University of Zurich[UZH]

Department of Geography
Geocomputation

GEO 620 – Master Thesis

Understanding changes in tourists' behavior in the last decade of tourism in Croatia using User-generated content

Date of Submission: 31.01.2020

Author

Tomislav Grcic

17-723-206

Supervisors

Ross Purves, Prof. Dr.

Rahul Deb Das, Dr.

Contact

Author

Tomislav Grcic
Uetlibergstrasse 111b,
8045 Zurich, CH
tomislav.grcic2@uzh.ch

Supervisors

Ross Purves, Prof. Dr.
University of Zurich
Department of Geography - Geocomputation
Winterthurerstrasse 190,
8057 Zurich, CH
ross.purves@geo.uzh.ch

Rahul Des Dab, Dr.
University of Zurich
Department of Geography - Geocomputation
Winterthurerstrasse 190,
8057 Zurich, CH
rahul.das@geo.uzh.ch

Abstract

Volunteered Geographic Information (VGI) and User-generated content (UGC), both being important features of the new era of the Internet – Web 2.0, provide a chance to explore society easily and with less cost than before. Using this data and gaining knowledge from it can be applied to research in tourism, one of the growing industries in the world, in which numbers have almost doubled in Europe and Croatia within the last decade. Croatian southern region Dalmatia is chosen as a case study of this research, mostly due to the increase of tourism figures. We used the data from a globally popular website for images and videos sharing, namely Flickr, which provided ca. 10 thousand photographs with geolocation suitable for our analysis, per year, in the period of 2009 to 2008.

By comparing this data with data from official authorities, it was shown that the VGI/UGC data can be a useful tool for rating popularity, and change of popularity, of destinations in the region. It has also shown that trajectories users make can show us connections between selected destinations. Furthermore, it can suggest how long tourists stay; how many destinations they visited, and how long is the path they made. Figures which can be compared to the data from official sources showed a rather high correspondence, as it detected up to eight most popular destinations each year; it showed correspondence with tourists visits and upload counts; as well as it relatively precisely calculated average stay of the tourists. On another hand, the correctness for figures not tracked by authorities can only be assumed, but can also be used for planning and recommendation in tourism. Additionally, by using a large amount of tags or titles of the photos, we presented user impressions and the change within two different periods we chose, addressing how the events in destinations affect their public image.

Keywords

Photographs, Points of interest, Trajectories, User-Generated Content, Flickr, Tourism, Dalmatia, Split, Dubrovnik

Acknowledgments

Writing a master thesis, a crown of one student's work is one of the biggest challenges in a person's professional career. And for me, it was not an exception. A year ago, when I was on a very beginning, I had a struggle on what to write about, how to begin, how to write in a scientific way. I was very lucky to have Mr. Des Dab, which had to listen to my ideas and from the very beginning directed me to develop the concept of my thesis.

I wish to express my deepest gratitude to Mr. Purves, whose expertise and advice led me to write the thesis without too much stress, as my questions were quickly answered and my progress was tracked and commented.

The writing process would also be impossible without the help from my friends. I am a truly lucky person to had an opportunity to meet some of the most amazing people on Earth. Thank you Sebastian, thank you Michael, thank you Yanis, Laura, Ivor, Mladen, Yuman, all Geo Team. Without your help, support, and time spent together, this process would be much more difficult.

I wish to thank my family for understanding me for moving far away. I miss them every day, they miss me every day. Special thanks to my sister Iva. You maybe did not help me much with the thesis, but you certainly cheered and annoyed me when I needed it. *Stari brat je uvijek ponosan na tebe. Nadam se, i ti na njega.*

List of Abbreviations

API -	Application Programming Interface
DBSCAN –	Density-based spatial clustering of applications with noise
DZS -	Drzavni zavod za statistiku (Croatian Bureau of Statistics)
GoT -	Game of Thrones
NP -	National Park
POIs –	Points of Interest
ROIs –	Regions of Interest
TF-IDF -	Term Frequency-Inverse Document Frequency
UGC –	User-generated content
UZH -	University of Zurich
VGI –	Volunteered Geographic Information
WGS -	World Geodetic System

Contents

Abstract
Acknowledgments
List of Abbreviations
List of Figures
List of Tables
Appendix
Literature

Chapter 1 – Introduction 1
1.1. Problem definition 3
1.1.1. Significance, Motivation, and Goal 3
1.1.2. Research questions 4
1.3. Thesis structure 6
Chapter 2 – Background 7
2.1. Movement analysis 8
2.2. Web 2.0 and social-media sites in research 10
2.3. UGC for movement analysis and the extraction of Points of interest 11
2.4. Route planning and tourist recommendation 15
2.5. Tourists' behavior concepts 18
2.6. Text retrieval 20
2.7. Bias types in the data 21
2.8. Effect of the filming and festivals on tourism 22
2.9. Research gap 23
Chapter 3 – Study area and data sources 25
3.1. Study area 26
3.1.1. Dalmatia 27
3.2. Dataset 31
3.3. Flickr website 32
3.3.1. Using Flickr 34
3.4. Data from official sources 35
3.4.1. Tourism figures 36

3.5. Game of Thrones series / Ultra festival 38
Chapter 4 – Methodology 39
4.1. Preliminary definitions 40
4.2. Software, Data extraction and pre-processing 42
4.2.1. Software 42
4.2.2. Data extraction 43
4.2.3. Data pre-processing 44
4.3. Points of Interest 46
4.3.1. Discovery of POIs 46
4.4. Trajectories 49
4.4.1. Discovery of trajectories 49
4.4.2. Patterns between destinations 51
4.5. Comparison with Authoritative data 54
4.6. User activity over time 55
4.7. Gaining knowledge using the metadata 56
4.7.1. Relative proportions and tag analysis 59
Chapter 5 – Results 60
5.1. Data Summary 61
5.3. Trajectories 68
5.3.1. General statistics 68
5.3.2. Frequency patterns between destinations 73
5.4. Figures of Users' temporal activity 78
5.5. Analysis of tags and titles 81
Chapter 6 – Discussion 85
6.1. The data 86
6.2. Points of interest, trajectories, and temporal change 87
6.3. Analysis of tags and titles 89
Chapter 7 – Conclusion 90

List of Figures

Figure 1 - Parameters of movement and their dimension 8
Figure 2 - One of the types of movement - within network space 9
Figure 3 - Simple example of tourist movement in space 12
Figure 4 - Trajectories in Western Australia made from the Flickr data 13
Figure 5 - Different ways of visualization of POIs/ROIs using DBSCAN 14
Figure 6 - Example of connecting two movements from two different tourists 17
Figure 7 - Croatia, its regions and neighbor countries 28
Figure 8 - Dalmatia – counties, important towns, islands, and rivers 28
Figure 9 - Split and Dubrovnik 30
Figure 10 - YFCC100M data visualized on the world map 32
Figure 11 - Drop in the number of photos on the website after changes in February 2019 33
Figure 12 - Uploading process in Flickr website 34
Figure 13 - Relative amounts of overnight stays within selected European countries and Dalmatia 37
Figure 14 - Example of a photo with its metadata 40
Figure 15 - Bounding box is determined by two points 43
Figure 16 – Steps of filtering (pre-processing) the data 45
Figure 17 - Heatmaps were the first step in Points of interest discovery 47
Figure 18 - Example of the clustering to discover POIs. 48
Figure 19 - Randomly selected user and their trajectory and part of the data used 49
Figure 20 - Trajectories within Dalmatia in 2009 50
Figure 21 - Selected sites for the analyzis 52
Figure 22 - Example visualization for 2009 – Daytime visualization, upload count graph 55
Figure 23 - Split with its borders ; green buffer is around Poljud Stadium 57
Figure 24 - Filming locations of Game of Thrones series 58
Figure 25 - Dubrovnik in its city borders 58
Figure 26 - Number of photos uploaded year by year 62
Figure 27 - Number of users which uploaded any number of photos 62
Figure 28 - Area around Split shows how data, without pre-processing, could lead to biased results 63
Figure 29 – Kernel density estimation and Heatmap visualization combined 64
Figure 30 - Clusters with the noise and top 10 destinations of the year 65
Figure 31 - Trajectory trajectory length per season 70
Figure 32 - Average number of days spent per visitor for each season 70
Figure 33 - Winter trajectories length year by year 71
Figure 34 - Spring trajectories length year by year 71
Figure 35 - Summer trajectories length year by year 72
Figure 36 - Autumn trajectories length year by year 72
Figure 37 - Visits between Split and Dubrovnik and other destinations 73
Figure 38 - Distribution of uploads, 2010 78
Figure 39 - Distribution of uploads, 2012 78
Figure 40 - Distribution of uploads, 2014 79
Figure 41 - Distribution of uploads, 2016 79
Figure 42 - Distribution of uploads, 2018 79

List of Tables

Table 1 - Summary of selected works on the topic of extraction of POIs, trajectories, and tourist recommendation 24
Table 2 - Croatia facts 26
Table 3 - Split and Dubrovnik facts 29
Table 4 - Tourism figures for Croatia, Dalmatia, Split, and Dubrovnik 36
Table 5 – Night spent figures for Croatia and Dalmatia from the official sources (2018) 37
Table 6 - Example of the data on user upload 45
Table 7 – Trajectories - example 50
Table 8 - Selected locations from the north to south 51
Table 9 - Visits between destinations in 2009, absolute figures 53
Table 10 - Example figures for 2009 and relations between Split and Dubrovnik 53
Table 11 - Data Statistics 61
Table 12 - Ranking of the destinations, according to our data 66
Table 13 - Most visited destinations according to the official data 66
Table 14 – Length results from trajectories 69
Table 15 - Duration of stay, year by year 69
Table 16 - Relative relations between destinations, 2009 and 2010 74
Table 17 - Relative relations between destinations, 2011 and 2012 74
Table 18 - Relative relations between destinations, 2013 and 2014 75
Table 19 - Relative relations between destinations, 2015 and 2016 75
Table 20 - Relative relations between destinations, 2017 and 2018 76
Table 21 - Proportion of photos with tags and titles 81
Table 22 - List of most common words within the city of Split 82
Table 23 - Absolute and relative proportions of photos within Split 82
Table 24 - List of most common words within the city of Dubrovnik 83
Table 25 – Absolute and relative proportions of photos within Dubrovnik 83

Chapter 1 –

Introduction

According to *Britannica*[1], tourism can be defined as „the act and process of spending time away from home in pursuit of recreation, relaxation, and pleasure, while making use of the commercial provision of services" and it is today considered as one of the most relevant industries worldwide, with a steady growth in last decades. For example, the number of tourists in 1995 was 527 million, and it generated 415 billion US dollars (Ponomarev, 2016). The worldwide number of tourists in 2009 was 920 million[2], which increased up to 1.4 billion in 2018[3]. Having this in mind, it is expected that a part of research on User Generated Content (UGC) and Volunteered geographic information (VGI), which focuses on changes in society, will cover topics related to tourism. An open and free approach to the great amount of data can suggest an understanding of society with more efficiency while reducing the cost and privacy issues that come with polls, interviews, or surveys, or use of tracking GPS devices. The data is created by Internet users and published on platforms social media platforms such as Facebook, Instagram, Twitter, Flickr and many more. There are several good reasons to use UGC in tourism and travel analysis. For example, it provides a high volume of easily accessible data and the results, because of its quantity, could complement a more traditional research approach in tourism. In theory, it can also provide data for any part of the world, from anywhere.

