a selection of those projects and provides an answer to the first research question: *What methods are available to detect spatial and temporal patterns from geosocial data?*

Based on the selected methods and definitions, chapter 3 describes the methods and techniques that we have implemented to harvest, pre-process, transform and mine a large dataset of geotagged photos. The chapter provides an answer to the second research question: *What methods need to be implemented to identify temporal distributions, spatial clusters and popular routes of tourists from the metadata of Flickr photos?*

The implemented methods revealed different spatial and temporal patterns of tourists in Amsterdam. Chapter 4 presents the project results and provides an answer to the third research question: *How well do the identified temporal distributions, spatial clusters and popular routes resemble the spatial and temporal behaviour of tourists?*

The final chapter presents the main conclusions, discussion and recommendations for future work. Each chapter in this report opens with a short outline of its content.

2 Related work

Over the last years, various scholars in the field of geo-information science have published articles about knowledge discovery from geosocial data. We have reviewed the state-of-the-art and provide an answer to the first research question: *What methods are available to detect spatial and temporal patterns from geosocial data?* The final section contains a selection of methods, which are suitable for this research project.

2.1 Visual analytics

Girardin, Dal Fiore et al. (2007) are amongst the first researchers that successfully used the spatial and temporal features of Flickr photos to study the behaviour of tourists. They analysed the density of tourists in the province of Florence in Italy by using a visual analytics approach and explored the temporal distribution of tourists at three different temporal granularities: days of the week, months of the year and days of the year. All geotagged photos that were taken in the province of Florence in 2006 were used to create an animation showing the monthly activity of tourists. Finally, they compared the locations visited by Americans with the locations visited by Italians in the northern part of central Italy. Their visualisations showed that domestic photographers visit many places all around Italy while Americans generally visit the well-known tourist destinations.

Jankowski, Andrienko et al. (2010) used geovisual analytic tools like dynamic maps and diagrams to explore people's activity in space and time. Instead of finding the most popular places in a city, the researchers were interested in locations that are potentially interesting. Their method subdivided the study area into Voronoi polygons based on the spatial distribution of photos. The centre of an area contains the highest photo density. Next, the timestamps of photos were used to create the temporal distribution of visitors per Voronoi shaped area. Based on the detected temporal distribution, areas were classified as potentially interesting or not and visualised on a map. Visitor flows between the Voronoi areas were calculated for different spatial and temporal scales and visualised with the flow mapping technique of Slocum, McMaster et al. (2009).

Sagl (2012) created a map with 3-dimensional squared clusters visualising the mobile phone activity on an average working day in the city of Udine in Italy. Other researchers that used this promising technique for visual analytics are Kádár and Gede (2013) who conducted a case study with tourists in Budapest. Inspired by Shoval (2008), they implemented a method to visualise grid-based clusters with the density of photographs in Google Earth. The interactive method was used to compare the photo densities of tourists with the photo densities of locals. Like Wood, Dykes et al. (2007), they demonstrated the usefulness of Google Earth for geovisualisations and visual analytics.

2.2 Spatial accuracy

Hauff (2013) investigated the spatial accuracy of geotagged Flickr photos. The researcher found a strong correlation between the popularity of a location and the accuracy of a geotag. Photos taken at popular locations are mostly geotagged with a high spatial accuracy while photos taken at less popular locations have a lower spatial accuracy. Hollenstein and Purves (2010) tested the location accuracy of geotagged photos in London and concluded that Flickr's accuracy value is useful for filtering out images that are placed at a wrong location. Their data exploration also revealed that some users upload many photos on a single location. These records have to be excluded before the

data is used for the detection of spatial patterns. The researchers used density maps to explore the distribution of images in the United States that are tagged with “downtown”. At city level, good results were obtained in mapping the borders of city centres with an algorithm that uses all photos that are tagged with terms like “downtown” and “centre”.

2.3 Classification of user nationalities

Many authors specifically focussed their research on the behaviour of tourists. But how do you classify a group of tourists from a large set of geosocial data? The literature describes two different approaches to identify tourists from the metadata of photography. This first approach uses the spatial and temporal characteristics of photographers. Amongst the researchers that applied this method are Girardin, Dal Fiore et al. (2007), Van Canneyt, Schockaert et al. (2011) and Sun, Fan et al. (2013). Photographers are classified as tourist if all their photos in the study area are taken in a short time period. All other users are classified as locals.

Several other scholars used an approach that makes use of the photographer’s home location. When a photographer specified his or her home location, it is attached as metadata to all their photos. Straumann, Çöltekin et al. (2014) demonstrated that the majority of the users in their study area had disclosed their location in their user profile. They have developed a semi-automatic method to geocode the country of residence of every user. Wood, Guerry et al. (2013) verified that the entered location information by Flickr users is generally correct. The researchers derived the photographers’ origin from the location that they specified in their Flickr profile and compared this with the nationalities of visitors recorded by immigration points. They found a strong correlation and concluded that geosocial data can be a useful source for studying the nationalities of visitors. The dataset with tourist photos was used to quantify the visitation rates of 836 recreational sites around the world. A comparison between the official number of visitors of recreational sites and the number of photographers taught them that geosocial data is a reliable source for the estimation of visitation rates.

2.4 Spatial clustering

For the detection of hotspots and landmarks in large geosocial datasets, spatial clustering methods are applied. Clustering can be defined as organising a collection of points into groups based on similarity (Jain, Murty et al. 1999). Point objects within a group share more similarities with each other than with points outside that group. Different spatial clustering methods have been used in geosocial research projects and two popular ones are mean shift clustering and DBSCAN. Researchers that have used mean shift clustering to identify popular places are Van Canneyt, Schockaert et al. (2011), Xin, Changhu et al. (2010) and Crandall, Backstrom et al. (2009). This method is often used in image segmentation. An advantage of this non-parametric algorithm is that the number of clusters doesn’t have to be specified in advance. Different clusters are formed based on the scale of observation. Crandall, Backstrom et al. (2009) grouped the points of unique photographers into different granularities: metropolitan-scale (100 km) and a landmark-scale (100 m). The algorithm locates maxima’s by shifting the kernel until the highest density is found. Their analysis showed that Amsterdam is the third most photographed city in Europe and the tenth most photographed city worldwide. Based on Flickr data, Amsterdam’s dam square is identified as the cities touristic highlight.

Another popular algorithm is DBSCAN: density-based spatial clustering for applications with noise (Ester, Kriegel et al. 1996). This method has been implemented by Lee, Cai et al. (2014), Sun, Fan et al. (2013) and Kisilevich, Krstajic et al. (2010). DBSCAN can detect clusters with different shapes and sizes and is not sensitive to noise. This makes the method very suitable for detecting spatial clusters in geosocial point data. Figure 2-1 visually compares the mean shift and DBSCAN clustering methods.

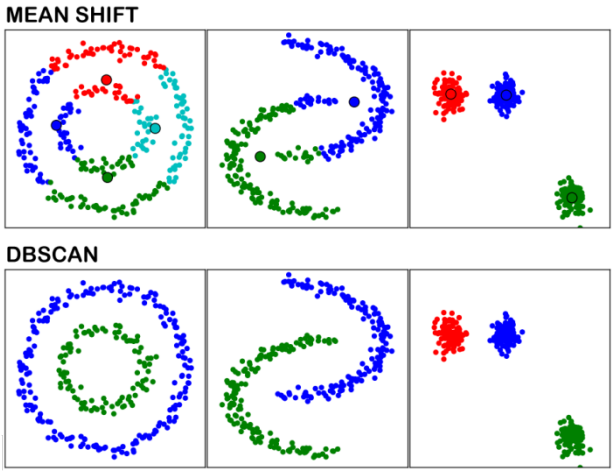


Figure 2-1: Visual comparison between Mean Shift and DBSCAN, adapted from Scikit-learn (2015)

2.5 Sequence patterns and route recommendation

Several scholars have mined geosocial data to create day itineraries for tourists (De Choudhury, Feldman et al. 2010, Xin, Changhu et al. 2010, Ali, Sas et al. 2013, Kurashima, Iwata et al. 2013, Okuyama and Yanai 2013). They focussed on the sequences and visit times of popular landmarks and the required travel times to get from one location to the next. Jankowski, Andrienko et al. (2010) discovered that the majority of user itineraries were located in a relatively small area in the centre of cities. Popescu and Grefenstette (2009) also focussed on mining trip related information at city level and noticed that tourists often don't take the shortest path between landmarks but visit other places in between.

This is confirmed by Sun, Fan et al. (2013). They explained that tourists like attractive routes that offer sightseeing, dining and shopping opportunities and found out that the number of photos on or near a road is a good indication for the popularity of a route. Their research has resulted in a system that recommends the most popular landmarks in a city, as well as the most attractive route between these landmarks. They have developed an algorithm that estimates the tourism popularity of a road based on the number of photos and landmarks in the roads proximity. The combination between the roads popularity and the length of the road determines the cost to travel over this road. Finally, Dijkstra's shortest path algorithm (Dijkstra 1959) is used to calculate the most touristic route between two landmarks.

Authors Kachkaev and Wood (2014) agree that the attractiveness of streets can be derived from photography. A key element of their research project is the implementation of some innovative data cleaning techniques. The authors reasoned that the photos on social media platforms were not created for classifying the attractiveness of urban streets. As a result, only a subset of the data is suitable for this purpose. For example, some of them contain faces while other photos are taken indoors. The filtered subset with photos is used to classify the attractiveness per road. To do so, the total amount of photos per road is counted and normalized with the road length. This normalized score is used to adjust the routing travel cost of the road. Like Sun, Fan et al. (2013), the researchers have used the shortest path algorithm to compute pedestrian routes. Their aim is to plan attractive walks in a city that take a specific amount of time. In other words, they are not looking for the most attractive route in the least amount of time but for an attractive route in a given time. The algorithm iteratively calculates paths between a start and endpoint until a route is found that requires approximately (+/- 5%) the given time. The result is suggested as an attractive route to tourists.

2.6 Selected methods and definitions

Geosocial data corresponds to Peuquet's (1994) characterisation of spatio-temporal data. It contains the three interrelated components *space* (where), *time* (when) and *objects* (what). In this research we regard an individual Flickr photographer as an object and this object is represented in space and time by a set of geotagged photos. Andrienko and Andrienko (2006) described the two-component relations between *objects and time* and *objects and space*. The *temporal distribution* refers to objects and their occurrences in time and the *spatial distribution* refers to objects and their positions in space.

In this chapter, a selection of state-of-the-art projects in the field of geo-information science and geosocial data are described. During the literature study, we identified various methods to detect temporal and spatial distributions from geotagged photography. Our work implements and extents a selection of suitable methods to provide planners and policy-makers valuable information about the spatial and temporal behaviour of tourists in a city. Inspired by work of Girardin, Dal Fiore et al. (2007), we will extract the temporal distributions of photographers for the granularities *days of the week*, *months of the year* and *days of the year*. In extension to this work, we will analyse the amount of photographers for several other temporal granularities and compare the temporal distributions of tourists and local residents to discover possible differences.

Other valuable information is the spatial distribution of tourists in public urban spaces. These results could for example help an urban planner to identify bottlenecks or to identify locations that have the potential and capacity to welcome more tourists. We have selected and improved the promising grid-based clustering methods of Sagl (2012) and Kádár and Gede (2013) to explore the spatial distribution of tourists in Amsterdam. The locations with the highest spatial tourist densities will be detected with the DBSCAN clustering algorithm that is applied by Kisilevich, Krstajic et al. (2010), Sun, Fan et al. (2013) and Lee, Cai et al. (2014). We improve the method by adding a pre-processing step that prevents that many photos of a single photographer could form a spatial cluster.

Subsequent photo capturing events of a single photographer (object) are defined as a form of event-based-movement by Andrienko, Andrienko et al. (2011). The photographer is the moving object and events are formed by the space and time components of a photo. Several publications showed that it is possible to study the linear flows of tourists in a city. Relating those tourist flows to a cities' pedestrian road network could help an urban planner to identify the locations where congestions occur. Researchers Sun, Fan et al. (2013) and Kachkaev and Wood (2014) proposed route recommendation systems for tourists. Their route planners calculate the most popular touristic path, defined by the number of photographs. We agree with those authors that the number of photos on or near a road is a good indication for its popularity. Their publications were discovered in the final stage of this research project. In contrast to this work, our work focuses on the most probable routes that tourists travelled between subsequent photo locations. We will aggregate all route calculations to create a route density map of tourists in Amsterdam.

All projects have in common that visual analytics was used to combine the analytical strength of humans with the data processing power of computers. This research project follows their example and makes use of geovisualisations, charts and tables to explore data, find patterns, and draw conclusions from its outcomes.

3 Methodology

This chapter describes the methods and techniques that we have implemented to harvest, pre-process, transform and mine a large dataset of Flickr photos into geovisualisations and charts. The results give insight in the *temporal distributions*, *spatial clusters* and *popular routes* of tourists in Amsterdam. Figure 3-1 provides a general overview of the different steps of the methodology.

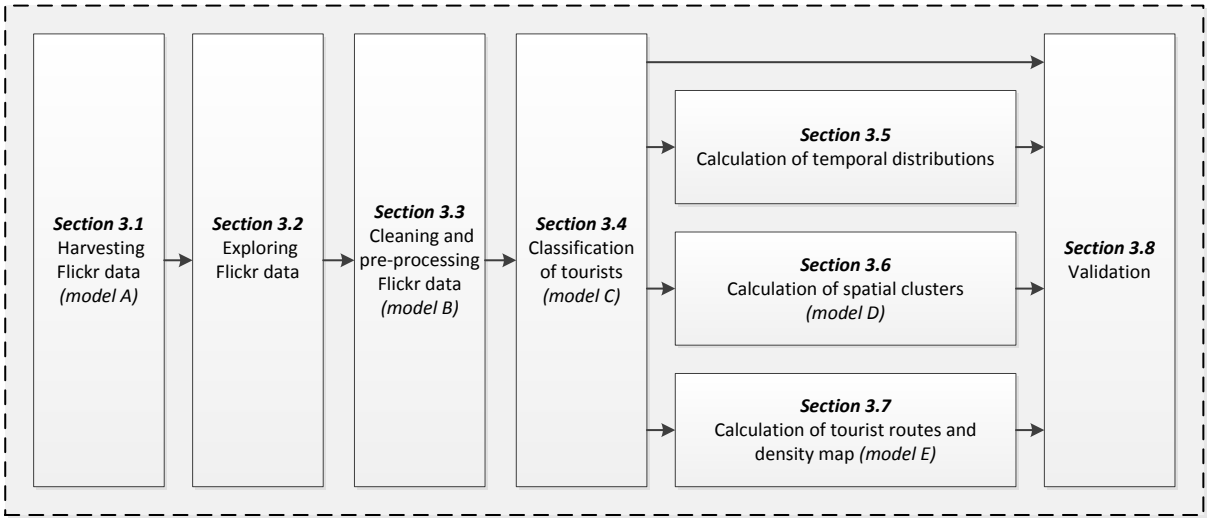


Figure 3-1: Flowchart of methodology

The first four sections of the methodology describe the data collection, exploration, pre-processing and tourist classification steps. Section 3.1 describes social media platform Flickr and the steps that are required to harvest metadata of geotagged photos from their server. A massive amount of data was collected and stored in a spatial database. The method that has been developed to explore subsets of geotagged Flickr photos in Google Earth is described in section 3.2. Thorough exploration revealed several invalid records. We have described the pre-processing steps that are required to exclude these records in section 3.3. Our research primarily focuses on tourists. A method was developed that classifies the country of residence of users with a home location in their Flickr profile. The implementation of this step is described in section 3.4.

Sections 3.5, 3.6 and 3.7 of the methodology provide an answer to the second research question: *What methods need to be implemented to identify temporal distributions, spatial clusters and popular routes of tourists from the metadata of Flickr photos?* A subset with photos is used to discover temporal distributions of tourists and locals at different granularities. The description of this step is given in section 3.5. Section 3.6 explains the implementation of the grid-based and density-based cluster algorithms. These algorithms are used to calculate the distribution of tourists and touristic hotspots in Amsterdam. Finally, we calculated the most likely routes that tourists travelled between subsequent photo locations. All route calculations are combined to create a route density map. The different steps that are required to create this density map are described in section 3.7.

Section 3.8 describes the approach to answer the third research question: *How well do the identified temporal distributions, spatial clusters and popular routes resemble the spatial and temporal behaviour of tourists?* The methods to validate the classification and temporal distributions of tourists are explained. Furthermore, we have described an expert judgement questionnaire that is used to validate the detected spatial clusters and route density map of tourists.

A mixture of techniques is used to process and visualise the data and results because traditional GIS tools are often not suitable for this type of data (Wood, Dykes et al. 2007). The presented methods make use of open source software. Among the used software packages are PostgreSQL, PostGIS, pgRouting, Java, QGIS, Python, R, Google Earth and OpenLayers.

3.1 Harvesting Flickr data

Flickr is a popular online platform that people use to store and organise photographs, share their work with the world and comment on the work of others. The website is free to use and all users are given 1 terabyte of space to store their images. We have selected the metadata of Flickr photos as the primary data source to detect the spatial and temporal patterns of tourists in Amsterdam. A set with geotagged photos is a very useful source because the timestamp and location attribute of a photo reveal the whereabouts of its capturer in space and time. It is especially suitable for tourism studies since there exists a very strong relationship between photography and tourism (Garrod 2008). Furthermore, we expect that the temporal interval between subsequent photos of a tourist is relatively small in comparison with other types of ambient geographic information like the data of Twitter. This makes photos more suitable for the detection of tourist routes.

The Flickr database stores actual images as well as the metadata of photos. This includes camera settings and the date and time when the photo was taken. A part of the photos also contain a location attribute. Friedland and Sommer (2010) discovered that approximately 4.3% of all photos uploaded in the first four months of 2010 were geotagged. Photo locations are either automatically captured by the camera or manually specified by placing the image on a map. Flickr automatically adds an accuracy value for all photos with a geographical location. The platform bases this value on the zoom level of the map when the image was geotagged. Values range between world (1), country (3), region (6), city (11) and street level (16). Many users mark the photos with various other forms of information including a title, description and a wide variety of textual tags. This metadata describes the photo and allows it to be found by other people via browsing or searching. Table 3-1 gives an overview and explanation of the metadata attributes that are relevant for our analyses.

Table 3-1: Flickr metadata attributes per photo

Attribute	Explanation
id	Unique identifier of the photo
secret	Unique key to get additional metadata of a photo
server	Server number where the photo is stored
user_id	Identifier of the photographer
user_name	User name of the photographer
user_real_name (optional)	Real name of the photographer
user_location (optional)	Home location of the photographer (city and / or country)
title (optional)	Title of the photo
media	Type of media (photo or video)
date_taken	Timestamp of photo capture
taken_granularity	Granularity of date taken timestamp (0: YYYY:MM:DD hh:mm:ss, 4: YYYY-MM, 6: YYYY, 8: Circa...)
tags (optional)	Set of labels that describe the content of the photo
latitude (optional)	Latitude coordinate of photo location (EPSG:4326)
longitude (optional)	Longitude coordinate of photo location (EPSG:4326)
accuracy	Accuracy of the photo location (World level is 1, Country is ~3, Region ~6, City ~11, Street ~16)
url	URL to the photo and profile of the photographer

An interesting feature of Flickr is the so-called Application Programmers Interface (API) that provides researchers free access to photo data for all non-commercial purposes. The process of gathering geosocial data is called harvesting. Figure 3-2 provides an overview of the different steps that are taken to harvest and pre-process the metadata of georeferenced Flickr photos. The first step (A1) is the collection of all unique photo identifiers that are present within a specific spatial and temporal extent. The detected photo ID's are used to harvest additional metadata per photo in the second step (A2). Both steps are executed with a Java application that has been developed for this research project. All results are stored in an object-relational database. The harvesting process is a very time consuming task because Flickr only allows users to send one request per second.

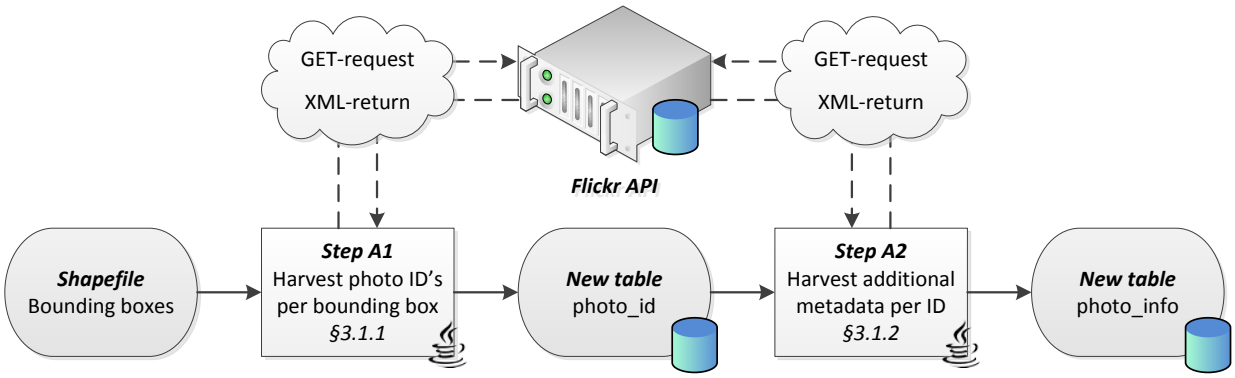


Figure 3-2: Flowchart of data harvesting (model A)

A client can request metadata from the Flickr server by using the HTTP-GET method. Which basically means sending an URL with the server's address and multiple parameters. One of these parameters is a personal API key that can be requested from Flickr. The other parameters specify the data you would like to receive. The server returns an XML or JSON file that contains all requested metadata. An example of an HTTP-GET request and XML response is given in code snippet 1.

```
https://api.flickr.com/services/rest/?
method=flickr.photos.geo.getLocation
&api_key=[your_api_key]
&photo_id=10005574224
&format=rest
```

➔

```
<?xml version="1.0" encoding="utf-8"?>
   <rsp stat="ok">
      <photo id="10005574224">
         <location latitude="52.373049"
                   longitude="4.893133"
                   accuracy="16">
         </location>
      </photo>
   </rsp>
```

Code Snippet 1: Example of HTTP-GET request and XML response

A Java application has been developed to harvest data from the Flickr API. This application makes use of the Java library Flickr4Java (2015) to make requests and interpret the returned XML-files. All results are stored in the open source database PostgreSQL (2015). The PostGIS (2015) extension adds support for geographic objects and is compliant with the "simple features for SQL" specification of OGC. Many GIS software packages support the use of this database, which makes it currently the most popular free and open source spatial database (Steiniger and Hunter 2013).

3.1.1 Photo ID's per bounding box

The GET-method flickr.photos.search can be used to get a list of photos that meet the given search criteria. We are interested in all geotagged photos within the Netherlands, taken between January 1, 2005 and January 1, 2015. The spatial search extent can be entered as parameter with a bounding box. A bounding box is defined by a comma-delimited list with the minimum longitude, minimum latitude, maximum longitude and maximum latitude. The date of photos can be specified with the parameters minimum date taken and maximum date taken in Unix timestamp format. Using one rectangular bounding box for the Netherlands would yield large amounts of photos in Belgium and Germany. Therefore, we have subdivided the Netherlands in 1550 bounding boxes. This will lower the amount of photos (only spatially relevant data will be returned), resulting in a shorter harvesting process and less load on the Flickr API. The bounding boxes are visualised in figure 1-2 on page 4.

The Flickr API returns a maximum number of 16 pages with 250 unique photographs per page, even if there are more photos available within the given bounding box. To deal with this limitation, we adjust the date taken parameters until the GET-method returns less than 4000 photos. The results (photo id, owner and secret) are stored in a PostgreSQL database table. Per unique photo, a Boolean value *isharvested* is added. This value specifies if additional metadata of the photo is harvested (step A2) and will be used in the next step. The pseudo code of the harvesting process per bounding box is given in Code Snippet 2. Pseudo code describes the code of a method in a way that is easy to interpret by humans. One line of pseudo code often represents many lines of Java or SQL code.

```
start_date <- January 1, 2005
stop_date <- January 1, 2015
bboxes <- read coordinates of all bounding boxes from shapefile

for every bbox in bboxes do
   search_from <- start_date
   search_to <- stop_date

   while search_from is before stop_date do
      pages <- search for number of pages in bbox between search_from and search_to date

      if number of pages is bigger than 16 then
         search_to <- (search_from plus search_to) divided by 2
      else

         for every page in pages do
            write all records on page to database table photo_id
         end for

      end if

      pause application for 1 second

   end while

end for
```

Code Snippet 2: Pseudo code of the method for harvesting photos per bounding box

3.1.2 Additional metadata per photo ID

Step A2 is executed after all photo ID's in the search area are harvested (step A1). The purpose of this step is to obtain additional metadata for the records that are stored in table *photo_id*. Method *flickr.photos.getInfo* requires the API key, photo ID and photo secret as parameters. Examples of additional information that is obtained in this step are the title, user ID, user location, timestamp, coordinates, tags, and the URL of the photo. The harvested values are stored in database table *photo_info*. After each successful request, the *isharvested* Boolean in table *photo_id* is set to true. This makes it possible to stop the harvesting process and continue at a later moment. The pseudo code of harvesting additional metadata per photo is given in Code Snippet 3.

```
photo_ids <- query the id's of all photos without additional metadata from database

for every photo_id in photo_ids do
   photo_metadata <- request photo metadata from Flickr server with photo_id
   write photo_metadata to database table photo_info

   pause application for 1 second

end for
```

Code Snippet 3: Pseudo code of the method for harvesting additional photo metadata

3.2 Exploring Flickr data

The next step is the preliminary exploration of the harvested dataset. Researchers Wood, Dykes et al. (2007) taught us that large spatio-temporal datasets can be best explored by means of visual analysis. Their project demonstrated that Google Earth is an excellent tool for visual exploration and interpretation of geosocial data. We have created a method that queries a subset of photos from the database and transforms the selected photos into a rich KML file. Keyhole Markup Language (KML) is an official open standard for the visualisation of geographic information (OGC 2008). Any type of query that is supported by the data and the database is possible. Examples of subsets are for example all tourist photos that are tagged with "*vondelpark*" or photos taken on king's day 2014. The KML file contains a placemark for every selected photo. After execution, the script automatically opens the file for exploration in Google Earth. Code Snippet 4 shows the pseudo code of our method.

```
photos <- query a subset of photos from the database
output_kml <- add kml code for the file header and placemark style

for every photo in photos do

   metadata <- extract required metadata from photo
   html_popup <- create html code with photo metadata

   placemark_kml <- add kml code for the header of a placemark
   placemark_kml <- add html_popup to placemark_kml
   placemark_kml <- extract coordinates from metadata and add to placemark_kml
   placemark_kml <- add kml code for style and footer of a placemark

   output_kml <- add placemark_kml

end for

output_kml <- add kml code for the file footer
write output_kml to file and open the file with Google Earth
```

Code Snippet 4: Pseudo code of the method that exports a photo subset to Google Earth

Figure 3-3 shows an example of a placemark with a popup. The selection consists of all photos that are tagged with the marketing slogan of Amsterdam: "iamsterdam". The statue that embodies the slogan is located near the Rijksmuseum and is a popular place for tourist photography. Photo details such as the title, tags, owner, capture date and the actual photo itself are revealed by the popup.

