TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology

D I P L O M A R B E I T

Discovering Context-Dependent Trajectory Patterns from Social Media

Ausgeführt am Institut für

Geoinformation und Kartographie

der Technischen Universität Wien

unter der Anleitung des Betreuers

Univ.-Prof. Mag. rer. nat. Dr. rer. nat. Georg Gartner

und unter Mitwirkung von

Proj.-Ass. Haosheng Huang, MSc

durch

Sau Chu Connie Kwok

Matrikelnummer 0728030

Promenadegasse 38/3, 1170, Wien

Wien, 17.09.2013
Ort, Datum

_____
Unterschrift (Studentin)

DECLARATION

This thesis contains no materials that have been accepted for an award of a higher degree or graduate diploma in any tertiary institutions. This thesis contains the individual work of mine based on the accepted scientific principles. It contains, to the best of my knowledge and belief, no materials that have been previously published, except where due references have been cited in the text.

Vienna, September 2013

_____

Sau Chu Connie, Kwok

PREFACE

As technology evolves, internet and social media have become integral parts of the everyday routine in our lives. Being an avid social media user myself, I am conscious about the danger of the social media platforms and concerned about the privacy issues. On the other hand, I have always wondered what information I could possibly deduce from these content-rich social media to constructively contribute to science. This is the reason why I have embarked on a study in this particular area. Keeping myself committed to the project has definitely turned out not to be easier than its commencement. I have encountered numerous challenges and difficulties which seem to be inevitable as with any scientific ambition. The news of my grandmother's death in the midst of my research has added to the frustration. With her support and encouragement in mind, I have; however, completed this thesis without hesitation.

This thesis is in partial fulfilment of the requirements of my Master study in Vienna, Austria. It is dedicated to my deceased grandmother. In the end, I hope that it contributes in some way to the scientific community.

ACKNOWLEDGEMENTS

I am very grateful to a number of people who have helped me in the process of my research. First and foremost, I would like to express my gratitudes to Prof. Dr. G. Gartner, my supervisor, for his guidance and valuable comments on this thesis. I am also thankful to H. Huang, MSc for his suggestions for improvements. Finally, I would like to specially thank my friend Günter for his sustained support throughout the ordeal, for proofreading the draft and giving his valuable advice. Their help makes this thesis a reality.

ABSTRACT

Thanks to the advances in Web development, internet usage is not limited to the static web browsing any more these days. With the bloom of social media, it is time to make good use of the freely accessible user-generated content to unmask new information. In recent years, many researchers have actually focused on user-contributed data in their studies. Very few of them have; however, investigated the trajectory patterns in contexts other than the spatio-temporal information. The work presented in this thesis explores the possibility of mining the trajectory patterns in the context of the Flickr user groups (according to "location", i.e., main residence) and seasons (summer versus winter). Studying the trajectory patterns in these contexts is novel. To attain this overall goal, the three largest Flickr user groups, who have visited the city of Vienna, Austria in a period of about 5 years (2007-2011), were selected. Then, the 10 most visited landmarks were sorted out for each group in each season. In the end, the trajectory patterns of each group visiting these top landmarks in each season were analysed. Although the landmarks and trajectory patterns of each group were overall similar, some interesting differences could be uncovered by considering the two contexts of user groups and seasons together.

KURZFASSUNG

Dank des Fortschritts der Web-Entwicklung ist die Internetbenutzung heutzutage nicht mehr auf statisches Web-Browsing beschränkt. Der zunehmende Einsatz von sozialen Medien eröffnet nunmehr die Möglichkeit, den frei zugänglichen Benutzer-erzeugten Inhalt zur Erforschung des Benutzerverhaltens näher zu analysieren. In den letzen Jahren haben sich viele Forscher tatsächlich auf Daten, die von Benutzer beigetragen werden, konzentriert. Jedoch haben nur sehr wenige von ihnen die Trajektorien in anderen Kontexten als den räumlich-zeitlichen untersucht. Die Studie, die in dieser Diplomarbeit präsentiert wird, erkundet die Möglichkeit der Bestimmung von trajektoriellen Mustern in den Kontexten von Flickr-Benutzergruppen (nach Weltregionen) und Saisonen (Sommer versus Winter). Die Untersuchung von trajektoriellen Mustern in diesem Zusammenhang ist neuartig. Um dieses Ziel zu erreichen, wurden die drei größten Gruppen von Flickr-Benutzern ausgewählt, die die Stadt Wien im Zeitraum von ungefähr 5 Jahren (2007-2011) besucht haben. Anschließend wurden die zehn meist besuchten Sehenswürdigkeiten für jede Gruppe in jeder Saison bestimmt. Schließlich wurden die trajektoriellen Muster von jeder Gruppe, die diese meistfrequentierten Sehenswürdigkeiten besucht haben, in jeder Saison analysiert. Durch die Verwendung der beiden Kontexte Benutzergruppen und Saisonen konnten interessante Unterschiede hearausgefunden werden, obwohl im Allgemeinen die trajektoriellen Muster jeder Gruppe ähnlich waren.

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION

With the introduction of World Wide Web (WWW) and the web-based technology (Web 2.0) in the late 20th century, internet users transmit tremendous information every second around the globe. Web 2.0 allows internet users to communicate, interact and stay connected to their counterparts in the virtual environment. In conjunction with the bloom of social media in the last decades, it is high time to make good use of the user-generated content in social media, which is readily accessible, to discover new information. It is prime time to consider and gather community wisdom to uncover new knowledge.

1.1 Research Statement and Objectives

The research investigates whether the trajectory patterns of different social media user groups can be mined using the user-generated content on the social media platform. The novelty of this research is that trajectory patterns are compared in the context of both social media user groups (in terms of users' location), and seasons. While quite a few scholars have investigated the top city landmarks or trajectory patterns using geo-tagged photos from social media, none of them have ever used the same context-based approach as in this study.

The ultimate goal of the study is to discover the travelling patterns of different social media user groups based on the spatial-temporal information of geo-tagged photos, users' geographical location, and season. Flickr is the social medium chosen for such purpose. 3 specific aims have been defined as follows (Further details are outlined in Chapter 4-6):

a.) Specific Aim 1 - To select the top Flickr user groups, grouped according to their geographical location, who have visited or in Vienna, Austria during the period of data collection and compare with the statistical information provided by the Statistik Austria for the same period of time.

b.) Specific Aim 2 - To find out the top 10 frequently visited landmarks for each user group for the entire data collection period and in seasonal context.

c.) Specific Aim 3 - To illustrate the top frequent trajectory patterns for each user group in connection with the findings of Specific Aim 2.

1.2 Definition of Terms

This section gives readers the definition on the key terms used throughout the thesis. Unless stated otherwise, the definition of these terms remains unchanged in this thesis as defined here. They are listed alphabetically as follows:

DEFINITION 1 *Continental Group* is a user group represented by any of the continents, namely, Asia, Europe, North America, South America and Australia (including New Zealand) but excluding the uninhabited Antarctica. This grouping follows the seven-continent model. For the purpose of the current research, only those places involved in the collected dataset are defined here. Country name in the attribute 'Location' of the dataset is the key to the categorization as it is the minimum information provided by the Flickr users (those records without this minimum information are eventually discarded as no categorization is possible). City/State names are not considered as a classifier even if they are provided. The rationale is two-fold.

a.) Not all Flickr users provide city/state names. They use country names instead. Even if, on the other hand, city/state names were used instead of country names, the corresponding country names could be derived easily, but not the other way around. Therefore country names are the only classifiers that are applicable to all records that contain information in the 'Location' attribute.

b.) Even if the city name were adopted as the classifier, the problem of transcontinental

boundaries would remain. For example, the city of Istanbul is located on both Europe and Asia.

Those places, involved in the dataset, require further definition for the present research are the Hawaii island, New Zealand, the two transcontinental countries, i.e. Turkey and Russia. For the transcontinental countries, as long as there is a part of the country connecting to the European continent, it is considered as a European country. New Zealand (NZ), strictly speaking, belongs to the submerged Zealandia continent or the Oceania region. However, it is also common to consider NZ as part of the continent of Australia, so is NZ defined in this thesis. Although the island of Hawaii (a state of North America) is located geologically on the Pacific Plate, which is a tectonic plate different from that of North America and it is geographically part of a chain of the Pacific islands, the country name in the dataset is used as the classifier for the reason mentioned above and for the consistency. North America is the country name provided by the Flickr users, therefore, the island of Hawaii is grouped into the continent of North America.

DEFINITION 2 *Geo-Location/Geographical Location* refers to the longitude and latitude of a location on earth. The longitude and latitude can be represented in form of (x, y) coordinates, where x is longitude and y is latitude. Geo-tagged photos are the photos with such *geo-location/geographical location* information, annotated by GPS coordinates (x, y). The geo-location indicates where the photo was taken.

DEFINITION 3 *Landmark* is a popular location frequently visited by users.

DEFINITION 4 *Owner/User* in the thesis refers to a Flickr user, who owns a photo on Flickr. This *owner* has uploaded his/her photo to Flickr and shares it with the community. A Flickr photo's *owner* is also a Flickr user. *Owner* or *User* is interchangeable in the thesis.

DEFINITION 5 *Trajectory* is a sequence of places visited by a user in a temporal order on the

same day[1].

DEFINITION 6 *Trajectory Pattern* refers to a sequence of landmarks visited by users where the frequency is no less than the minimum frequency. In the thesis, the minimum frequency is set as two. It means that every same sequence, repeating more than twice, forms a *trajectory pattern.*

DEFINITION 7 *Trajectory Pattern Frequency* is the number of users visiting the landmarks in the same sequence.

DEFINITION 8 *User's/Owner's Geographical Location* (or simply *Owner's/User's Location*), is defined in this thesis as user's residence or his/her centre of life during the period of data collection. This information is retrieved and assumed with the attribute 'Location', which is accessible with Flickr's Application Programming Interface (API).

1.3 Structure of the Thesis

After stating the research question, formulating the specific aims and defining terms in this introduction chapter, chapter 2 reviews relevant works in the literature. Chapter 3 describes the data collection process. It consists of the description of the study area, materials, the process and execution of data collection, the content of the datasets. Chapter 4 to 6 details the methodology, results and the discussion of the three specific aims. The last chapter concludes the research findings, states the limitations of the current research and suggests possible future work.

1 It is possible to define trajectory in a time span of multiple days; however, this thesis focuses the trajectory only on the same day.

CHAPTER 2 LITERATURE REVIEW

Over the past two decades, many attempts have been made to extract people's movement patterns in a city with different methods, from using conventional questionnaires to exploiting the innovative technology of Web 2.0. For the latter, research has also focused on tackling the semantic issues in the textual information of geo-referenced photos, dealing with photos of missing geo-reference, and missing annotations. A few research papers on developing the travel recommendation engine as an extension to mining the trajectory patterns have also been published. A review on these studies is outlined in the following sections.

2.1 Semantic Issue, Annotating, and Geo-locating Images

Semantics of textual attributes in online photo repositories has been studied in different respects. For example, Rattenbury et al. (2007, 2007 & 2009) introduced a method called Scale-structure Identification to automatically extract and distinguish event from place semantics in tags of Flickr photos. They compared their method with other well-known burst-analysis techniques for the same purpose and found out that their method outperformed those existing techniques. The scholars suggested also that their method could be applied to any text-based webpages, e.g. blogs, to extract useful information.

Yanai et al. (2009) took advantage of the textual and visual aspect of the geo-tagged photos from Flickr in a new dimension. They proposed the measure of entropy to describe the relation between word concepts (semantics) and geo-locations. High entropy represented a loose relationship and low entropy a close one. Interestingly, they found that word concepts with low image entropy tend to have high geo-location entropy and vice versa (Kawakubo and Yanai, 2009). e.g. if a photo was annotated 'flowers', the visual content of the photo featured only flowers, hence, the image had a low image entropy since the word concepts had a tight relationship with the image. However, the scholars found out that photos, tagged with this

word concept, were not taken at a specific geo-location. Instead, they were taken at any geo-locations over the world. These photos had a high geo-location entropy, representing a loose relationship between the word concept and the geo-location. Their approach helped detecting cultural differences of different countries.

Although a lot of scholars cannot deny the importance of those textual attributes for landmark identification, very few researchers steered their attention to the problem of missing annotations of photos in the image collection. For instance, Popescu and Moëllic (2009) presented a new automatic technique to annotate those unlabelled or untagged photos. Another method was presented by Kalogerakis et al. (2009). These scholars managed to geo-locate the non-geo-referenced Flickr images using image and temporal data. 6 million Flickr photos were used as the training dataset. They claimed that it was possible to geo-locate images without any recognizable landmarks to near-perfect accuracy.

2.2 Modelling Trajectory Patterns Using GPS Tracks, Videos, and Questionnaires

Mckercher and Lau (2008) analysed tourist movement patterns within a destination and categorized the resulting 78 discrete movement patterns into 11 movement styles. Lau and Mckercher (2007) discovered not only the fully independent individual travellers' movement patterns in Hong Kong, but also studied the factors affecting their preferences on itineraries. Lew and Mckercher (2006) modelled the tourist behaviours in connection with the urban transportation and highlighted the factors influencing their movements.

Xia (2007) proposed methods of modelling spatial-temporal movements of tourists at the macro and micro level. Questionnaires, interviews, and GPS tracks were used in the case studies as evaluation of the proposed modelling methods. The latter one was; however, used only at the micro level.

Millonig and Gartner (2008) aimed at developing a pedestrian typology by both qualitative

interpretation and quantitative statistical data. They analysed the spatio-temporal walking behaviours of moving objects, which were in their case the pedestrians, both indoor and outdoor. Other contexts, i.e. gender, age, education level, etc., were also used in the analysis of pedestrians' trajectory patterns.

Zheng et al. (2009) investigated and analysed interesting locations and tourist movement patterns within a given geo-spatial region using GPS trajectories. On top of studying the travelling patterns, Asakura and Iryo (2007) studied the topological characteristics of the travel routes among different tourists by clustering analysis. They used the GPS-enabled devices e.g. mobile phones to track the travelling routes instead of geo-tagged photos.

Another interesting application is the similarity-based retrieval of moving object trajectories in real life or in videos, as described in the paper of Chen et al. (2005). In other words, conventional questionnaires and GPS tracking are not the only options for investigating trajectory patterns. However, data embedded in a remote sensing system are usually noisy. Therefore, thorough data cleaning is needed prior to any further analysis. In the paper, the authors introduced a novel distance function as an alternative to Euclidean distance function to make data cleaning easier.

The above-mentioned scholars investigated the trajectory patterns by using various approaches in other contexts, e.g. age; education; movement styles or the topological characteristics of the landmarks. The results of their investigation help further develop location based services or a personalized travel route suggestion system in the future.

2.3 Landmark Identification and Trajectory Pattern Mining Using Social Media

In recent years, interest has dramatically increased on utilizing the user-generated content from social media for identifying city landmarks and on mining trajectory patterns of moving objects, especially in the fields of computer science, tourism and hospitality, and geography.

For instance, Crandall et al. (2009) studied how to use the location information, visual, textual and temporal features of a huge collection of photos (~35 millions) for extracting representative images of the most popular landmarks in the world. Zheng et al. (2009 & 2009) developed a world landmark recognition engine with object recognition and unsupervised photo-clustering techniques from GPS-tagged photos (~20 to ~21.4 millions) and Web articles. Relevant works carried out by Kennedy et al. (2007 & 2008) focused on the automation of generating the most representative image from a pool of diverse views of the landmarks in the San Francisco Area.

Another research focusing on the visual content of the geo-tagged photos was performed by Avrithis et al. (2010). They proposed an image clustering scheme to cluster and compress the consistent images (representing the same landmarks). With the study of the visual content of the images, they aimed at distinguishing the landmarks (e.g. Parthenon) from the non-landmarks (e.g. a house or a graffiti on a wall, etc.).

Ji et al. (2009) chose yet another approach: they used photos collected from blogs to identify POIs for tourists. They investigated the influence of user location on the preferences for certain city landmarks. They differentiated mainly Asian from European bloggers.

Understandably, resources from the community-contributed social media are not limited to only landmark identification or place extraction. While undoubtedly strong attention has been put on identification of POIs in a specific area or worldwide with different approaches or techniques, some scholars focused more on the duration of visits to tourist spots, people's interaction with urban space using mobile geo-tagging services, and travelling/trajectory patterns.

Popescu and Grefenstette (2009) attempted to answer questions on how long it takes to visit a tourist attraction and what can be done in a city in a day. The automatically deduced duration times of visits were compared to the manual estimates to validate their research.

Humphreys and Liao (2011) used one of the nowadays available mobile location-based geo-tagging services called Socialight to examine how participants interact with their urban spaces in terms of communication about spaces and communication through spaces. With the 'sticky notes' provided by Socialight, participants can share their favourite places in town with their friends at any time anywhere. it could be as local as a coffee shop around the corner of their home or as touristy as any popular landmarks. If their friends who were residing somewhere else and had never been to that town, visited one day, those descriptions on the sticky notes could act as a personal travelling guide. This type of interaction with their urban spaces is, according to Humphreys and Liao (2011) communication about spaces. What the researchers meant by communication through spaces was place-based storytelling and self-presentation through spaces. For example, one could tell his childhood's story that inevitably relates to different places he has visited in the past.

For establishing trajectory patterns of moving objects in relation to city landmarks, numerous researchers in recent years have investigated the Flickr platform in particular. Jankowski et al. (2010), Choudhury et al. (2010) and Yin et al. (2011) have all used the pool of geo-tagged photos on Flickr to construct the photographers' movement in cities. The studies differ only in scope. Jankowski et al. (2010) limited themselves to the Seattle metropolitan area, whereas the other two teams extended their research to several major cities. Furthermore, Choudhury et al. (2010) distinguished tourists from residents in a city by the duration of their stay. Only geo-tagged photos of a tourist were considered in their study. Yin et al. (2011) additionally attempted to define representative trajectory patterns based on similarity as well as to describe how diversify two trajectory patterns could be.

Zheng et al. (2011) studied the regions of attractions (RoA) and tourist traffic flow in these regions by a Markov chain model using GPS-tagged photos on Flickr, while in a previous study they utilized GPS data (Zheng et al., 2009). As an extension of their previous research, Zheng et al. (2012) further investigated the topological characteristics of the travel routes, as in the paper of Asakura and Iryo (2007). Unlike Choudhury et al. (2010), Zheng et al. (2012) distinguished tourists from non-tourists with a probabilistic model, which was based on the

concept of Shannon entropy in Information Theory. Their research also included a statistical method to test the reliability of the travel patterns derived from the noisy collection of the geo-tagged photos.

Kadar (2013) used the geo-referenced photos from the photo-sharing platforms to study the space usage of tourists in two urban cities (Vienna vs Prague), and define the causes of the conflicts between tourists and locals.

Girardin and Blat (2008) carried out their preliminary research on assessing the user-generated content on Flickr from the user-center approach in 3 cities. The research focused preliminarily on 'day trippers'. They assessed the difference of the trajectory patterns between Italian and North American tourists visiting Rome, Italy. They also assessed the human contribution on the accuracy of geo-referencing and annotating photos in relation to the cultural background of the tourists, e.g. comparing German against Spanish for geo-referencing accuracy; North American against European for annotating accuracy in the preliminary stage of their research. Their approach focused on assessing the difference of the trajectory patterns, geo-referencing, and annotating photos between visitors of different selected countries.

Arase et al. (2010) mined the frequent travel trip patterns of different cities based on the trip themes with about 5.7 million geo-tagged photos. The team classified the photos under the trip themes: 'Landmark'; 'Nature'; 'Gourmet'; 'Event'; 'Business', and 'Local'. In the end, they suggested an example application, with which users are able to search for the frequent trip patterns in their interested cities.

Apart from the geo-tagged photos, another way to extract the spatial and temporal information from those location-sharing services like Facebook Places, Foursqaure, etc. is the user-driven footprints, i.e. 'checkins'. This was illustrated by Cheng et al. (2011). The team modelled patterns of human mobility. They made use of the temporal information provided by the 'checkins' to measure the frequency of the users' 'checkins' daily and weekly. They

investigated also the periodic behaviours of those 'checkins'. In terms of spatial travelling behaviours, they measured statistically how far the users travelled from one place to another, whether it was local or global. Factors influencing such behaviours were explored.

Tussyadiah and Fesenmaier (2007) interpreted the travellers' blogs/journals to map their spatio-temporal movement. On top of spatial and temporal information provided by the blogs, other valuable information was embedded describing the travellers' overall travelling experiences (good vs bad), critical moments, knowledge gain, preparation or the information needs. These contents gave additional information on their travelling behaviours at the destination. Based on the findings, the researchers suggested some features for developing and designing the mobile guides for tourists. Similarly, Leung et al. (2011) constructed the overseas tourist movement patterns from 500 trip diaries on 6 websites for the city of Beijing during the Olympic periods.

These researchers made use of different types of social media to identify popular landmark locally or globally and analysed the trajectory patterns in different contexts. These studies allow us to have a better understanding on how human beings interact with their urban space when they travel. This opens a new perspective in tourism and hospitality as well as in infrastructure and urban planning.

2.4 Travel Recommendation System Development

Some researchers analysed not only the trajectory patterns, but also made use of their findings further for the development of a travel recommendation system which suggests tourist destinations and travel routes upon user-supplied query.

Yoon et al. (2012) made such an attempt. They used the user-generated GPS trajectories from travel experts and active residents for landmark extraction and trajectory mining as the base of their landmark search engine, intended for inexperienced travellers and newcomers to a

city. Similarly, Zheng and Xie (2011) proposed a tour route recommendation system by mining GPS traces.

Chen et al. (2009) aimed at firstly finding the most representative view of images for different point of interests (POIs) in a defined geographical area with the computer vision technology and secondly, developing a system to suggest tourist destinations in extension to Kennedy's work (2007 & 2008). In order to do that, they developed an algorithm to automatically cluster geo-tagged photos collected from social media into groups based on the location information. With the tag metadata, the algorithm then identified popular tags for each cluster for landmarks in that area. Afterwards, it further utilized the set of photos of each POI to identify the most representative image for it. Based on the geo-tagged web photos and hence the representative tags and images of each POI, the system recommends tourist destinations upon user's query. Cao (2010) demonstrated the similar approach to develop a worldwide tourism recommendation system using Flickr photos.

The online travel recommendation service developed by Lu et al. (2010) helps its users, according to their preferences, to plan travel routes worldwide. The team leveraged 20 million photos from the photo-sharing platform Panoramio.com and 200,000 descriptive travelogues. Similarly, Yin et al. (2012) used the resources from the same platform to develop an itinerary recommendation system. In the same manner, Kurashima et al. (2010) designed a travel route recommendation system based on the Flickr photos. Using Web information. Zhang et al. (2008) and Kawai et al. (2009) developed a tour recommendation system, which focused on the visibility of scenic sights between two tourist spots.

Cheng et al. (2011) aimed at constructing a probabilistic personalized travel recommendation model. They utilized not only the spatio-temporal information of the Flickr photos, but also the demographic groups (by a total of nine people's attributes) to extract the landmarks and travel paths. The people's attributes included gender (female vs male), age (kid, teen, middle-aged or elderly), and race (Caucasian, African or Asian). People's attributes in the photos were automatically extracted by facial detector. The research was performed for eight major

cities; e.g. Hong Kong, with about four million photos.