There are many Social-media platforms whose significance has over-passed traditional sources of informing. The number of their daily users easily goes over dozens of millions. However, a small fraction of those sites have open Application-programming interface (API) which allows going deeper into the analysis of created content. One such site is Flickr, in which users are uploading their photos from everyday life, travel, and leisure. Despite the fact that a relatively small amount of users upload their photos with a geotag, because of the popularity of such media and a large number of photos, this can still give a representative amount of data. While a single photo could present what, when, and where one individual did, a series of photographs from one user, if posted with correct geotag and time, can provide us trajectories of approximate spatio-temporal movement of the individual (Cai et al., 2014). It is important to point out that the quality of a trajectory created from a series of photos is strongly influenced by the number of photos uploaded, and thus is much less representative than a GPS tracking. Furthermore, a large number of photos, uploaded with a geotags by many users, can provide us information on the popularity of places. Part of the user-set metadata, tags, and titles of photos, can be semantically analyzed to show travelers' impressions of places. The idea that the visualization and analysis of these findings, Points

[1] https://www.britannica.com/topic/tourism
[2] https://www.e-unwto.org/doi/pdf/10.18111/9789284413591
[3] http://www2.unwto.org/press-release/2019-01-21/international-tourist-arrivals-reach-14-billion-two-years-ahead-forecasts

of interest (POIs) created from user uploads, their trajectories, and semantic analysis, can show the change over time, e.g. year by year, is one of the main motives behind the thesis. Since tourism represents one of the most important branches of industry in the world, the motivation to understand such movement and change in popularity is even clearer. Potentially, this approach could support or even replace some of the official statistics, or expensive measurement techniques such as GPS tracking, polls or interviews.

As will be shown in the Background chapter, many of the researchers who observed UGC focused on the spatial component of tourist behavior. This thesis will try to see if the knowledge about tourists and their spatial decisions can be extracted from the data using a temporal component, that is, comparing different years in ten years.

1.1. Problem definition

1.1.1. Significance, Motivation, and Goal

Even though there is a rising amount of the research which tries to gain knowledge on tourism from VGI/UGC, as will be presented in the next chapter, most of them focused on the analysis of the spatial distribution of photographs uploaded and extraction of POIs, trajectories, or development of recommendation system. Little work included temporal components in such research. Our primary goal, therefore, is to find out if UGC data can offer a possibility to observe the temporal change in popularity of POIs, impressions from tags, and trajectories users make.

Another goal is to show that tourism worldwide can be compared using this kind of data, even though we will only stick to our study area. The data can be approached and extracted from anywhere, for any space and period photos are available. So, ideally, the approach used in this thesis can be applied for most of the world, even if the official statistic data on tourism is not available. Such knowledge could check trends in tourism and complement or replace more expensive and time-consuming methods like polls, questionnaires, GPS tracking and similar. A better understanding of tourist behavior in the destination through time and space can help in better destination and tourism planning, as well as „managing of the social, environmental, and cultural impacts of tourism“ (Khairi and Ismail, 2015).

Dalmatia, a southern region of Croatia, is chosen for a few reasons. The growth of tourism in the last decade is steady and sizable – the number of both tourists and overnight stays nearly doubled. Moreover, two of the most visited destinations, the city of Split and town of Dubrovnik, have either significantly larger growth or changed their image/brand (or even both). By reading into the literature, one can notice that there is lack of such research not only in this area but on any similar areas since most of the similar research concentrated either on big cities (like London, Sankt Petersburg, Taipei, Paris, Budapest, etc) or on larger or more populated countries (Australia, Taiwan). Additionally, the personal knowledge of the region, as well as knowledge in the tourism sector, of the author is motivation as well.

Given those reasons, we can summarise our motivation as follows:

- Temporal information of UGC data was not used enough in the research, at least according to the knowledge of the author of this thesis, after thoughtful research of the current state of the art
- Tourism is one of the most important branches of the economy worldwide and research using UGC could complement the data from the official sources; the approach used in the thesis can be used to explore trends of almost any parts of the world, as long as social media sites provide open access to their API
- The region of Dalmatia has not been a case study for any research where social media is used as a data source
- UGC data goes over regional or national borders and can be a useful source to check relations between international destinations

1.1.2. Research questions

By using VGI/UGC, the thesis, therefore, aims to answer if temporal and spatial behavior, as well as impressions of tourists, evolved in the period from 2009 and 2018 and if this is reflected in the properties of VGI/UGC. A concept of the behavior of tourists will be discussed to detain through the thesis. Temporal behavior could answer if visitors behave differently within a day (daytime vs. night time, for example), or within a year (if seasonality changed, that is, if the peak of tourist visits is not so strong during summer months). Spatial behavior can be described as a sum of mobility decisions and patterns of the tourists in the area visited. We can differ such mobility on micro and macro level, first being as „representation of the collection of spatial points (x, y) within a destination with a temporal component of hours, minutes or even seconds", and latter „a collection of more locations (such as destinations) with a temporal component of weeks, days, or more rarely hours" (Xia, 2007). Finally, impressions of tourists can be observed with the use of textual metadata of photos, such as tags, descriptions, and titles. With this information, we can look into if they are more or less interested in particular topics while in some destinations (spatial), and if such topics occur more in a certain time or during/after particular events (temporal).

As discussed before, the growth of tourism in the region in the period of 2009 – 2018 is constant and steep. In 2009 the number of tourists in Croatia was, including domestic tourists, 11 million. The number of tourists in the year 2018 peaked at over 18 million. Moreover, the number of overnight stays increased by similar rates, going from around 55 million to over 90 million. Around half of these figures are contributed by Dalmatia. Also, two of the most prominent destinations of this region, Split and Dubrovnik, changed both the structure and numbers of visitors and will be observed as a separate case-study areas in more detail.

Having this in mind, the thesis' goal would be to answer the following research questions:

RQ 1: Are changes in the behavior and impressions of tourists in Croatia (Dalmatia) reflected in the properties of UGC data?

RQ 2: What dimension of these changes could be extracted from UGC?

These questions come from the following knowledge or assumptions:

- The number of tourists in Croatia increased by over 60% in 10 years
- In the city of Split, the number of tourists increased around 4 times
- There are new trends in tourism which affect tourists' behavior and impressions
- Impressions can be partly extracted using the UGC data

Hypotheses would be:

H 1: Spatial behavior of travelers has changed in the last decade, and a part of the reason is the changes in tourism figures.

H 2: A data from UGC should complement the official data.

- It can be assumed that the aforementioned changes in tourism figures, primarily more tourists, and changes in the image of destinations, affect travelers' spatial decisions in the region of Dalmatia. In addition to the change of figures for destinations, we assume that trajectories can show different length in trajectories (movement) being made, or that duration of stay changed. Some of this can be supported by the data from official sources. The relevance (decided by the number of photos uploaded or trajectories crossed by) of a particular POIs extracted from the UGC data could correspond to the figures in tourism provided by authorities[4]. It is expected that such data cannot replace the traditional sources, however, it can offer information such as places tourists visit within cities, visits between destinations, or impressions of the places visited.

[4] Example of such data: https://mint.gov.hr/UserDocsImages/AA_2018_c-dokumenti/180213_DZS_2017.pdf

1.3. Thesis structure

This thesis is structured as follows:

The following chapter, *Background*, will provide an extensive overview of the literature published on this issue. We discuss the concept of movement in space, UGC/VGI data within Web 2.0, extraction of trajectories, Points of interest, and semantics and impressions from such data. In brief, we discuss the concept of behavior of tourists and explain the research gap.

Chapter 3 will present the study area and data sources. The brief background on the region and destinations which are chosen is presented. In order to understand the part of the motivation for the thesis, the growth of tourism, the statistics of the tourism of the region are shown. The website Flickr is introduced, as well as the dataset extracted from it.

Chapter 4 will present the methodology used in the thesis. Firstly, important definitions used in the thesis and its methodology will be given. Next, the extraction process of the most important Points of interest and validation with the data from official sources will be shown. Similar to this, user trajectories in the given study area will be extracted. Then, the analysis of tags joined by users to photographs will be presented.

Chapter 5 will present the results of research. Firstly, the processed dataset will be presented. Then, the comparison of different years and the popularity of particular POIs, compared with the official data, will be shown, followed by user trajectories. We continue with the presentation of findings from user tags and titles, and finally, upload rate over the years will be given.

Chapter 6 will open the discussion on the results of data, POIs, trajectories, etc, relate them to the works presented in Background, as well as speculate on the future work.

Finally, *Chapter 7* will give the conclusions of the thesis.

Chapter 2 –

Background

This chapter will provide a comprehensive overview of the research that has been done in the application of UGC data for travel and tourism. We structured the overview of the *Background* as follows: firstly, since an important part of the thesis is related to the movement of objects (tourists) in space (within destination), we will present some work that focused on movement analysis. Then, a general overview of Web 2.0. and the use of social media sites in research is presented, despite some of those research are only loosely related to the topic of the thesis. Then, we will present works which, as the main objective, had to discover Points of interest, as well as user-created trajectories. Next, the background on the recommendation system in tourism, generated from VGI/UGC data, will be shown. This section shows how the finding from the previously presented works can be used in practice. To explain our steps in the data pre-processing, we also discuss biases in VGI/UGC data. We will then continue to the discussion on tourists' behavior and how different authors approached it. Then, we discuss the impact of film-making and festivals on the destination image and tourism figures. We will also discuss text analysis, focusing on tags from our metadata. Finally, we will discuss and present the research gap which is one of the main motifs of the thesis.

2.1. Movement analysis

Movement and mobility exploration, in the context of using technologies such as GPS, and in the purpose of tracking and understanding individuals or groups' patterns in space, is a still relatively new field of study within GIScience. As the main way to present the movement of tourists will be from extraction and analysis of the trajectories they make in space, first, we want to clearly define what a trajectory is. We took the definition presented by Buchin and Purves (2013), namely, it is described as „sequences of the time-stamped geographic position of a moving object", or:

$$\text{Trajectory } (T) = ((x_1, y_1, t_1), \dots, (x_n, y_n, t_n))$$

where *x* and *y* present the position and *t* denotes the time. In this work, researchers discussed how trajectories can vary in sampling frequency, depending on the density of sampling data. If the movement has many changes in velocity, but sampling frequency is low, one trajectory is considered as sparse, which adds to its uncertainty. On another hand, more points make the trajectory more precise and meaningful.