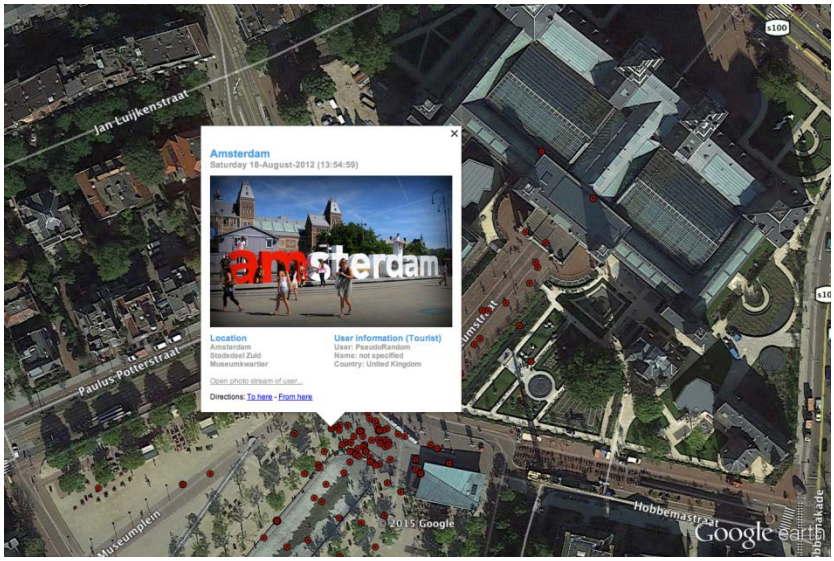


Figure 3-3: Google Earth screenshot of photos that are tagged with "iamsterdam"

3.3 Cleaning and pre-processing Flickr data

Exploration of the raw dataset revealed a substantial amount of invalid records. Examples are photos with a low spatial accuracy value or a capture timestamp that only specifies the year. Before the data can be used for further analysis tasks, data cleaning and pre-processing is needed. Data cleaning is defined as "detecting and removing errors and inconsistencies from data in order to improve the quality of data" (Rahm and Do 2000). The basic cleaning steps are included in this section. Additional data cleaning is implemented in other models of our project. The process of data cleaning and pre-processing is illustrated in figure 3-4.

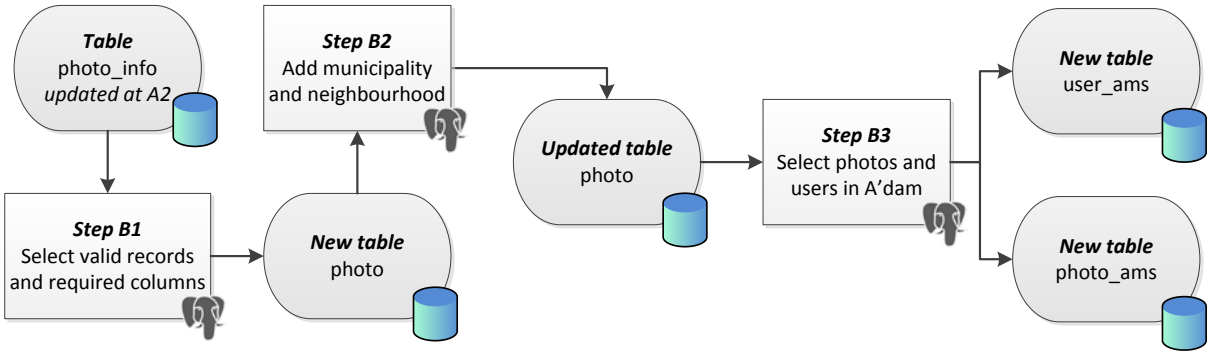


Figure 3-4: Flowchart of data cleaning and pre-processing (model B)

This first step (B1) of the process is the selection of all valid photos and metadata attributes that will be used in our analysis. Table *photo_info* contains many different attributes per photo and not all of them are relevant for our study. This step copies the necessary columns of table *photo_info* to the new table *photo*. The latitude and longitude values are used to make a PostGIS point geometry. This enables location queries and mapping of the photos. The coordinates of the point geometries are transformed to the Dutch projected coordinate system RD new (EPSG:28992). The attributes are described in table 3-1.

Hollenstein and Purves (2010) tested the location accuracy of geotags and reported that Flickr's accuracy value is useful for filtering out images that are placed at a wrong location. We are only interested in photos with the highest accuracy level (16: street level). All records that don't meet this condition are rejected. Other photos will not be copied to the new table because of a low taken granularity, a wrong media type or the absence of additional metadata. Step B1 is executed with a single SQL statement that copies all valid records and necessary attributes to a new table.

The next step (B2) is the enrichment of all photo records with the name of the location where they were taken. For this purpose, a spatial dataset with municipality and neighbourhood borders (CBS 2014) is added to the database. The spatial query that adds a municipality and neighbourhood label to every photo is given in Code Snippet 5. The municipality attribute makes it easy to create the tables *photo_ams* and *user_ams* in step B3. The first table contains all photos within our study area, the municipality of Amsterdam. The second table contains all unique users within our study area.

```
UPDATE photo
   SET municipality = cbs.gm_naam, neighbourhood = cbs.bu_naam
  FROM cbs_buurt AS cbs
 WHERE st_within(photo.geom, cbs.geom) AND water = 'NEE'
```

Code Snippet 5: SQL code for adding a CBS municipality and neighbourhood label to photos

3.4 Classification of tourists

The first step after obtaining the Flickr data is the classification of tourists. As described in section 2.3, there are two different approaches to identify tourists from raw Flickr data. This first approach uses the spatial and temporal characteristics of photographers. Photographers are classified as tourist if all their photos in the study area were taken in a short time period. All other users are classified as locals. This approach has several disadvantages. Tourists who visit the study area more than two times or longer than two weeks will be wrongly classified as locals and residents who only uploaded a few photos will be wrongly classified as tourists. Furthermore, this method does not reveal the country of residence of the photographer.

The second approach makes use of the photographer's home location. When a photographer specified his or her home location, it is attached as metadata to all their photos. Several researchers discovered that the majority of users disclose this information on social media platform Flickr. We are aware that this method excludes all Dutch tourists. Nevertheless, this solution seems to be the most promising one for this research project. Especially because Wood, Guerry et al. (2013) demonstrated that the entered location information by Flickr users is generally correct.

A model was developed that semi-automatically classifies the origin of Flickr users based on the home location that they have specified in their profile. An overview of the different steps is given in figure 3-5. The model makes use of SQL regular expressions and the geographical database GeoNames (2015). This open source database contains over 8 million place names that can be used to geocode user locations. Unfortunately, it is not possible to automatically detect the country names of all users due to spelling errors and unusual location names (e.g. A'dam instead of Amsterdam). The country names of these users are manually added to the database.

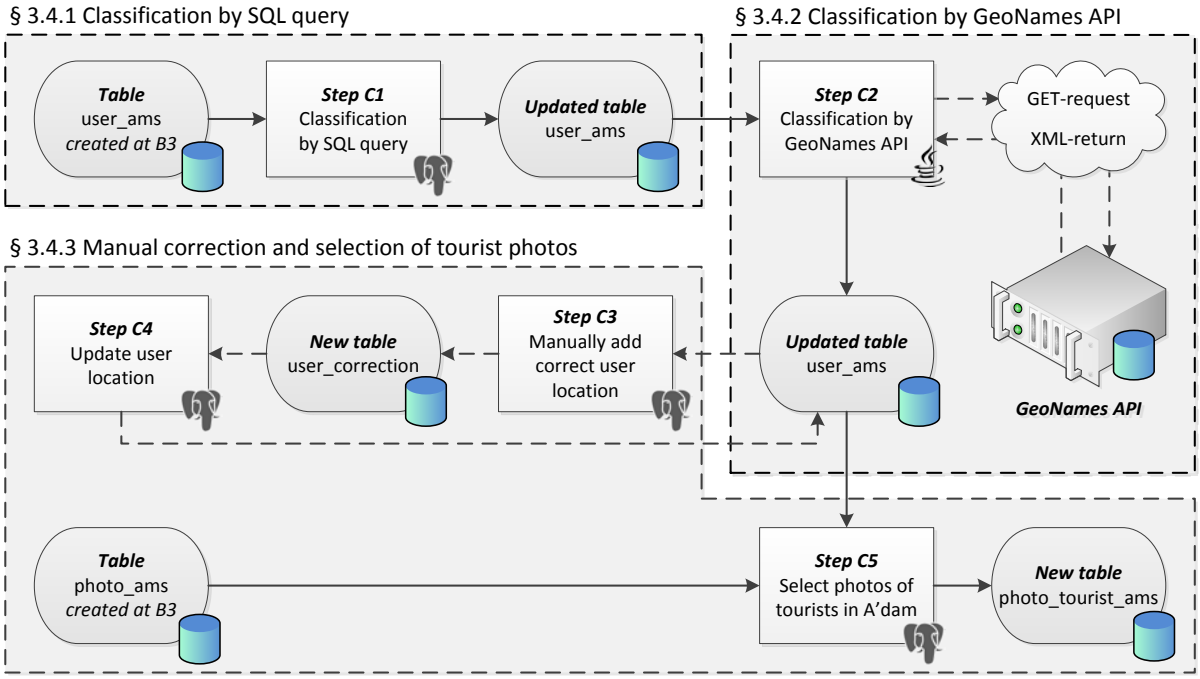


Figure 3-5: Flowchart of tourist classification (model B)

3.4.1 Classification by SQL query

The first step (C1) of the classification process makes use of SQL queries. This step is developed to efficiently classify all users with a common location value like *Amsterdam, Netherlands* and *United Kingdom*. The advantage of SQL over geocoding with the GeoNames API is that it is much faster and does not require an external database. Furthermore, we experienced better classification results with SQL. An example is the user location "*South-west of the Netherlands*". This location is correctly classified as the Netherlands by SQL because it contains the term "Netherlands". The external geocoding service GeoNames is not able to classify this location.

The columns *usercountry, istourist* and *classification* are added to the table *user_ams*. The first column specifies the detected country of the user. The second column (Boolean true or false) specifies if the user is a tourist. All users that don't live in the Netherlands are regarded as tourists. The third column stores the classification method. All records that are classified in this step (B1) get the value "*SQL*". An example of a classification query is given in Code Snippet 6. This SQL query classifies users with a *userlocation* attribute containing the phrase *holland*, *netherlands*, *nederland* or *amsterdam* as Dutch residents. Similar queries have been prepared and executed for 43 foreign countries.

```
UPDATE user_ams
SET usercountry = 'Netherlands', istourist = 'False', classification = 'SQL'
WHERE usercountry IS NULL AND userid IN
   (SELECT userid
      FROM user_ams
     WHERE userlocation ~* '\y(holland|netherlands|nederland|amsterdam)\y');
```

Code Snippet 6: SQL code for classifying the country name of Flickr users

3.4.2 Classification by GeoNames API

The second step (C2) uses the geographical database GeoNames to classify the user locations that are not classified by SQL. A Java method was developed that queries all users from our local database that are not classified by SQL but do have a location value in their Flickr profile. The method makes use of GeoNames' Java library. All unclassified user locations were geocoded with the public API and successful responses are stored in database table *user_ams*. The pseudo code of this step is given in Code Snippet 7.

```
users <- query all users without a classified origin from the database

for every user in users do

   user_id <- extract user_id from user
   user_location <- extract user_location from user
   geocode_result <- geocode user_location by sending a request to the geonames server

      if geocode_result contains at least one location then
         first_location <- extract first location from geocode_result

         if first_location is 'netherlands' then
            istourist <- 'true'
         else
            istourist <- 'false'
         end if

         write first_location and istourist to record user_id in table user_ams
         print 'location user_location of user user_id is geocoded as first_location'

      else
         print 'location user_location of user user_id cannot be geocoded'
      end if

end for
```

Code Snippet 7: Pseudo code of the method for geocoding user locations

3.4.3 Manual correction and selection of tourist photos

The automatic classification methods that make use of SQL regular expressions and the GeoNames database were unable to correctly classify the location of 144 users due to spelling errors and unusual location names. We have manually interpreted the locations of these users in step C3 and added the correct classification to the new database table *user_correction*. The table *user_ams* is updated with the correct values in step C4. After classifying the country of residence of all users with a user location attribute, the photos of tourists in Amsterdam are selected and stored in the new table *photo_tourist_ams*.

3.5 Calculation of temporal distributions

PostgreSQL queries are used to aggregate photos in Amsterdam per temporal interval. Subsequently, the amount of unique users is counted per interval. Inspired by work of Girardin, Dal Fiore et al. (2007), we have extracted distributions for the granularities *days of the week*, *months of the year* and *days of the year*. In addition, the amount of photographers for the granularities *hours of the day* and *hours of the week* were analysed. Moreover, we have extracted the distributions of tourists as well as local residents to discover potential differences. Code Snippet 8 shows the SQL code to extract the temporal distribution of unique tourists per hour. The results are given in section 4.3.

```
  SELECT EXTRACT(HOUR FROM datetaken) AS hour, count(DISTINCT photo.userid) AS users
    FROM photo, user_ams
   WHERE municipality = 'Amsterdam'
     AND photo.userid = user_ams.userid
     AND user_ams.istourist = 'True'
GROUP BY hour
ORDER BY hour
```

Code Snippet 8: SQL code to extract the temporal distribution of unique tourists per hour

3.6 Calculation of spatial clusters

Our analysis of related work identified several methods and techniques that can be used to spatially cluster geosocial data. Two methods are especially suitable to organise a collection of geotagged photos into groups based on their spatial relationship: grid-based clustering and density-based clustering. The implementation of the grid-based method is described in section 3.6.1 and will be used to explore Amsterdam's tourist densities in Google Earth. Section 3.6.2 explains the implementation of density-based clustering. This method identifies the major hotspots in the city. A flowchart with the different steps to calculate spatial clusters of tourists is given in figure 3-6.

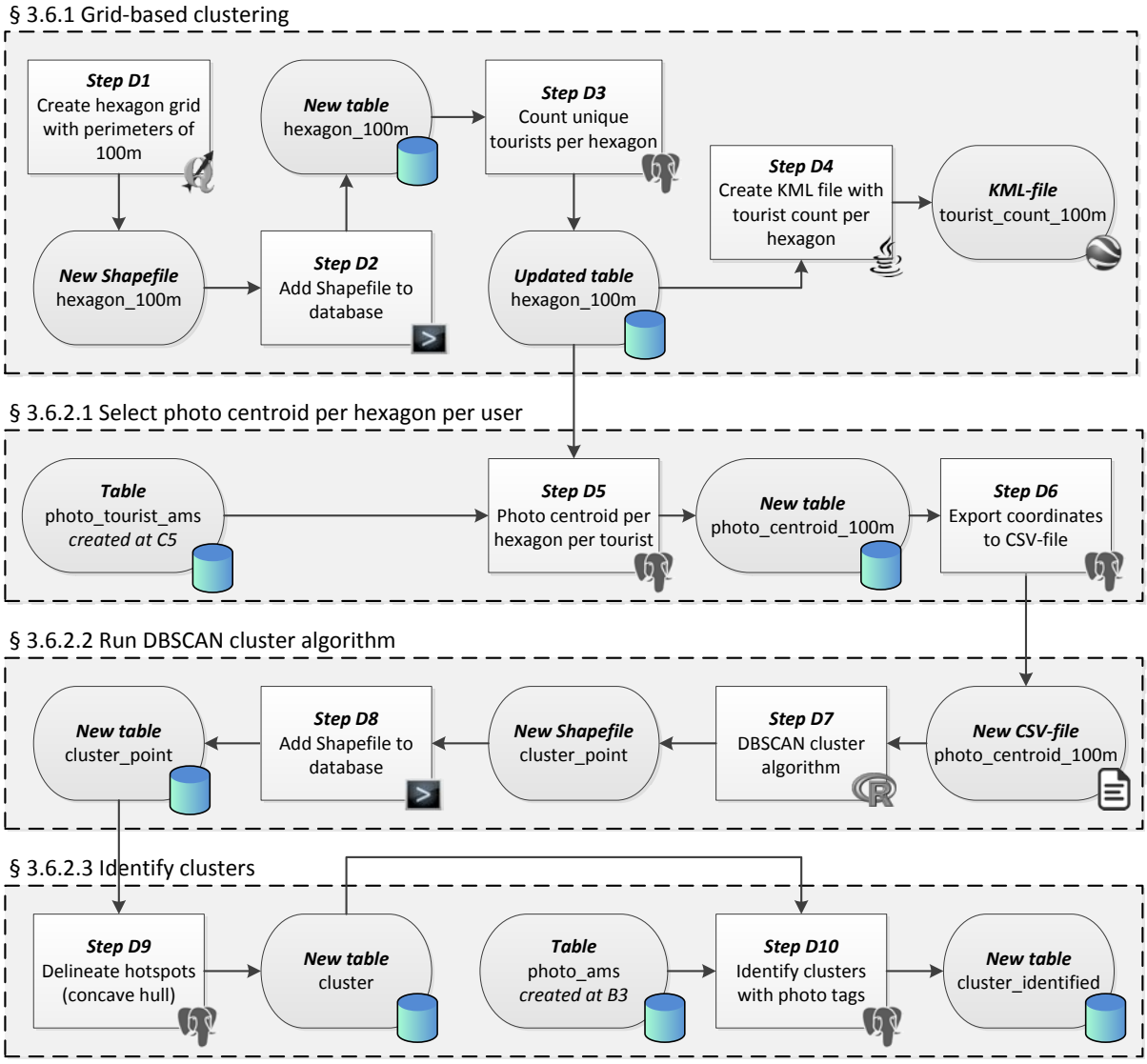


Figure 3-6: Flowchart to calculate spatial clusters of tourists (model D)

3.6.1 Grid-based clustering

Inspired by work of Sagl (2012) and Kádár and Gede (2013), we implemented a grid-based clustering method to explore densities of tourists in Google Earth. We have improved their method by making use of hexagonal grid cells and a colour classification that makes the results easier to interpret. The approach consists of three steps:

- Create a grid structure that covers the study area's spatial extent (step D1 and D2)
- Calculate the density of unique tourists per grid cell (step D3)
- Visualise the results (step D4)

The first step is the creation of a grid structure that covers the complete study area. The mentioned authors have used squared grid cells for their analysis. We have implemented a method that makes use of hexagonal shaped grid cells. Besides being visually attractive, hexagons have several advantages over squares. For example, Boots and Tiefelsdorf (2000) mentioned that hexagon tessellations represent the irregular topology of real world locations, on average, more accurately. A hexagon has more edges than a square, a shorter boundary length and smaller distances to the centre of the hexagon. Furthermore, hexagons are more suitable to visualize connectivity and movement paths because they have six neighbours at topological similar positions (Birch, Oom et al. 2007). This makes hexagons more suitable for the visualization of touristic corridors in a city.

Step D1 of the model uses the function *create grid layer* of QGIS' plugin MMQGIS to create a hexagon grid with cell perimeters of 100 meter (approximately 722 square meters). The width of one hexagon roughly matches the width of a large street in the centre of Amsterdam. Step D2 adds the shapefile containing the hexagon grid to the spatial database by using the PostGIS command *shp2pgsql*. Subsequently, the number of unique tourists is counted per hexagon and added to the table *hexagon_100m*. Code Snippet 9 shows the SQL query of step D3.

```
UPDATE hexagon_100m
   SET tourist_count = subquery.count
  FROM (
        SELECT grid.gid, count(DISTINCT userid)
          FROM hexagon_100m AS grid, photo_tourist_ams AS photo
         WHERE ST_CONTAINS(grid.geom, photo.geom)
      GROUP BY grid.gid
    ) AS subquery
 WHERE hexagon_100m.gid = subquery.gid;
```

Code Snippet 9: SQL code for counting the number of unique tourists per hexagon

We have developed a Java method that queries the geometry and tourist count of all hexagons from the database and transforms this to a KML file. Hexagons are visualised if at least 5 different tourists took a photo within its boundaries. The tourist count value is used to extrude the hexagons. Naturally, more tourists result in a greater shape height. Jenks (1967) natural breaks classification is implemented in our code to sort the shapes in classes. These classes are visualised with different colours that were obtained from ColorBrewer (2015). The pseudo code of this method is included in code snippet 10 and the results are given in section 4.4.1.

```
hexagons <- query all hexagons with tourist_count >= 5 from database table hexagon_100m
output_kml <- add kml code for the file header
output_kml <- add kml code for 5 different hexagon colors
jenks_breaks <- classify the tourist_count of hexagons into 5 classes

for every hexagon in hexagons do

   xy_coordinates <- extract xy_coordinates from hexagon
   tourist_count <- extract tourist_count from hexagon
   z_coordinate <- tourist_count * 3

   placemark_kml <- add kml code for the header of a polygon placemark
   placemark_kml <- add xy_coordinates and z_coordinate to placemark_kml
   placemark_kml <- add kml code for style based on tourist_count and jenks_breaks
   placemark_kml <- add kml code for footer of a polygon placemark

   output_kml <- add placemark_kml

end for

output_kml <- add kml code for the file footer
write output_kml to file and open the file with google earth
```

Code Snippet 10: Pseudo code of the method that exports photo densities to Google Earth

3.6.2 Density-based clustering

The major hotspots in Amsterdam have been detected with the DBSCAN clustering algorithm that is applied by Kisilevich, Krstajic et al. (2010), Sun, Fan et al. (2013) and Lee, Cai et al. (2014). We have improved the method by adding a pre-processing step that prevents that many photos of a single photographer could form a spatial cluster.

The process consists of pre-processing the data (section 3.6.2.1), running the cluster algorithm (section 3.6.2.2) and identifying the clusters with Flickr's photo tags (section 3.6.2.3). The different steps are illustrated in the flowchart in figure 3-6.

3.6.2.1 Select photo centroid per hexagon per user

The first step (D5) is the selection of one photo location per tourist per hexagon. If a tourist took more photos within the boundaries of a hexagon, the weighted centre point of all photo locations of this tourist within the hexagon is used instead of the photo location. This pre-processing step prevents that many photos of a single user can form a cluster. The photo centroids are selected using an SQL statement and copied in the new table *photo_centroid_100m*. The coordinates of all records in this new table are exported to a CSV-file in step D6.

3.6.2.2 Run DBSCAN cluster algorithm

As described before, the DBSCAN algorithm detects clusters with different shapes and sizes and is not sensitive to noise. This makes the method very suitable for detecting hotspots in geosocial point data. Two parameters are required: the radius around a point (*Eps*) and the minimum number of points within this radius (*MinPts*). Running the algorithm results in three types of points. Core points of clusters are points that have at least *MinPts* in their neighbourhood. Border points are all points that fall within the radius of a core point but do not have *MinPts* in their neighbourhood. All other

points are classified as noise. Detailed information about the algorithm can be found in the publication of Ester, Kriegel et al. (1996).

Hennig (2014) implemented the DBSCAN algorithm in programming language R and published package *fpc*. Step D7 is executed with an R-script that makes use of this package. The script first reads the CSV-file with photo centroids per hexagon, which was created in step D6. Then, the DBSCAN cluster algorithm is executed. Finally, the clustered points are exported to a shapefile. Step D8 uses the command line tool *shp2pgsql* to import the shapefile in the spatial database.

3.6.2.3 Identify clusters

Now that all points are classified as noise or as part of a cluster, we can delineate the clusters by making use of concave hulls. We have used concave hulls because they better represent real world cluster boundaries compared to the more commonly used convex hulls. This step (D9) is executed with PostGIS function *ST_ConcaveHull*. The result is stored in the new table *hotspot*.

The final step (D10) is the identification of the detected clusters by using photo tags. Flickr photographers use tags to describe the contents of their photos. Researchers Crandall, Backstrom et al. (2009) and Xin, Changhu et al. (2010) demonstrated that these tags often reveal the location of a photo. Our method first selects all photos from table *photo_ams* within a cluster boundary. Table *photo_ams* contains all photos in Amsterdam including the photos of users with an unclassified country of residence and users that live in the Netherlands. Their tags are just as valuable for the identification of an area. After selecting the photos, the tags are extracted from the photos. These tags are grouped per unique tag and the number of tags per group is counted. The method excludes all common tags like *square, street and tree.* The three tags that most often occur in a cluster are used to identify this cluster. Together with the cluster geometry, the tags and tag count are exported to the new table *hotspot_tagged.* The tags are used to label the clusters. The result of our density-based clustering method is included in section 4.4.2.