Another research carried out by Majid et al. (2013) focused on developing a context-aware personalized landmark recommendation system. They extracted the travelling preferences of Flickr users from their geo-tagged photos which allowed the system to suggest landmarks to target users with similar preferences. The landmark recommendation of the proposed system was based on the user-specific travelling preferences augmented with the current context (i.e. weather via online weather Web services) of the target city.

The studies mentioned in this section focused on developing a travel recommendation system to suggest popular landmarks or travel routes for tourists. The travel recommendation systems enable efficient trip planning, which was in the past a time-consuming task, especially, when travel criteria are complex.

To sum up, only few scholars in the literature investigated the trajectory patterns from social media in connection with contexts. Although some of them did analyse the trajectory patterns in the context of trip themes (Arase et al., 2010), selected countries (Girardin and Blat, 2008), and gender, age and race (Cheng et al., 2011) using geo-tagged photos (see Sections 2.3 and 2.4), none of them has studied the trajectory patterns with both contexts of Flickr user-groups and seasons. Combining both contexts of Flickr user-groups according to their geographic location, and seasons to investigate the trajectory patterns is the novelty of the present study.

CHAPTER 3 DATA COLLECTION AND PROCESSING

This chapter presents

- the scope of the current research,
- the brief description of the study area and the materials utilized for the thesis,
- the execution of data collection on Flickr photos and information of Flickr users, and
- the statistics of the collected data.

These are the basis of achieving the three Specific Aims, including the overall goal of the thesis, to discover the trajectory patterns in context of user groups and seasons.

3.1 Scope

The present research relies on spatial-temporal information as well as the textual attributes of photos uploaded by Flickr users to identify the top city landmarks and the trajectory patterns of those visitors. Textual attributes refer to tags, descriptions and titles the Flickr users input for their photos. Geo-location information, which are the latitude and the longitude, of the photos is considered as the dominant factor for landmark identification, while useful textual attributes are used as verification. As long as both geographical information and textual attributes of the photo indicate a specific city landmark, the photo would then be counted as one instance of that city landmark.

It is not the scope of the research to either study the semantic issues of textual attributes or the intention of the users for these attributes. The visual content of the photo is also not the consideration of the research, unlike the researches from Crandall et al. (2009), Kennedy et al. (2007 & 2008) and Chen et al. (2009). Hence, for instance, if a photo was taken at Stephansdom Cathedral, it was, nonetheless, about a bird standing in front of it, it would be

classified in the corresponding group of 'Stephansdom'.

The research studies the trajectory patterns of the Flickr users in a single day. Although investigating the trajectory patterns in a time span of multiple days is possible, it is out of the scope of the research. The same is true for how long those visitors stay at each place.

3.2 Testing Area and Materials

The city of Vienna in Austria is the area of interest. OpenStreetMap (OSM) (www.openstreetmap.org), which is a freely editable map founded in the UK, was used to extract the approximate geographical coordinates of Vienna. One of the features OSM offers is the map export with user-defined boundaries in JPEG, PNG, SVG and PDF formats. From this export function, the approximate maximum and minimum values of both latitude and longitude were obtained.

The selected social medium is Flickr (www.flickr.com), which was developed by Ludicorp Research & Development Ltd. and is owned by Yahoo! Corporation. It was first launched in February, 2004 and is now presented in over 10 different languages. By August 2011, Flickr had a host of 6 billion images, as reported by the site. Apart from the functionality of uploading, storing and managing photos, blog is one of its features. Besides, Flickr provides tools to assist the software developers to develop their own application with photos and data. For this research, the geo-tagged photos of Vienna uploaded on Flickr by its users during the period of 1 January, 2007 to 31 August, 2011[2] were used (see Section 3.3.1 for details).

Python 2.7 was used for data collection. It is a high-level programming language developed by Python Software Foundation as an open source programming language. It can be used as a scripting language or in a non-script context. Python can be accessed by its interpreters, which are compatible with many operating systems. Notepad++ were used for writing the

2 This period of data collection is addressed as '*the entire period of time*' throughout the thesis.

scripts and processed by the interpreters.

A simple XML-to-CSV converter has been used to convert the collected metadata of Flickr photos in Extensive Markup Language (XML) data format to Comma-Separated Values (CSV) format for further data management and analysis. The database management system called PostgreSQL 8.4 was used for joining the dataset of Flickr photos and that of the photos' owners together.

Waikato Environment for Knowledge Analysis version 3.6.9 was used for clustering the Flickr photos for city landmark identification. Other examples of WEKA applications, can be found in Bogorny et al. (2006, 2007 & 2009), Oliveira and Baptista (2012), and Oliveira et al. (2012). WEKA is machine-learning software, developed at the University of Waikato, New Zealand and written in JAVA programming language. It is an open source software, made available under the GNU General Public License. The software supports a few standard data mining tasks, namely, data preprocessing, clustering, classification, regression, visualization and feature selection with a user-friendly graphical interface. WEKA also allows connection to Structured Query Language (SQL) databases with Java Database Connectivity.

WEKA's main user graphical interface (GUI) is called 'Explorer', which was used for the current study. As an alternative, the software can also be accessed by an elevated command prompt with command line. Its graphical interface offers several components, called '*Preprocess*', '*Classify*', '*Associate*', '*Cluster*', '*Select attributes*', '*Visualize*'. Although WEKA's native file format is Attribute Relationship File Format (ARFF), which is a text file format, '*Preprocess*' allows users import a CSV file or from a database. '*Classify*' provides both classification and regression algorithm. '*Associate*' offers association rule learners to identify the inter-relationship between attributes in the datasets. '*Cluster*' features various clustering algorithms, such as, k-means, DBSCAN, OPTICS, etc. '*Select attributes*' gives the most predictive attributes in the dataset. '*Visualize*' presents each pair of attributes in the dataset with a scatter plot matrix. Users can possibly choose individual scatter plot matrix for visualization. For the purpose of this research, only '*Preprocess*', '*Cluster*' and '*Visualize*' have

been used.

Quantum GIS (or simply QGIS) version 1.7.4 is another open source desktop software being used for the present research. It has the functionality of data viewing, editing, managing and analysing It supports both vector and raster data formats, e.g. MapInfo or shapefiles, etc. for vector file formats whereas bitmap, jpg or tiff, etc. for raster file formats. On top of that, Web Map Service and Web Feature Service equipped in QGIS supports data from external sources. QGIS also allows access to other database systems, e.g. PostgreSQL. QGIS integrates with other GIS packages, such as, GRASS, PostGIS and MapServer. Plug-ins, written in Python, extend the functionalities of QGIS further, e.g. plug-ins of geo-code and reverse geo-code using Google Geo-coding/Reverse Geo-coding API. In this research, the major use of QGIS was to visualize the clusters of Flickr photos after being processed in WEKA and for graphical presentation. Reverse geo-coding and OpenLayers Plug-ins, and Web Map Service were also utilized for getting the location details, overlaying google street map as base map and obtaining the boundary of Vienna city respectively.

Apache OpenOffice (AOO) 3.4.1, which is distributed under the Apache License, was chiefly used for its spreadsheet component. It features also a presentation application (Impress), a word processor (Writer), a formula editor (Math), a drawing application (Draw) and a database management application (Base). It has its own native file formats for all of its components. However, import and export allows the interchange of Microsoft's file formats. The spreadsheet functionalities of data interpretation, data editing, data filtering, data analysis and data presentation were utilized (in charts or tables).

For the Specific Aim 1, statistics of arrivals, indicating the overnight accommodation has been referenced with Statistik Austria via its website (www.statistik.at). Statistik Austria is an official statistical office, providing statistical data in various aspects of public matter. They are in the realm of 'People & Society', 'Economy', 'Energy, Environment, Innovation, Mobility', 'Wealth & Progress' and 'International' issues. The relevant statistical data for the Specific Aim 1 can be found in the section of 'Economy' and its sub-section called 'Tourism',

under which a sub-section titled 'Accommodation' leads users to its statistical database (www.statcube.at). There are two major ways to interpret the accommodation statistical data, one is by year whereas another is by season (see Section 4.1 for further details).

3.3 Data Collection

3.3.1 Spatio-Temporal Information and Textual Attributes of Flickr Photos

Obtaining the metadata of Flickr photos was done basically by Flickr API. In order to use the service, registration of an account (free) with his/her Yahoo!ID, Facebook ID or Google ID is required. In 'The App Garden', comprehensive API methods for different domains are listed, for instances, 'People', 'Place', 'Photos' or 'Blogs', etc. Details can be found in the API documentation via http://www.flickr.com/services/api/. This research mainly used the APIs 'flickr.photos.search' and 'flickr.people.getInfo' (see Sec. 3.3.2) for collecting the necessary metadata.

The API method of 'flickr.photos.search' was used to retrieve the photo's ID, tags, latitude and longitude at which the photo was taken and its timestamp, etc. A number of parameters has been defined prior to data collection. They are listed in the following table (Table 1).

Parameter	Value	Description
bbox	16.18,48.115,16.58,48.325	Bottom left corner and top right corner of the boundary box of the study area. They are in the order of min. longitude (x); min. latitude (y); max. longitude & max. latitude
min_taken_date	2007-01-01	Photos with a taken date later than this date is retrieved. It is in the order of year, month & date
max_taken_date	2011-08-31	Photos with a taken date earlier than this date is retrieved. It is in the order of year, month & date
max_upload_date	2011-09-08	The latest date the photo is uploaded
media	photos	Photos vs videos
sort	date-taken-desc	The order in which the returned photo is sorted, according to date-taken or date-uploaded in descending or ascending order.
per_page	500	The number of the returned photos per page. Default is 1. The maximum value is 500.
extra	description, geo, tags, url_z	Description, tags, geographical location (latitude & longitude), Uniform Resource Locator (URL) and the size of the photo

Table 1 Parameters for the API method 'flickr.photos.search'

The maximum and minimum value of the photo-taking date are equivalent to the period of data collection, which is about five years. A defined value for the maximum value on the photo-uploading date was necessary to refrain from any inconsistent photo counts. 'url_z' presents the photo of a medium size of 640 pixels (on the longest side) on a specific URL. With this set of parameters, only geo-tagged photos within the boundary box were retrieved. Photos in Blogs on Flickr were not considered as well.

'API Explorer' was used for the execution of the API method with the above-mentioned parameters. An example of the search results is attached as Appendix A. The search results were first collected in XML format, whereas JSONP, JSON and PHP serial were also available. The corresponding search results are provided with a URL underneath the 'API Explorer' as well. With the URL the first 500 results of the returned photos were presented in a separate window explorer in XML format and saved. The XML file of the first 500 results was then converted to the CSV format using a simple XML-to-CSV converter. The results could then be read by any regular text processors or spreadsheet. The purpose of this conversion process was to present the data in a more manageable way. Further analysis on the dataset could be done with spreadsheet thereafter.

As mentioned before, the limit of each call for results is 500 photos per page and there are a total number of more than 150,000 photos during the designated period. Hence, the same process has been executed repeatedly until all relevant photos were retrieved. The data were stored in five different files of manageable size according to the year. They were then converted to CSV files and eventually saved in spreadsheet format for further data management and analysis.

3.3.2 Location Information of Flickr Users

Metadata of photos and those of users are stored on Flickr separately. The retrieval of the metadata of the photos with the API method 'flickr.photos.search', as described above, does not provide any detailed information of their owners, except the owner ID. In order to retrieve other information, (especially the location information of the photos' owners), another API method called 'flickr.people.getInfo' has been utilized. Unlike the API method 'flickr.photos.search' for photos, 'flickr.people.getInfo' returns only 1 result at a time with an owner's ID. With more than 150,000 collected photos, it would be undoubtedly laborious to collect the information of their owners one by one. So as to collect the owners' metadata more efficiently, a programming script, written in Python 2.7, have been used (Appendix B).

Besides the codes, three packages, namely, 'flickrapi', 'xlrd' and 'xlwt' have been downloaded and installed for the programming script. 'flickrapi' allows access to the metadata on Flickr through Flickr API. 'xlrd' reads data from the spreadsheet files row by row whereas 'xlwt' writes data to a new spreadsheet file. In other words, these three packages enable reading the metadata from the previously collected Flickr photos in the spreadsheet files, fetching the metadata of each corresponding owner via the Flickr API method 'flickr.people.getInfo' and writing the returned results to a new spreadsheet. A Flickr API key was required for the purpose and obtained through Flickr's website.

In order to get the information of the owner of each photo with the script, a better understanding of the Flickr data presented in XML data format was indispensable. XML document is both human-readable and machine-readable. It is a textual data format, which presents the data basically in a hierarchical way, called ElementTree. Each entity or object is presented as an element and each element can have several attributes. 'elementName.text' can be used to get the value of an element. The following is a simple example to illustrate this:

```
<person nsid="12037949754@N01" ispro="0" iconserver="122" iconfarm="1">
        <username>bees</username>
        <realname>Cal Henderson</realname>
        <mbox_sha1sum>eea6cd28e3d0003ab51b0058a684d94980b727ac</mbox_sha1sum>
        <location>Vancouver, Canada</location>
</person>
```

Example 1 ElementTree of XML data format (extracted Flickr data)

'person', 'username', 'location' and likewise are elements, while 'nsid' is an attribute of the element of 'person'. 'username.text' or 'location.text', etc. is used to extract the value of the corresponding element, which is 'bees' or 'Vancouver, Canada' in this case, respectively. Further details on the programming codes with Python and XML will not be discussed here. Further information can be found under http://docs.python.org/2/library/xml.etree.elementtree.html

Some special situations of collecting the owners' information has been taken into account in

the scripts. Not that every owner of a photo input a complete set of their data, in other words, when an owner did not input a value, say, his/her real name, the XML element was omitted during the data retrieval process. Another issue was that not that every photo had a matching returned result for its owner although every photo has an unique owner's ID in the photo dataset. The reason was due to the privacy setting of the owner's profile. If the owner has set its profile to 'private', no record could be retrieved through the Flickr API. 'User not found' was returned as the result. The script was tailored to handle both exceptions. For the latter issue, those photos without returned owners were eventually considered as irrelevant for further analysis and discarded.

The dataset of Flickr photos and the dataset of the owners were then joined in the database management system PostgreSQL 8.4. It allows import of CSV files. With the executions of some Structured Query Language (SQL) statements, two datasets were combined into one and exported in spreadsheet format for further analysis.

3.3.3 Statistics on Collected Datasets

There were a total number of 154,343 Flickr photos collected in the photo dataset. However, 25 of them were duplication. These duplicated photos have been eliminated. The following table (Table 2) gives the statistics on the number of collected photos and unique owners for five periods of time.

Period of Time	Number of Photos	Number of Unique Owners
2007-01-01 to 2008-02-29[3]	29980	1012
2008-02-29 to 2009-01-05	30939	991
2009-01-05 to 2010-01-18	31520	1094
2010-01-18 to 2011-01-31	39327	1425
2011-01-31 to 2011-08-31	22552	918
Total	154318	--

Table 2 Statistics on the number of collected photos and unique owners from 1 Jan, 2007 to 31 Aug, 2011

The total number of unique owners of the photos, presented in Table 2, is the total number of unique owners of the photos for *each* period. It is not useful to sum up the unique owners in Table 2 since they could have uploaded photos in different periods. The total would only give the duplicated counts of some photo owners for this entire period. The total number of the unique owners is presented separately in Table 3.

After joining both datasets of photos and owners, there were a total number of 153,608 records, among which there were 10 photos with invalid values of latitude and longitude, e.g. '0,0'. Further elimination of photos was consequently necessary. The discrepancy of the number of records after joining the datasets compared to that of Table 2 is due to the inaccessibility of owners' profile, as mentioned in the previous section. Table 3 compares the resulting total number of photos and unique owners to that of the raw data for the period of 1 Jan, 2007 to 31 Aug, 2011.

3 The irregular period of time aims to make the file size more manageable and at the same time contain as many records as possible so that it would not end up with too many files.

	Total number of photos	Total number of unique owners
Raw data (via Flickr API)	154318	4286
Data after elimination of invalid records	153598	4242

Table 3 Comparison on the number of photos and unique owners in raw datasets and after data management

The table reveals that there were a total number of non-duplicated Flickr photos collected, corresponding to 4286 unique owners. After the deletion of some invalid entries, 153598 photo records remained. Again, the deletion of invalid photos involves 710 records, corresponding to 44 unique owners, of which the information of the owners could not be accessed with Flickr API, plus 10 photo entries with invalid latitudes and longitudes

The values of the boundary box used for data collection gave a rectangular extent, but the boundary of the city of Vienna is irregular. Those data, which fell out of the boundary of Vienna, were removed by the geo-processing tool 'clip' in QGIS. Subsequently, 149,489 total number of photos belonging to 4,199 unique photo-owners remained in the dataset.

CHAPTER 4 SPECIFIC AIM 1: USER GROUPING

This chapter presents

- the methodology in achieving the Specific Aim 1,
- the ranking of the top user groups (according to the users' location) and comparison with the statistics of Statistik Austria, and
- the discussion of the results.

4.1 Methodology

The goal of the Specific Aim 1 is to select the top user groups (tourists and locals inclusive, see also Section 6.3 and Chapter 7) out of five continental groups plus the group of Austria. Data were compared to that of Statistik Austria. Only top user groups of significant group size were used for further analysis.

The dataset, consists of both photos and their owners' metadata, was categorized according to the photo owners' location, which was indicated by the attribute 'location' in the dataset (Definition 8). The attribute 'location' specifies both city name and country name in most cases. The data were first categorized into six groups, namely, Asia, Europe (without Austria), North America, South America, Australia & New Zealand and Austria. Africa was not represented in the dataset. An extra group called 'invalid' handled any data with either no values or ambiguous values. Considering visitors or residences from Austria, who visited or were in Vienna, might exhibit different travelling behaviours than visitors from the rest of Europe. Therefore, Europe was divided into Europe (without Austria) & Austria. Further distinction between Austria and Vienna was unfavourable for three major reasons, listed as follows:

a). There were 3627 records of Austria without specific city name.

b). The greater Vienna was considered as the 'grey zone' area. There were a few records of the greater Vienna, e.g. Korneuburg, Klosterneuburg, Baden etc. Since these locations are within commute distance, the residences could have travelled daily to Vienna for work. This crowd might have a different travelling pattern in Vienna. It would be unjustified to assign them to either of the groups.

c). The sample size of 'Austria without Vienna' was not large enough, even if Austria would be divided into the groups of 'Austria without Vienna' and 'Vienna' in the first place. The data showed that there would only be 96 unique owners with 4423 photos, which was quite similar to the sample size of 'Asia', consequently, the group would have been disregarded for further analysis.

The three geographical overland boundaries, involving five continents, are defined as follows:

- between Asia and Africa (dividing Afro-Eurasia into Africa and Eurasia): at the Isthmus of Suez
- between Asia and Europe (dividing Eurasia): along the Bosphorus, the Caucasus and the Urals (historically also north of the Caucasus, along the Kuma–Manych Depression or along the Don River)
- between North America and South America (dividing the Americas): the Isthmus of Panama

The categorization of user groups was basically done in the spreadsheet with filtering functions. During the categorization, the issue of different foreign languages arose (see Chapter 7). Fortunately, the majority of the country names, which is the key to categorizing user groups, was in English. After the categorization, the groups were ranked by the total number of unique owners, but not by the total number of photos.

The statistics on the number of arrivals (with an overnight accommodation) are arranged by Statistik Austria according to season or calender year from 1974 to April, 2013 (as of writing

the thesis) for 84 countries of origin as well as the 9 states of Austria. Statistics for each month of the years are also available. Season includes two half-years. According to Statistik Austria, summer season starts from the first of May to the end of October, while winter season begins from the first of November to the end of April (see also Section 5.1). For the purpose of the Specific Aim 1, the calender years and months were referenced and grouped in the exact way to match the periods of time of the dataset. Since the statistics on the number of arrivals (with an overnight accommodation) are given in terms of country instead of continent, regrouping the data from Statistik Austria to align with the dataset was necessary prior to comparison.

4.2 Results

The aim of this Specific Aim is first to sort out the most represented Flickr user groups and compare the results with the dataset from Statistik Austria. The valid Flickr data were aggregated into 6 user groups in accordance with the location information of the Flickr-Photo owners. Table 4 summaries the data distribution of the Flickr dataset after the grouping, in absolute number and in percentage.

	Flickr Dataset (1 Jan, 2007 – 31 Aug, 2011)					
Group	Number of Photos			Number of Unique Owners		
	absolute	in % (up to 2 d.p.)	Ranking	absolute	in % (up to 2 d.p.)	Ranking
Asia	2382	2.80	4	84	4.14	4
Australia & New Zealand	643	0.76	5	31	1.53	6
Austria	46187	54.23	1	628	30.94	2
Europe (w/o Austria)	18546	21.78	2	976	48.08	1
North America	16921	19.87	3	277	13.65	3
South America	485	0.57	6	34	1.67	5
Subtotal	85164	100		2030	100	
'Invalid'[4]	64325			2169		
Total	149489			4199		

Table 4 Summary on the data distribution of the Flickr dataset

Only 85164 photos of 2030 unique owners were considered. About 42% of the photos, corresponding to about 52% of unique owners, were not useful for the grouping. It is attributed to the facts that the required location information was either left empty in the owners' profiles or non-comprehensive, e.g. somewhere on the earth or in the space, etc. and the photo owners' profile is not accessible. The continental groups of Africa and Antarctica are not presented in the dataset.

The top three ranks in terms of number of uploaded Flickr photos and unique owners are held by the groups of 'Austria', 'Europe excluding Austria' and 'North America'. While the Austrian group has uploaded most of the photos (46,187 equivalent to 54,23%), it is ranked only second for the number of photos' owners In comparison, 'Europe excluding Austria' has about 2.5-fold less number of photos, but the total number of unique owners is the highest among the study groups, which amounts to 976 (48.08%). 'North America' ranks the third both in terms of number of photos and number of unique owners. Figure 1 depicts the data

4 'Invalid' are those data with either empty or ambiguous values of the 'location' attribute of the Flickr users.

distribution of these top three groups graphically in QGIS.

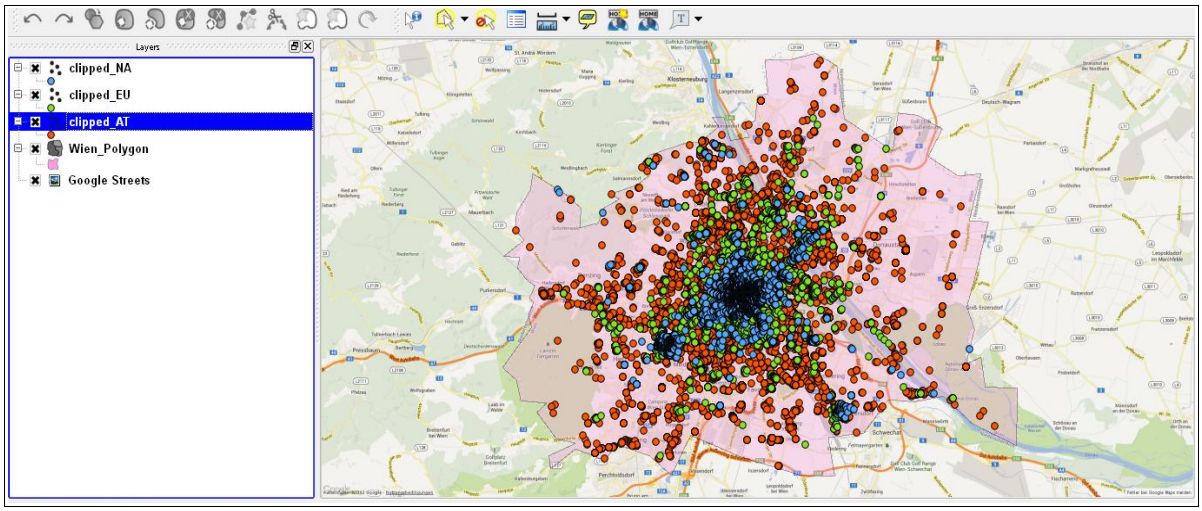

Figure 1 Data distribution of the 3 most represented user groups in QGIS

Each photo record was plotted with its longitude (x-coordinate) and latitude (y-coordinate) in QGIS. Different colours were assigned to different groups. The most prominent feature is that the North American photos (in blue) are concentrated in the inner city of Vienna, whereas the photos taken by Austrians (in orange) are much more scattered and can be found in greater distance from the inner city. The European distribution pattern lies in between.