It is also important to define the concept of the space. Actual space can be explained as the area that accommodates activities (Khairi and Ismail, 2015) of "moving objects", entities whose position changes over time. They are aimed to be tracked and are not limited to people, but comprise also vehicles or animals (Dodge et al., 2008). Their trajectories can be extracted, visualized, and analyzed. As can be seen in Figure 1, the authors extracted three major groups of movement parameters. Primitive parameters consist of only the position of the object and temporal component, which includes an instance (point in time) and interval (temporal sampling rate) (Dodge et al., 2008). Other type of parameters, primary and secondary derivatives, are complex and require data which UGC mostly cannot provide, but are rather collected with GPS or similar tracking technologies. They also specified different types of paths, *continuous* with regular, predictable moves between steps, and *discontinuous* with irregular moves between stops.

Parameters / Dimension	Primitive	Primary derivatives	Secondary derivatives
Spatial	Position *(x,y)*	Distance *f(posn)*	Spatial distribution *f(distance)*
		Direction *f(posn)*	Change of direction *f(direction)*
		Spatial extent *f(posn)*	Sinuosity *f(distance)*
Temporal	Instance *(t)*	Duration *f(t)*	Temporal distribution
	Interval *(t)*	Travel time *f(t)*	Change of duration *f(duration)*
Spatio-temporal (x, y,t)	—	Speed *f(x,y,t)*	Acceleration *f(speed)*
		Velocity *f(x,y,t)*	Approaching rate

Figure 1 - Parameters of movement and their dimension, taken from Dodge et al., 2008

Further discussion on the space, as well as advances in tracking technology, were on the focus of the work by Gudmundsson et al. (2014). They discussed how spatiotemporal information from the localization such as „GPS, wireless communication, mobile computing, and environmental sensing technologies“ result in large information volumes of great socio-economic relevance. To present such movement, different conceptual data models can be used to present the space - 2D Euclidean space, 3D Space-time cube, network space, irregular tesselation. Since most of the movement of humans is restricted to some network, which usually consists of roads and streets, the *network space* type of conceptual data model is the most appropriate for our work. An example of such a model is presented in Figure 2.

Figure 2 - One of the types of movement - within network space. Taken from Gudmundsson et al. (2014)

Both works of Gudmundsson et al. and Dodge et al., however, did not cover or discuss possible extraction of information on movement using the data this thesis is using, that is, UGC or VGI data. Grossenbacher (2014) stated that „in contrast to conventional data sources, (…) such trajectories are difficult to model, have no information on speed or direction, and may vary from user to user, depending on the frequency of updates“. As our data will later show, Grossenbacher was partly wrong, as ideally – if the data is correct, the direction can be extracted and the speed can be somewhat presented (in a rare cases where uploads of photos are in a real-time), as such data offers an information both on the location and time of the individual points/stops. However, there is no discussion that a great part of the data from VGI is incorrect or imprecise by its nature.

2.2. Web 2.0 and social-media sites in research

Social media sites can be defined as „Internet-based tools that facilitate communication, content exchange and collaborate in multiple ways“ (Sharma and Godiyal, 2016). They include social networking sites, consumer review sites, content community sites, wikis, Internet forums and location-based social media (Zeng and Gerritsen, 2014). The number of sites, whose content is primarily produced voluntarily by the users (such as Facebook, Twitter, Instagram, Flickr, YouTube, Tumblr, TripAdvisor), is growing, as well as their user count[5]. Because of knowledge on society that can be gained from such websites, many of them were in the focus of researchers in the last decade or more. Work from Liu et al. (2017) claims that in the period between 2008 and 2014 there were over 10 thousand papers published related to this topic. Such social media websites are an important feature of a new age of the Internet, so-called Web 2.0. This *phase* of the Internet is also called “the wisdom Web, people-centric Web, participative Web, and read/write Web” and promotes “more social interaction, collaborative manner, collective intelligence” (Murugesan, 2007). Wilson et al. (2011) gave a more elaborated definition – “(…) second generation of the Web, wherein interoperable, user-centered web applications and services promote social connectedness, media and information sharing, user-created content, and collaboration among individuals and organizations”. Akram and Kumar (2017) covered the positive and negative impact of these sites on society, predominantly children and youth. They named risks such as frauds, scams, and cyberbullying, while not discussing possible issues with sharing locations.

Zeng and Gerritsen (2014) discussed the growth of social media and similar information communication technologies influence many aspects of the tourism industry. It changes the source of the information for tourists, the way they „search, find, read and trust, as well as collaboratively produce information about tourism suppliers and tourism destinations“. It affects the marketing of the section and how customers and business owners communicate. They also discussed on the trustworthiness of UGC, limitations of using only English as a language in analysis of UGC content, lack of solid evidence of the link between social media and tourism figures.

Most social media sites do not offer open access to Application-Programming-Interface (API) and the information from them cannot be fully exploited. One of the rare sites which have fully open API and therefore will be used this thesis is Flickr. Many of the works presented in the next subchapters used information gained from Flickr API.

5 Source: https://www.oberlo.com/blog/social-media-marketing-statistics

2.3. UGC for movement analysis and the extraction of Points of interest

An overview of literature related to the humanistic framework within the website Flickr was done by Spyrou and Milonas (2016). They first presented general information on the website. Then, they explained its social community aspect, meaning that the people are not just publishing, but also actively commenting other people's photos, which is one of the reasons this website is among relevant social media sites. They presented its tagging feature – adding descriptive keywords to photos, and possibilities of a text retrieval from site. They presented Flickr's API and its use in the research. The rest of the work presents an extensive comparison of different works for research using Flickr, mostly in the tourism field. Since this work uses the data from Flickr, we will present it in more detail in section 3.4.

The researchers used different methods to determine most visited points within a destination, as well as the movement of tourists, by using the data from the Internet. One of the most common one is the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm which supports clusters with self-selected shapes. Such a method was used by Zheng et al. (2012), Zeng et al. (2012), or Huang (2016). Using DBSCAN, Zheng et al. (2012) extracted so-called Regions of Attractions (ROAs) and developed a scheme which analyzes tourist movement patterns between them. Then, they analyzed the topological characteristics of travel routes. Like other similar works, to make trajectories, they primarily used information on user and their location, as well as temporal data of photos. To analyze movement between two or more destinations, they used the Markov chain model, normally used to analyze the trends in spatiotemporal movements. This model describes how the event probability depends on the previous events. They developed also a model that gives a statistical significance of travel paths, which tells how many unique places a tourist visited. Simple visualization is presented in Figure 3. They aimed to manually separate tourists from non-tourists based on the visual content of the photos. They discovered in total 80 RoAs within 4 cities they observed (London, Paris, New York, and San Francisco), finding one of them to be „false“, as it is created the Pride Parade (an event), rather than is related to the point of an attraction within the city. Each of the cities has around 2000 person-day trips, which shows data to be large in quantity. Finally, their findings are as follows: tourists usually flow from several ROAs to the central (centric) one; tourists tend to visit RoAs in a particular pattern; the number of ROAs visited during one trip is mostly around 3.5 on average. Additionally, the top three destinations of each city observed were visited by 20.3% (lowest visit) up to 43.6% (highest visited destination). They did not use data from the official sources to support these findings. Zeng's and Huang's work are presented later on.

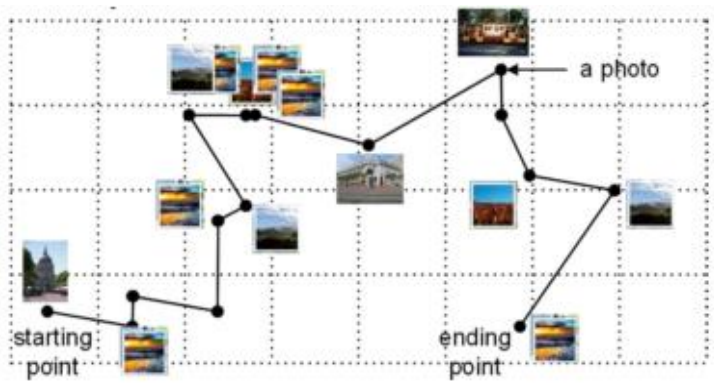


Figure 3 - Simple example of tourist movement in space, presented by Zheng et al. (2012). Each point represents an uploaded photo with geo-tag and time stamp – a series of such points connected presents simplified movement of a photographer

Another method to extract visited points was presented by Kuo et al. (2018). They extracted POIs from the Flickr data on the area of Taiwan (namely cities of Taipei and Tainan). Same as most of the research on the topic, they used a bottom-up method, which means that the POIs and Regions of interest (ROIs) are extracted from raw data, using clustering method - spatial overlap algorithm (SO Algorithm). They also presented a useful way to remove the noise, i.e. data with meaningless information, from the raw data. Eventually, they concluded that POIs/ROIs extracted using UGC data correspond well to official data (POIs/ROIs selected by officials). They also discussed how noise, biases from active users, and identifying clusters with a local maximum in dense areas present a challenge within work with UGC data.

Another work that focused on the ROIs discovery and visualization of trajectories from UGC data was done by Cai et. al (2014). They presented the application of the Trajectory Pattern Mining (TPM) algorithm to discover trajectories and ROIs in Western Australia from the Flickr dataset. As their first task, they did pre-processing steps necessary with User-generated content (UGC) data. They, namely, removed or fixed uploads with the incorrect time, duplicates, or spatial outliers (extreme longitude and latitude). They did not, however, touch the subject of removing outliers or distinguishing tourists and locals. In order to determine popular ROIs, they needed to set three input parameters: *MinSup*, which determines a threshold of support for RoI and trajectory patterns, *CellSize*, which is a grind size for RoI, and *Time Tolerance*, which is an acceptable range of similarity of time annotation (Figure 4, taken from their work, represents their rather simple visualization). They extracted a few ROIs in Queensland, Australia. They also extracted frequent patterns between them, by connecting points individual users make while on the trip and publishing photos. One of the conclusions was that the most frequent patterns are nearby movement, day trips, compared to longer, more distant movement. Another finding is that the city centers are more probable to be hubs for photo takers. They also extracted trajectories on the area of entire Australia, making an unsurprising concluding how they are denser in more populated areas. However, no quantitative results were presented; their visualization lacks cues which would support their findings.