3.7 Calculation of tourist routes and density map

As stated in section 2.6 on page 8, Andrienko, Andrienko et al. (2011) defined subsequent photo capturing events of a user as a form of event-based-movement. This form of movement data consists of a set of objects where each object is positioned in space and time. In our case, the photographer is the moving object and events are formed by the space and time components of a photo. This section describes the method that we have developed to calculate the most probable route that a tourist takes between subsequent photo events. Figure 3-7 illustrates the concept of relating a tourist trajectory with Amsterdam's pedestrian routing network. This route is not necessarily the shortest path in terms of distance. Road popularity, based on the number of photos in its neighbourhood, is taken into account. The calculated routes of all tourists are combined to create a road density map of Amsterdam's city centre. This area includes all neighbourhoods encircled by the city centres ring road and the areas around the major museums and the *Vondelpark*. The city centre is highlighted with a dark blue transparent fill in the right map of figure 1-2 on page 4.

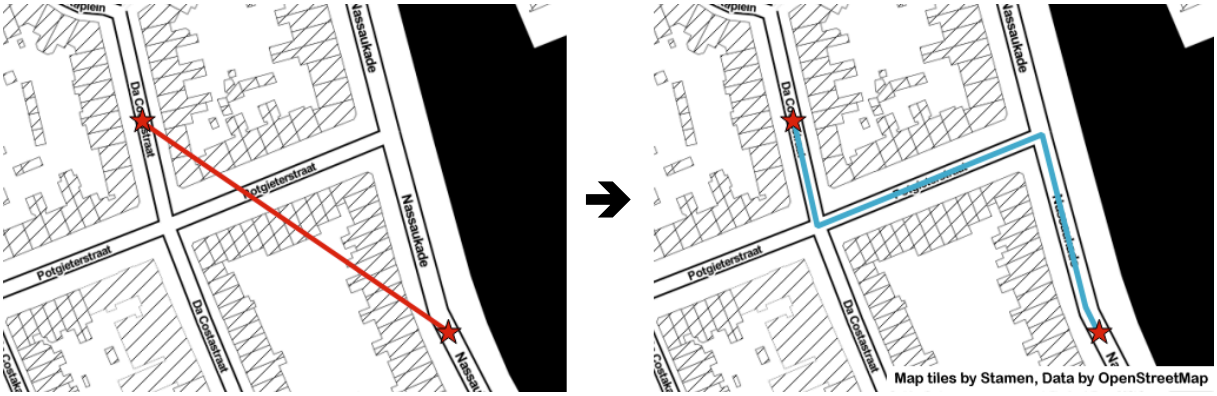


Figure 3-7: Relating tourist trajectories to Amsterdam's pedestrian routing network

Figure 3-8 visualizes a flowchart with the different steps that have been implemented to create a road density map of tourists in Amsterdam. The process starts with the creation of a pedestrian routing network. This is explained in section 3.7.1. Section 3.7.2 describes our method to calculate the travel cost per road segment based on the number of tourist photos in its neighbourhood. The network is prepared for routing purposes in section 3.7.3. Time-ordered photo pairs are selected and filtered based on speed, time and distance thresholds in section 3.7.4. Furthermore, the closest start and end node in the routing network is calculated for every photo pair. Section 3.7.5 calculates the most probable touristic route between the photos in a photo pair. Finally, all calculated routes are aggregated to create a road density map of tourists in Amsterdam.

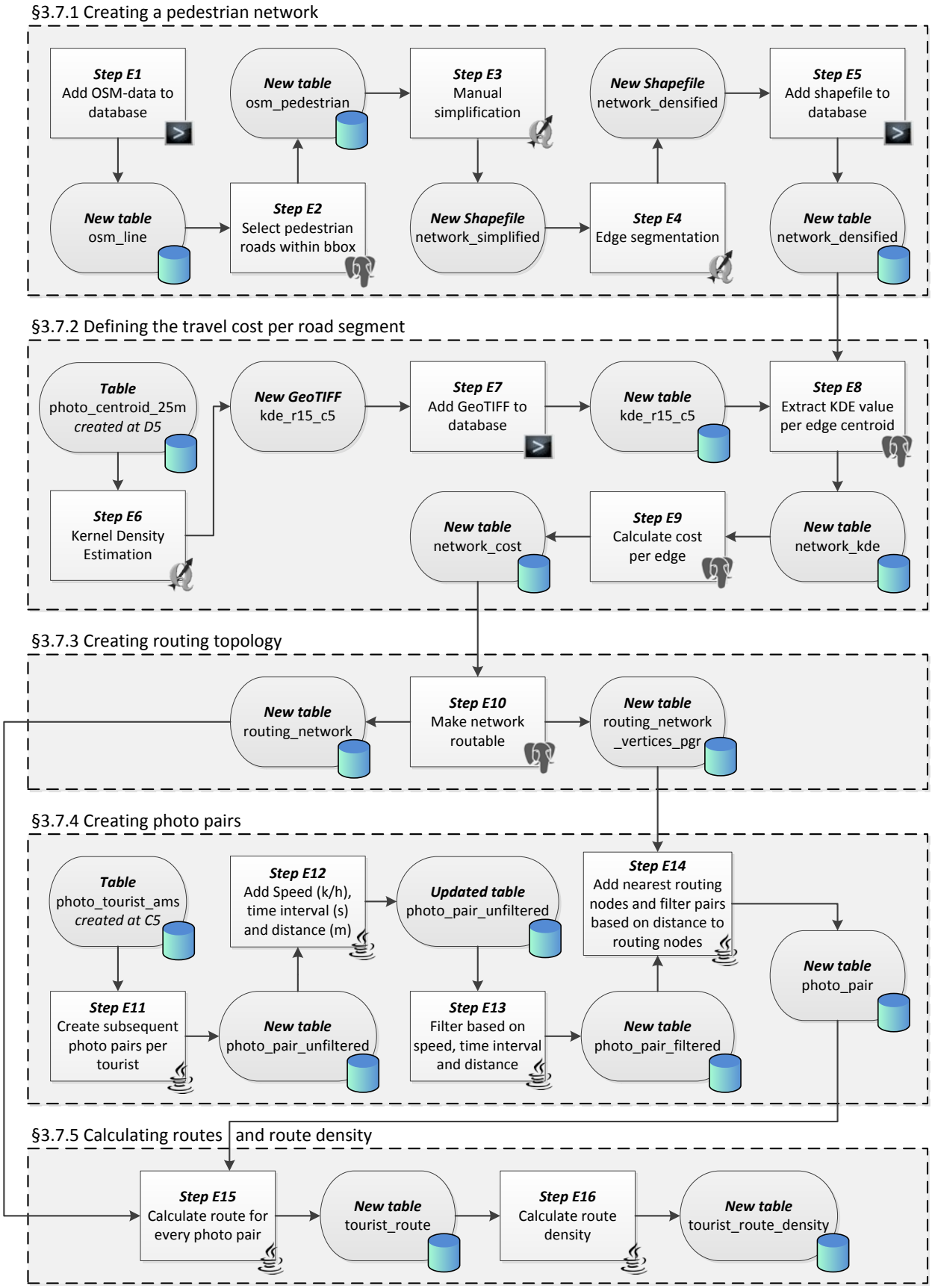


Figure 3-8: Flowchart of tourist route calculation (model E)

3.7.1 Creating a pedestrian network

The first steps of the process are designed to create a pedestrian network in the city centre of Amsterdam. Our network is based on OpenStreetMap data. Users are free to "copy, distribute, transmit and adapt" this data source (OSM 2015) and Haklay (2010) reported a good accuracy in comparison with official ordnance survey datasets. The data was downloaded from Geofabrik (2015) and added to the database in step E1 using the command line tool *osm2pgsql*. The OpenStreetMap project maps a wide variety of objects, including many different road classes. Step E2 is implemented to select all roads in our study area that are accessible by pedestrians. A polyline that represents a road is called an edge in routing terminology. Some streets are represented by a single edge, classified as a road for slow traffic. These edges are included in our selection because they are also accessible for pedestrians. The selection is stored in the new table *osm_pedestrian*.

The first image of figure 3-9 visualises a street junction with original network data from OpenStreetMap. It reveals that one street can consist of multiple edges. For example, two pedestrian links and two links for slow traffic. If we would use the network in its original state, a tourist route will make use of one of the four edges. Depending on the location of a photo, the route of another tourist could make use of a different edge. This complicates the execution of the final step of our model: the aggregation of all touristic routes into a road density map. Therefore, we manually simplified the routing network so that every pedestrian road is represented by a single edge (step E3). An example of a simplified road is visualised in the second image of figure 3-9.

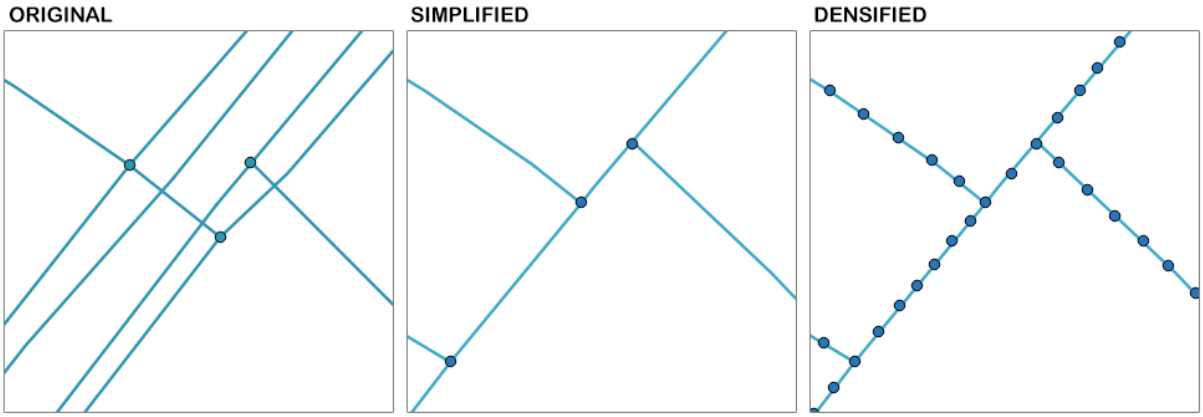


Figure 3-9: Steps to create the pedestrian network (E2, E3 and E4)

A routing algorithm creates a route between two nodes of a routing network. Nodes are located at the start and endpoint of an edge. Edges that touch each other share the same node at this specific location. Figure 3-9 visualises all network nodes with a blue dot. Our method calculates routes between two photo locations and these are usually not located on a network node. Therefore, the closest nodes need to be selected before the route can be calculated. Because many long streets consist of just one edge and two nodes, a node could be located very far from a photo location while the distance to the edge is very small. Ideally, a route would be calculated from the nearest location anywhere on an edge instead of the closest node. Unfortunately, the algorithm does not support this. To deal with this limitation, we introduced step E4 that segments every edge into segments with a length of maximum 5 meters. This step makes use of QGIS's functions *v.clean.snap, v.clean.break* and *v.split.length.* The result of this step is visualised in the third image of figure 3-9.

3.7.2 Defining the travel cost per road segment

We have selected Dijkstra's shortest path algorithm (Dijkstra 1959) to calculate routes between consecutive photo locations. The algorithm calculates the least cost path between a start and end node in a routing network. The total cost of the shortest path is equal to the sum of all edge costs that belong to this path. The cost of a single edge is often equal to its length but can be adjusted with parameters such as the average driving speed. Our method makes use of this possibility. As described before, tourists generally take attractive routes that offer sightseeing, dining and shopping opportunities. We expect, in agreement with Sun, Fan et al. (2013) and Kachkaev and Wood (2014), that the number of photos on or near a road is a good indication for its popularity. We have incorporated the road popularity with a method that adjusts the cost of an edge based on the number of tourist photos in its vicinity.

The first step (E6) is the creation of a photo density raster. To create this raster we make use of the commonly used kernel density algorithm (Parzen 1962). This function creates a smooth continuous surface based on the distribution and density of points. The photos in table *photo_centroid_100m* are used as input data for the algorithm. This table was created in step D5 of section 3.6.2.1 and contains a maximum of one location per tourist per hexagon. Using the points in this table prevents that a specific location becomes very popular if a single tourist took many photos in a small area. The KDE algorithm requires a kernel bandwidth (search radius) and an output cell size. We have selected a search radius of 15 meters so that the diameter of a search area roughly matches the width of a large street in the centre of Amsterdam. The cell size of the created density grid is 5 meters. This is equal to the road segment length. Once the photo density grid is created, we add it to the database by using the command line tool *raster2pgsql* in step E7.

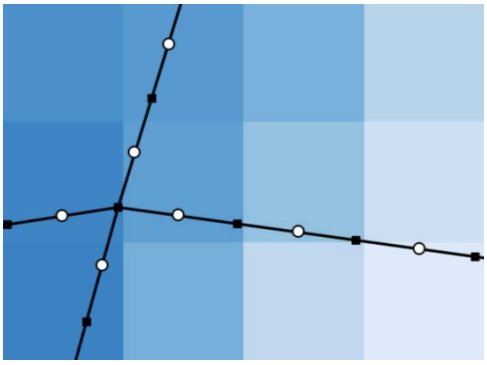


Figure 3-10: Extraction of density values (E8)

The next step (E8) is extraction of the photo density value per network edge. We execute this step using PostGIS' function *ST_Value*. This function returns the value of a raster at a given point. Our given point is the midpoint per network segment. These points are visualised with a white dot in figure 3-10. The blue cells represent the values of the density raster. The midpoints per edge are calculated using PostGIS' function *ST_Centroid*. All edges and density values are stored in the new table *network_kde*.

As described before, the amount of photos in the neighbourhood of a road is regarded as a good estimation for its popularity. It is more likely that a tourist travels over a road where a lot of photos were taken then over a road without photos. The calculated photo density is used to reduce the travel cost per edge. We have used a heuristic approach to define the amount of cost reduction. The cube root of the photo density gave the most promising results. Appendix D contains several sub results that were created while defining the edge reduction. The cost calculation per edge is given in the following equation:

$$Edge_{cost} = Edge_{length} \times \frac{1}{\sqrt[3]{Edge_{kde\ value} + 1}}$$

The cost per edge is calculated in step E9 and stored in the new table *network_cost.*

3.7.3 Creating routing topology

Now that the cost per edge is calculated, the network can be prepared for our routing algorithm. We have selected pgRouting (2015) to plan touristic routes between consecutive photo locations. This open source project adds routing functionality to PostgreSQL and PostGIS and includes a function to calculate shortest paths with Dijkstra's algorithm. Step E10 uses the pgRouting function *pgr_createTopology* to create the network topology that is required for the route calculation. The function assigns a source and a target ID to each edge. This ID corresponds with the ID of network nodes that are also created in this function. The edges with a column for source and target node are stored in table *routing_network* and the nodes are stored in table *routing_network_vertices_pgr.*

3.7.4 Creating photo pairs

The touristic routes will be calculated between time-ordered photo locations of tourists. Step E11 calculates pairs of subsequent photo locations. Step E12 calculates the speed (km/h), time interval (s) and distance (m) between the photos of every photo pair and stores the results in table *photo_pair_unfiltered.* Both steps are implemented in a small Java method of which the pseudo code is given in Code Snippet 11.

```
create table photo_pair_unfiltered in database
users <- query all tourists with more than 1 photo from database table photo_tourist_ams

for every user in users do

   photos_start <- query all photos from user, order by date taken, exclude last photo
   photos_end <- query all photos from user, order by date taken, exclude first photo
   photo_pairs <- join photos_start and photos_end

   for every photo_pair in photo_pairs do

      distance <- calculate distance between coordinates of start and end photo
      time_interval <- calculate time_interval between timestamp of start and end photo
      speed <- distance divided by time_interval

      write photo_pair with distance, time_interval and speed to table photo_pair_unfiltered

   end for

end for
```

Code Snippet 11: Pseudo code of the method to create pairs with subsequent photos of tourists

The travelled route of a tourist on foot can only be estimated with confidence if the start and end location of a pair is proximate in both space and time. If the spatial or temporal distance between two locations is high, it is very unlikely that a tourist actually took the route that has been calculated. We have introduced distance (m), time interval (s) and speed (km/h) thresholds to exclude all photo pairs that would produce an unreliable route calculation. The threshold values are given in table 3-2. The photo pairs that meet the distance, time interval and speed criteria are copied to the new table *photo_pair_filtered* in step E13.

Table 3-2: Threshold values for speed, distance and time interval between subsequent photos

	Greater than	Smaller than
Distance between start and end location of photo pair	50 m	750 m
Time interval between capture time of start and end photo		600 s
Speed of photographer between start and end photo	1 km/h	5 km/h

As described, Dijkstra's algorithm requires the ID's of two network nodes to calculate a least cost path. That means that the closest node needs to be calculated for both the start and end location of a photo pair. Step (E14), implemented as a Java method, selects the closest nodes and calculates the distances to these nodes. All photo pairs that have their start and end node at a distance smaller than 25 meters are added to the new table *photo_pair.* The distance threshold excludes all photos that are located too far from a road (*see* figure 4-11 on page 42).

3.7.5 Calculating routes and route density

The previous steps prepared the pedestrian routing network and selected all suitable photo pairs and their closest network nodes. Step E15 uses the created datasets as input for the calculation of the most probable touristic routes. Code Snippet 12 contains the pseudo code of the Java method that calculates a route for every photo pair. The edges that together form the least cost path of one photo pair are stored as separate records in table *tourist_route*. The edges of all least cost path calculations are added to this table.

```
create table tourist_route in database
photo_pairs <- query all photo_pairs from database table photo_pairs_filtered

for every photo_pair in photo_pairs do

   node_start <- extract node_start from photo_pair
   node_end <- extract node_end from photo_pair
   route_id <- extract route_id from photo_pair
   route <- calculate route between node_start and node_end with pgrouting

   write route_id and route to table tourist_route

end for
```

Code Snippet 12: Pseudo code of the method that calculates a route for every photo pair

The final step (E16) of our method uses the database functions *group* and *count* to aggregate all overlaying edges. A higher count naturally means that more tourists walked over this road. The results are stored in table *tourist_route_density* and used to create a road density map of tourists in Amsterdam.

3.8 Validation

Several methods have been developed to pre-process, transform and mine the large dataset with Flickr photos. But how reliable are the outcomes of our methods? Are the classified tourists really tourists? How well do the spatial densities of tourists resemble their use of the public space? And how valuable are the results for planners and policy-makers? Section 3.8.1 and 3.8.2 describe the methods that are used to validate the classification and temporal distributions of tourists. Because comparable quantitative data is not available, a qualitative approach has been used to validate the detected spatial clusters and route density map. We have interviewed eight tourism experts of the municipality of Amsterdam to see how their knowledge about the city resembles our findings. More information about the questionnaire is given in section 3.8.3.

3.8.1 Tourist classification

We have automatically classified the origin of photographers by using the location that they have specified in their user profile. To validate the outcomes of our classification method, 50 tourists and 50 locals were randomly picked from our set with classified users. The samples are stored in database table *random_tourist_50* and *random_local_50*. We have manually checked the classification of all users in our sample by investigating their Flickr user profile. A profile page contains albums, a photo stream, textual expressions and other indicators that easily disclose if the user is a tourist or not. The results are added to the tables. The automatic classification and manual classification results are reported in a confusion matrix and used to calculate the overall accuracy, precision and recall. We also analysed how different countries are represented on social media platform Flickr. Therefore, we compared the nationalities of Amsterdam's tourists in 2013 (O+S Amsterdam, 2014) with the origin of Flickr users that took photographs in this year.

3.8.2 Temporal distributions of tourists and locals

Section 3.5 describes how to extract temporal distributions of tourists and locals from the metadata of Flickr photos. We have compared the relative amount of tourists with relative amount of locals per temporal interval. Because there are no comparable statistics available, the temporal distributions are difficult to validate. The most suitable data that we obtained are the monthly number of hotel guests in Amsterdam in 2012 and 2013, provided by Statistics Netherlands (CBS 2015). These numbers are compared with the relative number of foreign photographers in the same time period.

During the exploration of Flickr photos, several incorrect timestamps were observed. This might be explained by incorrect camera times, for example caused by tourists from other time zones that did not change the time settings. We have assessed the timestamps from locals and tourists by inspecting photos that contain a real clock. An example of such a photo is given in Figure 3-11. Our initial selection contains all photos of tourists and locals tagged with the term "clock" and all photos taken near the central train station. Out of this subset, we have manually selected all images with clearly visible clocks. The time of the clock is compared to the time of the photo timestamp. Results are presented in a histogram and the standard deviation is calculated for the timestamps of tourists and the timestamps of locals.

Figure 3-11: No time to bike (Flickr 2013)

3.8.3 Spatial clusters and route density map of tourists

Amsterdam's spatial distribution of tourists is calculated by using grid-based and density-based clustering methods. We have exposed the major tourist hotspots in the city and photo tags are used to identify the places. Further, we calculated the most likely routes that tourist travelled between subsequent photo locations. All route calculations are combined to create a density map of touristic routes in the city centre of Amsterdam.

We aimed to validate the outcomes of the cluster and routing methods with official statistical data about the density of tourists in Amsterdam's city centre. Many departments of the municipality of Amsterdam were approached to acquire a suitable data source but there is unfortunately no comparable quantitative data available. Therefore, a qualitative approach was developed to validate the spatial clusters and route density map. The results were presented to 14 employees of the municipality of Amsterdam of which 8 are experts in the domain of tourism. The tourism experts were interviewed to see how their knowledge about the city resembles our findings. The questionnaire that is used is included in appendix G. The presentation is included on the DVD that is provided with this report.

To obtain unbiased results, the experts completed the first part of the questionnaire before we presented the study results. Participants were asked to provide personal background information and to rate their expertise about the city centre of Amsterdam. Next, maps of six different locations were presented. Every location contains two or three highlighted roads. All participants were asked to choose the most touristic road per location and rate the crowdedness of tourists at each highlighted road on a scale from 1 to 5. Finally, the experts gave a mark for their confidence on the given answers.

After the presentation, participants were asked how well the study outcomes resemble the real world situation on a scale from 1 to 5. The participants also rated the usefulness of the study outcomes for them and their department. Finally, the experts were asked to specify for what purposes they could use the study outcomes and if they had any other comments or suggestions. The results of the questionnaire are given in section 4.6.

4 Results and validation

We have developed and implemented several methods to derive spatial and temporal patterns from the metadata of Flickr photography. This chapter presents the project results and provides an answer to the third research question: *How well do the identified temporal distributions, spatial clusters and popular routes resemble the spatial and temporal behaviour of tourists?*

The first section of this chapter provides a description of the dataset with geotagged Flickr photos that were harvested at the start of this research project. Section 4.2 provides the outcomes of our method in which we classified users as "tourist" or "local". The origin of 50 classified tourists and 50 classified locals was assessed and reported in a confusion matrix. Section 4.3 presents the temporal patterns of photographers in our study area and the results of a comparison of photo timestamps with the true capturing times of photos. We have implemented grid-based and density-based clustering algorithms to explore Amsterdam's tourist densities in Google Earth and to identify the major hotspots in the city. The results are presented in section 4.4. Section 4.5 presents the results of the route calculations between subsequent photo locations and the route density map of tourists in Amsterdam. The last section of this chapter presents the results that were obtained by interviewing eight tourism experts of the municipality of Amsterdam.

4.1 Data collection

The metadata of 2,849,261 geotagged photos in the Netherlands was downloaded from Flickr in a time period of approximately 5 weeks. From this dataset, 786,943 records were excluded in the data-cleaning step. An overview of the causes and number of invalid records is given in table 4-1. All photo locations are enriched with the name of the municipality in which they were taken. Table 4-2 contains the top 5 most photographed municipalities and neighbourhoods in the Netherlands for the time period 2005-2015. All neighbourhoods in the top 5 are located in Amsterdam. After data cleaning, a total number of 2,062,318 photo records remained. 393,828 photos are located in the municipality of Amsterdam and uploaded by 15,992 unique users.

Table 4-1: Cause and number of invalid Flickr records

Cause	Number of records	Percentage of total
The accuracy is not 16 (street level)	774,008	27.2%
The taken granularity is not 0 (YYYY:MM:DD hh:mm:ss)	13,812	0.5%
The media is not photo but video	5,474	0.2%
The photo does not have additional metadata available	1,657	0.1%

Table 4-2: Top 5 most photographed municipalities and neighbourhoods in the Netherlands

#	Municipality	Users	Photos
1	Amsterdam	15,992	393,828
2	Rotterdam	4,660	179,797
3	The Hague	3,529	84,294
4	Utrecht	3,118	74,280
5	Haarlemmermeer	2,976	27,445

#	Neighbourhood	Users	Photos
1	Burgwallen-Nieuwe Zijde	6,449	50,257
2	Burgwallen-Oude Zijde	5,105	39,206
3	Nieuwmarkt/Lastage	3,617	25,168
4	Grachtengordel-West	3,337	18,926
5	Museumkwartier	3,302	23,396

4.2 Tourist classification

4.2.1 Results

Table 4-3 presents the results of our classification method. The origin of users was classified based on the home location that is added to a Flickr photo as metadata. Because we were only interested in the behaviour of tourists, 60,9% of all users and 72,8% of all photos were filtered out in this step. The data of a high percentage of users was discarded due to the absence of a user location.

Table 4-3: Results of tourist classification

	Number of users	Number of photos	Photos / user
Locals (17,6%)	2,821	154,599	54,8
Tourists (39,1%)	6,257	107,016	17,1
Unclassified users (43,2%)	6,914	132,213	19,1
Total	15,992	393,828	24,6

We have utilized the user's profile information to classify their country of origin. The countries in our dataset allow making a comparison with statistics. We have counted the number of photographers per foreign country that took a photo in Amsterdam in 2013 and divided these numbers by the total number of foreign photographers. These relative numbers are compared with tourism statistics provided by the statistics department of the municipality of Amsterdam (2014). The results are given in Figure 4-1. This comparison cannot be used for validation purposes because the popularity of social media platform Flickr differs between countries. It is meant to explore how Amsterdam's tourists are represented on social media platform Flickr. The chart shows that especially the United States is overrepresented in the Flickr dataset. A list with the countries of residence of Flickr users in Amsterdam in the time period 2005-2015 is given in appendix B.