The data of these three largest user groups were further analysed to define the top 10 landmarks in the Specific Aim 2 (Section 5.2). The data are composed of 81,654 photos (95.88% of all valid photos) corresponding to 1,881 users (92.66% among those valid users) (Table 5). Comparatively, the rest of the user groups are of insignificant sample size. Their total number of photos and unique owners are all less than 5%. These records were discarded.

	absolute	in % (up to 2 d.p.)
Number of Photos	81,654	95.88
Number of Unique Owners	1,881	92.66

Table 5 Number of photos and unique owners for the Specific Aim 2

The second goal of this Specific Aim is to compare the result of the Flickr dataset to that of Statistik Austria for the same period (1 Jan, 2007 – 31 Aug, 2011). Table 6 presents the comparison in both absolute number and in percentage and ordered by the absolute number of arrivals (with an overnight accommodation).

Group	Number of Arrivals (from Statistik Austria)		Number of Flickr-Photo Owners	
	absolute	in % (up to 2 d.p.)	absolute	in % (up to 2 d.p.)
Europe excluding Austria	12,004,592	58.41	976	48.08
Austria	5,043,112	24.54	628	30.94
Asia	1,622,653	7.90	84	4.14
North America	1,269,731	6.18	277	13.65
South America	271,093	1.32	34	1.67
Australia & New Zealand	248,830	1.21	31	1.53
Africa	92,505	0.45	--	--
Total	20,552,516	100	2,030	100

Table 6 Comparison of the photo-owner distribution of the Flickr dataset and the number-of-arrival distribution of Statistik Austria

According to Statistik Austria (Statistik Austria, 2003), 'Arrival' is defined as every person who checks-in to an accommodation establishment for at least one night. No matter how many nights he/she spends in that establishment, this 'Arrival' will only be counted once. Due to the prerequisite of an overnight accommodation, day-trip visitors are excluded. It is worth noting that the same person could possibly have multiple check-ins to different accommodation establishments during his/her visit. Such 'Arrival' would be counted repeatedly instead of being regarded as an individual guest. The number of arrivals is hence

an over-estimate. Another limitation of the Statistik Austria dataset is that 26% of visitors to Austria were not included in the statistics. 23% of them stayed with their relatives, friends or acquaintances, while the remaining 3% stayed in their secondary residence in Austria. Although documented for the whole Austria, a similar scenario can be assumed for our study area, the city of Vienna.

The ratio between the number of Flickr photo owners to the number of arrivals in Statistik Austria is about 1: 10,000. In spite of this large difference in absolute numbers, the two datasets are comparable. Table 6 reveals that more than 50% of the arrivals in Statistik Austria were from Europe (excluding Austria), while almost 50% of the Flickr-photo owners represented that group. Almost 25% of the arrivals in Statistik Austria and just over 30% of the Flickr-photo owners were from Austria. There were about 1.3-time more (equivalent to about 350,000 arrivals) arrivals from Asia than from North America according to Statistik Austria. In contrast, the Asian group in the Flickr dataset was about 3-fold less represented than the North American group. The presence of the African group in the dataset of Statistik Austria is yet another distinction between these two datasets.

In the Statistik Austria dataset, Europe (without Austria) ranked first and Austria ranked second in terms of number of arrivals to the city of Vienna (Table 7). The same ranking was found in the Flickr dataset for the number of photo owners, while the the third rank is different between the datasets. 'North America' is the third largest represented group in the Flickr dataset while 'Asia' is for Statistik Austria.

Ranking	Photo Owners in Flickr Dataset	Arrivals in Statistik Austria
1	Europe excluding Austria	Europe excluding Austria
2	Austria	Austria
3	North America	Asia

Table 7 Comparison of ranking for the most represented continental groups between the Flickr dataset and Statistik Austria

4.3 Discussion

In this Specific Aim, 6 user groups were identified based on continental affiliation. It is evident that the group size of Europe excluding Austria, North America, and Austria is much larger than that of Asia, South America, and Australia and New Zealand in absolute number (Table 6). Only the three largest groups were selected for further analysis.

The Flickr groups cannot be regarded as statistical samples and this type of social media might only be used by some more defined groups of people, e.g. younger generation, technology-oriented geeks or photographers, etc. Therefore it might not well represent the underlying population of visitors of Vienna. In order to investigate this further, the Flickr dataset was compared with the comprehensive dataset of Statistik Austria containing overnight-arrivals to Vienna during the study period. Interestingly, the three top continental user groups in Flickr represented the equivalent continental groups of Statistik Austria by the same ratio of 1:10,000. Asian visitors were clearly under-represented in the Flickr dataset and outnumbered by North Americans. On the other hand, in Statistik Austria, they constituted the third largest visitor group. This difference might reflect the less popularity of Flickr in Asia as a photo-sharing platform as compared to North America.

Having these differences to the underlying population in mind, it is not the purpose of the thesis or this Specific Aim to predict the trend for a larger population based on the Flickr user groups. For example, it should not be assumed that the distribution of the entire Flickr user community in each group is the same, as depicted in Table 6, if the study period would be extended.

CHAPTER 5 SPECIFIC AIM 2: LANDMARK IDENTIFICATION AND RANKING

This chapter presents

- the methodology in achieving the Specific Aim 2,
- the top 10 visited city landmarks, and
- the discussion of the results

5.1 Methodology

The goal of the Specific Aim 2 is to find out the top 10 frequently visited city landmarks of each group selected from the Specific Aim 1 for the entire period and two seasons, summer and winter. The QGIS software was used for visualizing the data point and the clustering results. The data for each user group was imported with the function of 'adding delimited text layer'. The CSV file of each group, which was the result of the Specific Aim 1, was added into QGIS as an individual vector layer. There were a total number of three vector layers.

During the process of data import, there was a problem with the inconsistency of the data format of latitudes and longitudes in the raw dataset. This issue was then handled by changing the data type to 'text' instead of 'number'. The data were read as text instead of number during data import and QGIS presented the extent of the study area properly.

In order to get a base map for the study area, the plug-in named 'OpenLayers plug-in' was utilized. This plug-in offers a few open source map layers to be added to QGIS, e.g. OpenStreetMap layers, Yahoo map layers or Google map layers, etc. Google streets layer was added as the base map so that the resolution at street level was attained. The Google Coordinate References System (CRS), Google Mercator EPSG: 900913 has been added in 'Project Properties'. The capability of on-the-fly CRS transformation was enabled to allow real time CRS transformation to align with the CRS of the dataset, which is WGS84 EPSG:

4326.

The delineation of city landmarks was based on the data clustering process, as described in the paper of Yin et al. (2011) and Majid et al. (2013). An alternative approach is the pre-definition of landmark areas (Choudhury et al., 2010), which has not been used for the present research (see Section 5.3). Instead, the data clustering algorithm called Density-Based Spatial Clustering of Applications with Noise (DBSCAN) was applied (used also by Majid et al., 2013). It is a density-based algorithm which aggregates the data into spatial clusters based on the estimated density distribution of the corresponding nodes. two mandatory parameters, namely, epsilon (ε) or eps and minPts, has to be defined before the data clustering. Epsilon defines the radius of the circle of search from a data point, while minPts defines the minimum number of data points in a cluster. OPTICS is a variation of the algorithm DBSCAN, which omits the requirement of the user-defined epsilon parameter. It replaces the parameter with a maximum radius of search. The following diagrams (Figure 2 & 3) illustrate the basic idea of the data clustering of the algorithm (Ester, et al., 1996).

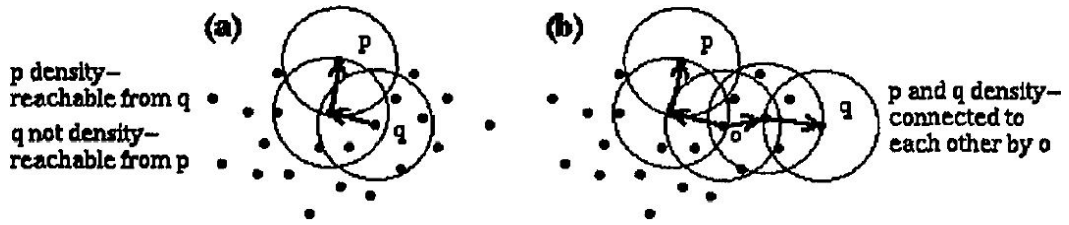

Figure 2(a) & (b) Density-reachability and density-connectivity

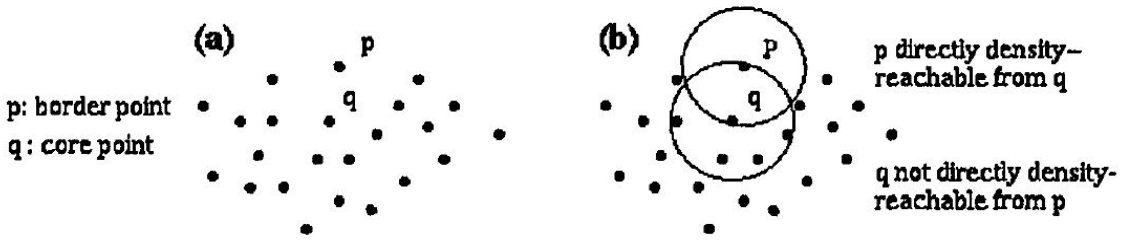

Figure 3(a) & (b) Core points and border points

The algorithm is based on the concept of density-reachability of data points. If points, say, *p* and *q* are density-reachable and the cluster fulfils the requirement of the minimum points, these points will be in the same cluster. However, this is not the same as the meaning of *directly* density- reachable. A point, *p*, is *directly* density-reachable by a point, *q*, if and only if *p* lies within the given distance (radius of circle of search) from the core point *q* (Figure 3b). From the figure 2 (a), it is obvious that the point *p* and *q* are density-reachable from *q*, but not *directly* density-reachable. A cluster will be formed only if the minimum point in the circle of search is reached. Another characteristic of the algorithm is that the density-reachability is asymmetric. If the algorithm starts with a random point, *p*, which lies on the edge of a cluster (from the point *q*), no cluster will be formed for the core point *p* if there are not enough data points around it. However, the point *p* is not necessarily classified as noise, because it can be density-reachable from the data point *q* (Figure 3). The density-connectivity is symmetric (Figure 2b). The core points *p* and *q* are density-connected by the core point *o*.

The algorithm kicks off at an arbitrary point, which has been visited and searches for the ε-neighbourhood, if there are as many points as the minPts requires, then a cluster will be formed. Otherwise, the points will be classified as 'noise'. However, this 'noise' might still be part of a cluster, as discussed above. If a point is the dense part of another cluster, then its entire ε-neighbourhood will be added to this cluster and forms a larger cluster. The process continues until the density-connected points are thoroughly found. The process repeats again with another unvisited point and continues in the same way to form clusters or noises until every data point in the dataset has been visited.

Like many other clustering algorithms, there are both advantages and disadvantages. DBSCAN is of no exception. The following table lists the advantages against the disadvantages of the algorithm (Table 8).

Advantages	Disadvantages
- It does not require the definition of the number of clusters, unlike another popular clustering algorithm, *k-means*	- It cannot deal with the huge differences in data density of the dataset well enough. The epsilon-minPts combination is not sufficient for all clusters.
- It allows any arbitrary-shaped clusters (Figure 4).	- The quality of the clustering results depends on the distance function of the epsilon parameter. Euclidean distance is used in DBSCAN for distance measure. Note that this distance measure is almost useless when dealing with high-dimensional datasets (The dataset of this research deals with two dimensional data: latitude & longitude).
- It deals with the issue of noises. The total number of noises is presented also as a result	
- It requires only two parameters	
- It is not sensitive to the ordering of the data points in the database. However, those data points at the edges of clusters might be affected by the ordering of the data points in the dataset and therefore, it gives rise to the swapping of the membership of the clusters.	
- It can handle a spatial database of a huge data volume (more than a few thousand data points)	

Table 8 Advantages and disadvantages of DBSCAN

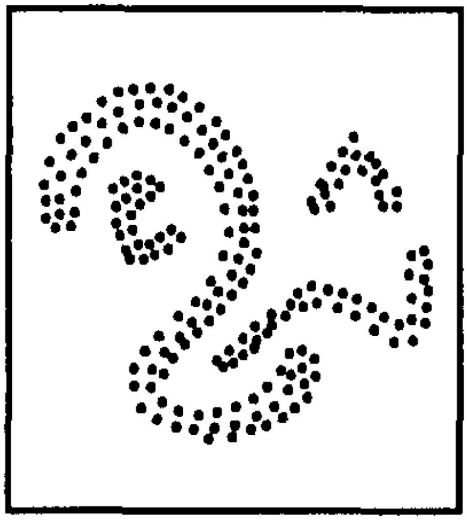

Figure 4 Arbitrary shape of data clusters (M. Ester, et al., 1996)

WEKA 3.6.9 was employed for the implementation of DBSCAN clustering algorithm. After the installation of the software, the 'classpath' of the environmental variable had to be set to the path of the directory of WEKA. Unlike many other software, WEKA can only be accessed by an elevated command prompt instead of directly double-clicking on its icon. In the command prompt, the directory has to be changed to the path of the directory of WEKA and 'weka.jar' is the command line to start the GUI of WEKA. The WEKA GUI can be initiated every time in the way. Figure 5 shows the imported data for Austrian group (with the crucial attributes only) prior to clustering.

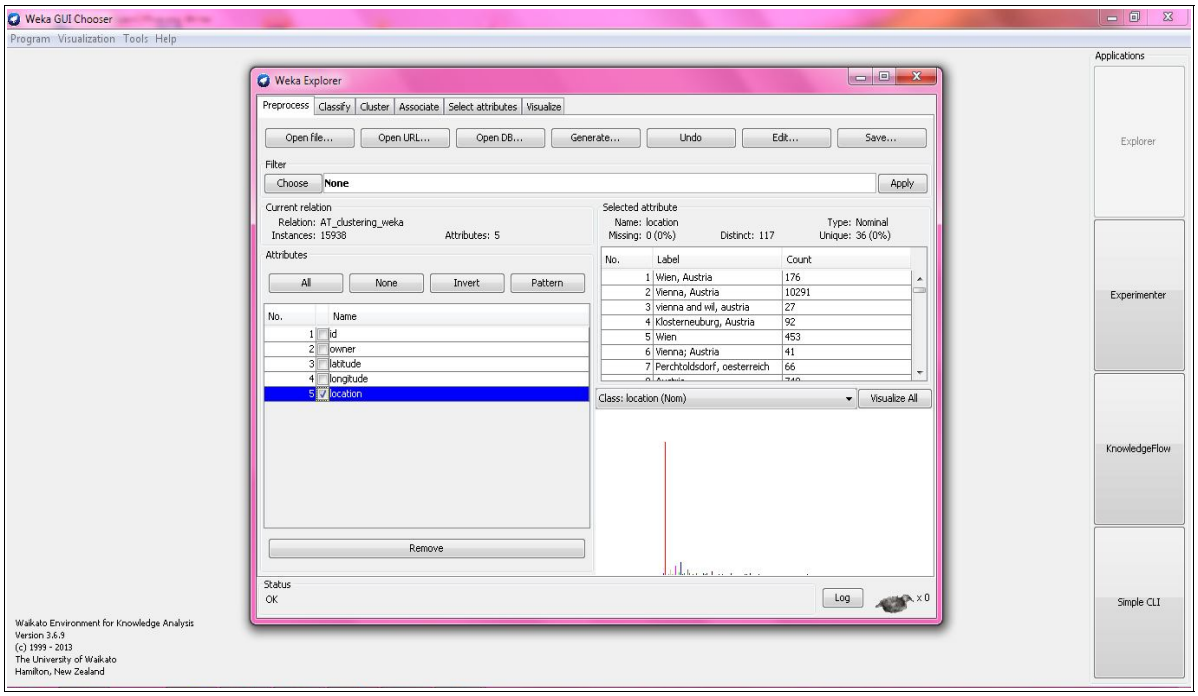

Figure 5 Imported data for Austrian group in WEKA

The window on the left shows the attributes associated with the imported CSV files. Simple statistics will be shown in the upper window on the right. The bottom right window depicts the data values with a graphical view.

In the 'Cluster' panel, DBSCAN clustering algorithm was selected and the test dataset was set. Afterwards, the two parameters, epsilon and minPoints were defined for the dataset. The attributes 'longitude' and 'latitude', which indicate the geo-location where each photo was taken, were used for the clustering process. The process commenced by clicking on the 'Start' button. Figure 6 illustrates how the two critical parameters for DBSCAN clustering algorithm can be set in WEKA.

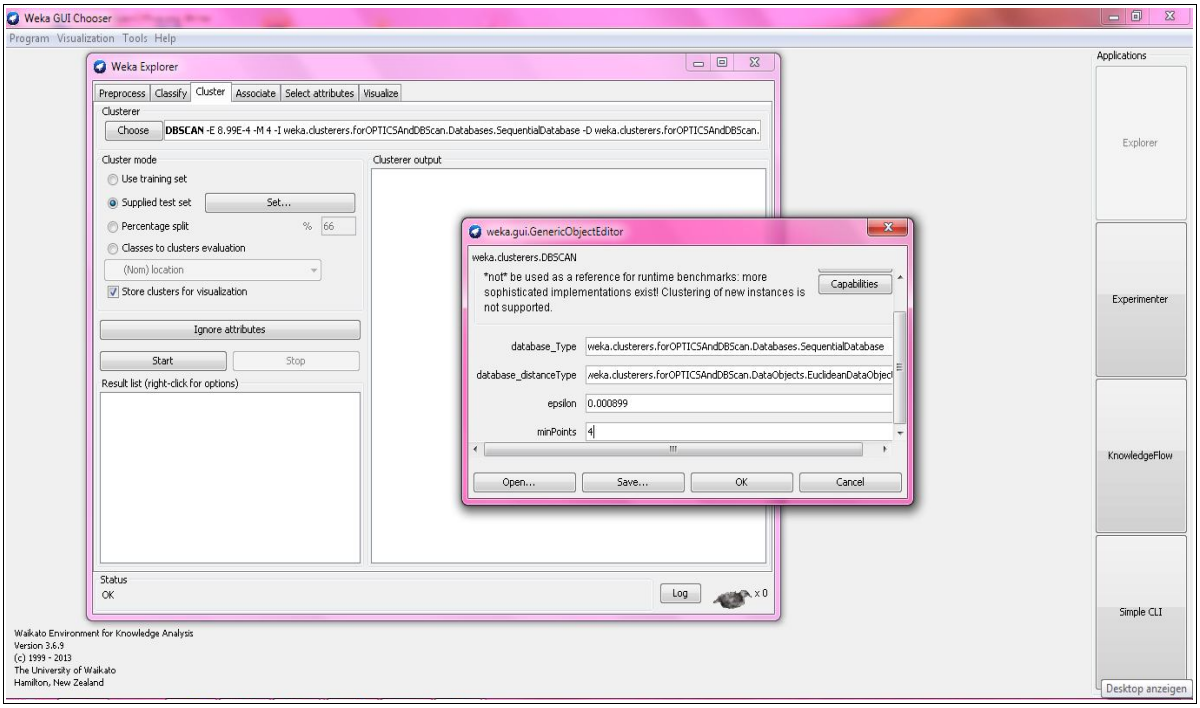

Figure 6 Setting of the two parameters, epsilon and minPoints for DBSCAN in WEKA

As for choosing an optimal ε-value, empirical tests were performed. As the starting point, 150m was considered as the radius of circle for a spatial cluster. Nonetheless, the geo-location values of the dataset, which are the key for data clustering, are based on the WGS84 datum. They are presented in terms of decimal degrees. Conversion of the distance in Cartesian Coordinate System to that of WGS84 was unavoidable. Example 2 demonstrates such conversion.

Given the equatorial radius (R) of the Earth = 6378.1km
The perimeter of the Earth = 2*π*R where π = 3.141593 (up to 6 d.p.)
= 40,054,468m

1° = 3,600"
If presenting 360° in seconds, 360° = 3,600" * 360° = 1,296,000"
Hence, at equator, 1" represents 30.9m on earth (40,054,468m/1,296,000")

Now let the radius (ε-value) of the circle be 150m, then
150/30.9 = 4.854369" (up to 6 d.p.)
It means that 4.854369 seconds represent approximately 150m for the radius of circle.

Since the latitudes and longitudes are presented in decimal degrees and 1° = 3600"
Converting 4.854369 seconds to decimal degrees is equal to
4.854369/3600 = 0.001348°

This will be the value for the ε-parameter of DBSCAN algorithm. If another value for the radius of circle is chosen, 150m is then replaced by that value for the calculation.

Example 2 Conversion of the distance in the Cartesian Coordinate System to the distance in WGS84

As discussed in the disadvantages of DBSCAN, the epsilon-minPts combination might not work for all clusters, if there is a huge difference in the density of the dataset. That was the case for the datasets in question. As for empirical tests, different sets of values for the epsilon-minPts combination have been attempted, for instances, if ε-value was set as 100m (0.000899°), the inner city of Vienna could be reasonably clustered, while for the sparsely dense areas, e.g. parks or cemeteries, etc. a lot of data points were classified as noises. In this way, valuable data points were disregarded for further analysis too early. The same situation occurred when choosing the values for the minPts for the datasets[5]. It does not make sense to set the minPts to one because then every point would be classified as a cluster. On the other hand, if it was set to six or more, data points which were supposed to be in a cluster were classified as noises.

5 The raw dataset was divided into three sub-datasets, each for each top user group from the results of the Specific Aim 1.

Because of the disadvantage of the algorithm and the results of the empirical tests, different sets of values of the epsilon-minPts combination were selected for different parts of Vienna. The same sets of values for parameters were applied to all datasets. Table 9 presents the values of the parameters for the data clustering.

Area of Vienna	epsilon (ε)	minPts
The inner city	0.000899 (=100m)	4
Outside the inner city, except parks, cemeteries or other larger areas	0.001348 (=150m)	2
Parks, cemeteries, or any other large areas	0.004495 (=500m)	2

Table 9 Values of parameters used for data clustering for all user groups

It should be emphasized that different sets of values of parameters had to be considered due to the known disadvantage of the DBSCAN algorithm. This had also been discussed in the literature (Jankowski et al., 2010 & Androgen and Androgen, 2010). Each user group adopted totally the same set of values for these two parameters. However, this set of values of parameters was never perfect. In some areas, e.g. larger parks like Prater park, data were aggregated by the algorithm into several spatial clusters instead of just one for the whole area. With also reference to the textual attributes in addition to the geo-locations, these clusters were grouped into one in the end.