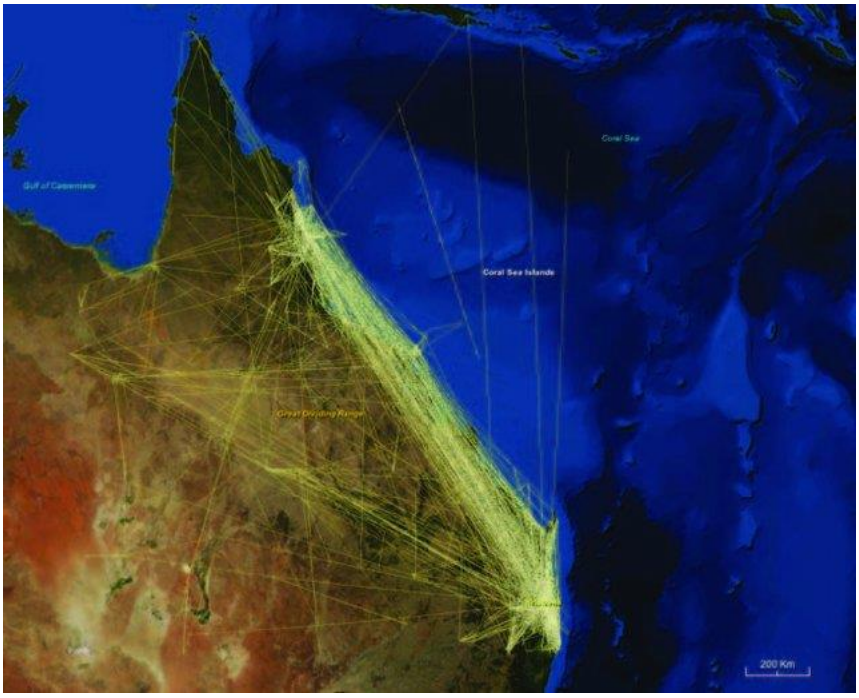


Figure 4 - Trajectories in Western Australia made from the Flickr data, as done by Cai et al. (2014)

Fisher et al. (2019) also used UGC to map tourist patterns and to access their preferences for cultural and natural landscapes. They compared how photos from Flickr, which include a geotag as part of its metadata, as well as *tweets* from Twitter and mobile communication[6], approximated visits at 36 sites where visitors were counted (by tickets sold or at entry gates). Such sites include, among others, parks and protected areas, museums and other cultural attractions. To get accurate estimations, they digitized polygons to represent the boundaries of each tourist site, using OpenStreetMap. They visualized average annual photo user-days, average annual twitter user-days, and *annual mobile population* (the number of mobile users). They found that Flickr upload count, Twitter posting, as well as mobile communication rates have positive correlations with visitation rate estimates. They also found that a great majority of the uploads come from tourists, while for mobile communication the results are expectedly opposite. They concluded that UGC data can lead to a better understanding of visitor preferences.

By using various visualization methods, Kadar and Gede (2013) presented uploads of Flickr photos in Budapest, Hungary. They separated the uploads from locals and tourists, using the method previously presented by Girardin et al. (2008), which „calculate the difference between the time-stamps of the users' first and last photos taken". They visualized those two different groups separately. As for uploads of the domestic population, they analyzed the change in upload count and spatial distribution in two periods, between 2000-2008 and 2009-2012. They pointed out that on three locations of primarily national significance, namely National Museum, CET (Whale) building, and Kopaszi-gát peninsula in Budapest, there are noticeably more photos uploaded in the period between 2009-2012. They explained how those thee locations had restoration at the end of

[6] Provided by mobile provider Sun Kyung Telecom (SKT)

the decade and became open, or more attractive, for the public. However, they did not compare uploads from tourists in a given period, nor they made user trajectories, which would be easily done considering the quantity in the rather small area.

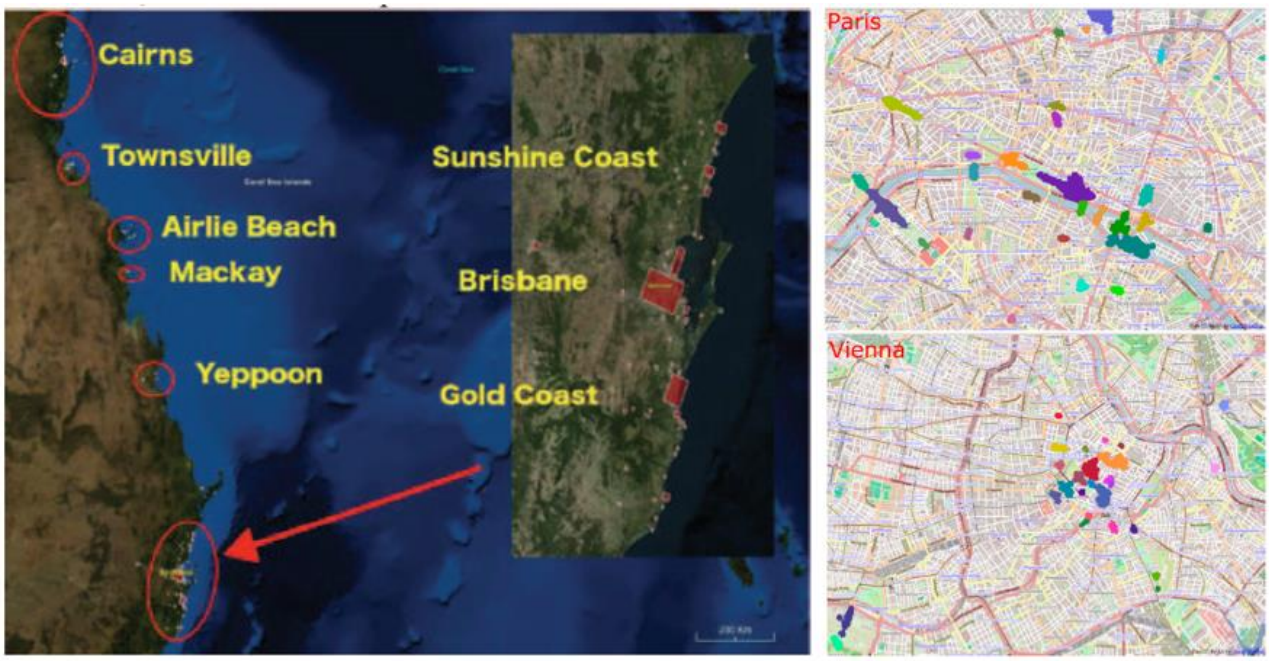


Figure 5 - Different ways of visualization of POIs/ROIs, with the same methodology - using DBSCAN. Left is taken from the work of Cai et. al (2014) and shows ROIs within western Australia, and the right examples are from the work of Huang (2016) and show POIs of Paris and Vienna

2.4. Route planning and tourist recommendation

Recommendation systems has a goal of providing users with a list of items that could meet their interests, based on either newly set preferences or previous searches. Such systems can be used for many platforms and are recently largely implemented by online shopping companies (such as eBay[7], Zalando[8]), streaming providers (such as Netflix[9], Amazon[10]), video sharing platforms (YouTube[11], Dailymotion[12]), etc. It is no surprise that this concept can be used in tourism, more specifically, for destination choice, route planning, and general recommendations in a newly visited destination. Using the information from platforms such as TripAdvisor[13], Flickr, Instagram[14], and other, researchers focused to find a way to make efficient methods for suggesting destinations or a route within a destination, and in some cases, to make it rather personalized.

Combining the information from TripAdvisor and Instagram, in order to understand touristic walking routes constructions, was in the focus of the work by Mukhina et al. (2018). They discussed the problem of uniting the different sources from the Internet correctly (namely TripAdvisor and Instagram). They applied the ant colony optimization algorithm (ACO) which, in this case, should provide the best path between two arbitrary points. For this, they used Google Directions API[15], and they based it on previous paths made by several tourists. Eventually, they present 3 different routes in the city center of Sankt Petersburg, depending on the planned length of stay of a visitor. They were also only concentrated on the walking tours while dismissed the possibility of using public transport or rented cars.

If social media sites can provide useful information for tourism was discussed in work by Dhiratara et al. (2016). They used information from Instagram and TripAdvisor to extract top locations within Paris. They compared these findings with data from official sources, and it showed matching to some extent, exception being that outdoor sights (such as Eiffel Tower) are more represented on social media than in official figures. They also presented temporal change in popularity for such sites, again finding how Eifel Tower has the peak of popularity on New Year's Eve, while other attractions have less of variability in popularity. They finally discussed the limitation that presents the API of Instagram. They did not discuss the use of other traveling-related social media sites.

One of the first works which took a large amount of geotagged photos from social media, namely from Flickr, and use them for recommendations was done by Cao et al. (2010). They wanted to develop a method that would be easy to use, intuitive, and with minimal effort. To use the recommendation system they made, a user has to provide either a photo of a desirable scenery or a keyword which describes their interest. They organized geotagged databases and extracted representative photos for future use. Then, they did a mean shift clustering algorithm to divide the

[7] https://www.ebay.com
[8] https://www.zalando.ch
[9] https://www.netflix.com
[10] https://www.amazon.com
[11] https://www.youtube.com
[12] https://www.dailymotion.com
[13] https://www.tripadvisor.ch
[14] https://www.instagram.com
[15] https://developers.google.com/maps/documentation/directions/start

Earth area into regions, based both on coordinates and the distribution of geotagged photos. After this, they found an appropriate photo and tag which represent each cluster. For query using photos, the results are shown based on the similarity level, while for a word query, locations are shown based on the similarity of query and tags.

Memon et al. (2014) used geotagged photos from Flickr and combined them with historical weather data to „derive their weather context, for recommending context-driven personalized semantic tourist location". Firstly, they cleaned the data by removing photos (1) that had spatial context which did not match to geographical context and (2) those with incorrect temporal context. They developed a method which, by giving a travel history of a person, can predict or suggest this person's preferences and locations in a new-visited city. They showed how to group photos by using „associated geo-tags to sense semantically meaningful tourist locations where the photos were taken". Lastly, they claim that their recommendation method showed better results than similar methods since it was able to predict tourists' preferences in new cities. By using their method, they concluded that it is easier to predict the preferences of people with short and targeted visits than from tourists with longer stays. Weather information, as well as geotagged photos from Flickr, were also used by Huang (2016). He used Weather Underground API to retrieve the weather context of visits. He filtered out uploads attributed to locals, putting a threshold for 5 days. He applied clustering methods (context-aware collaborative filtering approach) to detect touristic locations and extracted travel histories from geotagged photos (Figure 5). In addition, to include visiting context to recommendations, his proposed methods outperformed similar methods used for recommendation.

Lu et al. (2010) also targeted their work to discover automatic travel route planning. Using geotagged photos from Panoramio, they mapped footprints of tourists, setting a goal to plan a trip for users, that is, „which popular destinations to visit, the visiting order of destinations, the time arrangement in each destination, and the typical travel path within each destination". To do this, they first needed to: (1) extract popular destinations and (2) find popular paths. (1) They discovered popular destinations by clustering of geotagged photos, after which they did the following steps: destination naming, discovery of destination image, and its popular visiting time. (2) To find paths, they explained how the ideal path would have enough points represented by photos. The larger the distance is between the first and last photos, the more points are required. Additionally, they visualized how strong the connection between the duration of stay and the complexity of trajectories is. In the study area of Forbidden City in Beijing, China, they showed how the visitors which stayed up to two hours will have fewer points which eventually make a simple and straightforward path from the entrance to the exit. On another hand, those who stayed over five hours have complex paths and have visited more points. The authors also discussed on fragment merging - which user-made paths should be merged and how to logically do it (Figure 6). The idea behind this that the movement of two or more users could be merged as one ideal movement in one space, despite different time periods. Finally, they presented their Travel Route Suggestion (TRS) algorithm which tends to find the optimal path withing the destination in the time set by the user.