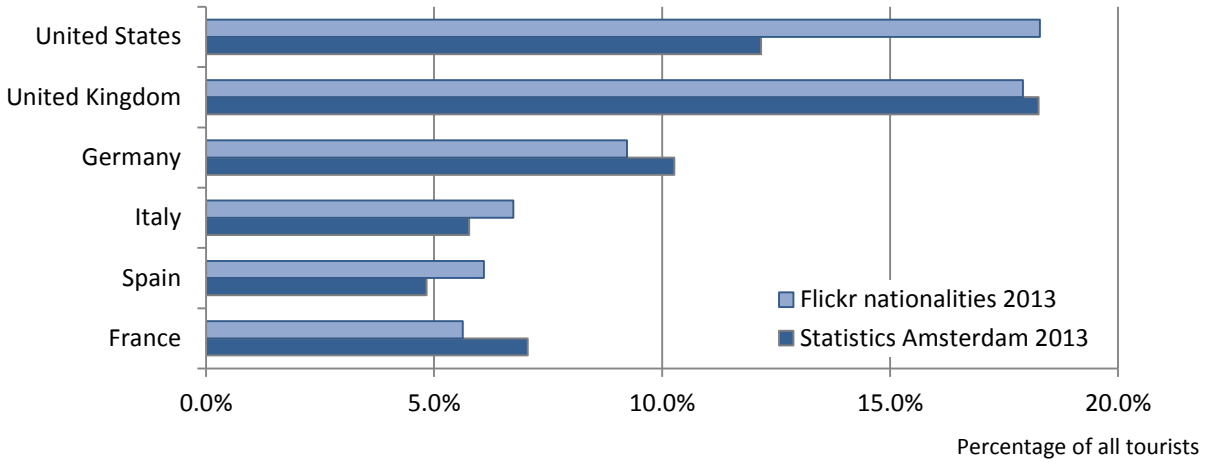


Figure 4-1: Comparison of photographer's country of residence with official statistics (2013)

4.2.2 Validation

We have automatically classified the origin of photographers by using the location that they have specified in their user profile. To validate the outcomes of our classification method, 50 tourists and 50 locals were randomly picked from our set with classified users. We have manually checked the classification of all users in our sample by investigating their Flickr user profile. A profile page contains albums, a photo stream, textual expressions and other indicators that easily disclose if the user is a tourist or not. Table 4-4 reports the classification results in a confusion matrix. The table also provides the precision and recall of tourists and locals. Out of our sample with 50 tourists and 50 locals, just one user was incorrectly classified as tourist. This user is a professional photographer from Belgium who visited the Netherlands to take photos at an event. All other users were real tourists or locals. The subset with classified tourists was used to conduct our data analysis, which makes the class precision of tourists (98%) the most relevant value for our research.

Table 4-4: Confusion matrix of tourist classification based on location in user profile

		True identity		**Classified users**	**Class precision**
		Tourist	**Local**		
Classified as:	**Tourist**	49	1	50	98%
	Local	0	50	50	100%
True identity of users		49	51	100	
Class recall		100%	98%		

Section 3.4 of the methodology explained that several researchers used the spatial and temporal traces of photographers to classify them as local or tourist. A photographer was for example classified as tourist if all his photos in the study area were taken within a temporal interval of 30 days. The method classifies all other users as locals. We have tested the temporal interval method on our datasets with randomly picked tourists and locals. Instead of only using the photos in Amsterdam, we used all 2,849,261 photos in the Netherlands to make sure that all Dutch Flickr users would be classified as locals. The results are given in Table 4-5. Comparing both classification methods revealed that the method based on the location in a user's profile outperforms the method based on a temporal interval of 30 days.

Table 4-5: Confusion matrix of tourist classification based on temporal interval

		True identity		**Classified users**	**Class precision**
		Tourist	**Local**		
Classified as:	**Tourist**	41	9	50	82%
	Local	8	42	50	84%
True identity of users		49	51	100	
Class recall		83,7%	82,4%		

4.3 Temporal distributions

4.3.1 Results

The first step after the tourist classification was the extraction of temporal distributions from the timestamps of Flickr photos. Figure 4-2 compares the relative amount of tourists and locals for three different temporal granularities: *hours of the day*, *days of the week* and *months of the year*. The relative amounts are based on the number of unique tourists and local photographers in Amsterdam per temporal interval. So, if one user takes many photos within one interval, it still counts as one.

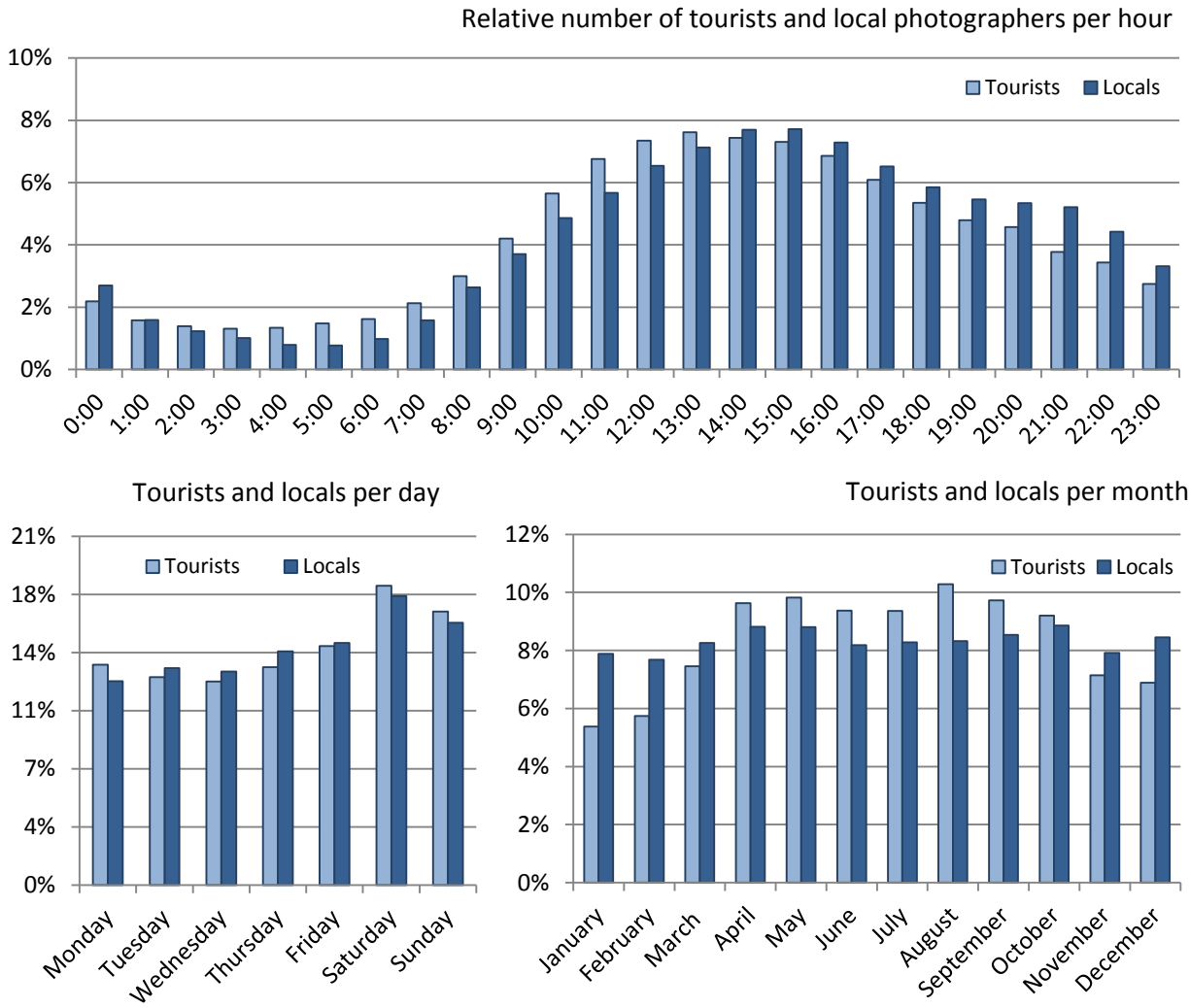


Figure 4-2: Relative number of tourists and locals per hour, day and month

All charts are based on the timestamps of tourist photos that were automatically assigned to the photos at the moment of capturing. The chart with the hourly tourist distribution reveals that most images of tourists and locals are taken during daylight. However, the chart also shows that a relatively high amount of tourist photos are captured during the night. Furthermore, most tourists took photos around 13:00 while most locals took photos around 14:00-15:00. The temporal profile of tourists is approximately one and a half hours shifted compared to the temporal profile of local photographers. This might be explained by incorrect camera times, for example caused by tourists from other time zones that did not change the time settings. We have validated the capture time at the end of this paragraph.

The second chart in figure 4-2 shows the relative amount of tourists and locals per day of the week. It reveals that Saturday is Amsterdam's busiest day in the week. This is confirmed by a report of the municipality that analysed the bustle in the city (Berge and Jakobs 2013). The third chart compares the relative number of tourists and local photographers per month. It shows that the monthly distribution of tourists and locals is very different. The majority of Flickr tourists took their photos from April to October while locals captured photos all year round.

Figure 4-3 compares the relative number of unique tourists and relative number photos per hour. We have extracted those distributions to analyse the influence of daylight on the number of photos. The chart shows that there were relatively more images captured per tourists during daytime and fewer during nighttime. Photographers took, on average, more photos per hour during the day.

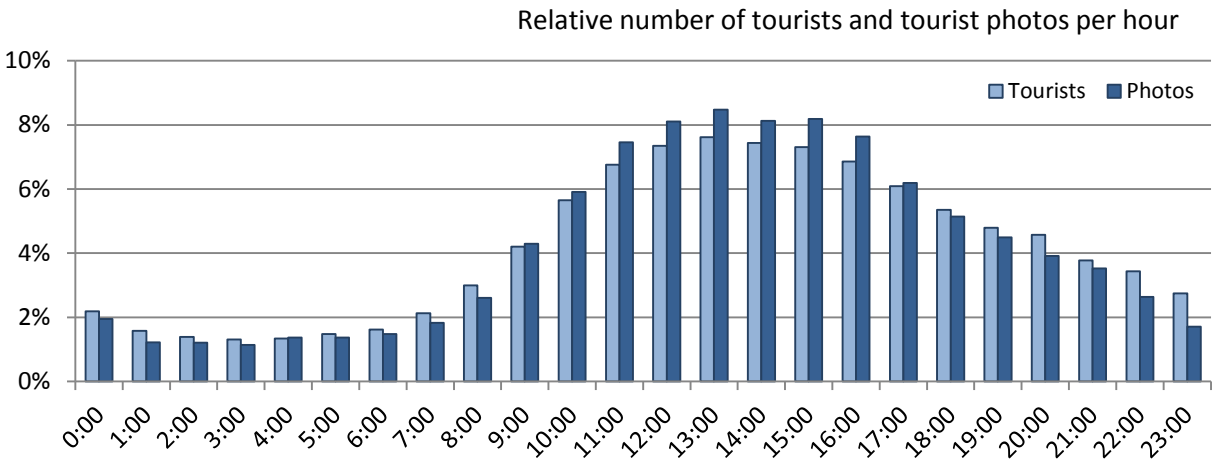


Figure 4-3: Relative number of tourists and tourist photos per hour

We have compared the relative number of photographers per month in 2012 and 2013 with the relative number of foreign guest in Amsterdam's hotels according to Statistics Netherlands (CBS 2015). The results are presented in Figure 4-4. A noticeable difference between the number of Flickr tourists and the official number of tourists is visible in the months April and September. The Flickr tourists are in the majority in these months. The tourists in Amsterdam's hotels outnumber the Flickr tourists from November until February.

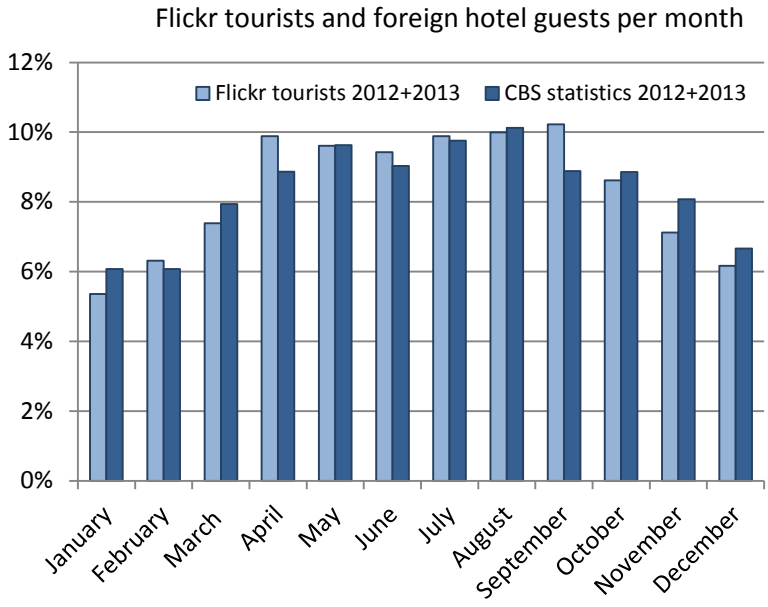


Figure 4-4: Relative number of Flickr tourists per month compared to hotel guests

Finally, we extracted the unique number of Dutch and foreign photographers in Amsterdam per *day of the year*. The results are presented in Figure 4-5. The chart shows a distinctive peak at Queen's Day. New Year's Day and the Gay Pride are also popular events for photography. The data reveals that Dutch citizens take fewer photos during the national festivity of Sinterklaas and Christmas Day.

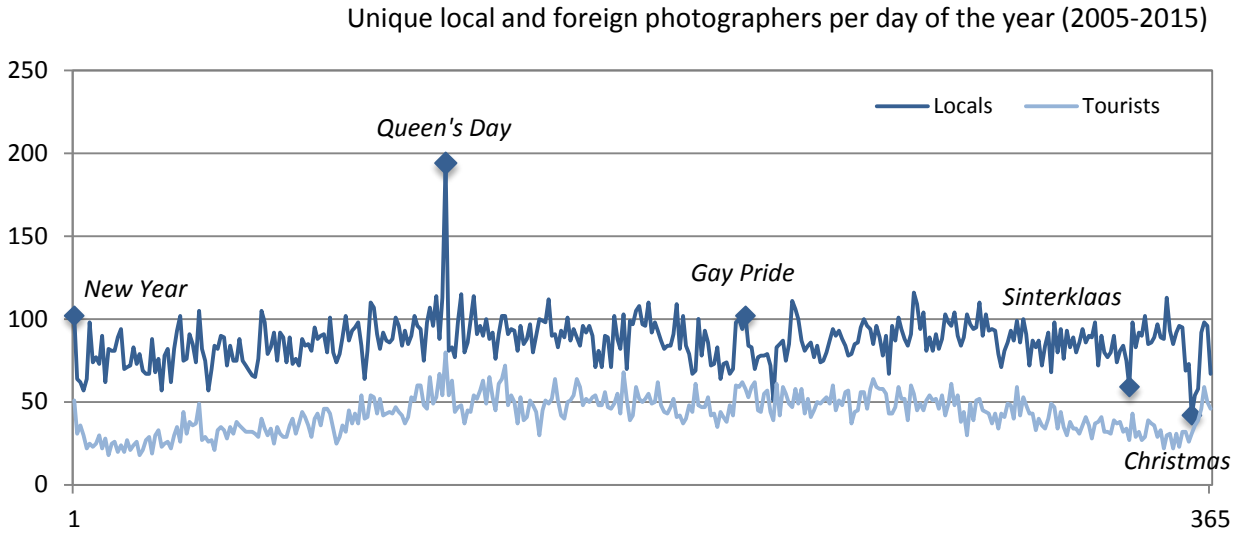


Figure 4-5: Temporal distribution of unique local and foreign photographers per day of the year

4.3.2 Validation

Exploring photos in Google Earth revealed several photos with incorrect timestamps. Moreover, a peculiar deviation between the relative amount of tourists and locals per hour was observed. This might be the result of people with incorrect camera time settings. We have assessed the timestamps of local and tourist photos by inspecting photos that contain a real clock. All photos of tourists and locals tagged with the term “clock” and all photos taken near the central train station were selected. This resulted in 1134 tourist photos and 1032 local photos. Out of this selection, 70 tourist photos and 50 local photos contain a clock that was suitable to read the time from. Each photo was captured by a different user. The photos are stored in database table *time_validation*, together with the real capture time that is derived from the photographed clock in the image.

Figure 4-6 shows a histogram with the time differences between photo timestamps and real photo times. Time differences are grouped per temporal interval of one hour. The chart shows that the time differences of locals are normally distributed. Approximately 72% of the local photos have a time difference of less than 30 minutes and most other photos have a time difference of plus or minus one hour. The difference between the mean capture time and mean photo time is less than a minute. The standard deviation of the local photos approximately 40 minutes.

The sample with tourist photos shows a different distribution of time differences. Just 52% of the photos have a time difference less than 30 minutes. Several photos, contributed by tourists from the United States, Canada and Brazil, have a time difference between minus 10 and minus 4 hours. The time difference roughly corresponds with the time zone difference. The photos with a time difference of plus 6 and plus 7 hours were contributed by tourists from Taiwan and Australia. The mean time difference is minus 1 hour and 22 minutes. The standard deviation of tourist photos is approximately 3 hours and 22 minutes.

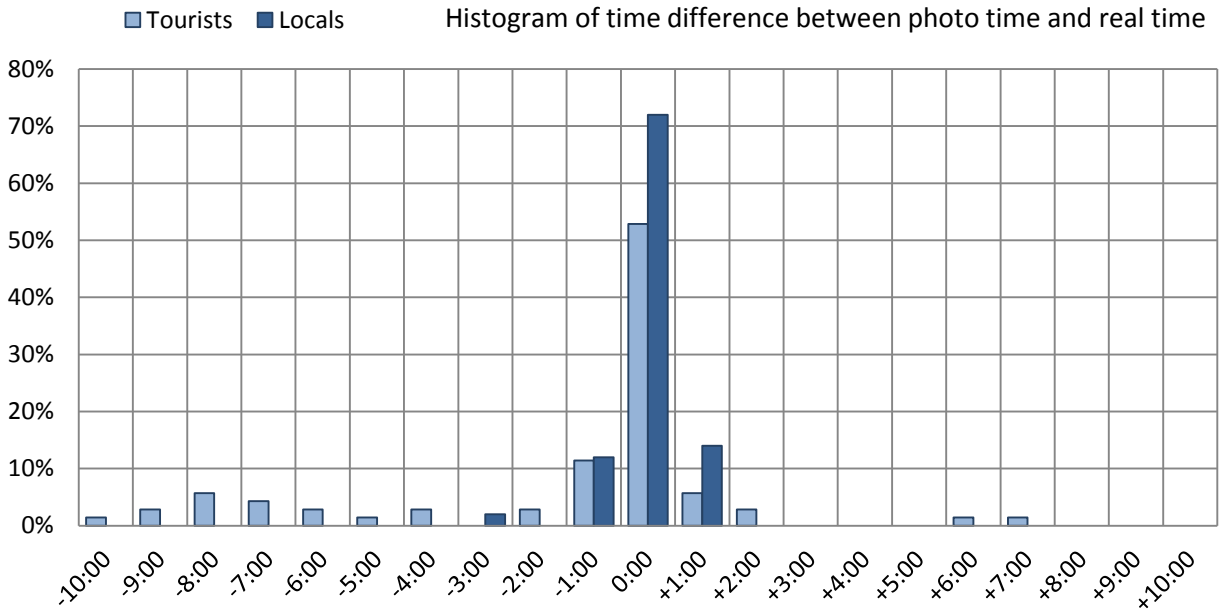


Figure 4-6: Histogram of time difference between photo time and real time

4.4 Spatial clusters

This section presents the results of the cluster algorithms that were implemented to detect the tourist distribution and major hotspots in Amsterdam. The results were presented to tourism experts of the municipality of Amsterdam. Their expert judgement of the results is reported in section 4.6.

4.4.1 Grid-based clusters

We have implemented a grid-based clustering method that derives tourist densities from geosocial data and automatically creates an easy to explore geovisualisation in Google Earth. The method makes use of a grid with hexagonal shaped cells where each cell has a perimeter of 100 meters. The height of a hexagon is based on the number of photographers multiplied by 3. We have divided the results in 5 classes by using Jenks natural breaks classification method. Each class was assigned a colour between red and yellow where red indicates a high amount of tourists and yellow a relatively low amount of tourists. Figure 4-7 visualises a bird's eye perspective of the results in our study area.

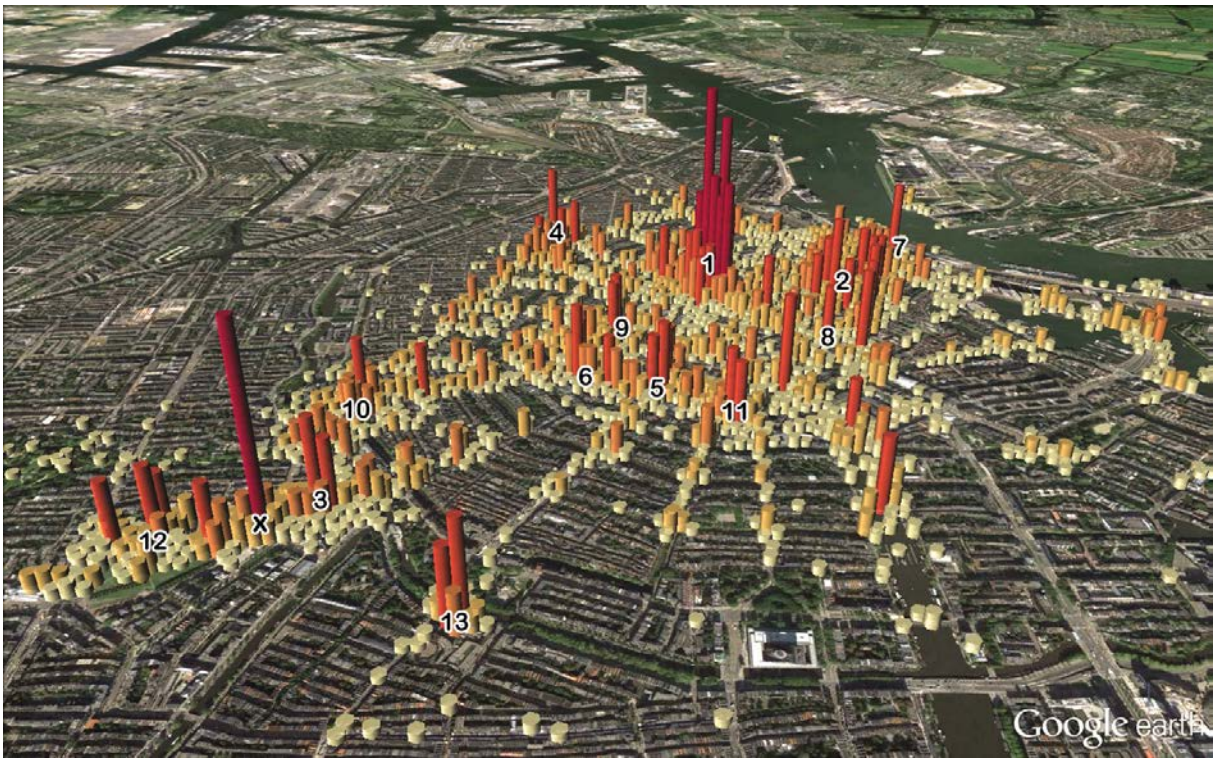


Figure 4-7: Tourist densities in Google Earth

The figure contains numbers for a selection of popular locations in Amsterdam. These locations correspond to the detected and identified hotspots that are reported in Table 4-6 of the next section. The visualisation identifies *Dam Square* (1) as the area with the highest tourists density. Between 2005 and 2015, 224 unique tourists took a photo on the location of the tallest hexagon. Another popular tourist area consists of the *Red Light District* (2), the nearby *Nieuwmarkt* (8) and Amsterdam's *Central Station* (7). Many tourists find their way to the *Rijksmuseum* (3), *Stedelijk Museum* and *Van Gogh Museum* (12) on the Museum Square. The red hexagon near the *Rijksmuseum* reveals the location of the iconic *iamsterdam* statue (x) that is illustrated in Figure 3-3 on page 14. The itinerary of many tourists also includes the *Anne Frank House* (4), *Begijnhof* (9) and *Flower Market* (6). Some other well-visited places are *Leidse Square* (10), *Rembrandtplein* (11) and *Heineken Experience* (13).

4.4.2 Density-based clusters

We have implemented the DBSCAN cluster algorithm to detect density-based clusters from a large collection of geosocial data. The algorithm can detect clusters with different shapes and sizes and is not sensitive to noise. We have used the hexagon grid that is created for the grid-based clustering method to pre-process the data prior to running the cluster algorithm. The purpose of this step was to select a maximum of one photo per tourist per hexagon and to prevent that many photos of a single user can form a cluster. Out of a total of 107,016 tourist photos, 44,834 photo centroids were formed and used for the cluster calculation.

The DBSCAN algorithm requires two parameters: the radius around a point (*Eps*) and the minimum number of points within this radius (*MinPts*). There is no fixed approach to define the parameter values *Eps* and *MinPts*. Researchers use heuristic methods to define the parameter values. Naturally, the values depend on the study area, the input data and the domain knowledge of the researcher. We have tested different combinations of *Eps* and *MinPts* for this research project. The outcomes of this exploration are illustrated in appendix C. After evaluating the results, 50 meter has been selected as search radius (*Eps*) and 250 as minimum amount of tourists (*MinPts*).