The selection and ranking of the top 10 frequently visited landmarks depend only on the count of the unique owners at each landmark instead of the number of photos at each location. If a owner takes many photos at the same location, it will be counted as 1 data point as those duplicated data do not contribute to the ranking of the landmarks. Further elimination of unnecessary photos was out of question before the clustering process. This step was favourable as it streamlined the dataset and shortened the clustering process. The following table (Table 10) shows the statistics on the number of photos before and after the

deletion.

	Number of photos before deletion	Number of photos after deletion
North America (NA)	16921	3733
Europe without Austria (or simply *EU* thereafter)	18546	8129
Austria (AT)	46187	15938

Table 10 Statistics on the number of photos before and after deletion of unnecessary photos for data clustering

The clustering results were visualized in QGIS. Different clusters were presented in different colours to make the visualization easier. Table 11 provides the total number of photos used for further landmark identification and the count of their unique owners. The next step was to assign the clusters into the corresponding city landmarks/locations. Other than the personal experience, the plug-in called 'Reverse geo-code' for reverse geo-coding from Google Web Service[6] was adopted to guarantee a more objective assignment. While geo-coding gives the geo-location values of a place by the address, reverse geo-coding, on the contrary, gives the address (or name of the landmark) of the geo-location. A Google API key was required for the service. After obtaining the Google API key from its website, it was set in the corresponding plug-in in QGIS.

6 Yahoo! Web Service would have been another option before 2011 (The year of the termination of its Web Service).

	Number of photos used for landmark identification after data clustering	Number of unique owners for landmark identification after data clustering
NA	2934	244
EU	4829	746
AT	6426	419

Table 11 Number of Photos and unique owners used for landmark identification after the data clustering

The geo-locations of photos were the data points for data clustering with the DBSCAN algorithm. The clusters were afterwards associated to the places with the help of the plug-in 'Reverse geo-code'. Since the clustering results for three different user groups were different, it was necessary to group some places together as one landmark so that the landmarks were comparable between these three user groups. For instance, 'Kohlmarkt', 'Michealerplatz', 'Hofburg & Heldenplatz' and 'Burggarten' were individual clusters as the clustering result for the AT group, while these places were represented as a single cluster for the EU group. In such case, comparison could not be made between these landmarks for the AT group and the EU group because of the difference in the clustering result. Hence, the AT clusters representing 'Kohlmarkt', 'Michealerplatz' & 'Heldenplatz' were lumped into one as in the EU group for the purpose of consistency and comparability.. The same was true for some larger landmarks, e.g. 'Schönbrunn' or Prater Park. These much larger areas were not entirely represented by a single cluster, instead, numerous clusters. These clusters had to be combined into one to objectively represent the landmarks. Apart from the geo-locations, as for verification of the assignment of the places, textual attributes, e.g. tags, title of photos and descriptions of photos were taken into consideration (see Section of Scope). For each cluster, the informative textual attributes were checked to see whether the cluster matched the name of the places. Table 12 gives the number of identified locations[7] for each group.

7 Clusters of insignificant size were disregarded for the identification process.

	Number of identified locations
NA	55
EU	59
AT	65

Table 12 Number of identified locations for each group

Unfortunately, some of the textual attributes were non-informative, they were either left blank, or with other information that did not indicate any places' name, e.g. 'IMG084' (the numbering of the photo in the owner's camera) or 'a bird', etc. In such case, the geo-location of photos would be the dominant and only factor for the landmark identification (proof with visual content of the images is out of the scope for the thesis, see Section of Scope). Table 13 shows the statistics on the non-informative textual attributes for each user group.

	Total number of photos with non-informative textual attributes	Total number of photos used for landmark identification after data clustering	Percentage of photos with non-informative textual attributes
NA	254	2934	8.66
EU	701	4829	14
AT	346	6426	5.38

Table 13 Statistics on the non-informative textual attributes for each user group

The same procedures applied to each group. The places were ranked according to the total number of unique owners visited at each location and the top 10 most visited city landmarks were selected as part of the result of the Specific Aim 2. Next step was to divide each user group into two further sub-groups, according to the two seasons. Summer season refers to the period of the first day of May to the end of October of the year, whereas winter season refers to the period of the first day of November to the end of April of the next year. This definition follows the one defined in the documentation from Statistik Austria. The delineation of the

two seasons was arranged in the spreadsheet by the attribute 'datetaken' instead of the attribute 'dateupload'. The former one indicates the date on which the photo was taken, while the latter one gives the date on which the photo was uploaded. The attribute 'datetaken' was referenced for both identifying landmarks and depicting trajectory patterns (see Section 6.1) in different seasons. After further data manipulation in the spreadsheet, the top 10 landmarks for each season for each group were found out and ranked.

5.2 Results

5.2.1 Top 10 Landmarks of Each Selected Group for the Entire Designated Period, Summer and Winter Seasons

In this section, the results of the top 10 landmarks of the largest Flickr user groups, namely, Europe (excluding Austria), Austria and North America, are presented. For the entire designated period, summer and winter seasons, the statistics on each individual user group are shown. This section focuses on the presentation of the quantitative data, while the next two sections emphasize on the graphical presentation of the results.

The top 10 visited landmarks are of interest instead of the top 10 photographed landmarks. Therefore, they are ranked by the total number of visitors (equivalent to the non-duplicated count of photo-owners, i.e. 'Unique Owners') at each landmark rather than the total number of photos being taken at each landmark. The following table presents the results of the most represented group 'Europe excluding Austria' of the collected Flickr data during the entire designated period (Table 14). The percentage of unique owners is in respect of the total number of unique owners of this user group. Unique (Flickr-photo-) owners refer to the Flickr users as visitors to the landmarks. 'Visitors' is used in most cases when presenting the data of tables.

Ranking	Landmark	Number of Photos taken	Number of Unique Owners	Ratio of number of photos to number of unique owners	% of Unique Owners
1	Stephansdom, Graben & Peterskirche	856	332	2.6	34.0
2	Kohlmarkt, Michaelerplatz, Hofburg, Heldenplatz & Burggarten	742	239	3.1	24.5
3	Albertina, Oper & Hotel Sacher	291	156	1.9	16.0
4	Schoenbrunn	560	126	4.4	13.0
5	Rathaus & Burgtheater	268	109	2.5	11.2
6	Maria Theresia Platz, Kunsthistorisches Museum & Naturhistorisches Museum	198	101	2.0	10.4
7	Parlament	186	93	2.0	9.5
8	Karlsplatz & Karlskirche	118	74	1.6	7.6
9	Prater Amusement Park	101	61	1.7	6.3
10	Secession & Naschmarkt	102	59	1.7	6.0

Table 14 Top 10 landmarks of the user group 'Europe excluding Austria' for the entire designated period

The first top landmark 'Stephansdom, Graben & Peterskirche' was visited by 332 Flickr users with 856 photos taken during the designated period. That is about 34% of the group. In the second and the third positions, the landmarks are 'Kohlmarkt, Michaelerplatz, Hofburg, Heldenplatz & Burggarten' and 'Albertina, Oper & Hotel Sacher' respectively. These visitors constitute almost 24.5% and 16% of the group. The last top visited landmark in the ranking is 'Secession & Naschmarkt', which was frequented by 59 Flickr users. These users represent only 6% of the group. It is interesting to observe that although 'Stephansdom, Graben & Peterskirche' is the most popular landmark in Vienna for this group, the landmark 'Schoenbrunn' has the highest number-of-photos-to-number-of-unique-owners ratio.

Table 15 illustrates the statistics on the top 10 landmarks of the user group 'Austria' for the designated period. Overall, the variation on the number of visitors is not as large as the user group 'Europe excluding Austria'. The most popular landmark for this group is 'Kohlmarkt, Michaelerplatz, Hofburg, Heldenplatz & Burggarten' with 108 visitors during the designated period. An interesting phenomenon is that even the most frequently visited landmark was visited by only about 17% of this user group. With only two visitors less than the most

popular landmark, 'Stephansdom, Graben & Peterskirche' ranks the second. 'Schoenbrunn' occupies the third place with further 10 visitors less, coming down to 96 visitors in total. Although being in the third rank, number of photos taken in 'Schoenbrunn' is dramatically more than the two most visited landmarks with 8.4 number-of-photos-to-number-of-unique-owners ratio. Both 'Albertina, Oper & Hotel Sacher' and 'Parlament' are in the same rank, being the ninth most popular city landmark with 53 visitors. The 10th rank is thus omitted.

Ranking	Landmark	Number of Photos taken	Number of Unique Owners	Ratio of Number of Photos to Number of Unique Owners	% of Unique Owners
1	Kohlmarkt, Michaelerplatz, Hofburg, Heldenplatz & Burggarten	497	108	4.6	17.2
2	Stephansdom, Graben & Peterskirche	446	106	4.2	16.9
3	Schoenbrunn	807	96	8.4	15.3
4	Karlsplatz & Karlskirche	245	75	3.3	11.9
5	Secession & Naschmarkt	255	73	3.5	11.6
6	Prater Amusement Park	339	72	4.7	11.5
7	Museumsquartier	301	68	4.4	10.8
8	Rathaus & Burgtheater	301	65	4.6	10.4
9	Albertina, Oper & Hotel Sacher	156	53	2.9	8.4
	Parlament	183	53	3.5	8.4

Table 15 Top 10 landmarks of the user group 'Austria' for the entire designated period

With 121 Flickr users 'Stephansdom, Graben & Peterskirche' tops the list of popular landmarks for the last selected user group 'North America' (Table 16). Another popular landmark 'Kohlmarkt, Michaelerplatz, Hofburg, Heldenplatz & Burggarten' occupies the second place in the ranking, while 'Schoenbrunn' ranks the third with 47 visitors. For this group, the variation on the number of unique owners is larger for the first three most visited landmarks than the rest in the ranking. On average, North Americans took more photos in 'Kohlmarkt, Michaelerplatz, Hofburg, Heldenplatz & Burggarten' and 'Schoenbrunn' than in 'Stephansdom, Graben & Peterskirche'. The 5th rank belongs to both 'Albertina, Oper & Hotel Sacher' and 'Maria Theresia Platz, Kunsthistorisches Museum & Naturhistorisches Museum',

which have 41 visitors. The 6th rank is hence omitted. 'Museumsquartier' is the last of the top 10 landmarks, visited by about 11% of the group. Although 'North America' is the smallest group among the three selected Flickr user groups, it has the largest portion of the group (almost 44%) for the most popular landmark.

Ranking	Landmark	Number of Photos taken	Number of Unique Owners	Ratio of Number of Photos to Number of Unique Owners	% of Unique Owners
1	Stephansdom, Graben & Peterskirche	450	121	3.7	43.7
2	Kohlmarkt, Michaelerplatz, Hofburg, Heldenplatz & Burggarten	482	90	5.4	32.5
3	Schoenbrunn	401	47	8.5	17.0
4	Rathaus & Burgtheater	90	43	2.1	15.5
5	Albertina, Oper & Hotel Sacher	95	41	2.3	14.8
	Maria Theresia Platz, Kunsthistorisches Museum & Naturhistorisches Museum	89	41	2.2	14.8
7	Karlsplatz & Karlskirche	80	40	2.0	14.4
8	Secession & Naschmarkt	79	39	2.0	14.1
9	Parlament	88	35	2.5	12.6
10	Museumsquartier	68	30	2.3	10.8

Table 16 Top 10 landmarks of the user group 'North America' for the entire designated period

Apart from analysing the results of the entire designated period, the data were investigated in the context of seasons for each group as well, i.e. summer and winter. The seasons were defined in accordance with the Statistik Austria one-year touristic calender year which starts from November of the current year to October of the next year. Summer season refers to the period from the first day of May to the last day of October, while winter season refers to the rest of the months of the calendar year (Statistik Austria, 2003). Before discussing the data on the top 10 landmarks, Table 17 gives an overview on the total number of Flickr users as

visitors (non-duplicated count) for each group in the two seasons.

	Summer	Winter
Europe excluding Austria	577	471
Austria	455	449
North America	176	117

Table 17 Summary on the total number of Flickr users in each selected group for each season

The size of the groups 'Austria' and 'Europe excluding Austria' is about the same in both seasons with North Americans as the distant third. During summer, the number of Flickr users went up in in all three groups. One limitation of the data is that the sum of the total number of Flickr users in each group for these two seasons is not equal to the total number of Flickr users of the corresponding group. The differences could be attributed to travelling across the seasons or re-visiting, etc. (see also Section 5.3)

Table 18 summaries the data of the top 10 visited landmarks in the summer months for 'Europe excluding Austria'. The percentage of the unique owners is with respect to the total number of the unique owners in summer. The number of visitors ranges from 186 (32.2%) for the Top 1 landmark 'Stephansdom, Graben & Peterskirche' to 29 (5.0%) for the Top 10 landmark 'Volksgarten'. 'Kohlmarkt, Michaelerplatz, Hofburg, Heldenplatz & Burggarten' and 'Albertina, Oper & Hotel Sacher' ranks the second and the third respectively, with about 15% to 24% of visitors. Although the landmark 'Albertina, Oper & Hotel Sacher' has 1% visitors more than the landmark 'Schoenbrunn' does, number of photos taken at 'Schoenbrunn' is almost double.

Ranking	Landmark	Number of Photos taken	Number of Unique Owners	Ratio of Number of Photos to Number of Unique Owners	% of Unique Owners
1	Stephansdom, Graben & Peterskirche	489	186	2.6	32.2
2	Kohlmarkt, Michaelerplatz, Hofburg, Heldenplatz & Burggarten	442	137	3.2	23.7
3	Albertina, Oper & Hotel Sacher	167	86	1.9	14.9
4	Schoenbrunn	303	80	3.8	13.9
5	Maria Theresia Platz, Kunsthistorisches Museum & Naturhistorisches Museum	125	58	2.2	10.1
6	Rathaus & Burgtheater	115	53	2.2	9.2
7	Parlament	112	52	2.2	9.0
8	Karlsplatz & Karlskirche	68	43	1.6	7.5
9	Prater Amusement Park	72	41	1.8	7.1
10	Volksgarten	62	29	2.1	5.0

Table 18 Top 10 landmarks of the user group 'Europe excluding Austria' in the summer season

In general, the number of visitors of the Austrian group fluctuates less than the group 'Europe excluding Austria' over the popular landmarks, ranging from 72 (15.8%) to 32 (7%) (Table 19). The table shows that 'Kohlmarkt, Michaelerplatz, Hofburg, Heldenplatz & Burggarten', 'Schoenbrunn' and 'Stephansdom, Graben & Peterskirche' are the Top 1, Top 2 and Top 3 landmarks in summer respectively. Although being in the 7th rank, the landmark 'Prater Amusement Park' was photographed almost as many times as the landmark 'Stephansdom, Graben & Peterskirche'. Fairly obvious is that the 8th landmark 'Rathaus & Burgtheater' was shot over 200 times by 39 visitors. Its number-of-photos-to-number-of-unique-owners ratio is just behind that of the most photographed 'Schoenbrunn'.

Ranking	Landmark	Number of Photos taken	Number of Unique Owners	Ratio of Number of Photos to Number of Unqiue Owners	% of Unique Owners
1	Kohlmarkt, Michaelerplatz, Hofburg, Heldenplatz & Burggarten	311	72	4.3	15.8
2	Schoenbrunn	431	66	6.5	14.5
3	Stephansdom, Graben & Peterskirche	187	62	3.0	13.6
4	Secession & Naschmarkt	151	47	3.2	10.3
5	Karlsplatz & Karlskirche	150	46	3.3	10.1
6	Museumsquartier	151	44	3.4	9.7
7	Prater Amusement Park	186	42	4.4	9.2
8	Rathaus & Burgtheater	206	39	5.3	8.6
9	Parlament	118	35	3.4	7.7
10	Maria Theresia Platz, Kunsthistorisches Museum & Naturhistorisches Museum	84	32	2.6	7.0

Table 19 Top 10 landmarks of the user group 'Austria' in the summer season

The following table (Table 20) presents 11 locations for the 10 most favourable landmarks in summer for 'North America'. Both 'Museumsquartier' and 'Augustinerkirche, Josefplatz, Nationalbibliothek & Palais Pallavicini' occupy the 10th rank. The Top 3 landmarks are in order 'Stephansdom, Graben & Peterskirche', 'Kohlmarkt, Michaelerplatz, Hofburg, Heldenplatz & Burggarten' and 'Secession & Naschmarkt'. Although the range of the number of visitors in the ranking is similar to that of 'Austria', the percentage of visitors varies significantly, from 43.2% to 9.1% (Discussion in Section 5.3). Another prominent feature of the data is the number-of-photos-to-number-of-unique-owners ratio for 'Schoenbrunn'. The high ratio is attributed to the substantial amount of photos being taken at Schoenbrunn by relatively few visitors. The range of the number of photos taken at these 11 landmarks is as well drastic. It is also interesting to see that the third popular landmark 'Secession & Naschmarkt' was featured in only 55 photos, which were taken by 28 visitors.

Ranking	Landmark	Number of Photos taken	Number of Unique Owners	Ratio of Number of Photos to Number of Unique Owners	% of Unique Owners
1	Stephansdom, Graben & Peterskirche	235	76	3.1	43.2
2	Kohlmarkt, Michaelerplatz, Hofburg, Heldenplatz & Burggarten	262	58	4.5	33.0
3	Secession & Naschmarkt	55	28	2.0	15.9
4	Schoenbrunn	330	27	12.2	15.3
5	Karlsplatz & Karlskirche	50	26	1.9	14.8
6	Albertina, Oper & Hotel Sacher	47	25	1.9	14.2
7	Maria Theresia Platz, Kunsthistorisches Museum & Naturhistorisches Museum	51	24	2.1	13.6
8	Rathaus & Burgtheater	36	22	1.6	12.5
9	Parlament	33	18	1.8	10.2
10	Museumsquartier	41	16	2.6	9.1
	Augustinerkirche, Josefplatz, Nationalbibliothek & Palais Pallavicini	26	16	1.6	9.1

Table 20 Top 10 landmarks of the user group 'North America' in the summer season

The presentation of the 10 most visited landmarks in winter starts with the group 'Europe excluding Austria' (Table 21). 152 (32.3%) visitors have visited the Top1 landmark 'Stephansdom, Graben & Peterskirche' in winter. The second popular landmark 'Kohlmarkt, Michaelerplatz, Hofburg, Heldenplatz & Burggarten' was frequented by about 9% less visitors. Another landmark 'Albertina, Oper & Hotel Sacher' in the vicinity was preferred by 74 (15.7%) visitors, ranking the third. At the other end of the ranks, 'Secession & Naschmarkt', 'Karlsplatz & Karlskirche' and 'Museumsquartier' were visited by 5-7% of visitors. Being in the fifth position, 'Schoenbrunn' is still by far the most photographed location in spite of winter time.

Ranking	Landmark	Number of Photos taken	Number of Unique Owners	Ratio of Number of Photos to Number of Unique Owners	% of Unique Owners
1	Stephansdom, Graben & Peterskirche	367	152	2.4	32.3
2	Kohlmarkt, Michaelerplatz, Hofburg, Heldenplatz & Burggarten	300	108	2.8	22.9
3	Albertina, Oper & Hotel Sacher	124	74	1.7	15.7
4	Rathaus & Burgtheater	153	58	2.6	12.3
5	Schoenbrunn	257	54	4.8	11.5
6	Maria Theresia Platz, Kunsthistorisches Museum & Naturhistorisches Museum	73	43	1.7	9.1
7	Parlament	74	42	1.8	8.9
8	Secession & Naschmarkt	54	33	1.6	7.0
9	Karlsplatz & Karlskirche	50	31	1.6	6.6
10	Museumsquartier	35	22	1.6	4.7

Table 21 Top 10 landmarks of the user group 'Europe excluding Austria' in the winter season

Table 22 shows that 67 visitors of the Austrian group travelled to the Top 1 landmark 'Stephansdom, Graben & Peterskirche' during the winter season. As in other periods of time, only a small portion (about 15%) of this community shows interest in this landmark. Equally popular are the landmarks 'Kohlmarkt, Michaelerplatz, Hofburg, Heldenplatz & Burggarten' and 'Schoenbrunn' for the second position. Same situation can be observed for the 5th rank. Therefore, the Top 3 and Top 6 are skipped. The same phenomenon on the number-of-photos-to-number-of-unique-owners ratio for 'Schoenbrunn' as in the group 'Europe excluding Austria' despite the winter period.

Ranking	Landmark	Number of Photos taken	Number of Unique Owners	Ratio of Number of Photos to Number of Unique Owners	% of Unique Owners
1	Stephansdom, Graben & Peterskirche	259	67	3.9	14.9
2	Kohlmarkt, Michaelerplatz, Hofburg, Heldenplatz & Burggarten	186	57	3.3	12.7
	Schoenbrunn	376	57	6.6	12.7
4	Prater Amusement Park	153	43	3.6	9.6
5	Karlsplatz & Karlskirche	95	42	2.3	9.4
	Museumsquartier	150	42	3.6	9.4
7	Secession & Naschmarkt	104	37	2.8	8.2
8	Rathaus & Burgtheater	95	35	2.7	7.8
9	Albertina, Oper & Hotel Sacher	78	32	2.4	7.1
10	Parlament	65	26	2.5	5.8

Table 22 Top 10 landmarks of the user group 'Austria' in the winter season

Table 23 summaries the statistics on the 10 most popular landmarks for the last user group 'North America' in winter. The first position belongs to the landmark 'Stephansdom, Graben & Peterskirche', which about 39% of the group have visited. Even though the number of visitors varies little, an obvious sharp drop occurs in the number of photos taken from the second rank to the third rank. The landmark 'Kohlmarkt, Michaelerplatz, Hofburg, Heldenplatz & Burggarten' being in the second position in this case has the highest number-of-photos-to-number-of-unique-owners ratio, replacing by far the most photographed landmark 'Schoenbrunn'. Equally competitive are the landmarks 'Rathaus & Burgtheater' and 'Schoenbrunn', which were visited by 22 Flickr users. Thus, the Top 4 landmark is omitted. At the other end of the list, 'Secession & Naschmarkt' was visited by 9.4% of the group.

Ranking	Landmark	Number of Photos	Number of Unique Owners	Ratio of Number of Photos to Number of Unique Owners	% of Unique Owners
1	Stephansdom, Graben & Peterskirche	215	46	4.7	39.3
2	Kohlmarkt, Michaelerplatz, Hofburg, Heldenplatz & Burggarten	220	34	6.5	29.1
3	Rathaus & Burgtheater	54	22	2.5	18.8
	Schoenbrunn	71	22	3.2	18.8
5	Maria Theresia Platz, Kunsthistorisches Museum & Naturhistorisches Museum	38	18	2.1	15.4
6	Parlament	55	17	3.2	14.5
7	Albertina, Oper & Hotel Sacher	48	16	3.0	13.7
8	Karlsplatz & Karlskirche	30	15	2.0	12.8
9	Museumsquartier	27	14	1.9	12.0
10	Secession & Naschmarkt	24	11	2.2	9.4

Table 23 Top 10 landmarks of the user group 'North America' in the winter season

5.2.2 Comparison of Top 10 Landmarks of Different Periods for Each Group

In this section, the results of the 10 most favourable landmarks in Vienna are compared within the group in seasonal context, i.e. summer and winter as well as for the entire period. Due to the discrepancy in the total number of unique owners when combining both seasons together (for reasons mentioned in the previous section), the term 'the entire period' is used to avoid confusion. The analysed data in this period refer to the first day of January of 2007 to the last day of August of 2011, which is the period of all collected data. Summer and winter are as defined by Statistik Austria (see Section 5.2.1). In this section, the results are presented in form of graphics to maintain visual clarity. As for the numerical data, one should refer to the previous section (Section 5.2.1).