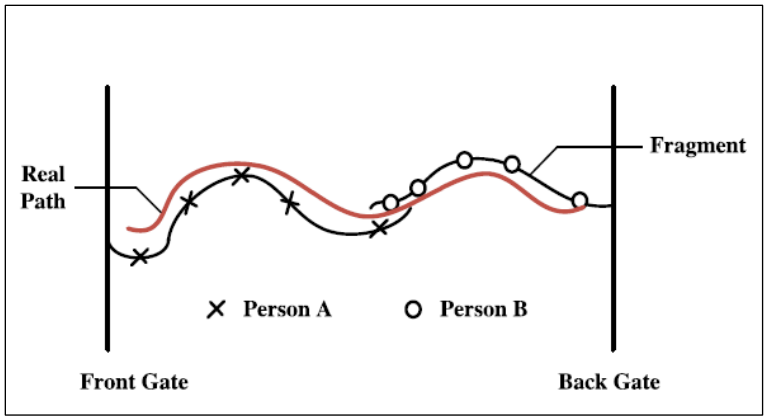


Figure 6 - Example of connecting two movements from two different tourists, which connected present an ideal and most recommendable path from A to B. Each point presents a user-uploaded geotagged photo. Taken from Lu et al (2010)

Zeng et al. (2012) focused on the analysis of the previous user- made or -uploaded trajectories and photos with a goal of „discovering high quality visiting paths for tourism sites". They used DBSCAN and SNN methods, concluding that DBSCAN achieves poorer clustering results when the data density is high, while SNN did not remove the noise (points not joined to a cluster). They used the HITS algorithm to identify the popularity of points of interest. They did not only extract existing trajectories but also rank them and suggested the best ones considering the time available, meaning that the more time tourist spent in the destination, the more complex the route was made.

To understand tourists' decisions at micro-level – an urban space, using GPS as a method, was covered in the work by Khairi and Ismaili (2015). However, they had only 13 eligible participants for further analysis. Because of this, their results will not be here presented, rather will confirm that while such tracking in order to understand tourists' consumption of space in destination can be rather in good detail (how many minutes were spent in a very specific location), it also has flaws of expensiveness and question of privacy and good sample size. They also suggest that despite advantages in the usage of GPS as a technology to investigate tourists, other traditional forms of investigation, such as interviews, questionnaires, and diaries, will stay as an option.

Understanding tourists' travel paths, by separating attractions in categories (Landmark, Nature, Event, Gourment, Business, Local), was done by Arase et al. (2010). The six categories they came up after they surveyed several Web sites of travel agencies, travel forums, and blogs. For each attraction observed, they ran their model for all of the categories to find the one that fits the best. For this, they used linear Support Vector Machines, for which to train they manually labeled a corpus of randomly selected 6000 photos. They also extracted typical descriptions of photos using the TF/IDF technique.

2.5. Tourists' behavior concepts

The concept of tourist behavior has been discussed differently by various disciplines (geography, economy, environmental science, psychology). Some of them, however, concentrated on the broader meaning of the term, explaining how the behavior of tourists is the part of general consumer behavior.

A comprehensive discussion on tourist behavior was presented by Juvan et al (2017). They pointed out that the behavior of tourists indicator of their future behavior, and can be an indicator of the behavior of others. It takes place in several phases, each of them contains the process of planning, decision-making, and purchase. They discussed the importance of motivation in the concept of behavior. While some questions, such as – who, where and how much – is rather simple to answer, the question of „why" is more complex to understand. They concluded the high importance of understanding tourist behavior for future planning. They also explained the time and money-consumes of traditional approaches to monitoring tourist behavior, such as polls, surveys, or interviews. One of such was conducted by Vuuren and Slabert (2011). They wanted to understand tourists' attitudes before, during, and after traveling affect their decisions where to travel. They conducted surveys in South African and showed motifs of traveling to this location.

Cohen et al. (2014) identified nine key concepts related to tourist behavior. Those are: (1) Decision making, (2) Values, (3) Motivations, (4) Self-concept and personality, (5) Expectations, (6) Attitudes, (7) Perceptions, (8) Satisfaction, and (9) Trust and Loyalty. For this thesis, the concepts of Decision making and Motivations are more important. The first one deals with the complexity of planned, unplanned, and impulse purchases and spatial decisions. The latter explains how tourists are pushed by their biogenic and emotional needs to travel and pulled by destination attributes (Yoon & Uysal, 2005 in Cohen et al., 2014). Their work, in general, focused more on the tourists being consumers, rather on their decisions in space. On the other hand, Pearce (1987) discussed the importance of the spatial behavior of tourists. In addition to this, he pointed out the importance of the temporal component of the movement. Such an approach to tourist movement is further explained by Xiu (2007), where he separates the movement of tourists to the micro and macro levels. Micro-level is approaches to the movement as „representation of the collection of spatial points (x, y)" within a destination with a temporal component of hours, minutes or even seconds, while a macro level is a collection of more locations (such as destinations) with a temporal component of weeks, days, or more rarely hours. Some of the works discussed if the behavior is different for tourists who visit destination for the first time, in and for those who have already visited the destination. It is concluded that first-time visitors are destination unaware and move widely, while repeat visitors are destination-familiar, with the degree of familiarity depending on the number of prior visits, and move more concentrated (McKercher et al, 2012).

Having all of the above in mind, and the fact that our data only provide movement patterns of tourists and their temporal activity within the day, we will define the behavior of tourists as sum of spatial activities within a destination at the micro (such as urban spaces) or larger level (on a regional level), with a focus on a temporal scale. Additionally, we can assume that tourist behavior

can be explained by their impressions. In the context of the UGC data, this refers to tags and titles given to the photos. In the following chapter, we present research that focused on text retrieval, having the goal to partly explain tourist spatial activities, impressions, and motives.

2.6. Text retrieval

A lot of information and knowledge of humans is stored in unstructured, textual form. Thus there are a large number of researches focused on text analysis and text retrieval, whit a goal of gaining knowledge and conclusions out of it. Unlike the data mining, where the source is usually structured, text mining focuses on unstructured text from emails, presentations, videos, social media and the Internet in general. The most used models for text retrieval are term frequency (TF), inverse document frequency (IDF), and TF-IDF approaches (Quaiser and Ali, 2018). The first approach, TF, simply presents the rate of occurrence of a term within a document or a word-corpus. IDF decreases the weight for commonly used words and increases for those less-used. TF-IDF is a ratio between two of those and shows how relevant a term is in a given document.

Adding tags for Flickr photos is often called social tagging or collaborative tagging, which is explained as the process of assigning keywords or tags to uploaded photos in order of organizing content and for future retrieval (Golder and Huberman, 2006). Flickr lets the user add up to 75 tags for a photo[16], but the user also can choose not to have any tags at all. Flickr also suggests tags once the photo is uploaded, and this is done by an automatic process. Additionally and optionally as well, a user can insert a title for a photo, which can have maximal 255 characters. Because of the nature of adding tags (optional and by voluntary users), they are unstructured by nature and quality is often rather low, as concluded by Rorissa (2010). He used two different collections to analyze user-created descriptions from social media (namely from Flickr) and professionally assigned indexing terms (from the University's[17] photo collection). Around a thousand photos and four thousand tags/terms were chosen for the research. Using Jörgensen's Twelve Categories (location, content/story, people, description, art historical information, and others), he concluded that location, content/story, and people are most represented in both collections. The Flickr collection also had more tags that were unique (52%), while University collection had 28% of unique terms. In general, there is not a large difference in how users from social media sites describe photos. Analysis of tags and titles was also done by Hollenstein and Purves (2010). Using tags from 8 million Flickr images, they wanted to observe the reliability of tags for describing geographical space, how urban space is described using tags, and how such an approach can gain knowledge of the collective understanding of the location. They presented popular tags within the city centers of Zurich, London, Sheffield, Chicago, Seattle, and Sydney. They showed how different terms of similar meaning (such as city, downtown, center/centre) are differently used within those cities. Both works from Rorissa and Hollenstein and Purves discussed problems related tags from social media, such as ambiguity, polysemy, or synonymy, or location-precision and outliers.

[16] https://help.flickr.com/tag-keywords-in-flickr-BJUJpQoyX
[17] University of St. Andrews Library Photographic Archive

2.7. Bias types in the data

When working with data from open source platforms and applications, it is expected that certain issues might occur. Some, for example, Fan et. al., 2014 and Seely-Gant and Freehill, 2015 claim that due to the nature of collection and mining, and the methods used therein, representativeness will be affected, and vulnerability will be increased, with sampling and other biases. For future references, we define bias as „any trend or deviation from the truth in data collection, data analysis, interpretation and publication which can cause false conclusions“ (Simundic, 2013). We present bias types relevant to our work.

Under the term *user bias* we can differ two related types – socio-demographic bias and participation user bias. The first one deals with the fact the users of social media sites usually do not represent the average population, including tourists in general. Social media primarily attracts the attention of the millennial generation, the population born between 1977 and 1992 (Joshi, 2015). Such bias can be described as selection bias, which, according to Seely-Gant and Freehill tells how some individuals or groups, by their use of social media platforms, can be overrepresented in the UGC data. Grossenbacher (2014) discussed in work how not only that social media does not represent the World's population in general, but contrary, it might only represent a fraction of the population of Western countries. Furthermore, he presented a concept of geodemography or profiling people based on where they live, the conclusion being that different socio-demographic groups are not equally distributed in space (Grossenbacher (2014) according to Harris et al., (2005)). Participation bias is also covered in his work. An example of such, firstly described by Nielsen (2006), is represented by formula 90-9-1, which claims that 90% of the users are simply „audience“ whit little to no contribution to content. Another 9% represents „editors“, or users which modify the content but rarely create a new one. Finally, 1% represents those who upload the high amount of content thus creating a possibility for false results if not excluded to some extent.

Spatiotemporal bias can be caused by unstable Wifi/Internet connection or rather complete lack thereof. Fisher et al. (2019) concluded how UGC data, in general, can be strongly influenced by this, especially during the travel. When it comes to platforms such as Flickr, they are more likely to mark their location incorrectly due to the lack of Internet connection, or simply because of not knowing the exact location of where the photo was taken.

Another bias is a platform bias, which comes from the fact that most of the data, due to the restrictions from API, usually comes from the same Internet platform. The popularity of such a platform can vary and go down causing the data, if temporarily compared, to lose its quality. However, to our knowledge, platform bias in this context was not discussed in any other works. As will be shown, our data is influenced by the aforementioned biases. In the following chapters, we will present examples within our data and solutions we applied to reduce the effect of them.