The detected clusters are highlighted with a red colour in Figure 4-8 and Figure 4-12. All light blue points are classified as noise. Concave hulls were used for the delineation of the clusters. Transforming points into a polygon made it possible to select all photos of tourists, locals and unclassified users within its spatial extent. The photo tags within these selections were mined to identify every cluster.

Figure 4-8: Map of detected and identified hotspots in Amsterdam

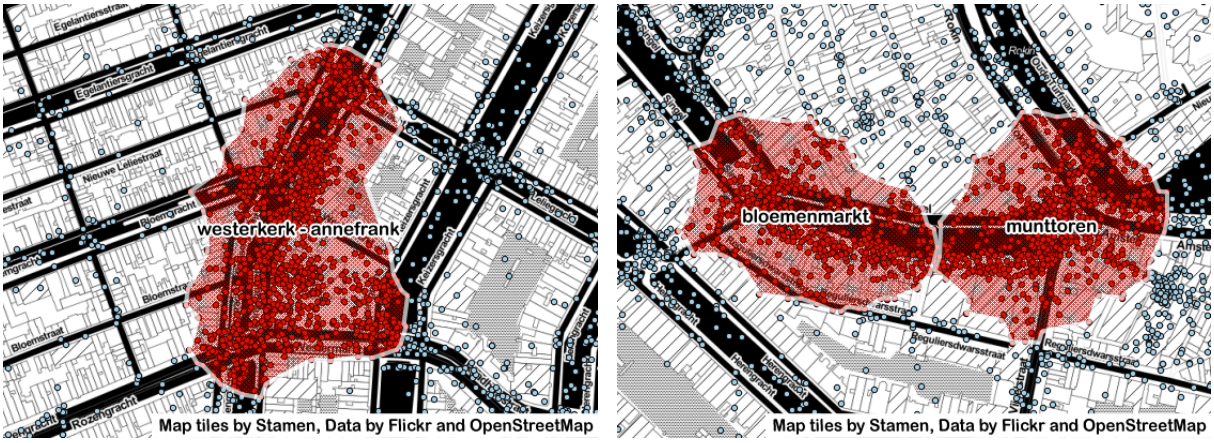


Figure 4-9: Zoomed in views of hotspot map

Table 4-6 contains all the clusters that were detected in the centre of Amsterdam. The hotspots are ordered based on the unique number of tourists that took a photo at this location. The first and second tag of every cluster is given in the table. Some clusters contain more attractions because they are located close to each other in a very dense photo area. An example is cluster four, illustrated in the left image of Figure 4-9. This area contains the Westerkerk as well as the Anne Frank House. Cluster five and six on the other hand, both contain a section of the Flower Market. A few more photos in the border area between the two clusters would have resulted in one single cluster. This location is illustrated in the right image of Figure 4-9.

We have compared the discovered hotspots with the *top picks* of popular travel website Lonely Planet (2015). The *top picks* are 17 places in Amsterdam that Lonely Planet recommends tourists to visit during their stay in the city. The selection of places is included in appendix E. Only the Central Station, Nieuwmarkt, Leidse Square and Rembrandt Square are not included in their list with top attractions in Amsterdam.

Table 4-6: Detected and identified hotspots in Amsterdam

#	Location	First tag	Second tag	Tourists	Lonely Planet
1	Dam Square	*dam*	*damsquare*	1605	Yes
2	Red Light District & Oude Kerk	*redlightdistrict*	*oudekerk*	921	Yes
3	Rijksmuseum & Museum Square	*rijksmuseum*	*museumplein*	718	Yes
4	Westerkerk & Anne Frank House	*westerkerk*	*annefrank*	662	Yes
5	Munt Tower & Flower Market	*munttoren*	*market*	473	Yes
6	Flower Market	*bloemenmarkt*	*flowermarket*	431	Yes
7	Central Station	*station*	*train*	416	No
8	Nieuwmarkt	*nieuwmarkt*	*waag*	396	No
9	Begijnhof	*begijnhof*	*spui*	380	Yes
10	Leidse Square	*leidseplein*	*stadsschouwburg*	362	No
11	Rembrandt Square	*rembrandtplein*	*rembrandt*	353	No
12	Van Gogh Museum	*vangogh*	*vangoghmuseum*	295	Yes
13	Heineken Experience	*heineken*	*beer*	239	Yes

4.5 Tourist routes and density map

This section describes the results of our method that calculated the most probable routes that tourists took between subsequent photo events. The calculated routes of all tourists were aggregated to create a road density map of Amsterdam's city centre. The results were presented to tourism experts of the municipality of Amsterdam. Their expert judgement is reported in section 4.6.

As described in the methodology chapter, calculated routes are not necessarily the shortest paths in terms of distance. Road popularity, based on the number of photos in its vicinity, is taken into account. Figure 4-10 shows the results of the shortest and most touristic path calculation between a start and endpoint. The black lines represent the pedestrian routing network that was created for our study area. The white and blue grid visualises the photo density map that is used to calculate the cost reduction per edge. More information about this method is given in section 3.7 of the methodology. The touristic path passes along the well-known flower market that is located at the south side of the *Singel*, a canal in Amsterdam. The shortest path passes the north side of the *Singel*. This side is relatively quiet and visited by much less tourists.

Figure 4-10: Shortest and touristic path on photo density map

Several pre-processing steps were executed prior to the calculation of the most probable touristic routes between subsequent photo locations. The first step was the creation of time-ordered photo pairs. This step processed 107,016 different photos of 6,257 tourists into 100,759 photo pairs. A travelled route of a tourist can only be estimated with confidence if the start and end location of a pair is proximate in both space and time. If the spatial or temporal distance between two locations is high, it is very unlikely that a tourist actually took the route that has been calculated. After applying a distance (m), time interval (s) and speed (km/h) filter, 8,103 photo pairs remained. Table 4-7 contains the numbers of photo pairs that were rejected per threshold. Many photo pairs were rejected by more than one threshold.

Table 4-7: Number of rejected photo pairs per threshold

Threshold	Number of Photo pairs
Time >= 600s	31,542 (*29.5%)*
Distance <= 50m	62,772 (*58.7%*)
Distance >= 750m	12,790 (*12.0%*)
Speed <= 1km/h	73,744 (*68.9%*)
Speed >= 5km/h	9,039 (*8.4%*)

After creating a filtered selection of time-ordered photo pairs, the closest start and end node in the routing network was calculated for every pair. A selection of photos and their closest routing node is given in Figure 4-11. Only photo pairs that have their start and end node at a distance smaller than 25 meters were used as input for Dijkstra's routing algorithm. The distance threshold excludes all photos that are located too far from a road. Photos and nodes that meet this condition are highlighted with a blue colour. Photos that were excluded in this step are visualised in red. From our initial selection of 8,103 photo pairs, 6,477 photo pairs have a start and end node at a distance less than 25 meters. Most pairs that were excluded in this step are located in water bodies, building blocks or outside the city centre of Amsterdam.

Figure 4-11: Connection of photo locations with closest node

The start and end node of the 6,477 selected photo pairs were used to calculate the most probable touristic routes with Dijkstra's routing algorithm. The algorithm outputs every road segment of the calculated path as separate polyline. The calculation resulted in 317,479 road segments. Overlaying edges were aggregated and counted to create a density map. A higher count naturally means that more tourists walked over this road. The results are visualised in Figure 4-11. The colour and width of roads represent the tourist density. The red polygons highlight the detected clusters.

Figure 4-12: Map of pedestrian tourist densities in Amsterdam

4.6 Expert judgement of spatial clusters and route density map

We used a qualitative approach to validate our spatial clusters and route density map by interviewing eight tourism experts of the municipality of Amsterdam. This section reports the answers of the experts and compares their knowledge about the city with the route density map. Table 4-8 contains general background information about the participants of the questionnaire. Each expert specified his or her knowledge about the city of Amsterdam on a scale from 1 to 5.

Table 4-8: Participants of the questionnaire

#	Gender	Age	City of residence	Neighbourhood of residence	Profession	*
1	M	53	Amsterdam	Centre	Policy Advisor Traffic & Public Space	5
2	F	26	Amsterdam	West	Data Analyst, Information en Statistics	5
3	M	43	Amsterdam	North	Senior Advisor Traffic Management	5
4	F	38	Amsterdam	Centre	Researcher, Information en Statistics	5
5	M	52	Baarn	-	Senior Advisor Traffic Research	3
6	M	57	Amsterdam	Centre	Urban Planner	5
7	M	59	Amsterdam	West	Urban Planner	5
8	M	48	Hilversum	Centre	Urban Designer	4

** How well do you know the city of Amsterdam on a scale from 1 to 5?*

As described in section 3.8.3, maps of six different locations were presented to the tourism experts. Each location contained two or three highlighted roads and participants were asked to choose the most touristic road per location. The maps that were presented are included with the questionnaire in appendix G. The same locations are visualised in the six figures on the next page and the calculated touristic route density is added to every map. Obviously, the routing results were not shown when the experts answered the questionnaire. The expert judgements of the most touristic road selection per location are given in Table 4-9. Apart from selecting the most touristic road per location, the experts rated the relative amount of tourists at each road on a scale between 1 and 5. The results of this question are given in Table 4-10. The last columns of both tables contain ratings between 1 and 5 that specifies the expert's level of confidence of their given answers.

In five out of six locations, the majority of the participants selected the road with the highest calculated tourist density. The only exception is the second location where five out of 8 participants selected *Paleisstraat* over *Mozes en Aäronstraat.* The experts are fairly sure about the first and third location, both selected by 75% of our test panel. The road with the highest calculated tourist density on the first location is *Damrak*, a touristic street that connects Amsterdam's central station with the most touristic area of the city, *Dam Square.* Based on calculations and expert judgement, *Kalverstraat* in location three is slightly more touristic than *Rokin*. Both the experts (*see* table 4-10) and the calculation results show a very small difference in tourist density between the locations.

There are two locations with a 100% match. The most touristic street in location four is *Oude Hoogstraat*. This street extents *Damstraat* and houses a lot of tourist related businesses. The south side of *Singel* is the most touristic road in location five. This street is home to Amsterdam's touristic *Flower Market*. The participant's confidence about these locations is confirmed by their judgement of the relative amount of tourists per road in table 4-10. Experts were able to choose between three roads at location 6. Five out of eight experts selected *Museumstraat*, the road that was calculated as the most touristic one.

Table 4-9: Answers questionnaire: most touristic road per location

Expert #	Location 1-1	Location 1-2	Location 2-1	Location 2-2	Location 3-1	Location 3-2	Location 4-1	Location 4-2	Location 5-1	Location 5-2	Location 6-1	Location 6-2	Location 6-3	Match [%]	*
1		●	●		●		●			●		●		*100%*	5
2		●		●		●	●			●		●		*67%*	3
3		●	●		●		●			●		●		*100%*	5
4		●		●	●		●			●			●	*67%*	4
5	●		●		●		-	-		●		●		*80%*	4
6		●		●		●	●			●		●		*67%*	5
7		●		●	●		●			●	●			*67%*	4
8	●			●	●		●			●	●			*50%*	4
Match [%]	*75%*		*38%*		*75%*		*100%*		*100%*		*63%*				

** How confident are you with your answers on a scale from 1 to 5?*

Table 4-10: Answers questionnaire: relative amount of tourists per road

Expert #	Location 1-1	Location 1-2	Location 2-1	Location 2-2	Location 3-1	Location 3-2	Location 4-1	Location 4-2	Location 5-1	Location 5-2	Location 6-1	Location 6-2	Location 6-3	*
1	4	5	5	4	4	4	5	2	3	5	3	5	3	5
2	4	5	3	4	4	5	5	3	3	5	4	5	3	3
3	4	5	5	4	4	4	4	3	3	5	3	5	2	5
4	5	5	4	5	5	5	3	2	4	5	3	4	4	4
5	5	4	4	3	5	3	-	-	3	4	4	5	4	4
6	3	5	2	5	3	5	4	3	3	5	3	5	4	5
7	4	5	3	4	5	4	4	3	2	5	4	3	2	4
8	5	5	3	4	5	4	5	2	4	4	3	3	2	4
Average	*4,3*	*4,9*	*3,6*	*4,1*	*4,4*	*4,3*	*4,3*	*2,6*	*3,1*	*4,8*	*3,4*	*4,4*	*3,0*	

** How confident are you with your answers on a scale from 1 to 5?*

After the presentation, participants were asked how well the presented spatial clusters and route densities of tourists resemble the real world situation. Each expert gave his or her judgement on a scale from 1 to 5. The participants were also asked to assess the usefulness of the study outcomes for them and their department. The results of both questions are given in Table 4-11.

Table 4-11: Validity and usefulness of presented spatial clusters and route densities of tourists

Expert #	Profession	Validity results [1-5] *	Usefulness results [1-5] **
1	Policy Advisor Traffic & Public Space	4	5
2	Data Analyst, Information en Statistics	4	4
3	Senior Advisor Traffic Management	4	4
4	Researcher, Information en Statistics	3	4
5	Senior Advisor Traffic Research	5	4
6	Urban Planner	5	5
7	Urban Planner	4	5
8	Urban Designer	4	5
		Average validity: 4.1	*Average usefulness: 4,5*

** How well do the study outcomes resemble the real world?*

*** Are the study outcomes useful for you or for your organization?*

Figure 4-13: Damrak (location 1)

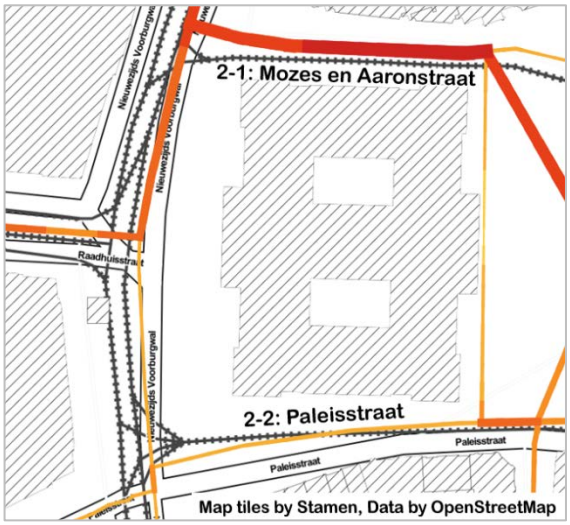


Figure 4-14: Mozes en Aäronstraat (location 2)

Figure 4-15: Kalverstraat (location 3)

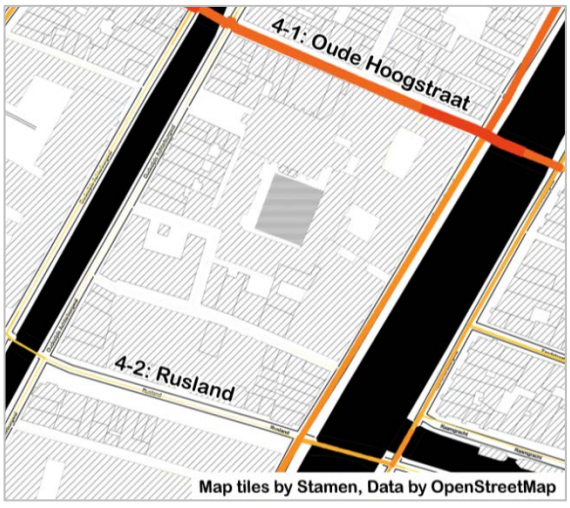


Figure 4-16: Oude Hoogstraat (location 4)

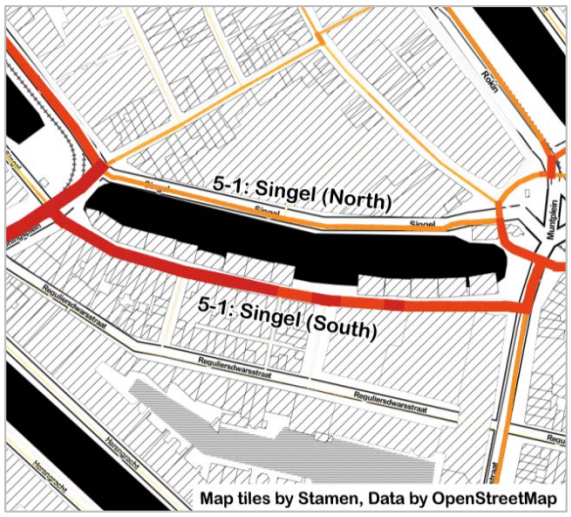


Figure 4-17: South side of Singel (location 5)

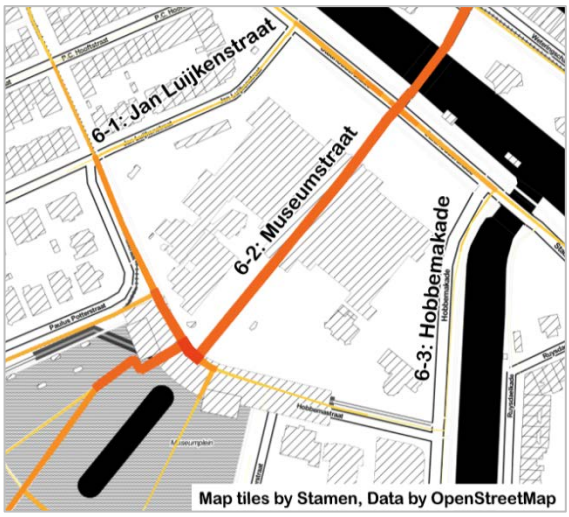


Figure 4-18: Museumstraat (location 6)

In last section of the questionnaire, experts were asked to specify for what purposes they could use the study outcomes and if they had any other comments or suggestions. The answers to these questions are given in appendix H.

5 Conclusion, discussion and recommendations

The first section of this final chapter presents the main conclusions of our research. Section 5.2 discusses the results and several recommendations for future work are given in section 5.3.

5.1 Conclusion

The objective of this exploratory research project was to develop, implement and test methods that reveal the spatial and temporal patterns of tourists from a large dataset of geotagged Flickr photos. To reach the objective, three research questions were proposed:

- The first research question (RQ-01) reviewed the current state-of-the-art and selected suitable methods for our case study in Amsterdam.
- The second research question (RQ-02) implemented and improved these methods to give insight in the *temporal distributions* and *spatial clusters* in our study area. Furthermore, a method was developed to create a *route density map* of tourists in Amsterdam.
- The third research question (RQ-03) used accuracy assessments and the opinion of experts to evaluate how well the detected patterns resemble the spatial and temporal behaviour of tourists.

For this research project, the metadata of 2,849,261 geotagged photos was downloaded in a time period of approximately 5 weeks. From this dataset, 393,828 photos were located in the municipality of Amsterdam. A literature study identified several methods to classify the countries of origin of photographers. In similar fashion as Wood, Guerry et al. (2013) and Straumann, Çöltekin et al. (2014), we utilized the user's profile information. Our semi-automatic classification method classified 39,1% of the users that contributed these photos as a tourist. The accuracy assessment of the classification method yielded a very high precision and recall for both tourists and locals. Our approach outperforms a commonly used method based on the photographers' visit time of a study area (Girardin, Dal Fiore et al. 2007, Van Canneyt, Schockaert et al. 2011, Sun, Fan et al. 2013). A comparison with official statistics showed that the tourists of several countries are overrepresented or underrepresented by the data. A dominant country in our dataset is the United States. Users originating from this country had a relatively strong influence on the spatial and temporal patterns that have been detected in this research.

Inspired by the work of Girardin, Dal Fiore et al. (2007), we extracted temporal distributions of photographers for the granularities *days of the week*, *months of the year* and *days of the year*. In extension to this work, we analysed the amount of photographers for several other temporal granularities and compared the temporal distributions of tourists and local residents to discover possible differences. A comparison between the distribution of locals and tourists per hour of the day revealed that the temporal distribution of tourists is approximately one and a half hours shifted compared to the temporal distribution of local photographers. We have developed a method that assesses the photo timestamps from locals and tourists by making use of photos that contain a real clock. The results indicated that the time shift is caused by incorrect camera times of tourists from other time zones. By comparing the relative number of unique tourists and photos per hour we revealed that relatively more images are captured during daytime and fewer during nighttime.

Amsterdam's spatial distribution of tourists is calculated by using grid-based and density-based clustering methods that were selected in the first research question. The results could, for example, help an urban planner to identify popular locations and locations that have the potential and capacity to welcome more tourists. We have selected the grid-based clustering methods of Sagl (2012) and Kádár and Gede (2013) to explore the spatial distribution of tourists in Amsterdam. Their method was improved by making use of hexagonal grid cells and a colour classification, which makes the results easier to interpret. All hexagons were visualized in Google Earth. The locations with the highest spatial tourist densities were detected with the DBSCAN clustering algorithm that was also used by Kisilevich, Krstajic et al. (2010), Sun, Fan et al. (2013) and Lee, Cai et al. (2014). The tags of photos located in the detected clusters were used to identify the location. We improved the method by adding a pre-processing step that prevents that many photos of a single photographer could form a spatial cluster. Further, we developed a method to estimate the most probable routes that tourists travelled between subsequent photo locations and aggregated all route calculations into a density map of touristic routes. The results could help policy-makers to identify locations where congestions might occur and provide suggestions for new business locations.

Due to the absence of comparable quantitative data, a qualitative approach was used to validate the study outcomes. The expert judgement of eight tourism experts of the municipality of Amsterdam was used to assess the detected spatial clusters and route density map of tourists. We found that their knowledge about the city bears a good resemblance with our study results. Moreover, our panel considered the temporal distributions, spatial clusters and route density map very useful for them and their organisation. Up till now, the municipality based most of its decisions on the statistics of controlled sites, tourist surveys and expert judgement. The proposed methods in this study provide new information about the spatial and temporal behaviour of tourists in public urban spaces. Some possible applications that were mentioned by the experts are crowd management, city marketing and the creation of new walking routes.

5.2 Discussion

We concluded, based on accuracy assessments and the opinion of experts, that our methods provide valuable insight in the *temporal distributions, spatial clusters* and *popular routes* of tourists in Amsterdam. However, it should be stated that the characteristics of the used geosocial dataset and our methods influenced the detected spatial and temporal patterns. This section discusses the data, our methods and the obtained results. A strong focus is given to the quality of the dataset and the developed methodology.

5.2.1 Characteristics of the data

First of all, let us put the sample of tourists in perspective. Naturally, the users of social media platform Flickr represent only a small part of all foreign visitors in our study area. According to Statistics Netherlands (CBS 2015), 9,4 million foreign visitors stayed in Amsterdam's hotels in 2012 and 2013 together. We regard this number as a good indication for the total amount of foreign visitors. In the same time period, 2,611 classified tourists took at least one photo in the city. That means that our dataset represents 0,03% of the foreign visitors. The sample is even smaller in earlier years when social media platform Flickr was less popular.

Furthermore, the results (Figure 4-1 on page 32) showed that tourists from the United States were overrepresented by the data. As a result, users originating from this country had a relatively strong influence on the spatial and temporal patterns that were detected in this research. The dominance of this group of people was understandable since social media platform Flickr originates from the United States. We also detected several other countries that are underrepresented and agree with Croitoru, Crooks et al. (2013) that geosocial data is demographically skewed. Moreover, Lo, McKercher et al. (2011) mentioned that people who post photos online "tend to be younger, better educated, and earn a higher income than those who do not" (p.725). We also suspect that the enthusiasm of Flickr users about new technology is on average higher than people who are not active on this social media platform. More research is needed to determine how the behaviour of tourists from Flickr differs from the behaviour of other tourists. It is also interesting to investigate how the behaviour of tourists from different countries differs. We hypothesise that different groups of users are interested in different sights and create different movement patterns. Despite the fact that our sample dataset is not a perfect representation of all tourists, we believe that the results are reasonably indicative for the spatial and temporal behaviour of tourists.

5.2.2 Temporal and spatial accuracy

During the exploration of Flickr photos, several incorrect timestamps were observed. We developed an approach that assessed the timestamps from locals and tourists by a comparison with the time read from photos that contain a real clock. An example of such a photo is given in Figure 3-11 on page 29. The time of the clock is compared to the time of the photo timestamp. We have only selected clocks that were easy to read and that have a high certainty to present the right time. All clocks that we were uncertain about are discarded. The DVD that is provided with this report contains a folder with our selection of clocks. The results indicate that tourists from other time zones occasionally have incorrect camera times. Apart from the detected temporal distributions, the time of the day does not influence the results of our methods.

Hauff (2013) investigated the spatial accuracy of geotagged Flickr photos and mentioned that photos taken at popular locations are often geotagged with a higher spatial accuracy than photos taken at less popular locations. In the exploration phase, several photos were detected that are indeed placed at a wrong location. Some of these records were revealed during the speed inspection of photo pairs. The results show that 1759 photo pairs (1,7% of total) have a speed that exceeds 50 km/h. The speed of 772 of those pairs even exceeds 200 km/h. It is impossible that a tourist travels through Amsterdam at this speed. Additionally, the coordinates of a photograph occasionally point to the location of the object being photographed instead of the location where the photograph was taken. Our algorithm filtered out photographs with a high speed. However, more research is required to develop a method that assesses the accuracy of geotagged Flickr photos.

Similar to the findings of Hollenstein and Purves (2010), we observed several users that uploaded many photos on a single location. This does not have a negative influence on the temporal distributions since the timestamps of photos appeared to be correct. Our pre-processing step in the methods to detect spatial clusters and tourist routes makes sure that just one photo per user per hexagon is selected. As a result, bulk uploads do not overly influence the estimated popularity of a location.