As shown in Figure 7, the 10 most popular landmarks (12 locations in total), visited by Flickr users in the European group (without Austria) (or simply EU) during different periods vary to

a certain extent. As could be expected, the landmark 'Prater Amusement Park' is not in the ranking during wintertime The same is true for 'Volksgarten', although it was one of the top 10 landmarks during summer times. In contrast, 'Museumsquartier' and 'Secession & Naschmarkt' were frequently visited in winter. They were, however, not one of the top 10 sites in summer. As for the entire period, 'Volksgarten' and 'Museumsquartier' remained out of the ranking, while 'Secession & Naschmarkt' were in the top 10. Overall, eight landmarks made it to the top 10 ranks in different periods of time. Most prominently, the landmark 'Stephansdom, Graben & Peterskirche' remains as the all-time-favourite. The second rank is held by the landmark 'Kohlmarkt, Michaelerplatz, Hofburg, Heldenplatz & Burggarten' with steady number of visitors all year round. Among these eight landmarks, the number of visitors are higher in summer, with only one exception: 'Rathaus & Burgtheater'.

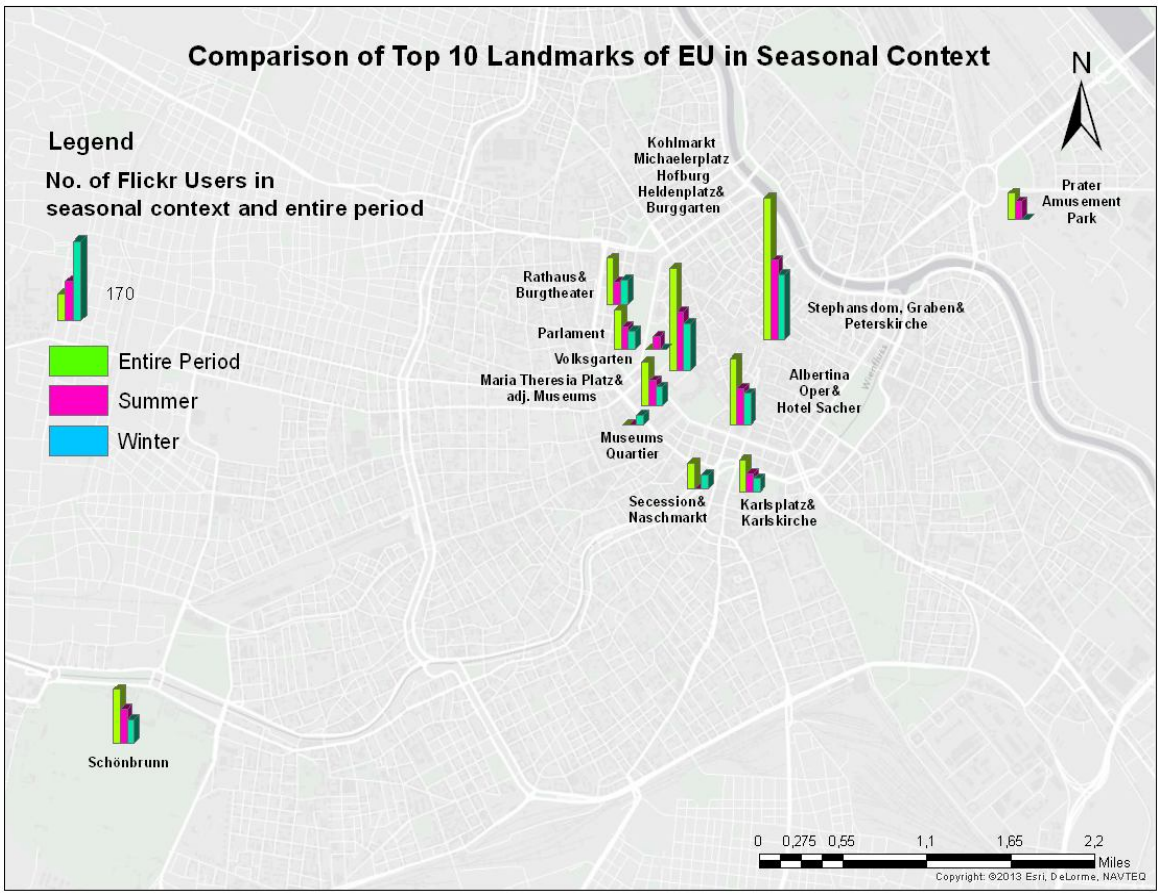

Figure 7 Comparison of Top 10 landmarks in different periods of time for the EU group

For the Austrian group (or simply AT), the landmarks 'Albertina, Oper & Hotel Sacher' and 'Maria Theresia Platz, Kunsthistorisches Museum & Naturhistorisches Museum' exhibited major differences across different periods, as shown in Figure 8. While 'Albertina, Oper & Hotel Sacher' was not in the summer list, 'Maria Theresia Platz' and its neighbouring museums were part of the favourites solely in summer seasons. Among the remaining nine landmarks (out of 11 locations), 'Stephansdom, Graben & Peterskirche', 'Kohlmarkt, Michaelerplatz, Hofburg, Heldenplatz & Burggarten' of the inner city and 'Schoenbrunn' in the south-west score the highest and 'Prater Amusement Park' and 'Stephansdom, Graben & Peterskirche' had more visitors in winter than in summer.

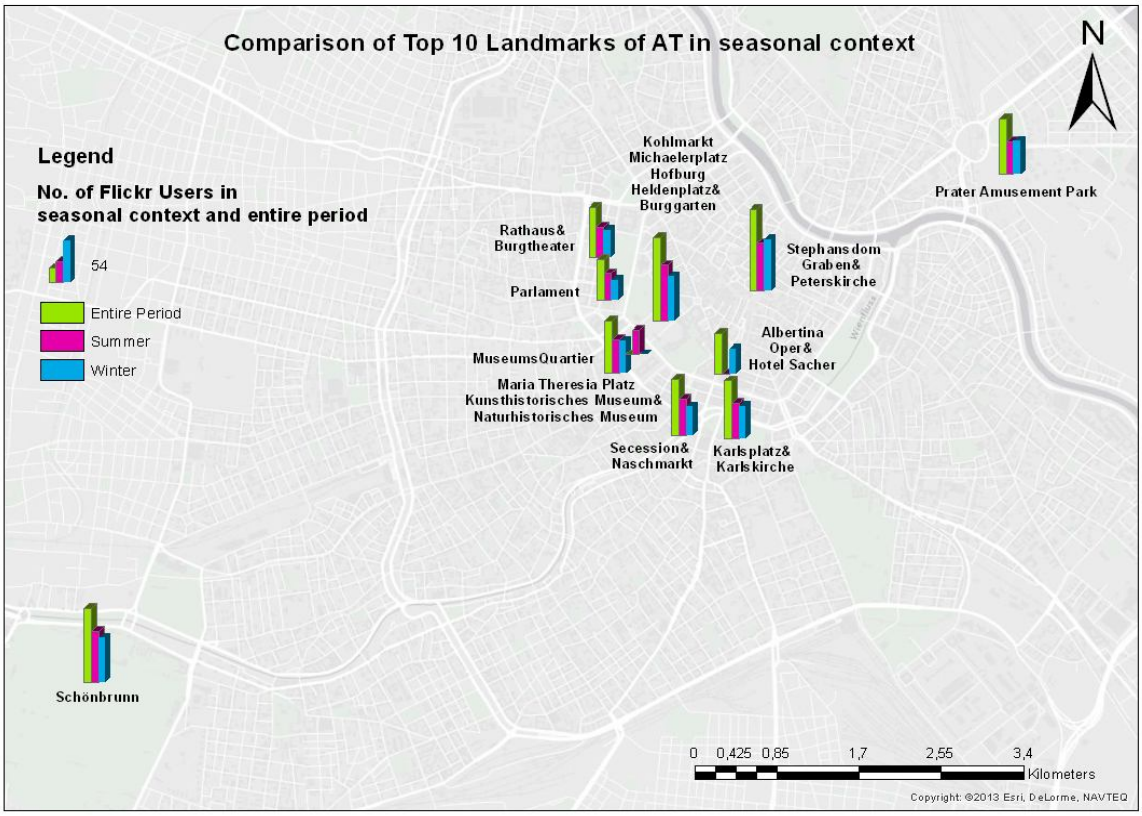

Figure 8 Comparison of Top 10 landmarks in different periods of time for the AT group

Figure 9 illustrates 11 hot spots of Vienna during different periods of time for 'North America' (or simply NA). The extra location is ascribed to the tied ranking in the 10th position (see Table 20). While the landmark 'Augustinerkirche, Josefplatz, Nationalbibliothek & Palais Pallavicini' was favourable only in summer times, the rest of the locations was popular at all times. 'Stephansdom, Graben & Peterskirche' and 'Kohlmarkt, Michaelerplatz, Hofburg, Heldenplatz & Burggarten' stood out to be the two most visited areas. 'Rathaus & Burgtheater', 'Parlament' and 'Museumsquartier' exhibited almost no differences in the number of visitors across seasons, whereas the remaining landmarks were more popular in summer than in winter.

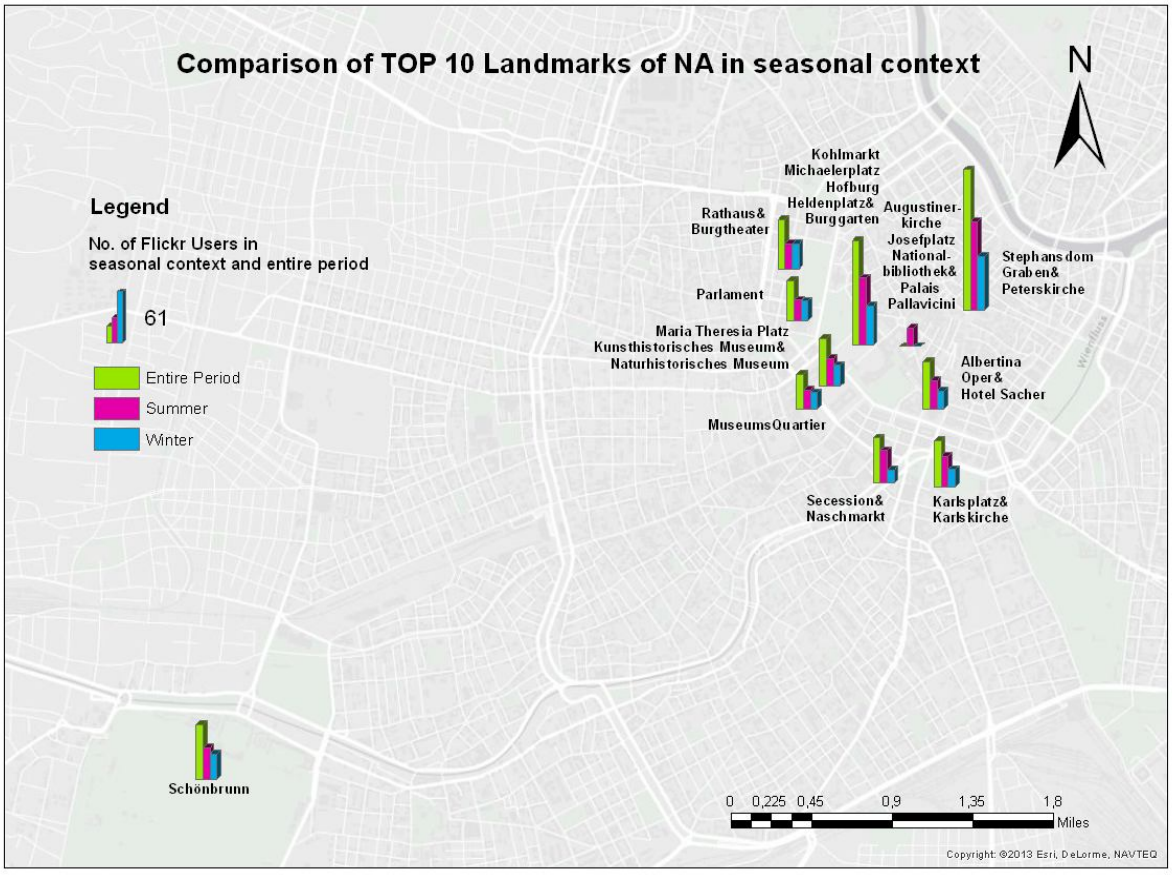

Figure 9 Comparison of Top 10 landmarks in different periods of time for the NA group

5.2.3 Comparison of Top 10 Landmarks of Same Periods for Different Groups

In this section, the favourable landmarks in Vienna are compared across the three groups for each designated period. As depicted in Figure 10, these three groups share eight identical landmarks in the specified period of about five years. Nonetheless, the three groups were distinct from each other at three locations, i.e. 'Prater Amusement Park', 'Maria Theresia Platz, Kunsthistorisches Museum & Naturhistorisches Museum' and 'Museumsquartier'. During this period, the North American group was not interested in 'Prater Amusement Park' as much as in 'Maria Theresia Platz', its adjacent museums and 'Museumsquartier'. In comparison, the Austrian group was not interested in 'Maria Theresia Platz, Kunsthistorisches Museum & Naturhistorisches Museum' and the European group was not interested in 'Museumsquartier', as much as its counterparts did.

Flickr users of all three groups most frequently visited the landmarks 'Stephansdom, Graben & Peterskirche' and 'Kohlmarkt, Michaelerplatz, Hofburg, Heldenplatz & Burggarten', albeit to a different extent. At most landmarks in common, EU outnumbered the other two groups, except at 'Prater Amusement Park' and 'Secession & Naschmarkt', where AT group exhibited a higher frequency. At 'Karlsplatz & Karlskirche', the proportion of EU visitors and AT visitors was almost the same. 'Stephansdom, Graben & Peterskirche' was the only landmark where the NA group outnumbered its counterparts.

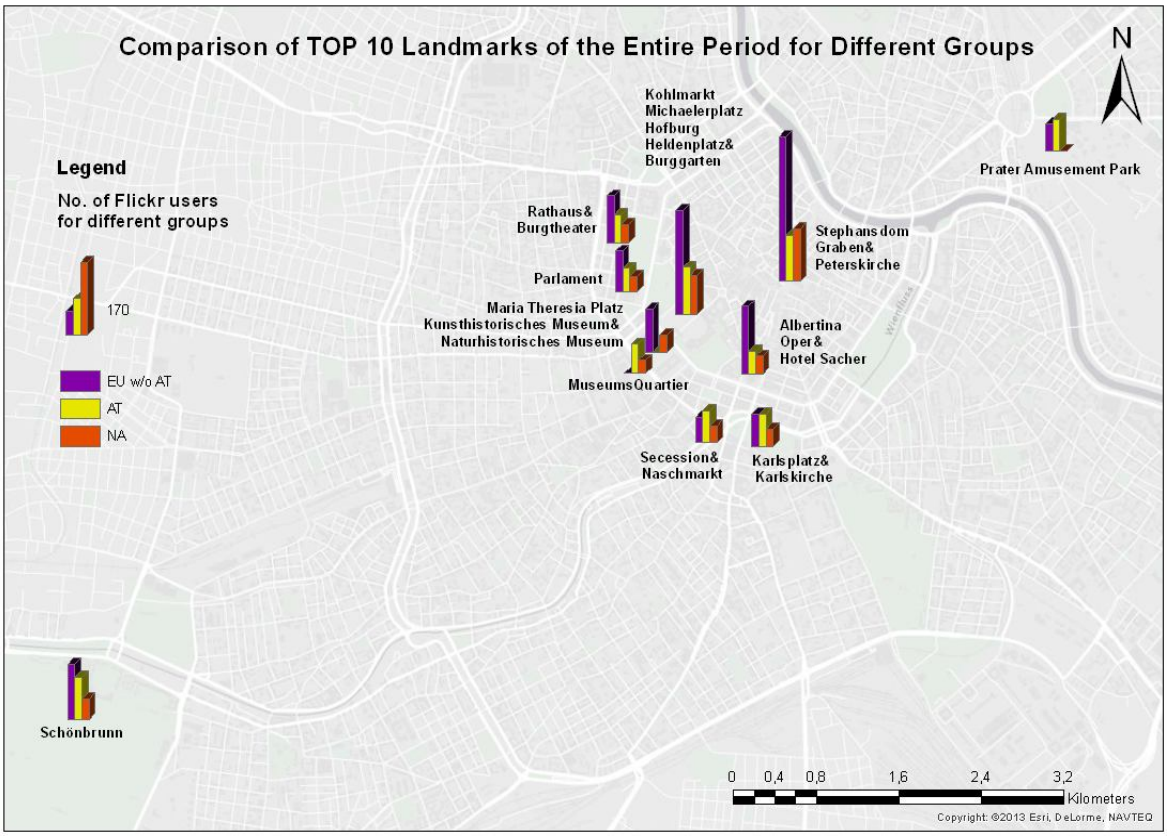

Figure 10 Comparison of Top 10 landmarks of the entire period for different groups

In summer, the variation in the distribution of the three user-groups around town was greater than that for the entire period (Figure 11). While the EU group was neither represented at 'Museumsquartier' nor at 'Secession & Naschmarkt', it was the only group at the landmark 'Volksgarten' that contributed to the top ranking. The NA group was in particular attracted to the landmark 'Augustinerkirche, Josefplatz, Nationalbibliothek & Palais Pallavicini' instead of 'Prater Amusement Park' on the other bank of the Donau Canal. 'Albertina, Oper & Hotel Sacher' did not appeal to the AT group as much as it did to the other two groups.

Figure 11 also shows that the landmarks in common had more visitors from the AT group rather than the NA group, except 'Stephansdom, Graben & Peterskirche'. While the number of visitors from the EU group outweighed that of the AT group at most landmarks in common, it

was quite the contrary at 'Prater Amusement Park' and 'Karlsplatz & Karlskirche'. 'Museumsquartier', 'Secession & Naschmarkt', 'Karlsplatz & Karlskirche' and 'Prater Amusement Park' were visited by more Flickr users of the AT group than those of the EU and NA groups in summer.

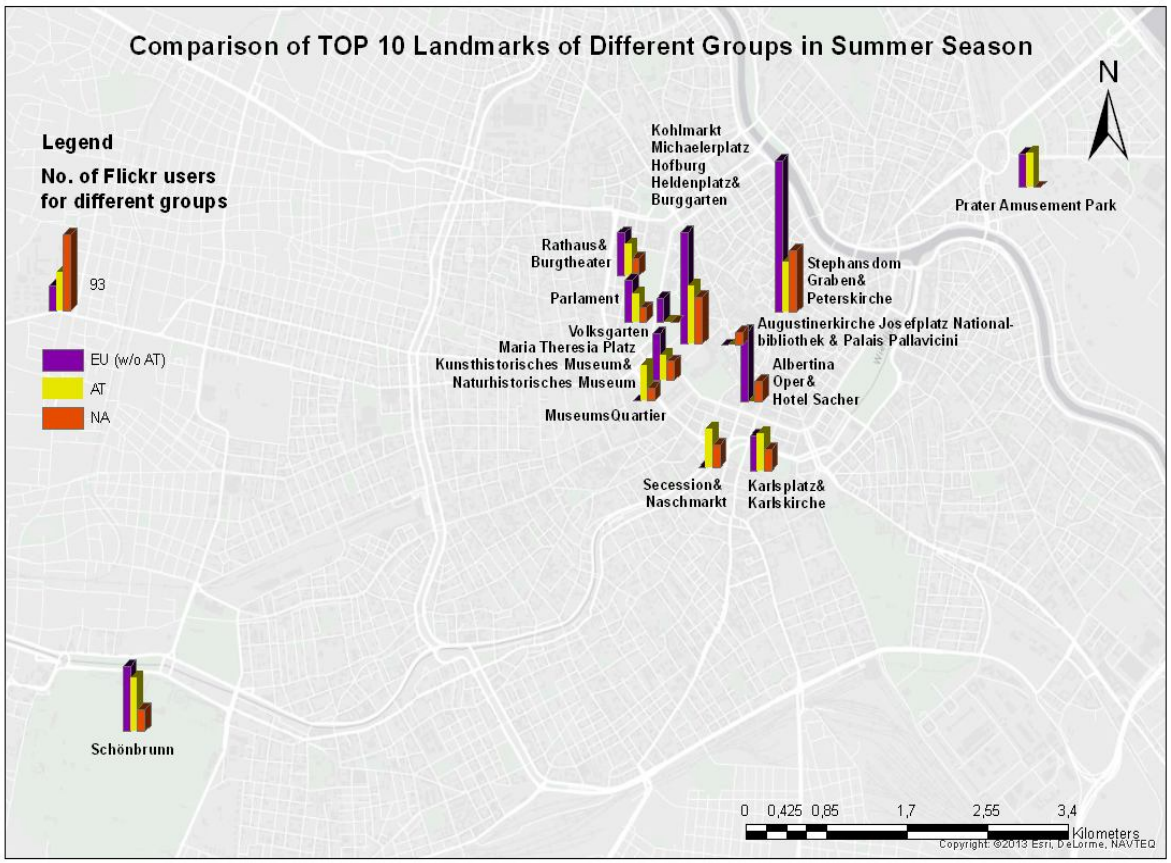

Figure 11 Comparison of top 10 landmarks of different groups in summer season

Resembling the summer season, 'Museumsquartier', 'Secession & Naschmarkt', 'Karlsplatz & Karlskirche' and 'Prater Amusement Park' had more visitors from the AT group than those from the other two groups in the winter season (Figure 12). In addition, 'Schoenbrunn' was also frequently visited by the Austrian group in winter.

Figure 12 illustrates that 'Stephansdom, Graben & Peterskirche' attracted most of the visitors

from all groups in winter and the number of visitors from EU doubled that from AT. In winter, the AT group stayed away from 'Maria Theresia Platz, Kunsthistorisches Museum & Naturhistorisches Museum', which had more visitors from the other two groups. In contrast, only the AT group favoured 'Prater Amusement Park'.