2.8. Effect of the filming and festivals on tourism

There is an increasing number of papers who focused on the importance of filming movies/ television shows in tourism. The general conclusion is that filming locations, serving as a background for such media, are well promoted, especially in the times of the new technologies in communication (Tkalec et al., 2017). In addition to the growth of tourism, which can be described as „indirect use to economy or Post Production Effects (PPE)" (Croy, 2004), the filming can bring the direct use for the local economy, in terms of new jobs or expanses to local film studios. There are several examples of the link between filming and tourism. The movie Braveheart, filmed mostly in Scotland, increased number of visitors of Wallace Monument by 300%; it is estimated that famous movie Captain Corelli's Mandolin increased visits to Greek island of Cephalonia by 50%; series of movies about Harry Potter again increased visits of filming locations by 50% (Hudson and Ritchie, 2006). One of the most prominent examples is the filming of Lord of the Rings in locations of New Zealand. Croy (2004) suggested that the film industry in New Zealand generated ca. 500 million NZ dollars, a stellar growth from 86 million in 1995. The national tourism office, Tourism New Zealand (TNZ), focused their promotion on LOTR trilogy. TNZ made a survey within visitors of the country and as Croy pointed out, around 8% of the visitors stated that the movie was one of the main reasons behind visiting this country and nearly all of the visitors knew about the filming of those movies within the country. While most of the visitors were not drawn by the movie itself, the study suggested a strong link between the image of New Zealand and the film. To observe the impact of the series Game of Thrones on tourism of Dubrovnik, Tkalec et al. (2017) used the synthetic control approach method. This means that they, by using the data on the 20 Croatian counties (minus the county where is Dubrovnik placed – Dubrovnik-Neretva county), made a synthetic Dubrovnik county, and compared tourism results of synthetic and Dubrovnik county. They concluded that from 2012 to 2015 the number of tourists increased for ca. a quarter of million due to the series. This corresponds to approx. 30% increase in arrivals, overnight stays, and sales of city walls admission tickets.

The link between music festivals and tourism is somewhat more direct. Festival is an event to provide joint entertaining or leisure-time experience of high quality for the audience, focusing on one or more topics, being organized regularly at one or more scenes, with cultural, art, gastronomical, sport or other programs (Nagy and Nagy, 2013). Duarte et al. (2018) surveyed visitors of festival WOMAD in 2013. They concluded that the festival, in addition to having a positive image among the audience, strongly affects the image of the destination, making also a loyal audience which tents to recommend both festival and the destination to others. However, the most obvious link is the tourist visits gained during the festival.

To our knowledge, there are no works that tried to link specific series, movies or festivals to tourism by using UGC data.

2.9. Research gap

As can be concluded from previous sections, a lot of work in tourism research used the data from UGC. A significant part focused on tourist movement in space and extraction of POIs, based on the upload count from the point the data is available. A small part of the research explored tags as part of the metadata and information which can be extracted from them. When it comes to methods, certain patters are used. However, to our knowledge, there is a lack of work which included the temporal component in research. This means that little work tried to compare if there are any changes in how tourists behave in space throughout time. Similarly, no work has compared the popularity of destinations based on the UGC data. Such an idea was partly used in the work of Kadar and Gede (2013). They – as is shown – compared uploads made by locals, avoiding using the same technique for tourists. Furthermore, it is important to remind that research in tourism can be of a high cost. This was discussed by Juvan et al (2017). They explained the high costs of traditional sources of knowledge in tourism, yet they missed to suggest the use of UGC data for such purpose. The metadata of UGC data can ideally provide movement, impressions, as well as popularity of some destinations, which is already valuable data in understanding their behavior and motivation.

We are also going to try to show trends in tourism from metadata of photos, expecting that popular events, such as the filming of series and well-visited festivals, can be reflected in UGC data. Finally, and as mentioned in the Motivation section, the region of Dalmatia, as well as similar European regions, were not chosen as a case study for such research. In most cases, as it is also shown in Table 1, researchers were mostly focused either on the world as a whole or rather big cities. Also, not too many works focused on removing biases from the data, as we presented it in future chapters.

To summarise, the following points are either less covered by the literature or present research gaps that inspired the work:

- Comparing the temporal component of UGC data has been done in a fairly small amount
- Biases of the UGC data is not always discussed and thus removed from the raw data
- Trajectories could reveal more knowledge than is presented
- Study area of the thesis, as well as culturally and geographically similar areas, was not covered by a such research

Such gaps are also visible on the Table on the next page (Table 1), where we extracted some of the most important elements of the work (such as POIs and trajectories extraction, tag analysis, and similar).

Table 1 - Summary of selected works on the topic of extraction of POIs, trajectories, and tourist recommendation

Work	Task(s)	Method(s) used	POIs / ROIs extraction	Trajectories / Routes	Temporal change	Tag analysis	Removing user bias*	Separate locals	Data quantity/ period of collection	Data Source	Study area
Mukhina et. al., 2018	- existed (suggested) route improvement; search for the optimal path	- ant colony optimization algorithm (ACO) - Google Direction API	Yes	Yes	No	No	No	No	Instagram: over 11 million posts, January 01, 2016 – July 01, 2017	Official city guide, Instagram, TripAdvisor	Sankt Petersburg, Russia
Memon et al, 2014	- prediction of tourist locations recommendation within famous places more precise than up-to-date methods	- collaborative filtering and context rank	Yes	No	No	No, except count	No	No	1,376,886 photos, January 01, 2000 - November 17, 2013	Flickr, weather historical data	Different cities in China
Huang, 2016	- derive personalized and context-aware location recommendations	- three CaCF methods - DBSCAN for clustering	Yes	No	No	No	No	Yes	Ca. 2.6 million photos, January 01, 2008 - December 31, 2013	Flickr, Weather Underground API	Amsterdam, Berlin, Paris, Prague, Rome, Vienna
Lu et al., 2010	- destination discovering - Internal path discovery - Customized trip planning	- Internal Path Discovering (IPD) algorithm	No	Yes	No	No	No	No	20 million photos and 200 thousand travelogues	Panoramio website	Worldwide, case study: Forbidden City, China
Cao et al., 2010	- recommendation system based on the representative tags and photographs	- mean shift clustering method	Yes	Yes	No	No	No	No	Ca. 1.1 million geotagged photos, -	Flickr	Various places worldwide
Zeng et al., 2012	- algorithm for precisely matching user-uploaded photos to tourism sites - a density-based clustering approach to identifying the point of interests inside tourism sites	- DBSCAN - SNN - HITS algorithm	Yes	Yes	Yes	No	No	No	17621 trajectories; 23,649 geo-tagged photos, a period of over 4 years	Geolife dataset [4], TripAdvisor	Beijing, several other destinations within China
Ponomarev, 2016	- interesting location extraction - lower the number of Flickr API calls in analysis	- split the area into smaller cells	No	No	No	No	No	No	-	Flickr	Sankt Petersburg, Tyumen (Siberia), Russia
Zheng et al., 2012	- analyzing tourist movement patterns in relation to RoAs	- DBSCAN	Yes	Yes	No	No	No	Yes	ca 769K geotagged photos	Flickr	Paris, London, San Francisco, and New York City
Kadar and Gede, 2013	- measure activity of tourists and locals in space and time	- visualization of the photo locations	No	No	Yes (upload figures)	No	No	Yes	2000-2008, 2009-2012	Flickr	Budapest, Hungary
Fisher, 2019	- mapping tourists' patterns - assessing people's preferences for cultural and natural landscapes	- divide the study area into smaller grinds - visualize mobile, Twitter and Flickr data	Yes	No	No	No	No, but discussed	Yes	Flickr: 2005-2014 (-), Twitter: ca 400,000 tweets, 2012-2014	Flickr, Twitter, mobile provider	Jeju Island, South Korea

*Note: Missing, or for the work irrelevant information, is presented by minus (-) sign; * as defined in section 2.7.*

Chapter 3 –

Study area and data sources

This chapter will introduce the country of Croatia and its region of Dalmatia, the study area for the thesis, as well as two of its major cities, Split and Dubrovnik. Then, we will present the Flickr website and the reasons behind the use of it. The dataset will be shortly presented since the code and preprocessing are explained in the next chapter, *Methodology*. We are presenting here sources for the official data, as well as figures for the region and Croatia. Lastly, we will introduce the TV-series Game of Thrones, which was filmed in the area, as well as the Ultra festival in Split.

3.1. Study area

For the study area, the Croatian region of Dalmatia is chosen, as well as the two most visited destinations of the region, Split and Dubrovnik. Croatia (Figure 7 and Table 2) is both Central European and Mediterranean country, with influences from its neighbors, namely Italy, Slovenia, Hungary, Serbia, Bosnia and Herzegovina, and Montenegro. The capital and the largest city is Zagreb. The country is geographically diverse, mostly divided into three different zones – coastline, mountains, and mostly flat central and east Croatia. Coastline consists of regions of Istria, Primorje (Croatian Litoral) and Dalmatia. The total length of the coastline of over 6200 km, which includes 1244 islands and islets, and has a diverse topography. Mountainous part includes regions of Lika and Gorski Kotar, with peaks not exceeding over 1800 meters over the sea level. Pannonian, mostly flat part, which consists of Central Croatia (around the capital), Zagorje, Slavonia, and Baranja, is industrially more and touristically less prominent. Economically, the country is among less developed EU countries, tourism being of very high importance, as it represents around 17% of GDP and supports much of the country's employment.

Table 2 - Croatia facts (data from 2008)

Area	56,542 km², additionally 31,030 km² of the sea area
Population	4.1 million
Population change	- 0.9% (2018/2009, anually)
Biggest cities	Zagreb (780 000), Split (180 000), Rijeka (110 000), Osijek (100 000)
Highest peak	Dinara (1831 m) near Knin
Coastline length	6,200 kilometers, of which islands encompass 4,320 km
Climate	Csa (coastal regions), Cfa (northern and northeastern areas), Df (mountains)
National Parks	8 - Plitvice, Krka, Brijuni, Paklenica, Mljet, Risnjak, Sjeverni Velebit, Kornati
Nature Parks	10
UNESCO Heritage sites	10
GDP	60.8 billion euros
Tourism in GDP	11.4% (direct), 16,9% (direct + indirect)

Sources: https://www.britannica.com/place/Croatia, https://www.weatheronline.co.uk/reports/climate/Croatia.htm

3.1.1. Dalmatia

Dalmatia is one of the oldest historical and geographical regions of the east Adriatic coast. Its territory and significance have changed over time and different rules (Roman, Byzant, Venetian, Austro-Hungarian)[18]. It stretches for ca. 370 km from the small village of Tribanj on the north-west to Cape Oštra to the south-east, and has an area of ca. 13,000 km², which roughly presents a bit over 20% of the total area of Croatia. It consists of 4 counties, the first-level subdivisions of Croatia (see Figure 8 and Table 3). The population comes close to 900 thousand in 2018 and follows the trend of Croatia in terms of losing population due to low birth rates and negative migration balance. The largest cities and towns are Split, Zadar, Dubrovnik, and Sibenik, also capitals of their counties (Split-Dalmatia, Zadar, Dubrovnik-Neretva and Sibenik-Knin County). Other important places include Knin, Makarska, Omis, Primosten, Imotski, Korcula, each counting around 15 thousand inhabitants.