5.2.3 Classification of tourists

As described, we developed a semi-automatic method that made use of the user's profile information to classify their country of residence. A similar method was used by Wood, Guerry et al. (2013) and Straumann, Çöltekin et al. (2014). An accuracy assessment of the results yielded a precision of 98%. The method performs significantly better than a commonly used method based on the photographers' visit time of a study area (Girardin, Dal Fiore et al. 2007, Van Canneyt, Schockaert et al. 2011, Sun, Fan et al. 2013). However, their method also has some advantages. For example, one relatively simple database query selects all users that fulfil the temporal requirements. Implementing the method is faster and does not require manual work. Their method also classifies the country of origin of users without a user location in their profile. However, before you can apply this method you should download all photos in a country that were ever posted. This requires many weeks of harvesting. Downloading all photos within the Netherlands for the time period 2005 – 2015 took approximately 5 weeks. A big advantage of our approach is that it can be applied to photo sets with restricted spatial and temporal extents.

The open source service GeoNames was used to geocode all user locations that could not be geocoded by SQL. We experienced that the service did not return results for several locations that should not be difficult to geocode correctly. For example, 'Ocean Beach, SF' does not return a result. Yet, the paid geocoding service of Google correctly identified the country of this user location as the United States. Several other geocoding requests returned a wrong result. We are sure that the implementation of a better geocoding service would improve the results. As a result, less tourist nationalities would have to be manually corrected.

5.2.4 Temporal distributions

By comparing the relative number of unique tourists and photos per hour we also revealed that relatively more images were captured during daytime and fewer during nighttime. The detected distributions are strongly influenced by daytime photography. This also has its effects on the detected spatial distributions and popular routes of tourists. Popular nightlife areas are possibly underrepresented in the study results while popular daytime areas could be overrepresented.

Figure 4-4 shows several differences between the number of Flickr tourists and the number of tourists that spend a night in one of Amsterdam's hotels. The Flickr tourists are in the majority in the months April and September while the tourists in Amsterdam's hotels outnumber the Flickr tourists from November until February. All Flickr tourists visited the city centre and took photos in this area. They can be regarded as tourists that visit the city for leisure purposes. The number of tourists according the Amsterdam's hotel statistics also contains people who visit Amsterdam for business purposes. We expect that the flower season in April and May draws many tourists, represented by a relatively high amount of Flickr users in these months. Furthermore, we expect that many tourists spend a weekend in Amsterdam just after the summer. The relatively high amount of people that spend a night in a hotel during the winter months is possibly explained by people who visit Amsterdam for congresses and business purposes. A detailed inspection of the temporal distribution by a tourism expert is required to confirm our assumptions.

5.2.5 Spatial clustering

To provide an overview of Amsterdam's spatial distribution of tourists, we implemented and improved existing methods for grid-based clustering and density based clustering. Our grid-based clustering method makes use of a hexagon shaped grid. Figure 5-1 shows the effect of the chosen spatial granularity on the level of detail of the final visualization. The left image was one of the first grid-based clustering results of this project. The image reveals the location of Amsterdam's touristic city centre but does not reveal a lot of extra details. The right visualization is specially created for an urban planner of the municipality of Amsterdam who requested our project results for his design study. Due to the smaller hexagon sizes, every single touristic sight in Amsterdam can be observed.

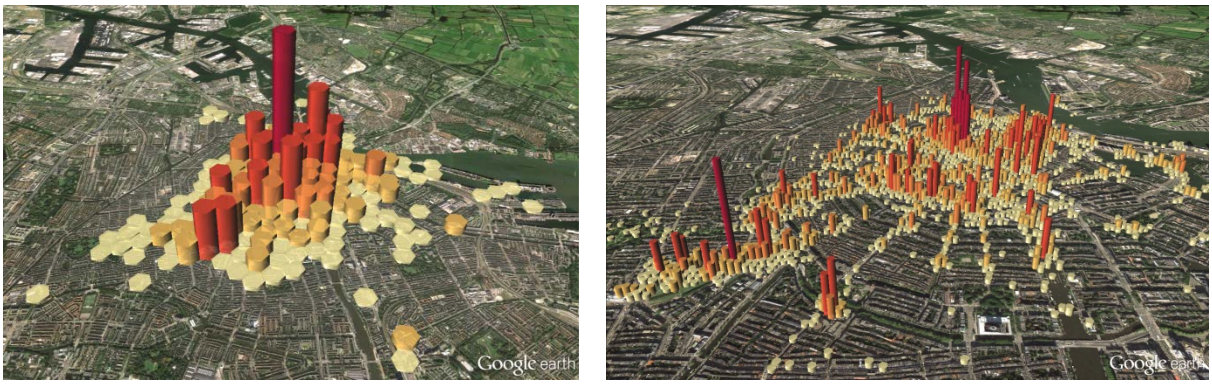


Figure 5-1: Effects of spatial granularity on level of detail

We compared our detected clusters with the suggested highlights of a tourist website Lonely Planet. Only the Central Station, Nieuwmarkt, Leidse Square and Rembrandt Square were not included in their list with top attractions in Amsterdam. It is logical that a central station draws many tourists but is not a typical touristic highlight in a city. The other locations are mainly home to bars and restaurants. They were not suggested as a *top pick* but they are included on the website.

The density-based cluster algorithm DBSCAN was implemented to detect Amsterdam's most touristic hotspots. Mean shift clustering, a different clustering algorithm, was used by Crandall, Backstrom et al. (2009) and their results identified Amsterdam's dam square as the cities touristic highlight (*see* section 2.4). Their findings correspond with the results that were found with our grid-based and density-based clustering methods.

The applied DBSCAN algorithm requires two parameters, *Eps* and *MinPts.* Using a heuristic approach, a value of 50 was selected for *Eps* and a value of 250 was selected for *MinPts.* To our knowledge, the algorithm produced very good results. Yet, detailed inspection of the clusters in cooperation with a tourism expert is needed to confirm that these values are appropriate for our dataset and study area.

5.2.6 Tourist routes and density map

The estimation of the most probable touristic routes between subsequent photo locations is the most explorative and experimental method of this research project. The presented method depends on arbitrary parameters and was not validated with objective quantitative data. More research is needed to refine and optimize the presented method.

Related work was published by Sun, Fan et al. (2013) and Kachkaev and Wood (2014), who proposed route recommendation systems for tourists. Their work, discovered in the final stage of this research

project, calculates the most popular touristic path based on the number of photographs. We agree that the number of photos on or near a road is a good indication for its popularity. Our work differs from their work since we estimated the most probable routes that tourists travelled between subsequent photo locations. All route calculations were aggregated into a density map of touristic routes.

Section 3.7.1 describes our approach to manually simplify OpenStreetMap's road network into a pedestrian routing network. More research is required to create an algorithm that executes this step based on a set of rules. In contrast to Kachkaev and Wood (2014), we argue that it is difficult to capture walking behaviour in a routing network consisting of edges and nodes. Objects like cars are restrained to the network while pedestrians walk on both sides of streets, take the inner curve and walk in many directions (especially on squares). All these decisions are of great influence on the length and position of a travelled route.

Although many photos were excluded in the different pre-processing steps of the methods, the selection of data that was used to detect the spatial patterns of tourists still contains several records that should have been rejected. Firstly, the chart in Figure 4-5 on page 36 shows distinctive peaks at Queen's Day, New Year's Day and the Gay Pride. Photos taken during those events most likely influenced the detected spatial clusters and touristic routes. Locations that host an event can become more popular than they are on an average day. The proposed method can be improved by filtering out all photos that are taken during events. Secondly, we did not implement advanced pre-processing steps to exclude photos inside buildings and portrait photos. Nevertheless, we do agree with Kachkaev and Wood (2014) that those photos are not describing the attractiveness of a street. Our method can be improved by implementing the advanced pre-processing steps that they proposed in their work. Thirdly, our study area contains many canals and other water bodies. Several photos in our dataset were taken from boats. These photos do not represent walking behaviour of a tourist. The threshold that selects all photos located within 25 meters of the pedestrian network excludes a selection of those photos. Nevertheless, not all the photos that were taken from boats are rejected and some of those photos were used to estimate touristic walking routes. It is relatively simple to exclude all photos that are spatially located within a water body. Yet, this would exclude many photos that are actually taken on land because of the spatial accuracy of the data.

We assumed that it is more likely that a tourist travelled over a road where a lot of photos were taken than over a road without photos. Kernel density estimation was used to create a grid with photo densities. These values were used as an input of a formula that calculated the cost per edge based on the local photo density. More information about the method is given in section 3.7.2 and appendix D. Based on a heuristic approach, the cube root of the photo density gave the best results. However, more work and quantitative data is required to improve the formula and method that estimates routes of tourists in public urban spaces.

Due to the absence of quantitative data about the behaviour of tourists in public urban spaces, expert judgement was used to validate the outcomes of this method. We prepared a questionnaire to validate the route density map. The results of the questionnaire indicate that the knowledge of our panel with tourism experts bears a good similarity with the calculated route densities. To optimize the method and enhance the results, a more detailed assessment of the study outcomes should be carried out by a panel of tourism experts.

5.2.7 Privacy

Flickr is a platform intended to share your work with the world. However, the privacy settings of the platform allow it to make your profile only accessible for friends and acquaintances. This research only used data from users who shared their photographs with the whole world. We explored the data in personal detail in several steps of this research project. Examples are the exploration of Flickr data in section 3.2, the validation of 50 random users classified as "tourist" and 50 random users classified as "local" in section 3.8.1 and the timestamp validation with photos that contained a real clock in section 3.8.2. However, we made sure that the final results of our methods only contain aggregated data and do not reveal any recognizable personal information of the dataset's creators.

5.3 Recommendations

The cited literature and methods that were proposed in this work revealed that geosocial data holds great potential for knowledge discovery. While working with the data and discussing the results, many ideas were developed. This section provides several recommendations for future work.

- Create a dataset that better represents the tourists in Amsterdam: Enlarge the dataset by incorporating additional sources like Twitter, Instagram, Foursquare and alternatives from countries like China and Russia. A combination of multiple sources could result in a better representation of Amsterdam's tourist population. Moreover, different platforms yield different types of information and provide answers to more tourism related questions.

- Correct or exclude photos with incorrect timestamps: The assessment of photo timestamps revealed that tourists from other time zones occasionally have incorrect camera times. Researchers that make use of the time of the day should be aware of those timestamp errors. We hypothesise that the results will improve by rejecting the photos of users that took many photos between 1:00 AM and 6:00 AM.

- Divide spatial distributions in seasonal and yearly intervals: Our work focused on spatial and temporal patterns. Combining the temporal distributions and spatial patterns could yield very valuable results. Dividing the spatial patterns in patterns per year could for example reveal the consequences of large urban projects and museum openings in Amsterdam on the spatial distribution of tourists. Further, these patterns can potentially be used to analyse the effects of city marketing on the spreading of tourists. It is important to take the increasing popularity of social media platform Flickr into account.

- Divide spatial distributions in hourly intervals: The municipality mentioned that they are interested in the parts of the city that are popular during daytime and the parts of the city that are popular during nighttime. Spatial patterns per hour of the day have the potential to reveal which parts of the city are popular at what times of the day. There are two remarks. Firstly, there are relatively more pictures captured during daytime. Secondly, the dataset contains several photos with time errors, mainly caused by tourists originating from different time zones.

- Compare the spatial distribution of locals and tourists: Our work is focused on the spatial distributions of tourists. Comparing these distributions with the distributions of locals could potentially reveal locations that are interesting for tourists but not yet discovered. Additional city marketing could promote the locations that are suitable for foreign tourists to draw them away from overcrowded locations.

- Divide spatial and temporal patterns for different nationalities: All the presented spatial and temporal patterns regarded tourists as one class of data. Yet, our semi-automatic classification method also revealed the origin of those tourists. It is possible to utilize this nationality and analyse which parts of the city are visited by which groups of tourists. The information could for example be used to inform specific groups of tourists about locations that are potentially interesting for them. Walking routes can be tailored for tourists of different nationalities.

- Enhance the route density method by adding the direction of movement: We have developed a method that aggregates all probable routes of tourists into a route density map. Due to a limitation of the pgRouting algorithm, the method does not reveal the direction of tourists. Yet, the direction of movement could be interesting information for planners, policy-makers and tourism marketing. We recommend enhancing the method so that the direction of tourist routes is revealed instead of only the density.

- Use the revealed patterns as input for an agent-based model: Our models look back into a history of photo taking events to identify temporal distributions, spatial densities and popular routes of tourists. Using this information as input data for an agent-based model has the potential to study the effects of future tourism policies and large-scale urban projects.

- Discover locations with tourism related problems with other types of geosocial data: Typical tourism problems that are often mentioned are overcrowding, high noise levels, litter and dangerous traffic situations. These problems are often related to negative sentiments, which are not captured by tourism photography. However, officials of the municipality mentioned that they would benefit from an overview of locations where these problems are reported. We recommend the use of Twitter's Tweets to analyse these sentiments. Wi-Fi, Bluetooth and camera tracking are suitable methods to study overcrowding and congestion problems.

5.4 Final words

While geotagged photos hold great potential for discovering knowledge about the behaviour of tourists, it is important to understand the limitations. Frequently mentioned tourism problems like high noise levels, litter and dangerous traffic situations deal with negative sentiments of people. These sentiments are usually not captured by tourist photos. The data can be used to identify popular touristic locations but is unsuitable to analyse overcrowding. Moreover, the data tends to be demographically skewed and several records have spatial and temporal inaccuracies. Despite the imperfections of geosocial data, the validated results prove that geotagged photos can provide meaningful insight into spatial and temporal patterns of tourists. Moreover, our panel with tourism experts regard the results as a valuable addition to traditional tourism studies.

6 References

Ali, A. e., S. v. Sas and F. Nack (2013). Photographer paths: sequence alignment of geotagged photos for exploration-based route planning. Proceedings of the 2013 conference on Computer supported cooperative work. San Antonio, Texas, USA, ACM.

Amsterdam, O. S. (2014). Amsterdam in cijfers, Jaarboek 2014, Gemeente Amsterdam, bureau Onderzoek en Statistiek**:** 334-343.

Andrienko, G., N. Andrienko, P. Bak, D. Keim, S. Kisilevich and S. Wrobel (2011). "A conceptual framework and taxonomy of techniques for analyzing movement." Journal of Visual Languages & Computing **22**(3): 213-232.

Andrienko, G., N. Andrienko, D. Keim, A. M. MacEachren and S. Wrobel (2011). "Challenging problems of geospatial visual analytics." Journal of Visual Languages and Computing **22**(4): 251-256.

Andrienko, N. and G. Andrienko (2006). Exploratory analysis of spatial and temporal data: a systematic approach, Springer Science & Business Media.

Ashworth, G. and S. J. Page (2011). "Urban tourism research: Recent progress and current paradoxes." Tourism Management **32**(1): 1-15.

Berge, J. t. and E. Jakobs (2013). Drukte in de binnenstad 2012, Gemeente Amsterdam, Bureau Onderzoek en Statistiek.

Birch, C. P. D., S. P. Oom and J. A. Beecham (2007). "Rectangular and hexagonal grids used for observation, experiment and simulation in ecology." Ecological Modelling **206**(3–4): 347-359.

Boots, B. and M. Tiefelsdorf (2000). "Global and local spatial autocorrelation in bounded regular tessellations." Journal of Geographical Systems **2**(4): 319-348.

CBS (2014). Wijk- en buurtkaart 2014, Centraal Bureau voor de Statistiek.

CBS. (2015). "Hotels: gasten, overnachtingen, woonland, sterklasse." Retrieved March 18, 2015, from http://statline.cbs.nl/StatWeb/publication/?VW=T&DM=SLNL&PA=82060NED&LA=NL.

ColorBrewer. (2015). "ColorBrewer 2.0 - Color advice for cartography." Retrieved March 12, 2015, from http://colorbrewer2.org/.

Crandall, D. J., L. Backstrom, D. Huttenlocher and J. Kleinberg (2009). Mapping the world's photos. Proceedings of the 18th international conference on World wide web, ACM.

Croitoru, A., A. Crooks, J. Radzikowski and A. Stefanidis (2013). "Geosocial gauge: a system prototype for knowledge discovery from social media." International Journal of Geographical Information Science **27**(12): 2483-2508.

Croitoru, A., A. T. Crooks, J. Radzikowski, A. Stefanidis, R. R. Vatsavai and N. Wayant (2014). Geoinformatics and Social Media: A New Big Data Challenge. Big Data Techniques and Technologies in Geoinformatics. H. Karimi. Boca Raton, Florida, CRC Press**:** 207-232.

De Choudhury, M., M. Feldman, S. Amer-Yahia, N. Golbandi, R. Lempel and C. Yu (2010). Constructing travel itineraries from tagged geo-temporal breadcrumbs. Proceedings of the 19th international conference on World wide web, ACM.

Dijkstra, E. W. (1959). "A note on two problems in connexion with graphs." Numerische mathematik **1**(1): 269-271.

Edwards, D., T. Dickson, T. Griffin and B. Hayllar (2010). "Tracking the urban visitor: Methods for examining tourists' spatial behaviour and visual representations." Cultural tourism research methods: 104-114.

Ester, M., H.-P. Kriegel, J. Sander and X. Xu (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. Kdd.

Estima, J. and M. Painho (2013). Flickr Geotagged and Publicly Available Photos: Preliminary Study of Its Adequacy for Helping Quality Control of Corine Land Cover. Computational Science and Its Applications. B. Murgante, S. Misra, M. Carlini et al., Springer Berlin Heidelberg. **7974:** 205-220.

Fayyad, U., G. Piatetsky-Shapiro and P. Smyth (1996). "From data mining to knowledge discovery in databases." AI magazine **17**(3): 37-54.

Flickr4Java (2015). Java library that wraps the REST-based Flickr API.

Flickr. (2011). "6 billionth photo!" Retrieved October 8, 2014, from http://blog.flickr.net/en/2011/08/04/6000000000/.

Flickr. (2013). "Photo: 9452043497 - User ID: 98251082@N00 - Title: "No time to bike"." Retrieved March 19, 2015, from https://www.flickr.com/photos/andrepmeyer/9452043497/.

Friedland, G. and R. Sommer (2010). Cybercasing the Joint: On the Privacy Implications of Geo-Tagging. HotSec.

Garrod, B. (2008). "Understanding the Relationship Between Tourism Destination Imagery and Tourist Photography." Journal of Travel Research **47**(3): 346-358.

Geofabrik. (2015). "OpenStreetMap Data Extracts." Retrieved March 12, 2015, from http://download.geofabrik.de/europe/netherlands.html.

GeoNames. (2015). "GeoNames geographical database." Retrieved March 10, 2015, from www.geonames.org/.

Girardin, F., F. Calabrese, F. D. Fiore, C. Ratti and J. Blat (2008). "Digital footprinting: Uncovering tourists with user-generated content." Pervasive Computing, IEEE **7**(4): 36-43.

Girardin, F., F. Dal Fiore, J. Blat and C. Ratti (2007). Understanding of tourist dynamics from explicitly disclosed location information. Symposium on LBS and Telecartography, Citeseer.

Goodchild, M. (2007). "Citizens as sensors: the world of volunteered geography." GeoJournal **69**(4): 211-221.

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): 682-703.

Haklay, M., A. Singleton and C. Parker (2008). "Web mapping 2.0: The neogeography of the GeoWeb." Geography Compass **2**(6): 2011-2039.

Hauff, C. (2013). A study on the accuracy of Flickr's geotag data. Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, ACM.

Hennig, C. (2014). R-package fpc: Flexible procedures for clustering.

Hollenstein, L. and R. S. Purves (2010). "Exploring place through user-generated content: Using Flickr to describe city cores." Journal of Spatial Information Science **1**(1): 21-48.

Jain, A. K., M. N. Murty and P. J. Flynn (1999). "Data clustering: a review." ACM computing surveys (CSUR) **31**(3): 264-323.

Jankowski, P., N. Andrienko, G. Andrienko and S. Kisilevich (2010). "Discovering landmark preferences and movement patterns from photo postings." Transactions in GIS **14**(6): 833-852.

Janowicz, K., M. Raubal, S. Levashkin, C. Keßler, P. Maué, J. Heuer and T. Bartoschek (2009). Bottom-Up Gazetteers: Learning from the Implicit Semantics of Geotags. GeoSpatial Semantics, Springer Berlin Heidelberg. **5892:** 83-102.

Jenks, G. F. (1967). The Data Model Concept in Statistical Mapping. International Yearbook of Cartography. **7:** 186-190.

Kachkaev, A. and J. Wood (2014). Automated planning of leisure walks based on crowd-sourced photographic content. 46th Annual Universities' Transport Study Group Conference. Newcastle, UK.

Kádár, B. and M. Gede (2013). "Where do tourists go? Visualizing and analysing the spatial distribution of geotagged photography." Cartographica **48**(2): 78-88.

Kaplan, A. M. and M. Haenlein (2010). "Users of the world, unite! The challenges and opportunities of Social Media." Business Horizons **53**(1): 59-68.

Keim, D., G. Andrienko, J. D. Fekete, C. Görg, J. Kohlhammer and G. Melançon (2008). Visual analytics: Definition, process, and challenges. Lecture Notes in Computer Science. **4950 LNCS:** 154-175.

Kisilevich, S., M. Krstajic, D. Keim, N. Andrienko and G. Andrienko (2010). Event-based analysis of people's activities and behavior using Flickr and Panoramio geotagged photo collections. Information Visualisation (IV), 2010 14th International Conference, IEEE.

Kurashima, T., T. Iwata, G. Irie and K. Fujimura (2013). "Travel route recommendation using geotagged photos." Knowledge and Information Systems **37**(1): 37-60.

Lee, I., G. Cai and K. Lee (2014). "Exploration of geo-tagged photos through data mining approaches." Expert Systems with Applications **41**(2): 397-405.

Lo, I. S., B. McKercher, A. Lo, C. Cheung and R. Law (2011). "Tourism and online photography." Tourism Management **32**(4): 725-731.

LonelyPlanet. (2015). "Top things to do in Amsterdam." Retrieved March 11, 2015, from http://www.lonelyplanet.com/the-netherlands/amsterdam/things-to-do/top-things-to-do-in-amsterdam.

Mennis, J. and D. Guo (2009). "Spatial data mining and geographic knowledge discovery—An introduction." Computers, Environment and Urban Systems **33**(6): 403-408.

Miller, H. J. (2010). "The data avalanche is here. Shouldn't we be digging?" Journal of Regional Science **50**(1): 181-201.

Miller, H. J. and J. Han (2009). Geographic Data Mining and Knowledge Discovery - An Overview. Geographic data mining and knowledge discovery. H. J. Miller and J. Han. Boca Raton, CRC Press (Taylor & Francis Group)**:** 1-26.

NBTC (2013). Toekomstperspectief Destinatie Holland 2025, NBTC Holland Marketing.

O+S (2014). Kerncijfers Amsterdam 2014, O+S, Research and Statistics, City of Amsterdam.

OGC (2008). Open Geospatial Consortium KML.

Okuyama, K. and K. Yanai (2013). A travel planning system based on travel trajectories extracted from a large number of geotagged photos on the web. The era of interactive media, Springer**:** 657-670.

OSM. (2015). "OpenStreetMap Copyright and License." Retrieved March 12, 2015, from www.openstreetmap.org/copyright.

Parzen, E. (1962). "On estimation of a probability density function and mode." The annals of mathematical statistics: 1065-1076.

Peuquet, D. J. (1994). "It's about Time: A Conceptual Framework for the Representation of Temporal Dynamics in Geographic Information Systems." Annals of the Association of American Geographers **84**(3): 441-461.

pgRouting (2015). pgRouting 2.0.0 — Open Source Routing Library.

Popescu, A. and G. Grefenstette (2009). Deducing trip related information from flickr. Proceedings of the 18th international conference on World wide web. Madrid, Spain, ACM.

PostGIS (2015). PostGIS 2.1 - Spatial and Geographic Objects for PostgreSQL.

PostgreSQL (2015). PostgreSQL 9.3 - Open source object-relational database system.

QGIS (2015). QGIS 2.6.1 Brighton - Open source Geographic Information System - OSGeo.

Rahm, E. and H. H. Do (2000). "Data cleaning: Problems and current approaches." IEEE Data Eng. Bull. **23**(4): 3-13.

Roick, O. and S. Heuser (2013). "Location Based Social Networks – Definition, Current State of the Art and Research Agenda." Transactions in GIS **17**(5): 763-784.

Sagl, G. R., B.; Hawelka, B.; Beinat, E. (2012). From Social Sensor Data to Collective Human Behaviour Patterns – Analysing and Visualising Spatio-Temporal Dynamics in Urban Environments. GI-Forum 2012: Geovisualization, Society and Learning.

Scikit-learn. (2015). "Overview of clustering methods." Retrieved March 12, 2015, from http://scikit-learn.org/stable/modules/clustering.html.

Shoval, N. (2008). "Tracking technologies and urban analysis." Cities **25**(1): 21-28.

Slocum, T., R. McMaster, F. Kessler and H. Howard (2009). Thematic Cartography and Geovisualization, Prentice Hall.

Stefanidis, A., A. Crooks and J. Radzikowski (2013). "Harvesting ambient geospatial information from social media feeds." GeoJournal **78**(2): 319-338.

Steiniger, S. and A. J. S. Hunter (2013). "The 2012 free and open source GIS software map – A guide to facilitate research, development, and adoption." Computers, Environment and Urban Systems **39**(0): 136-150.

Straumann, R. K., A. Çöltekin and G. Andrienko (2014). "Towards (Re)Constructing Narratives from Georeferenced Photographs through Visual Analytics." The Cartographic Journal **51**(2): 152-165.

Sui, D. and M. Goodchild (2011). "The convergence of GIS and social media: challenges for GIScience." International Journal of Geographical Information Science **25**(11): 1737-1748.

Sun, Y., H. Fan, M. Bakillah and A. Zipf (2013). "Road-based travel recommendation using geo-tagged images." Computers, Environment and Urban Systems(0).

Thatcher, J. (2014). "Big Data, Big Questions| Living on Fumes: Digital Footprints, Data Fumes, and the Limitations of Spatial Big Data." International Journal of Communication **8**: 1765–1783.