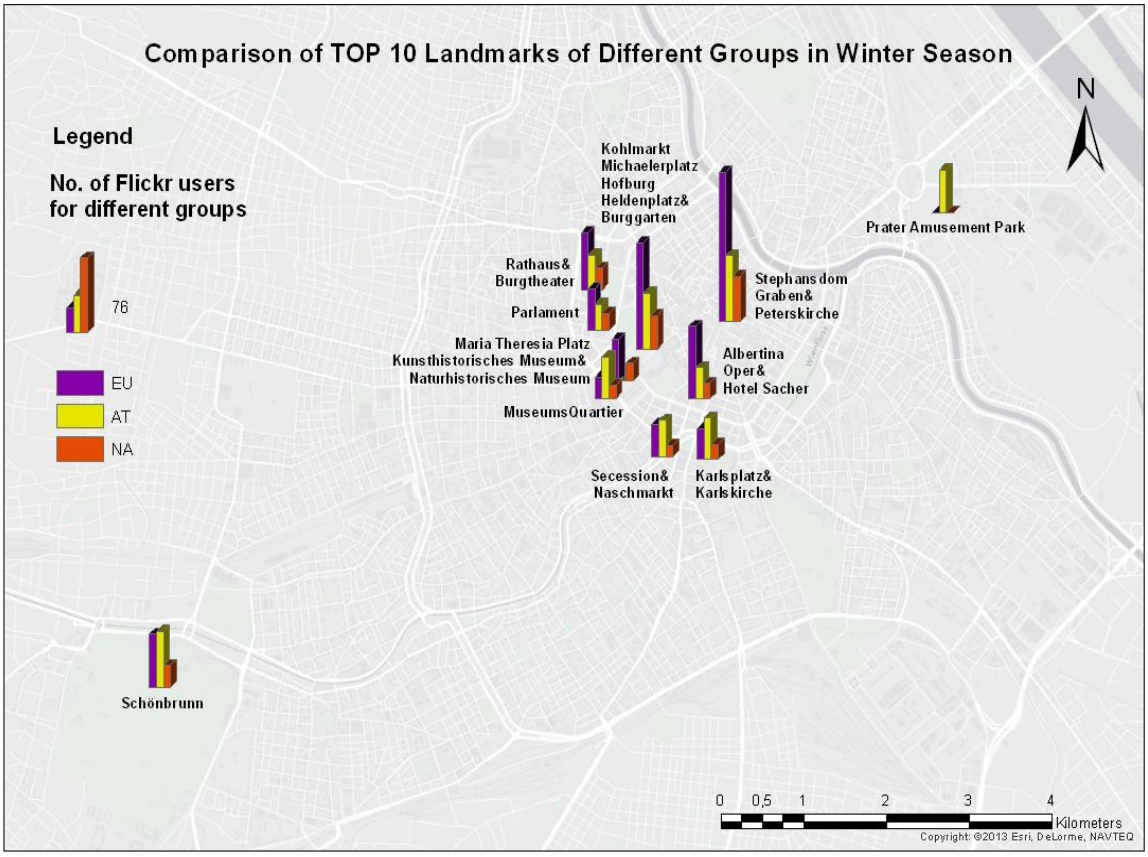

Figure 12 Comparison of Top 10 landmarks of different groups in winter season

5.3 Discussion

The data clustering algorithm DBSCAN was used for defining the city landmarks for the groups. This approach is similar to the one adopted by Yin et al. (2010). Although an alternative approach is to pre-define landmarks with reference to some travel resources, as Choudhury et al. (2010) did in their research, defining the city landmarks based on a data

clustering algorithm was more appropriate for the present study. In this way, we were not limited to a preselected list of landmarks, but could consider the differences in data clusters that were due to the preferences of the user groups.

The choice of the appropriate set of parameters is subject to the investigating area. Since the algorithm adopted in the present research is density-based, the data clustering process is sensitive to the density of data. The set of parameters chosen for Vienna (Table 9) might not be suitable for other cities. For example, Hong Kong is a densely populated big city. Most landmarks are all located in vicinity, esp. in the northern part of the Hong Kong island, and in Kowloon. Using the density-based algorithm, the data of the entire northern Hong Kong island, and Kowloon might be clustered into one. Hence, either another set of parameters or another data clustering algorithm should be applied.

In the present study, we could uncover distinct features of the landmark-visiting distribution between the user groups. In Figure 1, it is evident that the NA group focused on the landmarks of the inner city of Vienna, whereas the AT group went farther beyond the city centre and was disperse around Vienna. The EU group exhibited a distribution pattern in between. This phenomenon was also reflected by the numerical data. In the AT group, only 16% visitors visited the most popular landmark (Kohlmarkt, Michaelerplatz, Hofburg, Heldenplatz & Burggarten) in summer (Table 17 & 19). In contrast, 43% of the much smaller NA group visited the Top 1 landmark (Stephansdom, Graben & Peterskirche) in summer (Table 17 & 20). The numbers of the EU group lay in between. Similar percentages were applicable for the winter season and the entire period. In summary, these numbers again reflected the highest concentration toward the city centre for the NA group, the least one for the AT group.

As discussed in Section 5.2.1, the sum of the number of visitors (unique owners) in both seasons was not the same as the total number of visitors for the entire period. The reasons are: a.) visitors recurring in different seasons were counted only once for the entire period in order to avoid duplication, but they were counted twice, once in summer and once in winter, as

unique owners in both seasons This phenomenon became much more obvious when dealing with the AT group, in which the locals was involved, and b.) the delineation of the seasons was the end of April and the first of May, and the end of October and the first of November. Some visitors had the trip across this period. They were counted only once for the entire period and twice in seasonal context.

The EU group was the only group, which frequented 'Volksgarten' during summer season (Figures 7 & 11). The reasons could be that they enjoy particularly beautiful flowers in gardens. 'Volksgarten' is the only garden in the heart of the city, which offers beds and beds of beautiful blooming roses of different colours. 'Volksgarten' could also be used as a passageway to connect the area of the parliament and the city hall, and the 'Heldenplatz' area.

The EU group visited only the amusement park in Prater in summer, but not in winter (Figure 7). This seasonal difference could be attributed to the extended opening hours of the attractions within the amusement park in summer. According to the website of the amusement park (http://www.prater.at/Berichte/Ansicht.php?Id=5964), the main season for it is from mid-March to end of October. In contrast, the amusement park in Prater was not one of the most popular landmarks frequented by the NA group in both seasons. The reason could be that there is no lack of amusement parks of different scales in North America. They have two world famous theme parks Disneyland in the United States or the biggest amusement park called 'Canada's Wonderland' in Canada. Hence, the visitors from North America might focus on the attractions more specific to Vienna.

CHAPTER 6 SPECIFIC AIM 3: TRAJECTORY PATTERN ANALYSIS

This chapter presents

- the methodology in achieving the Specific Aim 3,
- the top trajectory patterns in context of both users' location (user groups) and seasons (summer vs winter), and
- the discussion of the results.

This Specific Aim is the overall goal of the thesis.

6.1 Methodology

The goal of this Specific Aim is to find out the top frequent trajectory patterns in relation to the top 10 landmarks of each group for the entire period of time and both seasons (summer and winter). The analysis started with handling each group's photos of the top 10 landmarks from the results of the Specific Aim 2 for the entire period of time.

Photos and their corresponding owners were considered as irrelevant or duplicated in three situations and eliminated. They are a.) Any owner with only one single entry for the entire period of time. No trajectory is possible in this case; b.) Owners who have more than one uploaded photos in multiple days, but only one photo on a particular day, this record will be deleted. Since the trajectory pattern analysis was based on a single-day trajectory, this entry was not useful. It will be relevant if analysis of multi-day trajectory patterns is of interest, and c.) Owners who have taken multiple photos at the same landmark in a single day and those photos were the only photos taken on that day. No trajectories are possible. The following table (Table 24) gives the statistics on the total number of photos and the unique owners before and after the elimination.

	Before elimination		After elimination	
	Number of Photos	Number of Unique Owners	Number of Photos	Number of Unique Owners
NA	1922	202	1386	93
EU	3422	628	2238	241
AT	3530	306	981	73

Table 24 Statistics on the total number of photos and unique owners before and after the irrelevant photo elimination

The table shows that Austria has a bit more photos than Europe but only about half of the size in the total count of the unique owners before the elimination. After the elimination of the irrelevant photos, the statistics on the photos and the unique owners are relatively proportional among the groups. It shows also that massive number of irrelevant photos were eliminated for Austria and it was attributed to the reason *c* stated above. Those Flickr users in the group of Austria might have visited only one landmark on a particular day at leisure pace or for social gathering, etc.

The analysis of trajectory patterns continued with the valid dataset of each group. First, the trajectory, ascendingly ordered by the date and time, of each unique owner on each day was extracted. If some owners have photos of different landmarks on multiple days, several trajectories of those owners for each day would be depicted. Second, the sequences of landmarks visited by the photo owners were extracted. Only frequency of the same sequence of larger than or equal to two was considered. Third, the total number of the unique owners visiting the landmarks in the same sequence was counted and ranked (see Section 6.2 for results).

Since the result of trajectory patterns was based on the count of unique owners, some patterns were considered as duplicated and eliminated during the trajectory pattern analysis. They were a.) The same owner visited the same landmarks in the same sequence on different days, and b.) The same owner visited the same landmarks in the same sequence on the same day. In

summary, every sequence of visiting landmarks will only be counted once for each unique owner.

Concerning seasonal differences, the analysis started with the corresponding top 10 landmarks (from the results of Specific Aim 2) of each group for each season, i.e. summer and winter. Due to the identity of the top 10 landmarks of North American group in Winter and those for the entire period of time, some steps previously mentioned could be skipped and data analysis commenced directly by extracting those winter months from the trajectories of the unique owners for the entire period of time. However, this was not the case for the summer season and the other two groups. Table 25 and Table 26 give the statistics on the data distribution of each group for the winter and the summer season respectively before and after the irrelevant photo elimination[8]. The same two issues, mentioned earlier, on the data distribution between the EU group and the AT group can be observed as in Table 24.

	Before elimination		After elimination	
	Number of Photos	Number of Unique Owners	Number of Photos	Number of Unique Owners
NA	782	83	678	37
EU	1487	298	888	112
AT	1561	200	433	41

Table 25 Statistics on the total number of photos and unique owners of each group for winter season before and after data manipulation

8 The sum of the numbers for winter and summer season will not be the same as the data shown in Table 24 because statistics on different top 10 landmarks were dealt with.

	Before elimination		After elimination	
	Number of Photos	Number of Unique Owners	Number of Photos	Number of Unique Owners
NA	1166	128	774	57
EU	1955	355	1333	135
AT	1975	210	464	48

Table 26 Statistics on the total number of photos and unique owners of each group for summer season before and after data manipulation

6.2 Results

The final goal of the thesis is to find out the trajectory patterns of each selected group in the seasonal context. In this section, trajectory patterns, trajectory pattern frequencies and the ranking of the trajectory patterns based on the frequency are presented graphically. First of all, trajectory patterns of different periods for each group are compared. Then, trajectory patterns of the same period of time across the groups are compared.

6.2.1 Trajectory Patterns of Different Periods for Each Group and Comparison

In order to keep the significance of the results, only the top 5 ranking of the trajectory patterns of different periods for each group is shown in this section. Figure 13 presents the trajectory patterns of the EU group in different periods of time, i.e. the entire period, summer and winter. The five involved landmarks, which are represented by the red circles, are 'Stephansdom, Graben & Peterskirche', 'Kohlmarkt, Michaelerplatz, Hofburg, Heldenplatz & Burggarten', 'Albertina, Oper & Hotel Sacher', 'Parlament' and 'Rathaus & Burgtheater'. The sequence of the trajectory patterns is in the length of two.

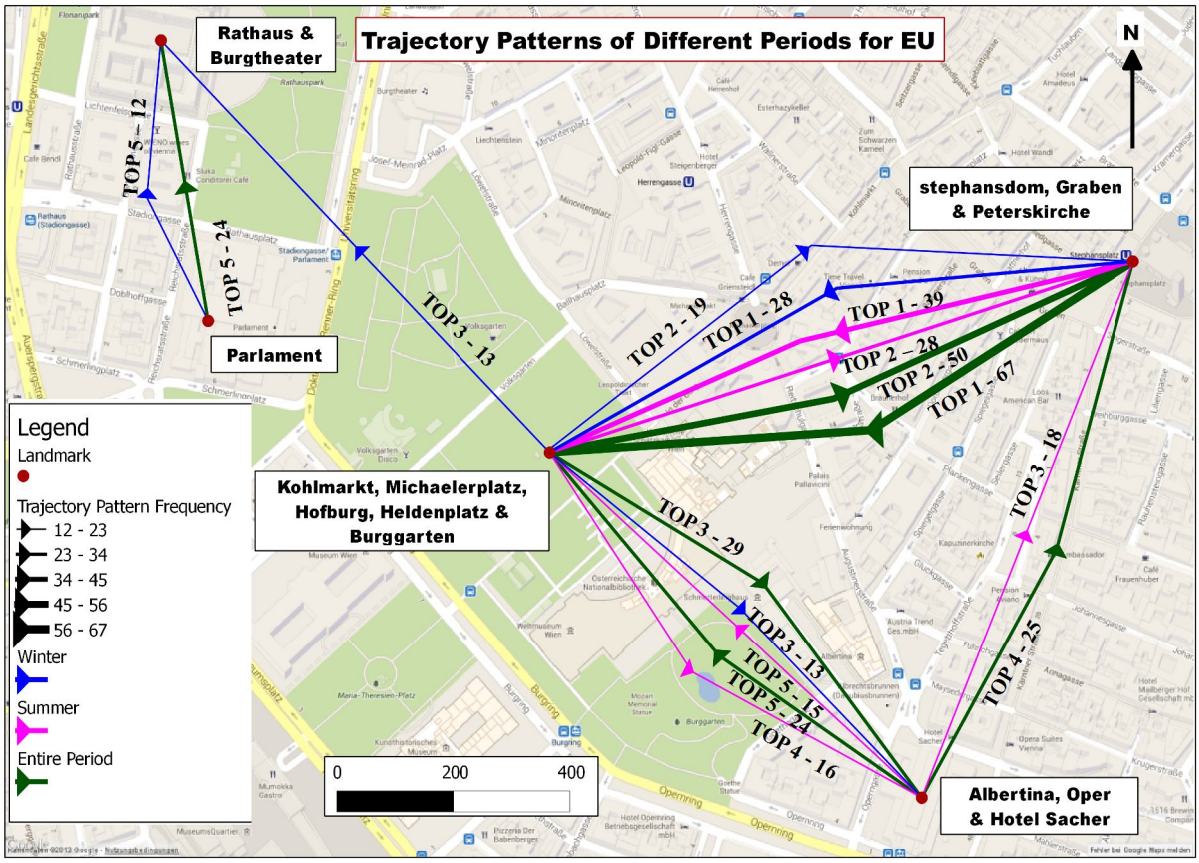

Figure 13 Trajectory patterns of different periods for the EU group

It was particularly popular for the visitors from the EU to travel from 'Stephansdom, Graben & Peterskirche' to 'Kohlmarkt, Michaelerplatz, Hofburg, Heldenplatz & Burggarten' and vice versa in all seasons. Travelling from 'Stephansdom, Graben & Peterskirche' to 'Kohlmarkt, Michaelerplatz, Hofburg, Heldenplatz & Burggarten' was, however, slightly more frequent than traversing from the latter one to the former one all year round. Also more visitors took this path in summer than in winter.

The second prevalent area to shuttle between for all seasons was from the Hofburg area to the area of the Oper. The journey in the reverse direction was only popular during summertime in the EU group. Equally popular were the sequences of visit from the Hofburg area to the Rathaus area in winter and from 'Albertina, Oper & Hotel Sacher' to 'Stephansdom, Graben &

Peterskirche' in summer. In winter, 12 visitors among the group visited the Parlament prior to the Rathaus area in the course of a day. Clearly depicted also in the figure, in winter, most visitors arrived at 'Rathaus & Burgtheater' either from the neighbouring parliament or directly from 'Hofburg' across the street.

Two additional landmarks in the trajectory pattern analysis for the AT group are 'Maria Theresia Platz, Kunsthistorisches Museum & Naturhistorisches Museum' and 'Museumsquartier', as shown in Figure 14. The sequence of the trajectory patterns for this group is in the length of two. The figure shows that the most popular trajectories for the AT group in summer were found between 'Stephansdom, Graben & Peterskirche' and 'Kohlmarkt, Michaelerplatz, Hofburg, Heldenplatz & Burggarten'. However, in winter, the most popular travelling sequence was from the Stephansdom area to the area of Hofburg, but not in the reverse order.

As popular as the travelling sequence from 'Kohlmarkt, Michaelerplatz, Hofburg, Heldenplatz & Burggarten' to 'Stephansdom, Graben & Peterskirche', 8 visitors among them also preferred stopping by 'Kohlmarkt, Michaelerplatz, Hofburg, Heldenplatz & Burggarten' before visiting 'Albertina, Oper & Hotel Sacher' in summer. With just one visitor (one count of frequency) less, the reverse sequence was the next popular trajectories in summer.

In winter, shuttling between 'Parlament' and 'Rathaus & Burgtheater' was popular among visitors of the AT group. In addition, visiting the 'Maria Theresia Platz' and its adjacent museums before heading to 'Museumsquartier' was also one of the favourite routes for nine of them. The statistics shows that the route from the parliament to the city hall was equally prevalent to the one starting from 'Maria Theresia Platz, Kunsthistorisches Museum & Naturhistorisches Museum' and ending at 'Museumsquartier' in winter.

In the summer season, visitors in this group tended to travel more often from the parliament to the landmark 'Kohlmarkt, Michaelerplatz, Hofburg, Heldenplatz & Burggarten', whereas in the winter season more visitors made their journeys in the other direction. The statistics also

shows that the route from the city hall to the parliament was as frequent as that from 'Kohlmarkt, Michaelerplatz, Hofburg, Heldenplatz & Burggarten' to the parliament during wintertime

Not surprisingly, the trajectory from the 'Stephansdom' area to the Hofburg area is in the first rank for the entire period, followed by the trajectory from the parliament to the Hofburg area. Since three equally popular trajectories are tied in the forth rank, a total number of six top trajectories is given for this designated period of time.

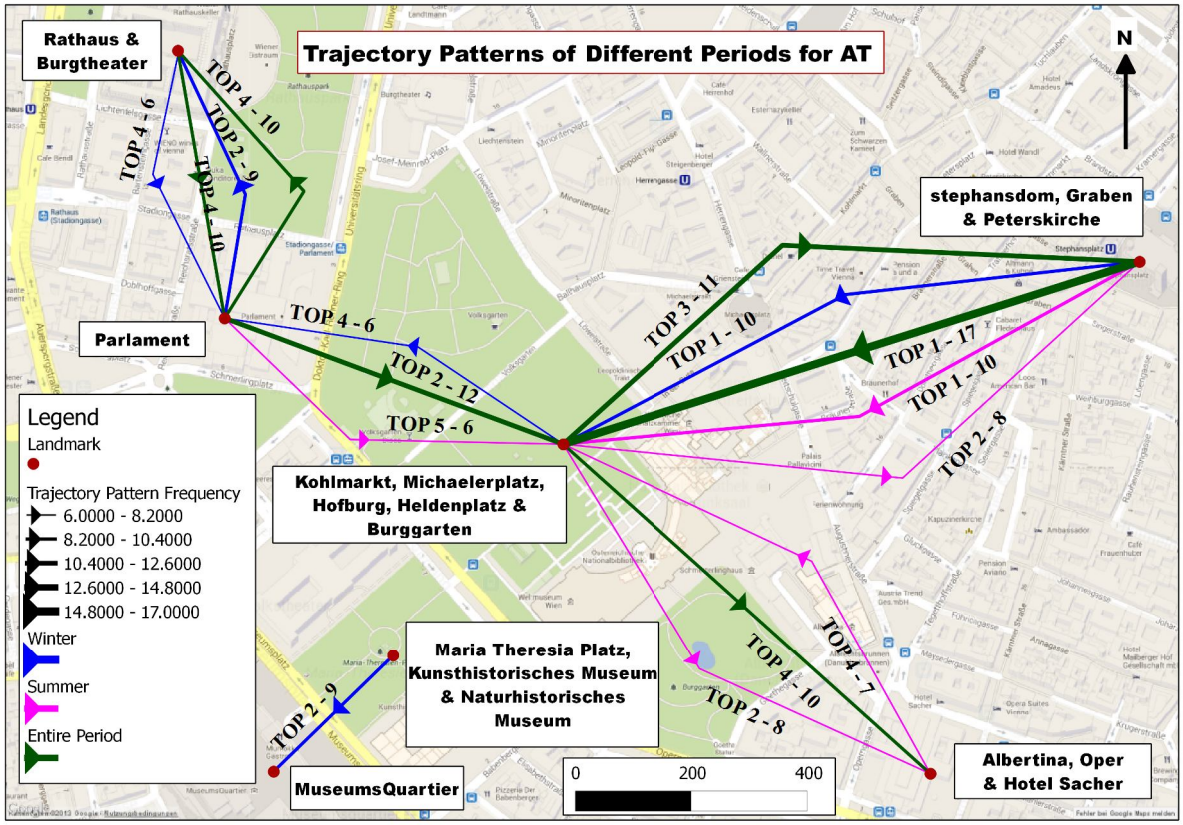

Figure 14 Trajectory patterns of different periods for the AT group

As for the other two groups, the most frequent trajectories for the NA group took place in the city centre of Vienna. The seven landmarks involved are very similar to those for the AT

group, with the exception that 'Museumsquartier' is replaced by 'Augustinerkirche Josefplatz Nationalbibliothek & Palais Pallavicini'. The sequence of the trajectory patterns is, again, in the length of two.

As shown in Figure 15, in summer, almost all the most frequent trajectory patterns originated from 'Kohlmarkt, Michaelerplatz, Hofburg, Heldenplatz & Burggarten', except the one in Top 2. These visitors of the AT group then travelled in all different directions to their next stop, i.e. the Stephen's cathedral, the city hall, 'Augustinerkirche Josefplatz Nationalbibliothek & Palais Pallavicini' and 'Maria Theresia Platz, Kunsthistorisches Museum & Naturhistorisches Museum'.

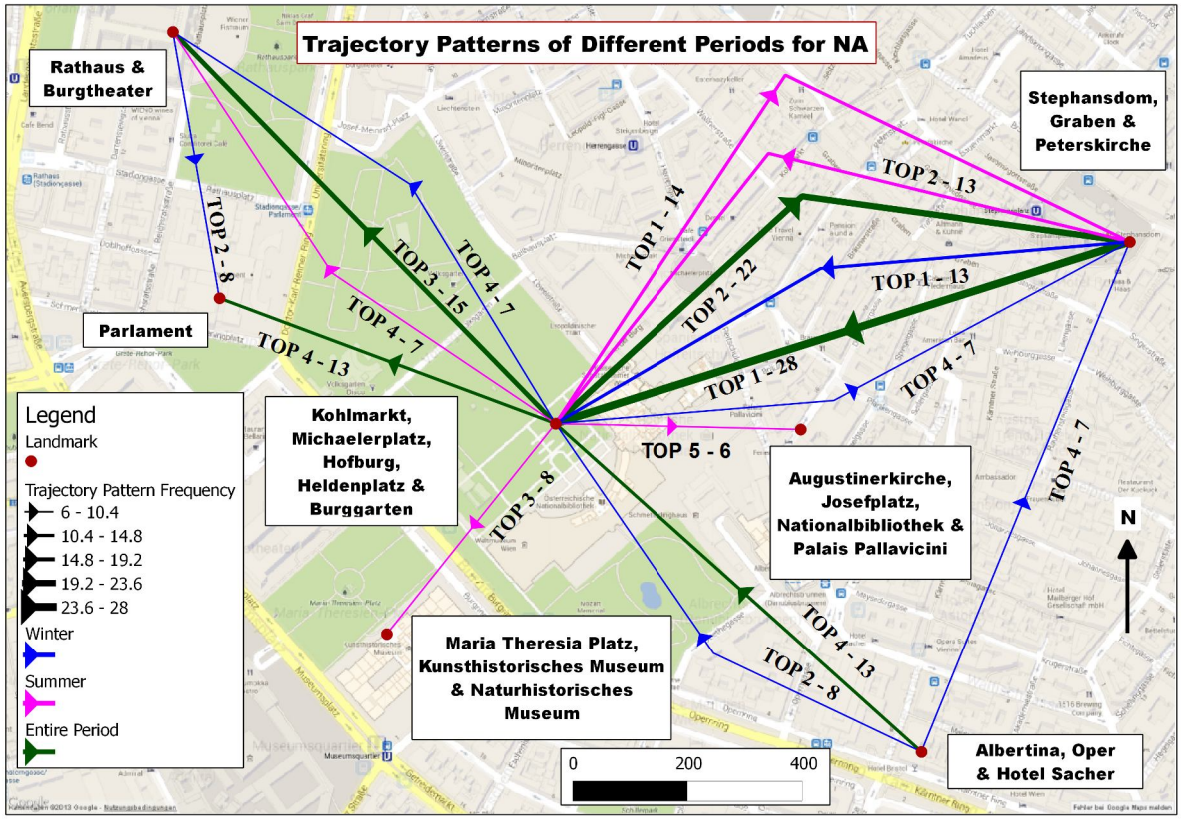

Figure 15 Trajectory patterns of different periods for the NA group

In winter, while travelling to the Hofburg area from the Stephen's cathedral was still the most popular route among the visitors of the AT group as in summer, traversing in the reverse direction, however, was less favourable (Top 4). The trajectory from the city hall to the parliament was as frequent as that from 'Albertina, Oper & Hotel Sacher' to 'Kohlmarkt, Michaelerplatz, Hofburg, Heldenplatz & Burggarten' (Top 2). The number of visitors arriving at the Stephen's cathedral from 'Kohlmarkt, Michaelerplatz, Hofburg, Heldenplatz & Burggarten' was as high as that coming from 'Albertina, Oper & Hotel Sacher'. Equally frequent in winter was the way to the city hall from the Hofburg area. Since there are three trajectories with equal frequencies in Top 4, the fifth rank is omitted, the total number of the most frequent trajectories in winter is six.