The region is geographically rather diverse, having inland mountains, numerous islands, karst rivers, and fields. In total, there are over 900 islands along the coastline, most of them are small (under 1 km²), uninhabited, and often categorized as islets. The biggest islands are Brač, Hvar, Korčula, Pag, Dugi Otok, Vis, Lastovo. Rivers are relatively short-length (up to 100 kilometers), longest are Cetina, Krka, Zrmanja, and Neretva, which is for the most part located in the neighbor region of Herzegovina. The rivers, however, are because of the scenic beauty of its canyons, waterfalls, clear water, and vegetation, well-visited by tourists. There are four national (Paklenica, Krka, Kornati, Mljet) and three nature parks (Telascica, Vransko lake, Lastovo) within the region[19].

Tourists are also attracted by cultural heritage. One of the prime examples would be UNESCO-protected sites. There are, in total, 6 within the region[20]: The Venetian Works of Defence in Zadar and Sibenik, St. Jacob Cathedral in Sibenik, Diocletian's Palace in Split, Stecci medieval tombstone in Zagora region (the site is shared with neighbor countries – Bosnia and Herzegovina, Montenegro, and Serbia), Starigrad plain in Hvar island, and Dubrovnik Old town.

As it will be shown statistically (Table 4), the area has a constant and high increase in tourist figures, especially in the period of the last ten years (from 2009 to 2018). It is noticeable that the region usually follows or surpasses the growth in tourism of Croatia. Both in Dalmatia and in Croatia the tourists stay on average around 5-6 days, which is also more or less constant during the decade observed.

[18] https://www.britannica.com/place/Croatia
[19] Opća i nacionalna enciklopedija, Vecernji List, Zagreb
[20] http://whc.unesco.org/en/statesparties/hr

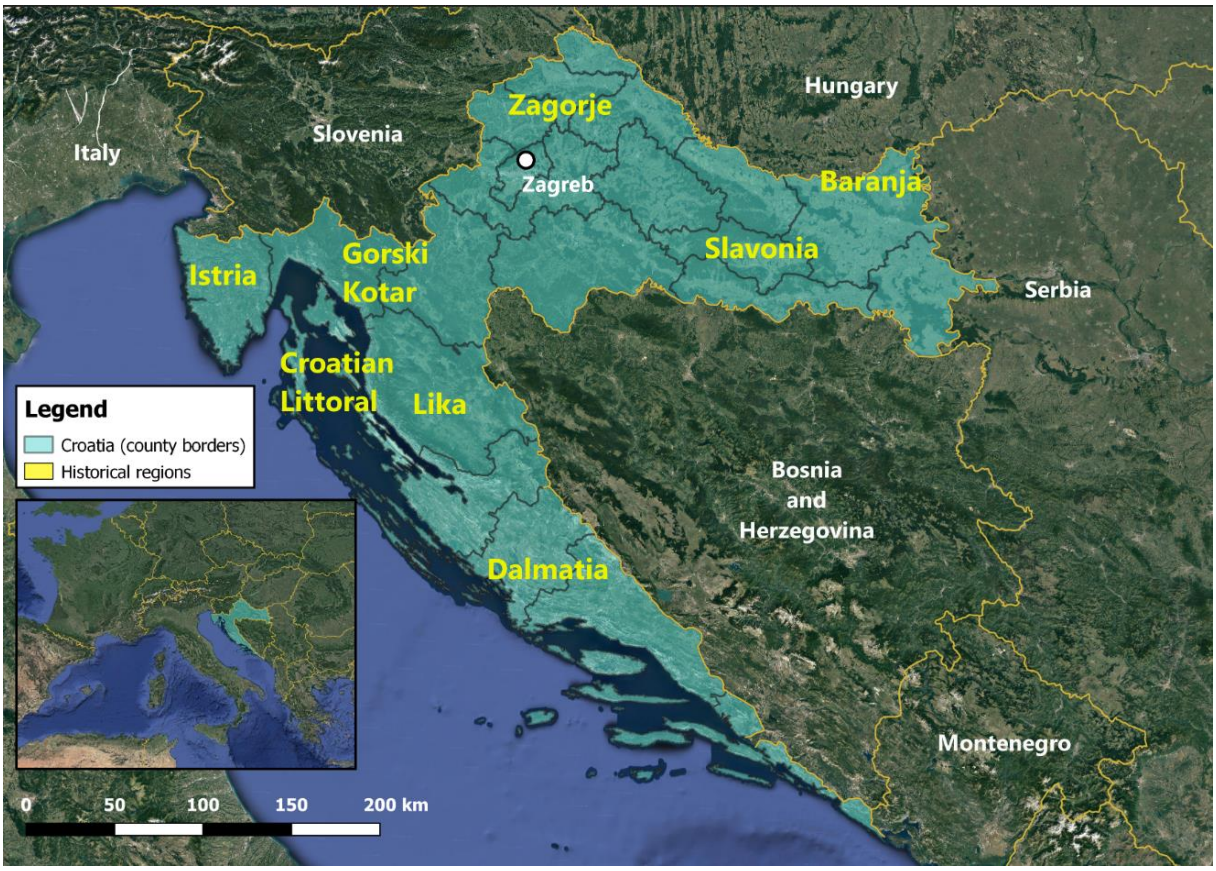


Figure 8 - Croatia, its regions and neighbour countries

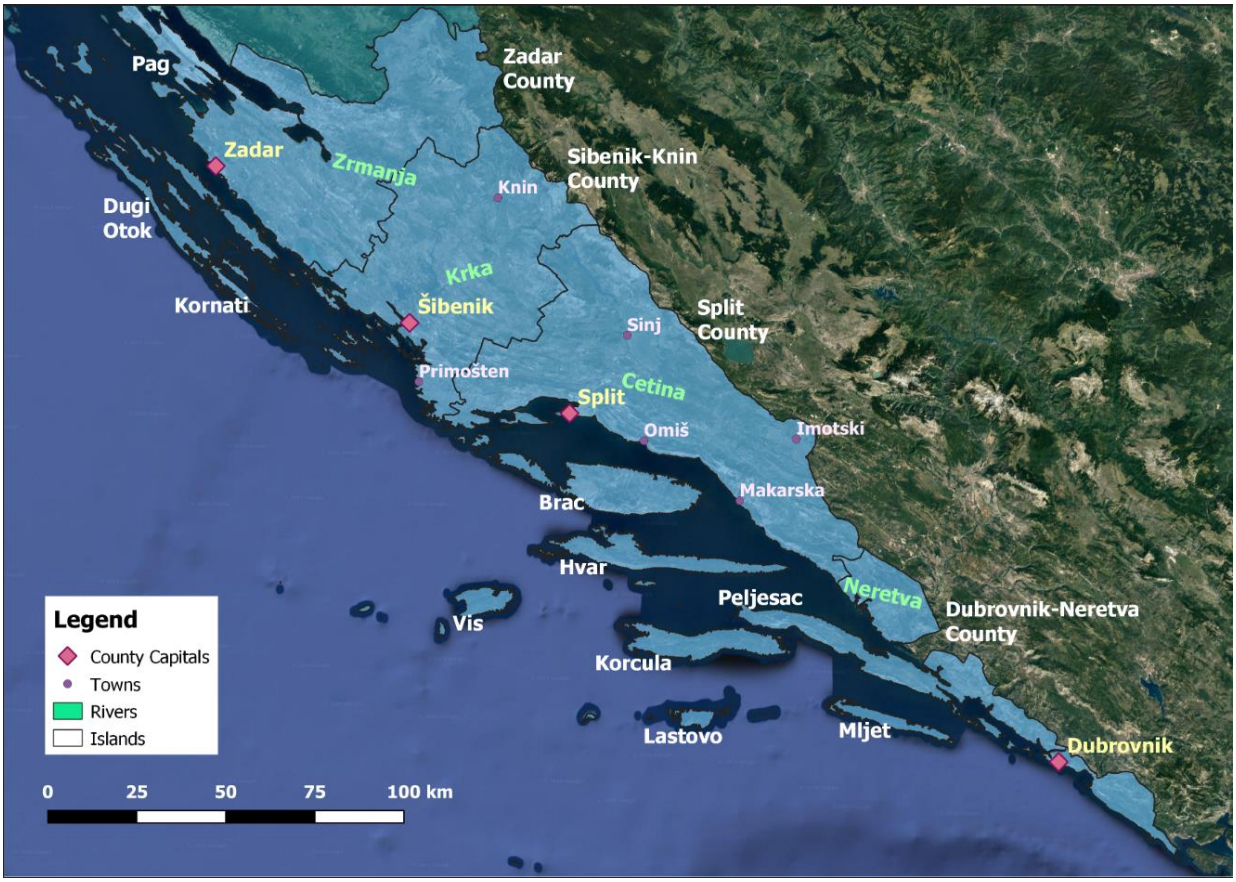


Figure 7 - Dalmatia – counties, important towns, islands, and rivers

3.1.2. Split and Dubrovnik

The city of Split (Figure 9, up) is the second-largest city in Croatia and the regional capital, with a population of around 170 thousand. Some of the most recognizable sightseeing points include Diocletian's Palace, Marjan Forest Park, Klis Fortress, and Riva Promenade[21]. Diocletian's Palace is on the UNESCO list of World Heritage sites from 1979. Part of its popularity is due to cultural events and festivals which occur during summer in the city. One example is Ultra Europe, whose first edition was in July 2013. Additionally, the area close to Split includes well known natural sites such as islands of Hvar, Brac, Vis and many others; National park Krka, Cetina river, as well as cultural sites such as towns of Trogir, Omis, or Sibenik. Having a strong increase in the number of tourists in the last decade, it transformed itself into one of the most visited destinations of the Adriatic sea[22].

Dubrovnik (Figure 9, down), while being famous for its famous walls and historical buildings, in 2011 became a filming location for „Game of Thrones", currently one of the most famous TV series[23]. There are several filming points within the town and tours offering visits to those points are often very visited[24]. Such locations are presented in **sec. 4.7.** The old town is also on the UNESCO list of World Heritage from 1979. Much like Split, the town has a busy sea- and airport. The area around Dubrovnik includes attractions such as NP Mljet, Ston, Konavle, Peljesac peninsula and others.

Table 3 - Split and Dubrovnik facts

	Split	**Dubrovnik**
Area of the city proper	79.3 km²	143.4 km²
Old Town Area	0.77 km²	0.41 km²
Population	167 thousand	42 thousand
Major attractions	Diocletian's Palace, Marjan Hill, Poljud Stadium	The old town, City walls, Srdj Hill
Nearby attractions (common as daily trip)	Trogir, Brac and Hvar Islands, Cetina river	Peljesac, Ston, Konavle, Elafiti islands

Note: population data refers to 2011., tourist figures to 2018. and 2009.