Twitter. (2014). "Twitter usage." Retrieved October 7, 2014, from https://about.twitter.com/company.

Van Canneyt, S., S. Schockaert, O. Van Laere and B. Dhoedt (2011). Time-dependent recommendation of tourist attractions using Flickr. Proceedings of the 23rd Benelux conference on artificial intelligence.

Wolf, J., R. Guensler and W. Bachman (2001). "Elimination of the travel diary: Experiment to derive trip purpose from global positioning system travel data." Transportation Research Record: Journal of the Transportation Research Board **1768**(1): 125-134.

Wood, J., J. Dykes, A. Slingsby and K. Clarke (2007). "Interactive visual exploration of a large spatio-temporal dataset: Reflections on a geovisualization mashup." IEEE Transactions on Visualization and Computer Graphics **13**(6): 1176 - 1183.

Wood, S. A., A. D. Guerry, J. M. Silver and M. Lacayo (2013). "Using social media to quantify nature-based tourism and recreation." Scientific Reports **3**(2976): 1-7.

Xin, L., W. Changhu, Y. Jiang-Ming, P. Yanwei and Z. Lei (2010). Photo2Trip: generating travel routes from geo-tagged photos for trip planning. Proceedings of the international conference on Multimedia. Firenze, Italy, ACM.

Appendix A: Table of contents DVD

The DVD that accompanies this thesis report contains the following information:

Documents

- Final thesis report (Word, PDF)
- All cited literature (Endnote library)
- Questionnaire municipality (PDF)

Presentations

- Midterm presentation (PDF)
- Municipality presentation (PDF)
- Final presentation (PDF)

Data

- PostgreSQL database tables (naming corresponds models thesis report)
- Flickr photos used for timestamp validation
- Hexagonal photo densities in Google Earth

Applications and scripts

- SQL-scripts (naming corresponds models thesis report)
- R-script for DBSCAN clustering (naming corresponds models thesis report)
- Java applications (naming corresponds models thesis report)

Figures and maps

- Figures report
- QGIS files for all maps used in report

Appendix B: Origin of Flickr users in Amsterdam (2005-2015)

Table B1 contains a list with the country of residence of Flickr photographers in Amsterdam in the time period 2005-2015. The rank, percentage, number of photographers and number of photos is given for all countries with a minimum of 20 photographers in Amsterdam.

Table B1: Countries with more than 20 photographers in Amsterdam (2005-2015)

Rank	Country	Percentage	Photographers	Photos
1	Netherlands	31.1%	2,821	154,599
2	United States	13.7%	1,243	26,960
3	United Kingdom	11.4%	1,035	17,982
4	Germany	5.2%	475	7,929
5	Italy	5.0%	450	3,708
6	Spain	4.8%	434	4,526
7	France	4.0%	360	4,999
8	Brazil	2.4%	215	3,289
9	Canada	2.3%	209	4,241
10	Belgium	2.1%	192	2,889
11	Australia	1.5%	133	2,807
12	Russia	1.3%	114	2,960
13	Switzerland	1.1%	100	2,999
14	Sweden	0.9%	79	747
15	Finland	0.8%	77	1,160
16	Denmark	0.8%	72	844
17	Norway	0.7%	66	797
18	Ireland	0.7%	61	631
19	Austria	0.7%	60	1,213
20	Portugal	0.6%	58	778
21	Japan	0.6%	54	1,121
22	Taiwan	0.6%	51	1,534
23	Poland	0.5%	46	1,726
24	China	0.5%	43	468
25	Greece	0.4%	37	1,960
26	Mexico	0.4%	37	223
27	Israel	0.4%	34	755
28	Hungary	0.3%	30	784
29	Singapore	0.3%	29	383
30	Chile	0.3%	28	842
31	Argentina	0.3%	28	95
32	Czech Republic	0.3%	27	466
33	India	0.3%	25	109
34	Malaysia	0.3%	24	402
35	Romania	0.3%	23	185
36	Indonesia	0.3%	23	127
37	Turkey	0.2%	22	415
38	New Zealand	0.2%	22	334
-	*Other countries*	*2.7%*	*241*	*3628*

Appendix C: DBSCAN parameter exploration

A heuristic approach has been used to define the *Eps* and *MinPts* parameters of the DBSCAN cluster algorithm. Figure C1 contains a selection of sub results of the parameter exploration. The black dots identify all photos that were classified as noise. The coloured dots are assigned to clusters.

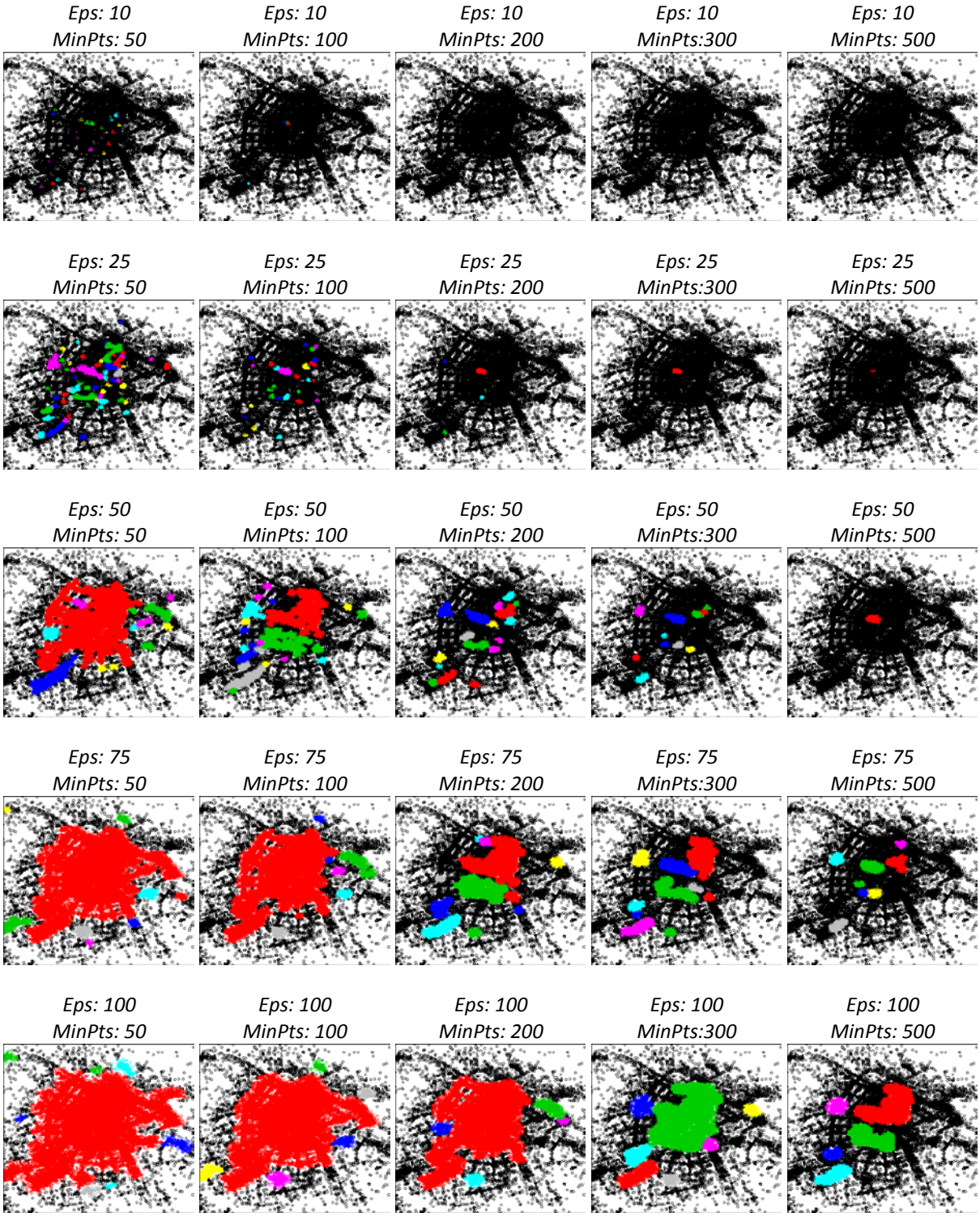


Figure C1: Sub results of DBSCAN parameter exploration

Appendix D: Route cost reduction based on road popularity

The amount of photos in the neighbourhood of a road is regarded as a good estimation for its popularity. It is more likely that a tourist travels over a road where a lot of photos were taken then over a road without photos. The calculated photo density is used to reduce the travel cost per edge. We have used a heuristic approach to define the amount of cost reduction. Figure D1 contains the results of a selection of reduction factors that were considered. The cube root of the photo density gave the most promising results. Simply using the photo density or the root of the photo density tends to create too many routes at streets with many photos.

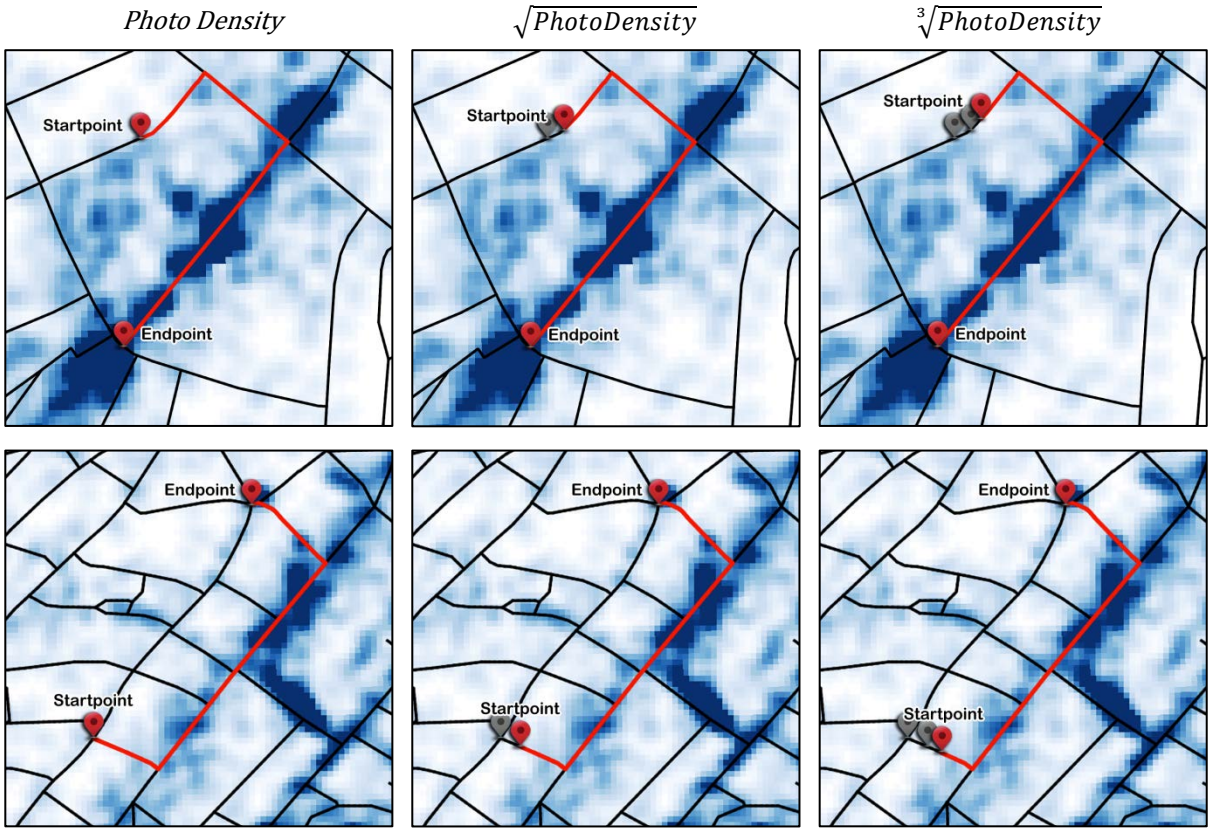


Figure D1: Effects of reduction factors on calculated routes Rijksmuseum and Damrak

Appendix E: Lonely Planet's *Top Picks* in Amsterdam

The detected and identified clusters were compared with the *top picks* of travel website Lonely Planet (2015). The *top picks* are 17 places in Amsterdam that Lonely Planet recommends tourists to visit during their stay in the city. The selection of places is included in figure E1.

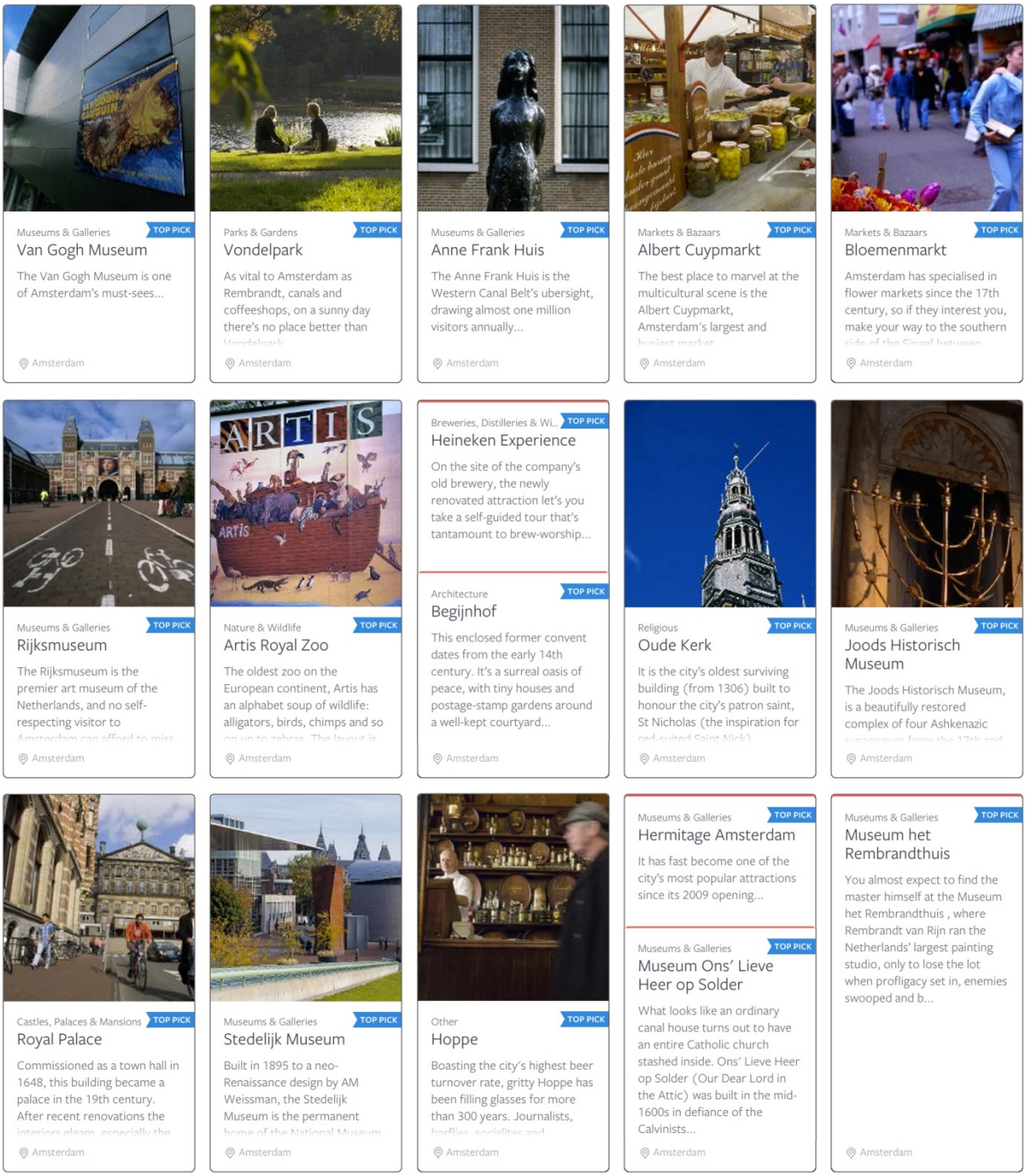


Figure E1: Top picks of travel website Lonely Planet (2015)

Appendix F: Temporal distribution of tourists per hour of the week

Several temporal distributions were extracted from the collection of tourist photos in the municipality of Amsterdam. Those distributions are presented in section 4.3.1 of the chapter results and validation. In addition, we extracted the unique number of tourists per *hour of the week* from the timestamps of tourist photos. The visualisation of this distribution is given in figure F1.

	Monday	*Tuesday*	*Wednesday*	*Thursday*	*Friday*	*Saturday*	*Sunday*
00:00	85	78	83	83	69	108	119
01:00	52	58	50	49	68	70	92
02:00	46	47	57	51	53	54	75
03:00	43	52	54	46	61	61	55
04:00	39	49	53	56	55	71	57
05:00	61	58	69	48	55	70	70
06:00	68	62	55	58	60	86	73
07:00	78	76	62	76	104	98	90
08:00	119	131	104	113	123	145	114
09:00	182	156	146	143	185	237	209
10:00	232	212	215	199	248	322	269
11:00	284	228	244	259	293	398	363
12:00	287	256	251	251	322	476	412
13:00	288	260	311	278	368	453	407
14:00	281	268	274	285	332	458	400
15:00	255	255	273	291	320	462	397
16:00	251	259	245	274	313	435	336
17:00	226	226	217	244	264	379	284
18:00	195	192	212	216	263	319	236
19:00	180	211	178	205	210	260	196
20:00	186	162	143	187	214	268	183
21:00	140	140	134	163	173	193	173
22:00	117	131	139	157	140	175	140
23:00	87	100	117	109	120	140	97

Figure F1: Distribution of unique tourists per hour of the week in Amsterdam

Appendix G: Questionnaire for expert judgement

The expert judgement of eight tourism experts of the municipality of Amsterdam was used to assess the detected spatial clusters and route density map of tourists. The questionnaire that was used is included on the next pages of this appendix.

QUESTIONNAIRE – PART 1

What is your gender?	MALE / FEMALE
How old are you?	…… years
In which city do you live?	……
In which neighborhood do you live?	……
For which organization do you work?	……
What is your profession?	……

How well do you know the city center of Amsterdam
1=not, 5=very well

1	2	3	4	5
☐	☐	☐	☐	☐

Which location is more touristic?

☐ 1-1: NIEUWENDIJK	☐ 1-2: DAMRAK	☐ Unknown
☐ 2-1: MOZES EN AARONSTRAAT	☐ 2-2: PALEISSTRAAT	☐ Unknown
☐ 3-1: KALVERSTRAAT	☐ 3-2: ROKIN	☐ Unknown
☐ 4-1: OUDE HOOGSTRAAT	☐ 4-2: RUSLAND	☐ Unknown
☐ 5-1: SINGEL (NOORDZIJDE)	☐ 5-2: SINGEL (ZUIDZIJDE)	☐ Unknown

☐ 6-1: JAN LUIJKENSTRAAT	☐ 6-2: MUSEUMSTRAAT	☐ 6-3: HOBBEMAKADE	☐ Unknown

How touristic are the following locations?

1 = not touristic, 5 = very touristic

	Unknown	1	2	3	4	5
1-1: NIEUWENDIJK	☐	☐	☐	☐	☐	☐
1-2: DAMRAK	☐	☐	☐	☐	☐	☐
2-1: MOZES EN AARONSTRAAT	☐	☐	☐	☐	☐	☐
2-2: PALEISSTRAAT	☐	☐	☐	☐	☐	☐
3-1: KALVERSTRAAT	☐	☐	☐	☐	☐	☐
3-2: ROKIN	☐	☐	☐	☐	☐	☐
4-1: OUDE HOOGSTRAAT	☐	☐	☐	☐	☐	☐
4-2: RUSLAND	☐	☐	☐	☐	☐	☐
5-1: SINGEL (NOORDZIJDE)	☐	☐	☐	☐	☐	☐
5-2: SINGEL (ZUIDZIJDE)	☐	☐	☐	☐	☐	☐
6-1: JAN LUIJKENSTRAAT	☐	☐	☐	☐	☐	☐
6-2: MUSEUMSTRAAT	☐	☐	☐	☐	☐	☐
6-3: HOBBEMAKADE	☐	☐	☐	☐	☐	☐

How confident are you with your answers?	1	2	3	4	5
1=not confident, 5=very confident	☐	☐	☐	☐	☐

QUESTIONNAIRE – PART 2

How well do the study outcomes resemble the real world situation?

1=no, 5=very well

1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Are the study outcomes useful for you or for your organization?

1=no, 5=very useful

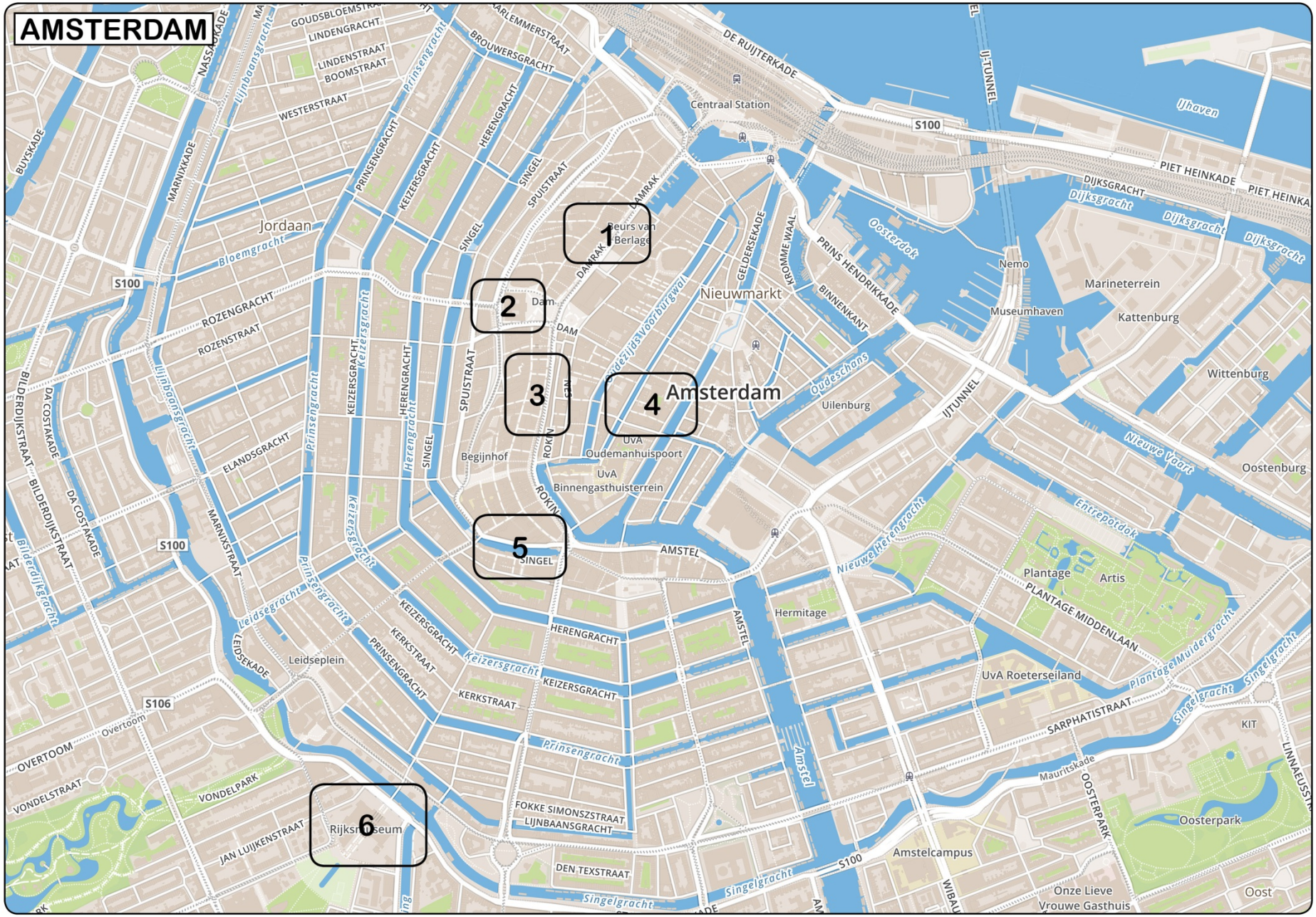
1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

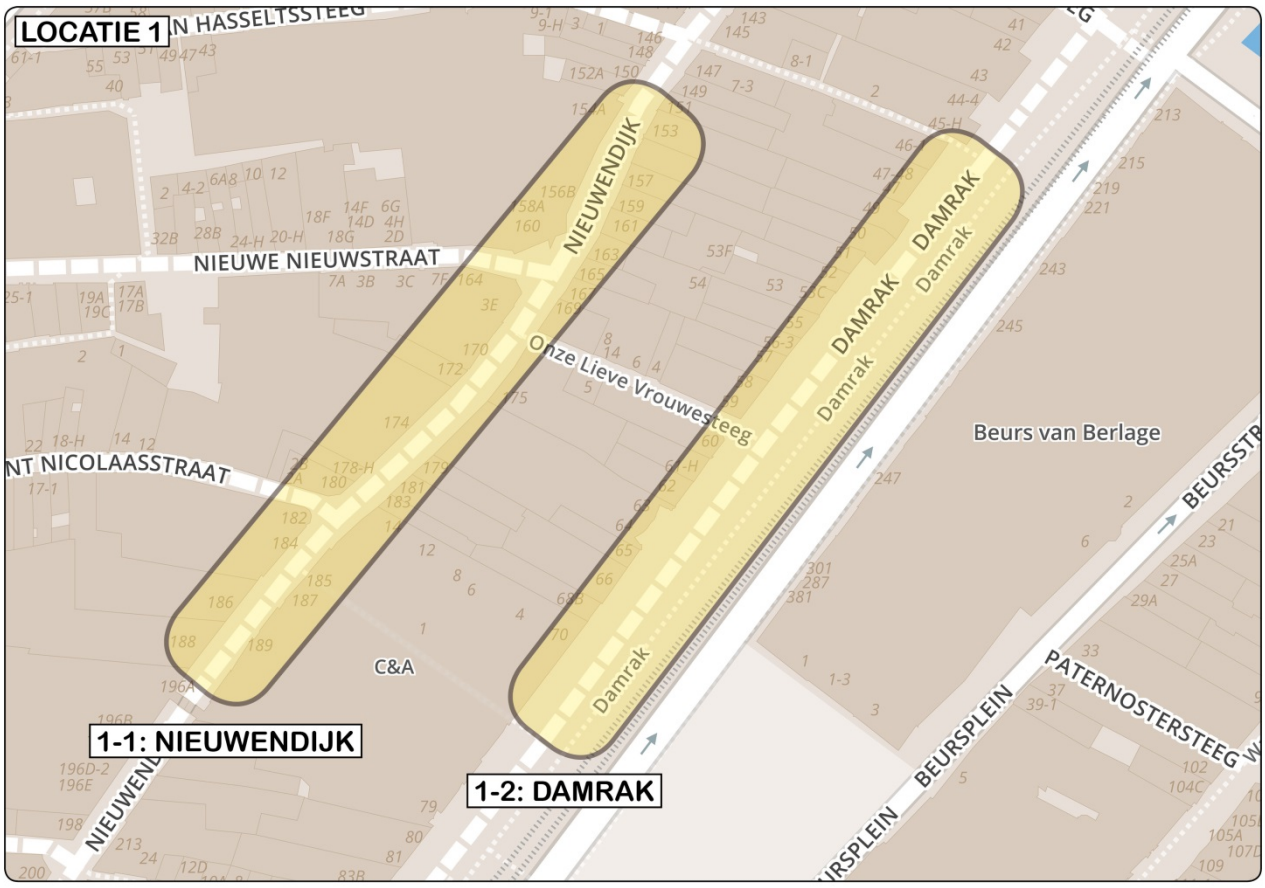
For what purposes could you use the study outcomes?