The figure also shows that the path from the Stephen's cathedral to 'Kohlmarkt, Michaelerplatz, Hofburg, Heldenplatz & Burggarten' was not the only one popular in both summer and winter. Travelling from the Hofburg area to the city hall was equally favourable in both seasons. The statistics shows that the frequencies of those popular trajectories are about the same in both seasons.

6.2.2 Trajectory Patterns of Same Periods for Different Groups and Comparison

After presenting the results of trajectory patterns of different seasons for each group in the previous section, this section describes and compares the trajectory patterns of different groups for each period of time. Due to the small size of the data, only the five most frequent trajectory patterns are presented for each case so as to keep the significance of the results. As seen in the previous section, the sequences of the trajectory patterns for all groups in all seasons are of the length of two.

Figure 16 gives the trajectory patterns of different groups for the entire period. Five landmarks were involved for the trajectory pattern analysis during this period. It depicts that the two most frequent trajectory patterns, being Top 1 and Top 2, took place between

'Stephansdom, Graben & Peterskirche' and 'Kohlmarkt, Michaelerplatz, Hofburg, Heldenplatz & Burggarten' for both EU and NA groups. Nonetheless, more visitors of the AT group stopped at the parliament before heading to the Hofburg area.

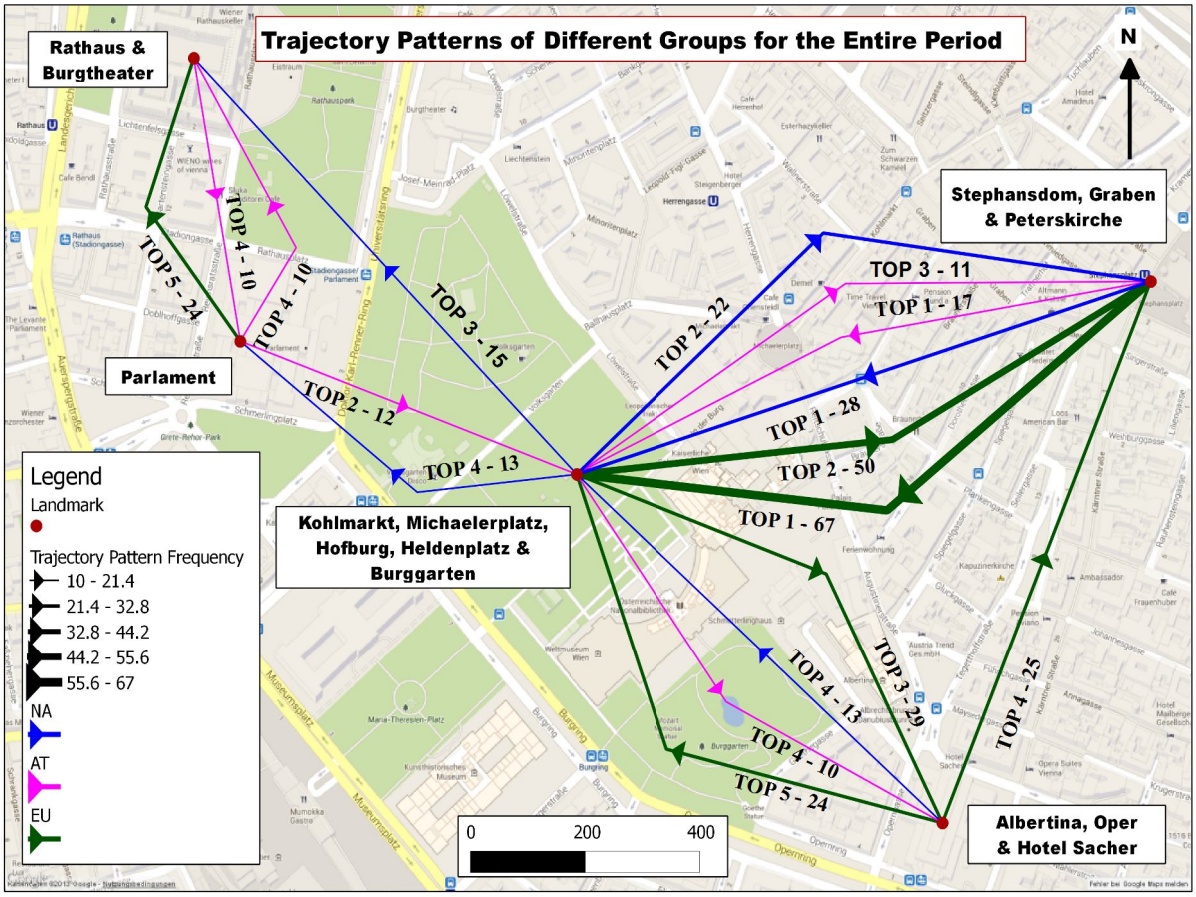

Figure 16 Trajectory patterns of different groups for the entire period

After visiting 'Kohlmarkt, Michaelerplatz, Hofburg, Heldenplatz & Burggarten', some North American visitors continued their journeys to the parliament or the city hall, whereas some visitors of both the AT group and the EU group made their stops at 'Albertina, Oper & Hotel Sacher'. For the AT group, going from the parliament to the city hall is as frequent as visiting these two places in the reverse order. This also occurred between 'Kohlmarkt, Michaelerplatz, Hofburg, Heldenplatz & Burggarten' and 'Albertina, Oper & Hotel Sacher'. Stopping at

'Albertina, Oper & Hotel Sacher' before 'Kohlmarkt, Michaelerplatz, Hofburg, Heldenplatz & Burggarten' was one of the favourite routes for the North American visitors. The opposite was true for the AT group. The majority of the visitors travelling to the Stephen's cathedral from 'Albertina, Oper & Hotel Sacher' was from the EU group.

The statistics shows that the range of the trajectory pattern frequency for the NA group is about twice as large as that for the AT group. That for the EU group is about six-fold larger than that for the AT group.

In summer, seven landmarks in the city centre of Vienna were involved in the analysis of the trajectory patterns of different groups (Figure 17). Going back and forth between 'Stephansdom, Graben & Peterskirche' and 'Kohlmarkt, Michaelerplatz, Hofburg, Heldenplatz & Burggarten' were the two most frequent exhibited trajectory patterns across the groups.

After visiting 'Kohlmarkt, Michaelerplatz, Hofburg, Heldenplatz & Burggarten', some visitors from the NA group stopped at the city hall, 'Augustinerkirche Josefplatz Nationalbibliothek & Palais Pallavicini' or 'Maria Theresia Platz, Kunsthistorisches Museum & Naturhistorisches Museum', apart from visiting the Stephansdom area. These travelling patterns were observed only in the NA group, but not with the other two groups. The route from 'Albertina, Oper & Hotel Sacher' to 'Stephansdom, Graben & Peterskirche' was frequently travelled only by the EU group, the route from the parliament to the Hofburg area across the street was only taken by the AT group.

Travelling from the Hofburg area to 'Albertina, Oper & Hotel Sacher' and vice versa was also a frequent travelling pattern, as exhibited in both the AT group and the EU group. The statistics shows that the range of the trajectory pattern frequency for the NA group is twice as large as that for the AT group, while that for the EU group is even six-fold as large as that for the AT group. This phenomenon is exactly the same as that for the entire period.

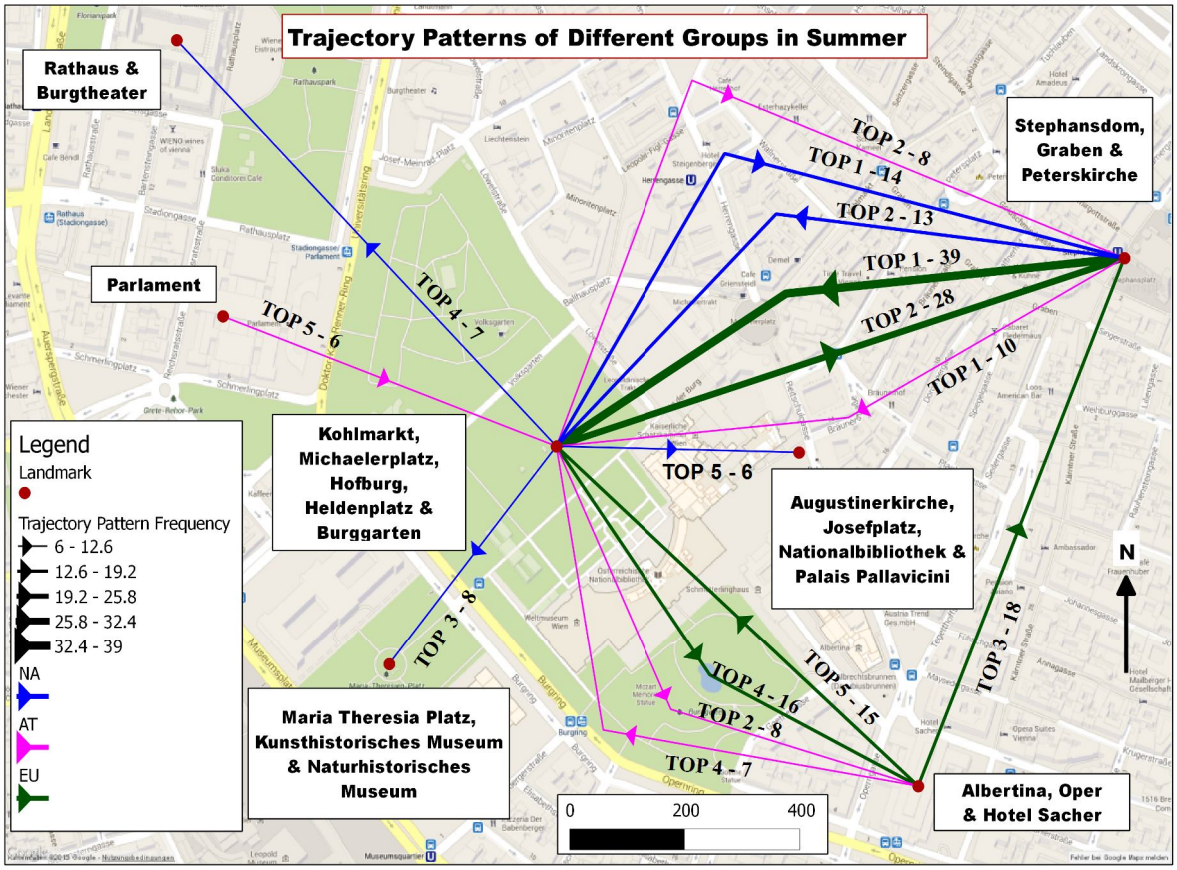

Figure 17 Trajectory patterns of different groups in summer

Figure 18 presents the trajectory patterns of different groups in winter. While the number of the landmarks involved for the trajectory pattern analysis remains the same for both seasons, 'Augustinerkirche Josefplatz Nationalbibliothek & Palais Pallavicini' is substituted by 'Museumsquartier' for the winter season.

While it was very popular for all three groups to approach the landmark 'Kohlmarkt, Michaelerplatz, Hofburg, Heldenplatz & Burggarten' after 'Stephansdom, Graben & Peterskirche', the route in the reverse direction was frequently travelled in winter by the EU group and the NA group only.

The NA group showed a particular trajectory pattern in winter, i.e. travelling from 'Albertina,

Oper & Hotel Sacher' to 'Stephansdom, Graben & Peterskirche' which the EU group interestingly exhibited in summertime Additionally, compared to the summer (Figure 17), more visitors of all groups, especially in the AT group, shuttled between the parliament and the city hall in winter. While the NA group stopped first at the city hall and then the parliament afterwards, the EU group exhibited the reverse trajectory pattern in that area.

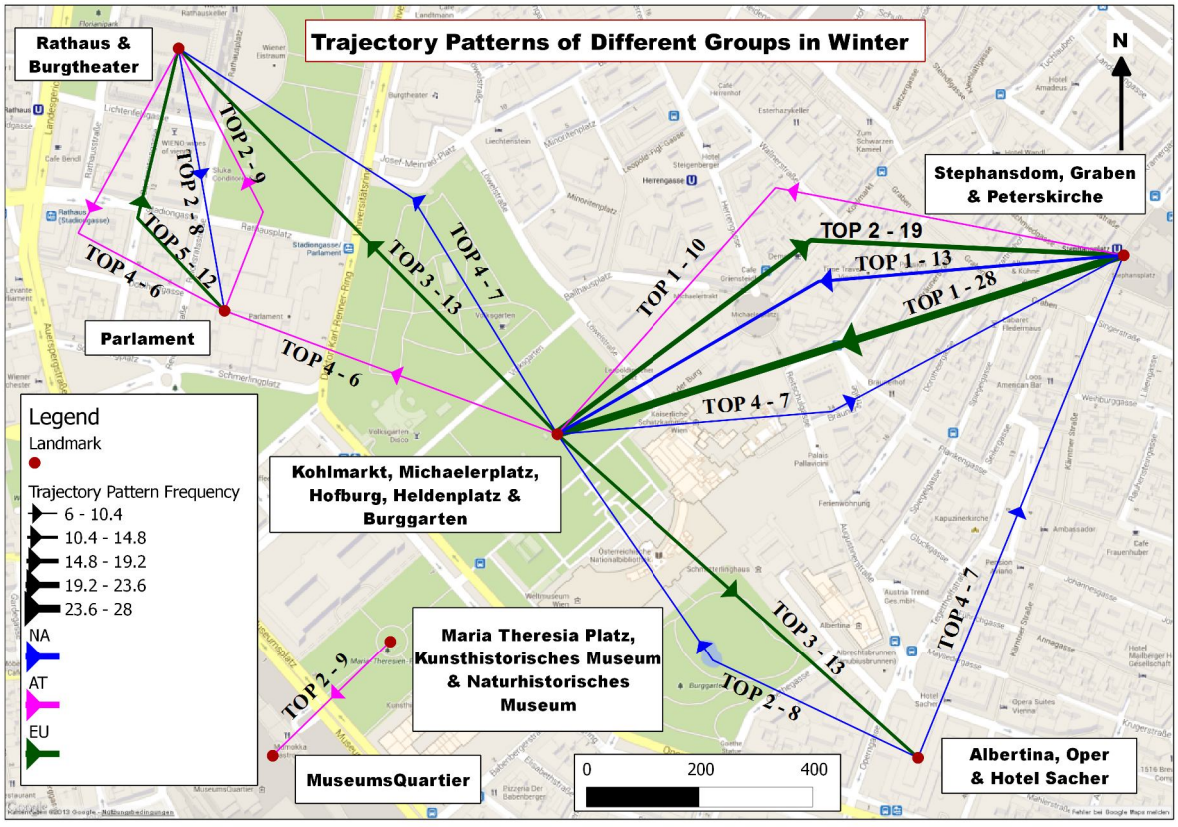

Figure 18 Trajectory patterns of different groups in winter

Interestingly, in winter all three groups travelled north-westwards after taking photos at the Hofburg area, either to the city hall and the theatre or the adjacent parliament. Specifically for the EU group, the visitors also preferred heading to 'Albertina, Oper & Hotel Sacher' from the Hofburg area. It was exactly the opposite for the NA group. A route taken more often by the AT group was crossing the street from the Maria Theresa's square and the neighbouring museums to the 'Museumsquartier'.

While the range of the trajectory pattern frequencies in winter was similar between the NA and the AT group, it was four-fold larger in the EU group. These ranges in winter were different from those of the entire period and those of summer. When in comparison to the AT group, they were twice as large in the NA group and six-fold larger in the EU group (see Page 75).

6.3 Discussion

In this thesis, the study groups were categorized according to the Flickr users' geographical location, i.e. where the Flickr users resided at the time of data collection (Definition 8). It is also possible to categorize the Flickr users into locals and tourists with the approach described by Choudhury et. al. (2010). Alternatively, we could have followed the approach of Choudhury et. al.(2010). These researchers have used the duration of their stay to filter the local residents from tourists. As long as a user took his/her first photo and the last one in a time span of no more than 21 days and visited at least two sites during the stay, he/she was considered as a tourist. This definition was based on the idea that the locals take photos over a longer period of time and they visit the places without time constraint. In that way, they could analyse the trajectory patterns of tourists.

This approach; however, has two major limitations. If this approach were adopted to the present study, short-term visitors to Vienna for business would have been categorized as tourists. However, daytime activities of business travellers are largely restricted by the schedules of the meetings, conferences or the training programmes, etc. Although they might still visit Vienna at night like a tourist, their trajectory patterns could be very different. For example, it is very unlikely for them to visit 'Schoenbrunn' after a full-day conference. Hence, it is not justified to put this crowd into the tourist-group and compare their trajectory patterns with all other regular tourists.

Another limitation is that this approach would not sufficiently take into account the recurring

visitors to Vienna. These visitors would most likely be misclassified as locals based on the large time span from their first photo to the last one, for instance a period of several years. Although it might be true that the trajectory patterns of these multi-time visitors could be different from the first-time visitors, they are not locals and would travel under pressure with time constraint as much as the first-time visitors do for each visit. Another more sophisticated method has recently been introduced by Zheng et al. (2012) who presented a probabilistic model to distinguish tourists from non-tourists.

Instead of categorizing the Flickr users into locals versus tourists, it is more interesting to compare the trajectory patterns of Vienna visitors from different regions of the world. Thus, Flickr users were grouped according to their location. The underlying assumptions are: a.) this location represents their centre of life, i.e. their residences at the time of data collection; b.) it is not changed in their profiles just because of travelling, it would be updated only when they really relocate to another city.

One of the factors that can affect the ranking of the trajectory patterns is the accuracy of the timestamps of the Flickr photos. There were two issues associated with the timestamps. Firstly, the setting of the date and time might not be adjusted by the Flickr users when they travelled to Vienna from another time zone, resulting in the deviation with the local time of Vienna. Secondly, if the information of date and time was missing when the users uploaded the photos to Flickr, the attribute 'datetaken' (the date the photo being taken) would be defaulted by Flickr to the date uploaded (attribute of 'dateuploaded'), as also mentioned by Choudhury et al. (2010). Solution was sought to resolve the latter problem on the default setting, the relevant records were further examined. The 'datetaken' was counter-checked with the 'dateupload'. The 'datetaken' is represented on Flickr in date format; however, the 'dateupload' is the number of seconds since 1970 that the photo was uploaded. Prior to comparison, conversion of the 'datetaken' was necessary. The comparison shows that all relevant records had different dates and times than the 'dateupload'.

The elimination process as outlined in Page 65 served another purpose: we were able to

estimate the number of local non-tourists. As evident in Table 24, the number of unique owners of the AT group dropped dramatically after photo-elimination, from 306 to 73. Such a significant drop in the number of unique owners could be attributed to the heavy involvement of the locals. The locals might visit the landmarks and took photos for a different purpose, e.g. gathering with friends or attending social events. The photo(s) taken in such case might involve just one particular venue and no trajectory patterns can be observed. Based on the rules (a) and (c), stated in Page 65, these records were deleted, resulting in the decrease of the number of unique owners.

As mentioned in Section 4.1, Vienna and the rest of Austria was grouped together at the beginning of the analysis for three major reasons. Firstly, there were 3627 records of Austria without specific city name. Secondly, the greater Vienna was considered as the 'grey zone' area. Lastly, the sample size of 'Austria without Vienna' was not large enough, even if Austria would be divided into the groups of 'Austria without Vienna' and 'Vienna' in the first place. It is still interesting to know the composition of the AT group in the results of the overall goal – Specific Aim 3. Table 27 gives the statistics of the composition of the AT group in the results of trajectory pattern analysis.

Location	Number of Unique Owners	%
Vienna	57	77.0
Greater Vienna	4	5.4
Other part of Austria	7	9.5
cannot be determined[9]	6	8.1
Total	74	100.0

Table 27 Composition of the AT group in the results of trajectory pattern analysis

It is clear that a large portion of the remaining Flickr-Photo owners constituting the AT group was from Vienna (77%). Hence, the trajectory patterns of the AT group were largely made up by those Flickr users, who resided in Vienna during the designated period.

9 Only country name was available (city/state name was not specified).

Another way to uncover the context-based differences in trajectory patterns was by subtracting the frequency of the least frequent one (“Top 5”) from the frequency of the most frequent trajectory pattern (“Top 1”). That gives us the ranges in trajectory pattern frequency. A larger range of the trajectory pattern frequency in the group implied a stronger preference of the group to travel in the Top 1 pattern. Overall, the Austrian group exhibited the smallest ranges, the EU group the largest ranges with the NA group falling in between with seasonal differences. We can conclude from these variations that the EU group had the strongest preference for the Top 1 trajectory pattern, the AT group the least, and the NA group fell in between with seasonal differences. Classifying the Flickr data in the context of both user groups and seasons made this type of findings possible.

Another distinct feature derived from the trajectory pattern analysis was that the landmarks 'Stephansdom, Graben & Peterskirche' and 'Kohlmarkt, Michaelerplatz, Hofburg, Heldenplatz & Burggarten' were often visited together. The route between these two popular landmarks was taken by all groups in all different periods of time. This homogeneity of travelling behaviour in groups reflected not only how popular these two places were, but also the topological relation of these two places. This result shows that the visitors did not travel randomly, but deliberately with plans based on the topological characteristic of the surroundings. This is in line with Zheng et al. (2012), and Asakura and Iryo (2007) who investigated the relation of the topological characteristics of the POIs to the tourist movement patterns.

In summary, some key interesting trajectory patterns of different Flickr user groups (according to their location) in different seasons were found. This makes the current research novel. We can see that the route from 'Maria Theresia Platz' and its neighbouring museums to 'Museumsquartier' was only frequented by the AT group in winter, while the path going to 'Augustinerkirche, Josefplatz, Nationalbibliothek & Palais Pallavicini' from the landmark 'Kohlmarkt, Michaelerplatz, Hofburg, Heldenplatz & Burggarten' was special to the NA group in summer. Another seasonal difference was that in summer there was no frequent trajectory patterns between the city hall and the parliament. That was quite the contrary in

winter (Figure 17 & Figure 18). This could be attributed to the popular events held in front of the city hall starting from late November to early March, e.g. 'Christkindlmarkt'; 'Silvesterpfad' & ice skating. Another possible reason is that the city hall is lit up earlier in winter because of the early sunset. The visitors could enjoy the beautiful night view of the city hall and have the photos of the lit-up city hall taken earlier.

CHAPTER 7 CONCLUSIONS, LIMITATIONS AND FUTURE WORK

7.1 Conclusions

This research explores the potential of mining trajectory patterns in the context of user groups and seasons using user-generated content inherent in social media. Geo-spatial and temporal information allows construction of Flickr users' travelling patterns. In order to reach this overall goal, three specific aims have been defined and followed in sequence: firstly, identification of the significant Flickr user groups and comparison with an official statistical data source - Statistik Austria; secondly, ranking of the 10 top landmarks, and finally, extraction of the frequent trajectory patterns. This yields interesting findings on the trajectory patterns of different user groups and seasons, although in general they are similar. The study confirms that it is possible to make use of user-contributed content inherent in social media to discover trajectory patterns in contexts. Section 7.2 highlights some challenges which have yet to be addressed in the future.

7.2 Limitations and Future Work

One of the limitations of this research is the questionable accuracy of attributes, in particular, the Flickr users' textual information. For example, it is possible that a user stood in front of 'Musikverein' and took a photo of 'Karlskirche', but mistakenly tagged the photo as 'Musikverein'. The photo features actually the 'Karlskirche' across the street, the photo would; however, be allocated to the landmark 'Musikverein' based on the geo-location and the textual attributes. In a more extreme case, both the geo-location and the tags could be incorrect. The photographed object could be something totally unrelated. Ways to resolve this issue are a.) to perform the data cleaning prior to the analysis, and b.) to further investigate the image content as the third factor in addition to the geo-location information and textual attribute.

Some textual attributes (tags; title & descriptions) were written in languages other than German or English. Fortunately, this issue was not a major concern, because the names of the landmarks were recognizable even when they were written in other languages. To systematically tackle this issue in the future, methods have to be implemented to automatically detect and translate the concerned records to the desirable languages.

In this research, the study grouping was based on the user's geographical location which can be reasonably assumed as the user's actual main residence. However, this did not take into account where the user originally came from, where he/she had grown up or had been educated. The arising problem can be exemplified by the following case: an Austrian national who is born in Vienna, but is actually living in North America. He/she is classified in this study as North American based on his/her geographical location that he/she put into his/her Flickr profile. However, he/she will exhibit a travelling pattern rather different from that of a native North American when he visits his home town Vienna. Most likely; however, the trajectories of this atypical visitor would have been treated as outliers and eliminated because of their rarity (trajectory pattern frequency < 2). Eventually, only the most frequent five trajectory patterns have been analysed in the current study.

This research investigates trajectory patterns in the context of user-groups (with reference to the attribute 'Location' in the users' profile) and seasons. The user-groups were constituted by both locals and tourists. There was no further distinction between them in the present research. With the current Flickr dataset, it is not favourable to further differentiate the user groups into tourists and locals. Otherwise, the 5 top trajectory pattern frequencies would be too low. If that were to be considered, it would be necessary to extend the period of the data collection so that sufficient data are available for analysis. In addition, prudent definitions of the entities 'local' and 'tourist' would need to be outlined prior to data analysis as has been proposed by Choudhury et. al., 2010 and Zheng et al. (2012).

The set of parameters chosen for the data clustering algorithm was by no means the only set of parameters applicable for the Flickr data. Neither should it be considered as the best.

However, this set of empirically tested parameters proved the validity of the approach for data clustering with the study area in question, and afterwards, landmark identification was possible. To obtain another better set of parameters, depending also on the study area, more extensive empirical experimenting on parameters should be carried out.

The present study used geo-location information of the tagged Flickr-photos as the primary source to determine the location of landmarks. This geo-location information originated either from the users' GPS-equipped cameras or from their manual input. The accuracy of this information has not been evaluated. Accuracy assessment on the geo-location might be required in the future. The same applies to the reliability of the resulting trajectory patterns. Statistical test could be performed, as described in Zheng et. al. (2012).

As mentioned in Section 5.3, time zone shift can be one of the factors affecting the outcomes of trajectory pattern analysis. Although it is true that during the course of a day, the sequence of a trajectory is of prime concern instead of the exact timestamps, this will cost a certain impact on the accuracy of trajectory pattern mining when dealing the stops at the borders of a day. Solutions have to be sought in order to tackle this issue.

The current research can be extended to study the multi-day trajectories. In that case, we would look at the trajectory patterns in successive days for each Flickr-photo owner. In addition, we could investigate when the trajectory patterns occurred during daytime; e.g. in the morning, in the afternoon or in the evening and at night. In that instance, the accuracy of the timestamps would have to be guaranteed with particular consideration of time zone shift. By analysing the timestamps, we could also estimate the minimum duration of a Flickr-photo owner's visit to each landmark. An itinerary suggestion system could then be developed to provide travel plans, tailor-made for the visitors of different countries of origin.

APPENDIX A AN EXAMPLE OF API SEARCH RESULTS IN XML FORMAT ON FLICKR

```
<?xml version="1.0" encoding="utf-8" ?>
<rsp stat="ok">
<photos page="1" pages="6" perpage="500" total="2848">
        <photo id="2396272657" owner="36777255@N00" secret="47e1a97630"
        server="2246" farm="3" title="verticalidad gótica desde la ventana" ispublic="1"
        isfriend="0" isfamily="0" dateupload="1207602791" datetaken="2008-02-29
        01:05:17" datetakengranularity="0" latitude="48.208348" longitude="16.371581"
        accuracy="16" place_id="G7TwxipUV7uNr80" woeid="551759"
        geo_is_family="0" geo_is_friend="0" geo_is_contact="0" geo_is_public="1"
        tags="vienna wien architecture hotel austria design arquitectura cathedral catedral
        viena osterreich diseño doco stsptephans"
        url_z="http://farm3.static.flickr.com/2246/2396272657_47e1a97630_z.jpg?zz=1"
        height_z="640" width_z="426">
                <description>Do&amp;amp;amp;co Hotel. Viena</description>
        </photo>
        <photo id="2397106824" owner="36777255@N00" secret="549b6f986d"
        server="2351" farm="3" title="Por la noche" ispublic="1" isfriend="0"
        isfamily="0" dateupload="1207602796" datetaken="2008-02-29 01:08:04"
        datetakengranularity="0" latitude="48.208348" longitude="16.371581"
        accuracy="16" place_id="G7TwxipUV7uNr80" woeid="551759"
        geo_is_family="0" geo_is_friend="0" geo_is_contact="0" geo_is_public="1"
        tags="vienna wien architecture club hotel austria design arquitectura cathedral
        catedral viena osterreich diseño doco ststephans"
        url_z="http://farm3.static.flickr.com/2351/2397106824_549b6f986d_z.jpg?zz=1"
        height_z="426" width_z="640">
                <description>ambiente muy pijo; con pose; sobre todo pose.</description>
        </photo>
</photos>
</rsp>
```

APPENDIX B PROGRAMMING SCRIPTS IN PYTHON FOR COLLECTING USER'S LOCATION DATA ON FLICKR WITH PHOTO RECORDS (IN EXCEL)

```
'''
Created on 19 Nov 2012
Based on beejs Flickr API
Produce a list of owner's info for a given
list of owner's ID on Flickr

'''
import flickrapi
import xlrd
import xlwt
import string
import datetime
import time

api_key = '7d8a2835eee80ca8bcc66315bc02883f'

wb = xlrd.open_workbook('Data2011b.xls')
sh = wb.sheet_by_name(u'Sheet1')
third_column = sh.col_values(2)

wbk = xlwt.Workbook()
sheet = wbk.add_sheet('sheet 1',cell_overwrite_ok=True)
head = ['nsid', 'username', 'realname', 'location']

colnum = 0
col_num = 0
row_num = 1

for i in range(len(head)):
        sheet.write(0,colnum,'%s' %(head[i]))
        colnum += 1
```

```
#fetch owner's Info

if __name__ == '__main__':

        flickr = flickrapi.FlickrAPI(api_key)
        for j in range(len(third_column)):
                try:
                        response_person =
                                flickr.people_getInfo(user_id='%s' %(third_column[j]))

                if response_person.attrib['stat'] == 'ok':
                                for person in response_person:

                                        nsid = person.get('nsid')

                                if not response_person.findall(".//person/username"):
                                        Uname = 'no element'
                                else:
                                        for username in
                                                response_person.findall(".//person/username"):

                                                Uname = username.text

                                if not response_person.findall(".//person/realname"):
                                        Rname = 'no element'
                                else:
                                        for realname in
                                        response_person.findall(".//person/realname"):
                                                Rname = realname.text

                                if not response_person.findall(".//person/location"):
                                        loc = 'no element'
                                else:
                                        for location in
                                        response_person.findall(".//person/location"):
                                                loc = location.text
```

```
                        info=['%s' %nsid,'%s' %Uname,'%s' %Rname,'%s' %loc]
                        print '%s' %info

                        for k in range(len(info)):
                                sheet.write(row_num,col_num,'%s' %(info[k]))
                                col_num += 1

                        row_num += 1
                        col_num = 0

                except:
                        pass

wbk.save('getFlickrData_2011b.xls')
```

BIBLIOGRAPHY

Androgen, N., & Androgen, G. (2011). Spatial Generalization and Aggregation of Massive Movement Data. *IEEE Transactions on Visualization and Computer Graphics*, *17*(2), 205-219.

Arase, Y., Xie, X., Hara, T., & Nishio, S. (2010). Mining People's Trips from Large Scale Geo-tagged Photos. *In Proceedings of the international conference on Multimedia. MM'10* (pp. 133-142).

Asakura, Y., & Iryo, T. (2007). Analysis of Tourist Behaviour based on the Tracking Data Collected using Mobile Communication Instrument. *Transportation Research Part A: Policy and Practice*, *41*(7), 684–690.

Avrithis, Y., Kalantidis, Y., Spyrou, E., & Tolias, G. (2010). Retrieving Landmark and Non-Landmark Images from Community Photo Collections Categories and Subject Descriptors. *In Proceedings of the International Conference on Multimedia. MM'10* (pp. 153-162). New York, New York, USA: ACM.

Bogorny, V., Kuijpers, B., Tietbohl, A., & Alvares, L. O. (2007). Spatial Data Mining : From Theory to Practice with Free Software. *In Proc. of WSL International Workshop on Free Software. WSL'07*.

Bogorny, V., Palma, A. T., Engel, P. M., & Alvares, L. O. (2006). Weka-GDPM – Integrating Classical Data Mining Toolkit to Geographic Information Systems. *In SBBD Workshop on Data Mining Algorithms and Applications WAAMD'06* (pp. 16-20).

Bogorny, V., & Wachowicz, M. (2009). A Framework for Context-Aware Trajectory Data Mining. *Data Mining for Business Applications* (pp. 225-239). Springer.

Cao, L., Luo, J., Gallagher, A., Jin, X., Han, J., & Huang, T. S. (2010). A Worldwide Tourism Recommendation System Based On geo-tagged Web Photos. *In Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing. ICASSP'10* (pp. 2274-2277).

Chen, L., Ozsu, M. T., & Oria, V. (2005). Robust and Fast Similarity Search for Moving Object Trajectories. *In Proceedings of the ACM SIGMOD International Conference on Management of Data.* (pp. 491-502). New York, New York, USA: ACM.

Chen, W. C., Battestini, A., Gelfand, N., & Setlur, V. (2009). Visual Summaries Of Popular Landmarks From Community Photo Collections. *IEEE Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers* (pp. 1248-1255). IEEE. doi: 10.1109/ACSSC.2009.5469962.

Cheng, A. J., Chen, Y. Y., Huang, Y. T., Hsu, W. H., & Liao, H. Y. M. (2011). Personalized Travel Recommendation by Mining People Attributes from Community-Contributed Photos. *In Proceedings of the 19th ACM International Conference on Multimedia. MM'11* (pp. 83-92). New York, New York, USA: ACM.

Cheng, Z., Caverlee, J., Lee, K., & Sui, D. Z. (2011). Exploring Millions of Footprints in Location Sharing Services. *In Proceedings of the 5th International AAAI Conference on Weblogs and Social Media* (Vol. 2010, pp. 81-88).

Crandall, D. J., Backstrom, L., Huttenlocher, D., & Kleinberg, J. (2009). Mapping The World's Photos. *In Proceedings of the 18th International Conference on World Wide Web - WWW'09* (p. 761). New York, New York, USA: ACM Press. doi: 10.1145/1526709.1526812.

De Choudhury, M., Feldman, M., Amer-Yahia, S., Golbandi, N., Lempel, R., & Yu, C. (2010). Automatic construction of travel itineraries using social breadcrumbs. *In Proceedings of the 21st ACM conference on Hypertext and hypermedia - HT'10* (pp. 35-44). New York, New York, USA: ACM Press. Doi: 10.1145/1810617.1810626.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *in Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining. KDD-96* (p. 226–231). AAAI Press.

Girardin, F., & Blat, J. (2008). Assessing Pervasive User-Generated Content To Describe Tourist Dynamics. *In Proceedings of First International Workshop on Trends in pervasive and Ubiquitous Geotechnology and geo-information*. New York, New York, USA: ACM.

Humphreys, L. (2011). Mobile geo-tagging : Reexamining Our Interactions with Urban Space. *Journal of Computer-Mediated Communication, 16*, 407-423. doi: 10.1111/j.1083-6101.2011.01548.x.

Iryo, T., & Asakura, Y. (2007). Analysis of Tourist Behaviour based on the Tracking Data Collected using Mobile Communication Instrument. *Transportation Research Part A: Policy and Practice, 41*(7), 684–690.

Jankowski, P., Androgen, N., Androgen, G., & Kisilevich, S. (2010). Discovering Landmark Preferences and Movement Patterns from Photo Postings. *Transactions in GIS*, *14*(6), 833-852. doi: 10.1111/j.1467-9671.2010.01235.x.

Ji, R., Xie, X., & Yao, H. (2009). Mining City Landmarks from Blogs by Graph Modeling. *In Proceedings of the 17th ACM International Conference on Multimedia. MM'09* (pp. 105-114). New York, New York, USA: ACM.

Kalogerakis, E., Vesselova, O., Hays, J., Efros, A. A., & Hertzmann, A. (2009). Image Sequence Geolocation with Human Travel Priors. *In Proceedings of 12th International Conference on Computer Vision* (pp. 253 - 260).

Kawai, Y., Zhang, J., & Kawasaki, H. (2009). Tour Recommendation System Based on Web Information and GIS. *In Proceedings of IEEE International Conference on Multimedia and Expo* (pp. 990-993).

Kawakubo, H., & Yanai, K. (2009). An Analysis of the Relation between Visual Concepts and Geo-Locations using geo-tagged Images on the Web. *In Proceedings of the 2009 IEEE international conference on Multimedia and Expo. ICME'09* (pp. 1644-1647).

Kennedy, L. S., & Naaman, M. (2008). Generating Diverse And Representative Image Search Results For Landmarks. *In Proceeding of the 17th International Conference on World Wide Web - WWW'08* (pp. 297-306). New York, New York, USA: ACM Press. doi: 10.1145/1367497.1367539.

Kennedy, L., Naaman, M., Ahern, S., Nair, R., & Rattenbury, T. (2007). How Flickr Helps us Make Sense of the World : Context and Content in Community-Contributed Media Collections. *In Proceedings of ACM Multimedia. MM'07* (pp. 631-640). New York, New York, USA: ACM.

Kurashima, T., Iwata, T., Irie, G., & Fujimura, K. (2010). Travel Route Recommendation Using geo-tags in Photo Sharing Sites. *In Proceedings of the 19th ACM International Conference on Information and Knowledge Management. CIKM'10* (pp. 579-588). New York, New York, USA: ACM.

Kádár, B. (2013). A Morphological Approach In Defining The Causes Of Tourist-Local Conflicts In Tourist-Historic Cities. Presented at the International RC21 Conference, Berlin (Germany), 29-31 August 2013. Session:17. Resistance and Protest in the Tourist City.

Lau, G., & Mckercher, B. (2007). Understanding Tourist Movement Patterns In A Destination : A GIS Approach. *Tourism and Hospitality Research, 7*(1), 39-49. doi: 10.1057/palgrave.thr.6050027.

Leung, X. Y., Wu, B., Wang, F., Xie, Z., & Bai, B. (2012). A Social Network Analysis of Overseas Tourist Movement Patterns in Beijing: The Impact of the Olympic Games. *International Journal of Tourism Research, 14*(5), 469–484.

Lew, A., & Mckercher, B. (2006). Modeling Tourist Movements: A Local Destination Analysis. *Annals of Tourism Research*, *33*(2), 403-423. doi: 10.1016/j.annals.2005.12.002.

Lu, X., Wang, C., Yang, J. M., Pang, Y., & Zhang, L. (2010). Photo2Trip : Generating Travel Routes from Geo-Tagged Photos for Trip Planning. *In Proceedings of the International Conference on Multimedia. MM'10* (pp. 143-152). New York, New York, USA: ACM.

Majid, A., Chen, L., Chen, G., Mirza, H. T., & Hussain, I. (2012). GoThere : Travel Suggestions using geo-tagged Photos. *In Proceedings of the 21st International Conference Companion on World Wide Web* (pp. 577-578). New York, New York, USA: ACM.

Majid, A., Chen, L., Chen, G., Mirza, H. T., Hussain, I., & Woodward, J. (2013). A Context-Aware Personalized Travel Recommendation System Based On geo-tagged Social Media Data Mining. *International Journal of Geographical Information Science*, *27*(4), 662-684. doi: 10.1080/13658816.2012.696649.

Mckercher, B., & Lau, G. (2008). Movement Patterns of Tourists within a Destination. *Tourism Geographies: An International Journal of Tourism Space, Place and Environment*, *10*(3), 355-374.

Millonig, A., & Gartner, G. (2009). Ways of Walking – Developing a Pedestrian Typology for Personalised Mobile Information Systems. *In Location Based Services and TeleCartography II* (pp. 79-94). Berlin Heidelberg: Springer.

Oliveira, M. G. D., & Baptista, C. D. S. (2012). Geostat – A System For Visualization , Analysis And Clustering Of Distributed Spatiotemporal Data. *In Proceedings XIII geo-info* (pp. 108-119).

Oliveira, M. G. D., Baptista, C. D. S., & Falcão, A. G. R. (2012). A Web-based Environment for Analysis and Visualization of Spatio-temporal Data provided by OGC Services. *In Proceedings of GEOProcessing 2012, the 4th International Conference on Advanced Geographic Information Systems, Applications, and Services* (pp. 183-189).

Popescu, A., & Grefenstette, G. (2009). Deducing Trip Related Information From Flickr. *In Proceedings of the 18th International Conference on World Wide Web - WWW'09* (pp. 1183-1184). New York, New York, USA: ACM Press. doi: 10.1145/1526709.1526919.

Popescu, A., & Moëllic, P. A. (2009). MonuAnno : Automatic Annotation of geo-referenced Landmarks Images. *In Proceedings of the ACM International Conference on Image and Video Retrieval* (p. 11). New York, New York, USA: ACM.

Rattenbury, T., Good, N., & Naaman, M. (2007a). Towards Extracting Flickr Tag Semantics. *In Proceedings of the 16th International Conference on World Wide Web* (pp. 1287-1288). New York, New York, USA: ACM.

Rattenbury, T., Good, N., & Naaman, M. (2007b). Towards Automatic Extraction of Event and Place Semantics from Flickr Tags. *In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 103-110). New York, New York, USA: ACM.

Rattenbury, T., & Naaman, M. (2009). Methods for Extracting Place Semantics from Flickr Tags. *ACM Transactions on the Web*, *3*(1), 1. doi: 10.1145/1462148.1462149.

Standard-Dokumentation Metainformationen Beherbergungsstatistik : Monatliche Nächtigungsstatistik Jährliche Bestandsstatistik. (2003). *Statistik Austria*. Retrieved from http://www.statistik.at/web_de/statistiken/tourismus/index.html.

Stvilia, B., & Jörgensen, C. (2009). User-Generated Collection-Level Metadata In An Online Photo-Sharing System. *Library & Information Science Research*, *31*(1), 54-65. doi: 10.1016/j.lisr.2008.06.006.

Tussyadiah, I. P., & Fesenmaier, D. R. (2007). Interpreting Tourist Experiences From First-Person Stories : A Foundation For Mobile Guides. *ECIS*, 2259-2270.

Xia, J. C. (2007). Modelling The Spatial-Temporal Movement Of Tourists. (Unpublished doctoral dissertation). RMIT University, Melbourne, Victoria.

Yanai, K., Kawakubo, H., & Qiu, B. (2009). A Visual Analysis of the Relationship between Word Concepts and Geographical Locations. *In Proceedings of the ACM International Conference on Image and Video Retrieval. CIVR'09*. New York, New York, USA: ACM.

Yin, H., Wang, C., Yu, N., & Zhang, L. (2012). Trip Mining and Recommendation from Geo-tagged Photos. *In Proceedings of IEEE International Conference on Multimedia and Expo Workshops* (pp. 540 - 545). doi: 10.1109/ICMEW.2012.100.

Yin, Z., & Cao, L. (2011). Diversified Trajectory Pattern Ranking in Geo-Tagged Social Media. *In Proceedings of the 11th SIAM International Conference on Data Mining. SDM'11* (pp. 980-991).

Yoon, H., Zheng, Y., Xie, X., & Woo, W. (2012). Social Itinerary Recommendation from User-Generated Digital Trails. *Personal and Ubiquitous Computing, 16*(5), 469-484.

Zhang, J., Kawasaki, H., & Kawai, Y. (2008). A Tourist Route Search System Based on Web Information and the Visibility of Scenic Sights. *In Proceedings of Second International Symposium on Universal Communication. ISUC '08* (pp. 154-161). doi: 10.1109/ISUC.2008.19.

Zheng, Y. T., Li, Y., Zha, Z. T., & Chua, T. S. (2011). Mining Travel Patterns from GPS-Tagged Photos. *In Proceedings of the 17th ACM International Conference on Advances in Multimedia Modeling - Volume Part I. MMM'11* (pp. 262-272). Berlin, Heidelberg: Springer-Verlag.

Zheng, Y. T,, Zha, Z. J., & Chua, T. S. (2012). Mining Travel Patterns from geo-tagged Photos. *ACM Transactions on Intelligent Systems and Technology, 3*(3), 1-20.

Zheng, Y. T., Zhao, M., Song, Y., Adam, H., Buddemeier, U., Bissacco, A., Brucher, F., Chua, T. S., Neven, H. (2009). Tour The World : Building A Web-Scale Landmark Recognition Engine. *In Proceedings of International Conference on Computer Vision and Pattern Recognition*.

Zheng, Y. T., Zhao, M., Song, Y., Adam, H., Buddemeier, U., Bissacco, A., Brucher, F., Chua, T. S., Neven, H., & Yagnik, J. (2009). Tour the World : a Technical Demonstration of a Web-Scale Landmark Recognition Engine. *In Proceedings of the 17th ACM International Conference on Multimedia. MM'09* (pp. 961-962). New York, New York, USA: ACM.

Zheng, Y., & Xie, X. (2011). Learning Travel Recommendations from User-Generated GPS Traces. *ACM Transactions on Intelligent Systems and Technology* (Vol. 2, pp. 1-29). doi: 10.1145/1889681.1889683.

Zheng, Y., Zhang, L., Xie, X., & Ma, W. Y. (2009). Mining Interesting Locations and Travel Sequences from GPS Trajectories. *In Proceedings of the 18th International Conference on World Wide Web* (pp. 791-800). New York, New York, USA: ACM.

LIST OF FIGURES

LIST OF TABLES

LIST OF EXAMPLES