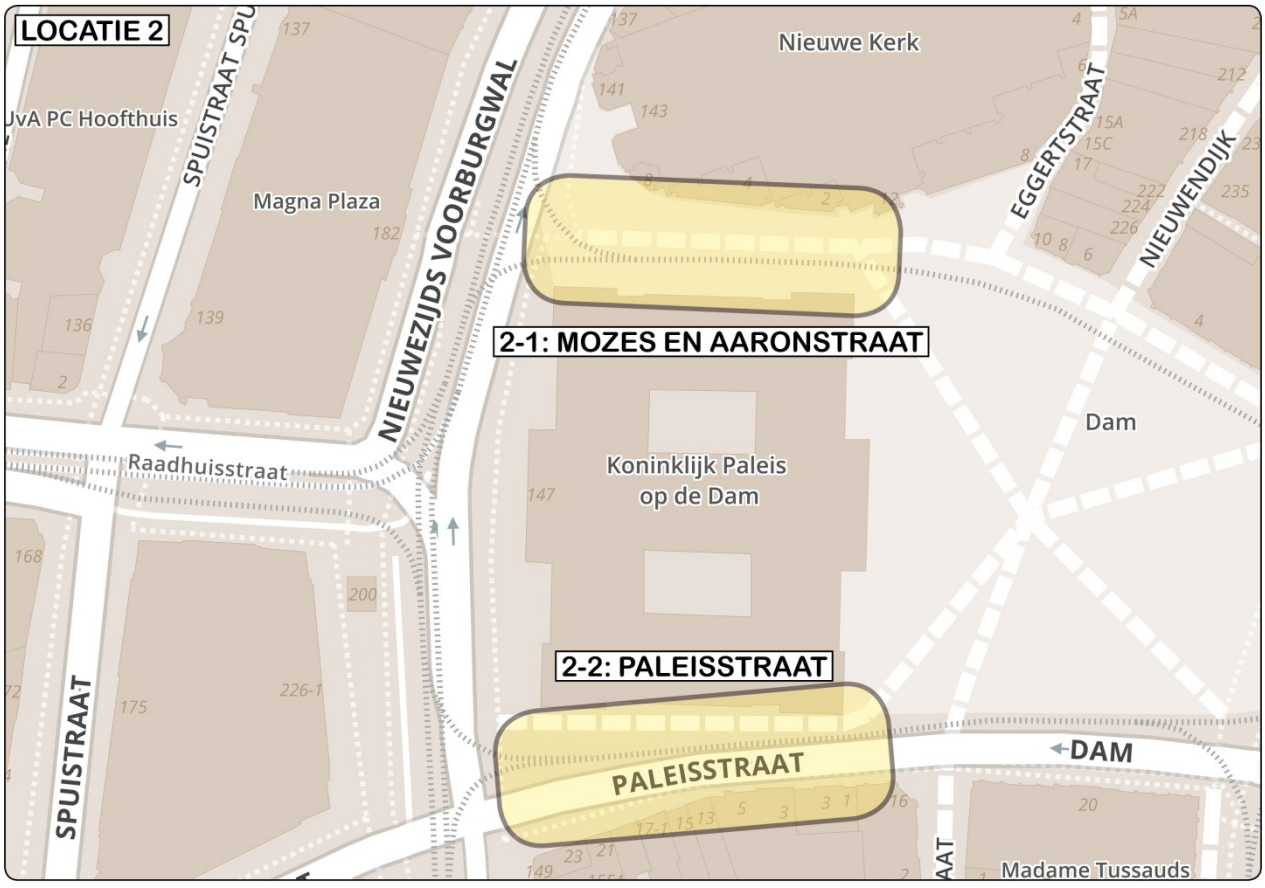
……
……
……
……
……
……
……
……
……
……

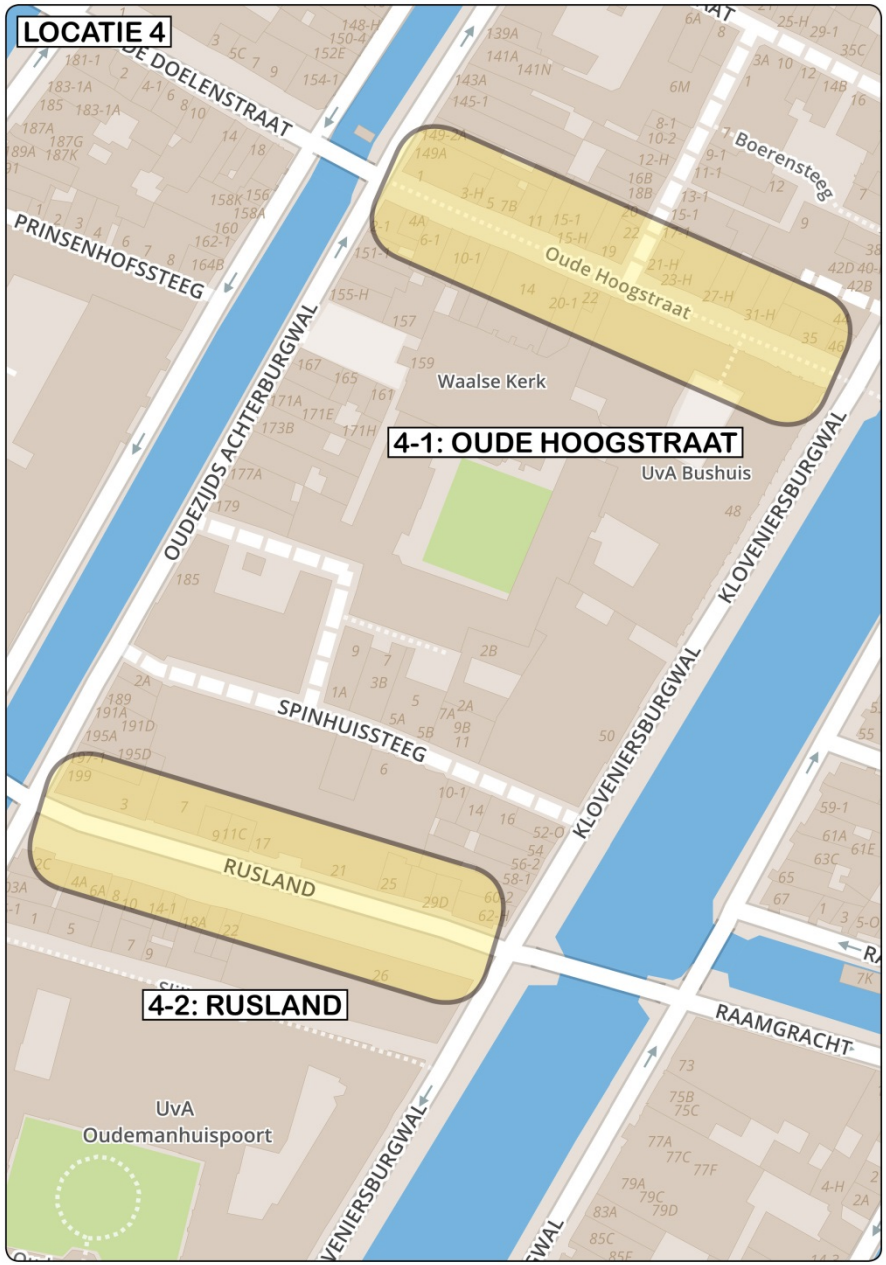
Do you have any other comments or suggestions?

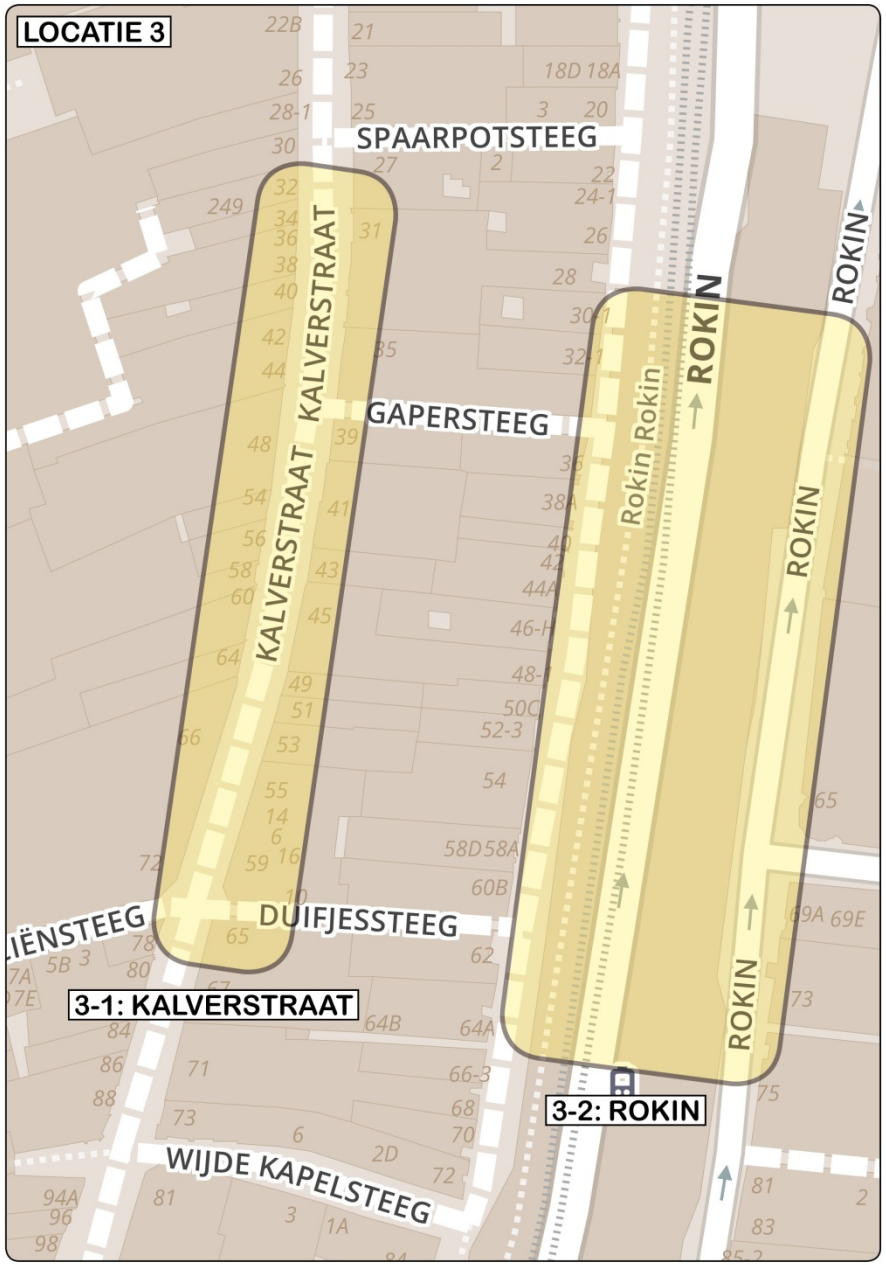
……
……
……
……
……
……
……
……
……
……

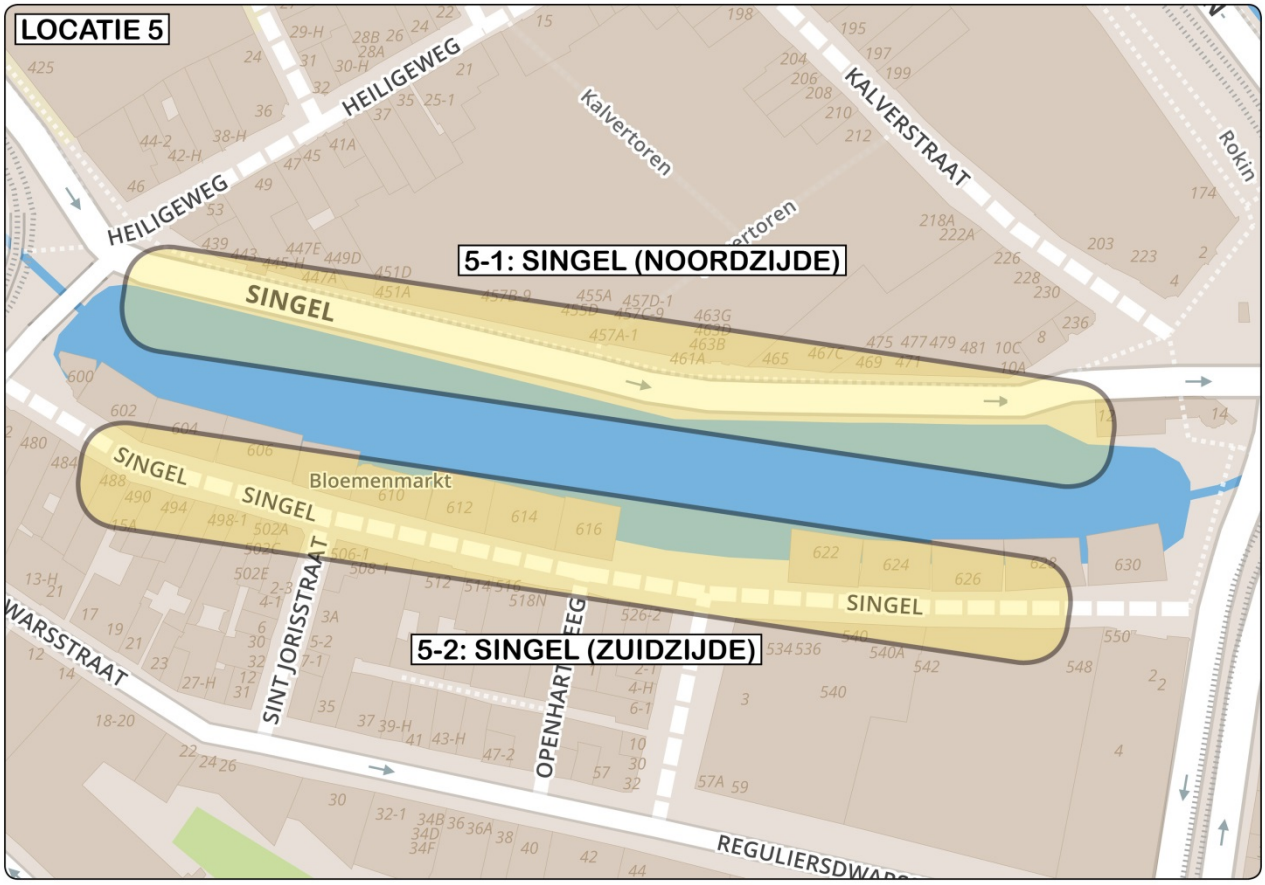

AMSTERDAM
1
2
3
4
5
6
Amsterdam
Centraal Station
Nieuwmarkt
Jordaan
Dam
Begijnhof
Oudemanhuispoort
UvA
Binnengasthuisterrein
Beurs van Berlage
Nemo
Marineterrein
Museumhaven
Kattenburg
Wittenburg
Oostenburg
Uilenburg
Oudeschans
Oosterdok
IJhaven
IJ-TUNNEL
DE RUIJTERKADE
PIET HEINKADE
DIJKSGRACHT
PRINS HENDRIKKADE
BINNENKANT
KROMME WAAL
GELDERSEKADE
DAMRAK
DAM
OUDEZIJDS VOORBURGWAL
SPUISTRAAT
SINGEL
HERENGRACHT
KEIZERSGRACHT
PRINSENGRACHT
ROZENGRACHT
ROZENSTRAAT
BLOEMGRACHT
ELANDSGRACHT
BROUWERSGRACHT
HARLEMMERSTRAAT
GOUDSBLOEMSTRAAT
LINDENGRACHT
LINDENSTRAAT
BOOMSTRAAT
WESTERSTRAAT
NASSAUKADE
BUYSKADE
MARNIXKADE
MARNIXSTRAAT
LIJNBAANSGRACHT
BILDERDIJKSTRAAT
BILDERDIJKGRACHT
DA COSTAKADE
AMSTEL
Amstel
Nieuwe Herengracht
Nieuwe Vaart
Entrepotdok
Singelgracht
Plantage
Artis
PLANTAGE MIDDENLAAN
Plantage Muidergracht
Hermitage
UvA Roeterseiland
SARPHATISTRAAT
KIT
LINNAEUSSTR
Mauritskade
OOSTERPARK
Oosterpark
Amstelcampus
Onze Lieve Vrouwe Gasthuis
Oost
Leidseplein
LEIDSEKADE
Leidsegracht
KERKSTRAAT
FOKKE SIMONSZSTRAAT
LIJNBAANSGRACHT
DEN TEXSTRAAT
Rijksmuseum
JAN LUIJKENSTRAAT
VONDELSTRAAT
VONDELPARK
OVERTOOM
Overtoom
S100
S106

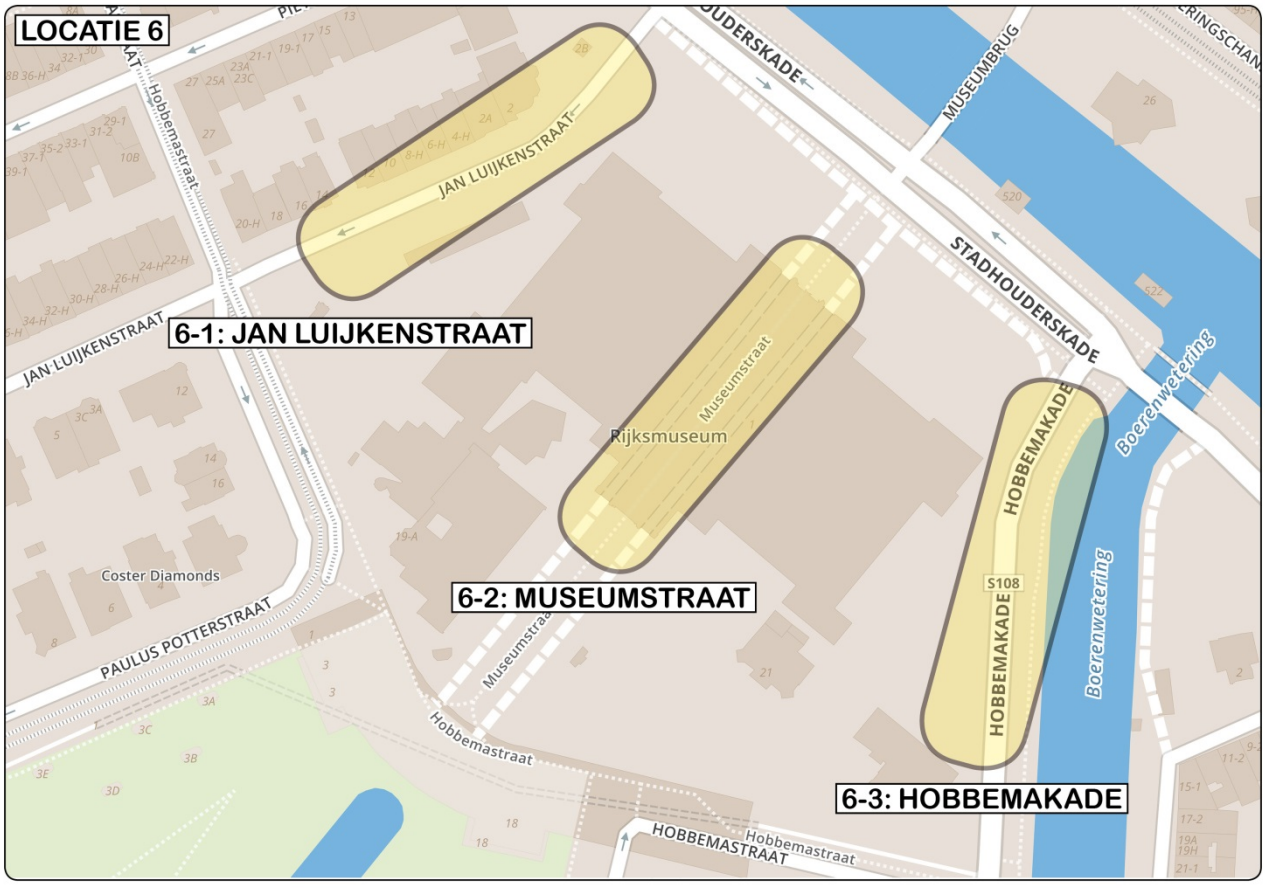

LOCATIE 1
NIEUWENDIJK
NIEUWE NIEUWSTRAAT
Onze Lieve Vrouwesteeg
DAMRAK
Damrak
Beurs van Berlage
C&A
1-1: NIEUWENDIJK
1-2: DAMRAK
BEURSPLEIN
PATERNOSTERSTEEG


LOCATIE 2
Nieuwe Kerk
UvA PC Hoofthuis
SPUISTRAAT
Magna Plaza
NIEUWEZIJDS VOORBURGWAL
EGGERTSTRAAT
NIEUWENDIJK
2-1: MOZES EN AARONSTRAAT
Dam
Raadhuisstraat
Koninklijk Paleis
op de Dam
2-2: PALEISSTRAAT
PALEISSTRAAT
DAM
Madame Tussauds


LOCATIE 4
DE DOELENSTRAAT
PRINSENHOFSSTEEG
OUDEZIJDS ACHTERBURGWAL
Boerensteeg
Oude Hoogstraat
Waalse Kerk
4-1: OUDE HOOGSTRAAT
UvA Bushuis
KLOVENIERSBURGWAL
SPINHUISSTEEG
RUSLAND
4-2: RUSLAND
RAAMGRACHT
UvA
Oudemanhuispoort


LOCATIE 3
SPAARPOTSTEEG
KALVERSTRAAT
GAPERSTEEG
Rokin
ROKIN
DUIFJESSTEEG
3-1: KALVERSTRAAT
3-2: ROKIN
WIJDE KAPELSTEEG


LOCATIE 5
HEILIGEWEG
KALVERSTRAAT
Kalvertoren
5-1: SINGEL (NOORDZIJDE)
SINGEL
SINGEL
SINGEL
SINGEL
Bloemenmarkt
5-2: SINGEL (ZUIDZIJDE)
SINT JORISSTRAAT
OPENHARTSTEEG
WARSSTRAAT
REGULIERSDWARS


LOCATIE 6
JAN LUIJKENSTRAAT
6-1: JAN LUIJKENSTRAAT
Hobbemastraat
STADHOUDERSKADE
MUSEUMBRUG
Museumstraat
Rijksmuseum
6-2: MUSEUMSTRAAT
HOBBEMAKADE
S108
Boerenwetering
6-3: HOBBEMAKADE
Coster Diamonds
PAULUS POTTERSTRAAT
Hobbemastraat
HOBBEMASTRAAT

Appendix H: Comments and remarks of tourism experts

In last section of the questionnaire, the experts were asked to specify for what purposes they could use the study outcomes and if they had any other comments or suggestions. The answers to these questions are given below.

For what purposes could you use the study outcomes?

- Crowd management (answered by 2 experts)
- Get an impression of crowds and hotspots in the city (answered by 2 experts)
- The project '*Balans in the Stad*'
- Spreading tourists and creating new walking routes (answered by 2 experts)
- Publication of the maps on the website maps.amsterdam.nl
- City marketing
- Using it as input for pedestrian traffic models

Do you have any other comments or suggestions?

- Divide the data in different temporal periods to study trends (answered by 3 experts). For example to study the effect of the opening of museums.
- Divide the results per country of origin.
- Enlarge the dataset by incorporating additional sources like Twitter, Instagram and alternative from China and Russia (answered by 2 experts)
- Investigate the reliability of the data
- Is the behaviour of Flickr users different compared to the behaviour of other tourists?
- Are Flickr user young people?
- Compare the temporal distributions per month with the touring car season
- The methodology is very interesting. It could make it easier/faster/cheaper to create tourist monitors
- Share your knowledge! I have setup a working group to gather knowledge about pedestrians. Can I share your presentation with this group?
- Very interesting research and a really nice presentation! My compliments!