For relatively small cities, tourism figures and growth are high and are represented in Chapter 3.4.1., **Table 4**. Tourism activity for both cities is mostly concentrated in their centers or *old towns*, while a large number of tourists take day trips to neighbor attractions by ferries or buses. Tourists spend their nights in hotels, hostels, and private accommodation.

[21] https://www.croatiatraveller.com/central%20dalmatia/Split.htm

[22] https://100posto.hr/news/tisuce-turista-u-gradu-pod-marjanom-hamburger-na-akciji-98-kuna-a-svaka-sobica-pretvorena-je-u-apartman (Croatian)

[23] https://www.imdb.com/search/title?title_type=tv_series&sort=num_votes,desc

[24] https://www.total-croatia.com/game-of-thrones-croatia

Figure 9 - Split (up) and Dubrovnik (picture down), with the old towns enlarged in the top right corner

3.2. Dataset

The data is extracted on 10.03.2019 using the code provided by the University of Zurich. The code uses open API of the Flickr website to collect all photos uploaded on the area set and the selected period. Our final datasets, 10 csv files, contain a different quantity of entries. Each entry, presenting a photo, has the following information: *lng* (longitude), *lat* (latitude), *owner* (code for the person who uploaded the photo), *title*, *tags*, and *taken* (date and hour the photo was taken). Part of the information which came with photo is selected as redundancy and removed. Details behind the code and pre-processing are presented in the Methodology section (sec. 4.2.), as well as some such as user contribution, yearly distribution, example photo with its metadata, etc (see Figures 14 and 19).

We presented some data biases in the previous chapter (sec 2.7.). As expected, our dataset does not lack of such issues. The one that affects the data the most is the case of (over)contributing users, which tend to produce different trends from reality. An example of this is shown in the Result chapter and can be read in Table 11. Other possible data issues can come from the fact that the users of Flickr come from predominantly Western countries, and cannot be representative of all tourists. This, however, to some extent does not apply for our work, since the majority of tourists in Croatia come from those countries where Flickr is well used. Also, the fact that Flickr loses its popularity in recent years addresses the platform bias we discussed.

Since we focused only on tourists and their behavior, we will also remove the uploads of locals, using the methods presented in the *Background*, and further discussed in the next chapter. As the numbers on uploads in Table 11 will show, this did not affect the data on a large scale. Part of the analysis is focused on gaining knowledge from the tags and titles of the photos. For this, we first wanted to check if the users have a habit of adding such information and if this changed over time. The method behind this is presented in section 4.7. and the results in sec. 5.5.

3.3. Flickr website

Flickr is considered to be one of the oldest and most popular images and video hosting websites in the last 15 years for the sharing and organizing photographs (Spyrou and Milonas, 2016). The site started with the work in February 2004, and has around 90 million active users daily, according to website Alexa[25] and other sources[26]. It is owned by Yahoo and is a globally recognized site. However, its popularity varies greatly in different parts of the world.

As there are no official statistics on the number of users depending on country of their origin, a clue how user numbers can vary can be shown by visualization of the large dataset provided by Yahoo, namely *YFCC100M*, which contains of over 100 million photos and videos, around half of which are geotagged[26] (see Figure 10). The visualization suggests that the majority of the users are placed in Europe, North America, and East Asia. This, naturally, also can suggest that those regions are in general more visited by tourists. The most dominant language used for titles and tags by users is English. Other well-used languages are Spanish, German, Italian, French, etc.

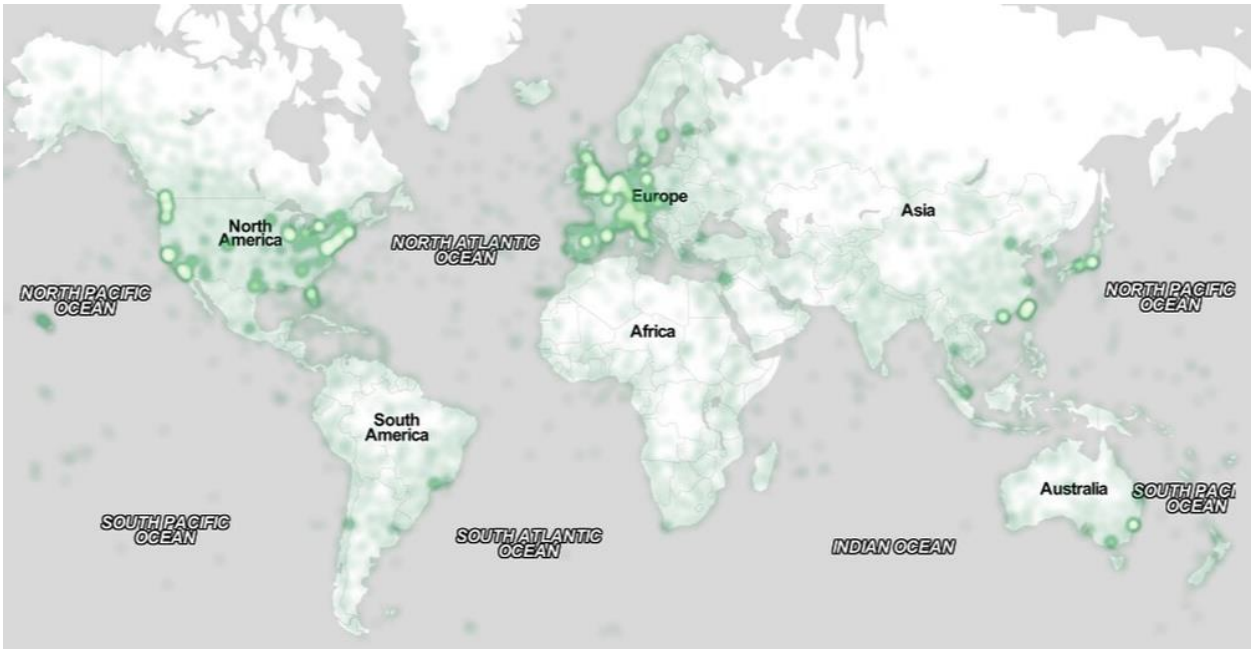


Figure 10 - YFCC100M visualized on the world map, the map shows the distribution of a million randomly selected photos

[25] https://www.alexa.com/siteinfo/flickr.com
[26] https://expandedramblings.com/index.php/flickr-stats/

There are different estimates on the number of registered Flickr users and hosted photos. Some estimations give the number of around 6.47 billion in January 2018, which decreased to 2.38 billion in July 2019[27]. This big drop in numbers comes from a new policy in February 2019 where an individual user is limited to upload a maximum of thousand photos, going from 1 terabyte as it was before[28]. The users, however, had an option to move to a pro version of Flickr, in which case all of their photos would be kept. Additionally, many uploads of users considered to be inactive were deleted (Figure 11). The amount of uploaded videos is noticeably smaller, as it takes only 0.3% of the total number of uploads. Despite the fact that only 18% of uploaded photos have some kind of geographic information and even much less have geotag (around 3%, according to Wider et al., 2013), Flickr is frequently used among researchers because of its open API, which is not the case with similar photo sharing and social media sites.

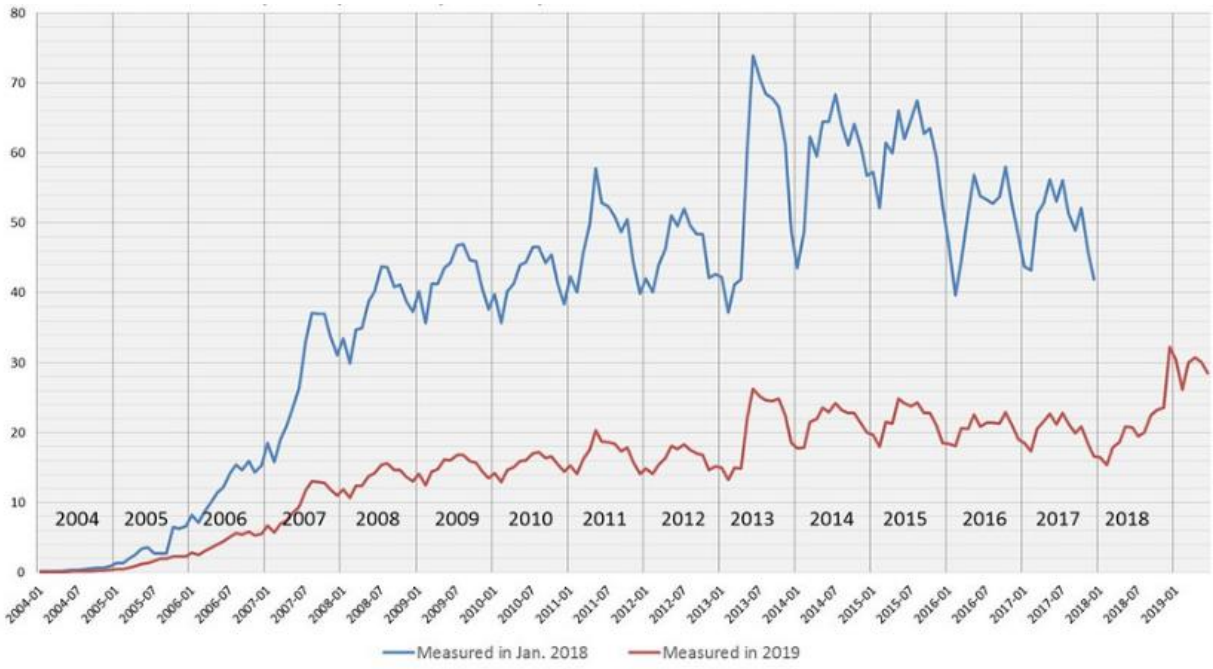


Figure 11 - Drop in the number of photos on the website after changes in February 2019

[27] https://www.flickr.com/photos/franckmichel/6855169886
[28] https://www.vox.com/the-goods/2019/2/6/18214046/flickr-free-storage-ends-digital-photo-archive-history

3.3.1. Using Flickr

The content of Flickr can be reached either by using a desktop site or the application. For users that are not logged, the front page usually just offer the possibility to search images by a query. The results are by default set to show „relevant“ photos first, but it can be changed to upload date. When the user is logged in, the front page offers more personalized content, based on interests and previous searches, as well as groups the user joined as a member.

The users can upload their photos again on the desktop site or by using the app. On the desktop site, there is a possibility to upload a photo from the computer or to use the „drag and drop“ method. If the user uses the app, it is possible to either upload a photo from a device or to take one while using the app. After the upload, the user inserts the metadata (see sec. 4.2.1). This includes tags and titles related to the content, as well as description and location, all of them being optional[29]. Some GPS-enhanced cameras can automatically provide the information on location, however, most of the users need to self-set the location by zooming on the map (see Figure 12), either before or after the upload. The app also provides the option to use the current position of the user as the location of the photo. The photos uploaded with the exact GPS location should have the best correctness in contrast to other upload ways. It is expected that a large proportion of photos uploaded have incorrect location inserted.

Users can set their photo albums either to be public or private[30]. If public, photos are subjected to both query and API search.

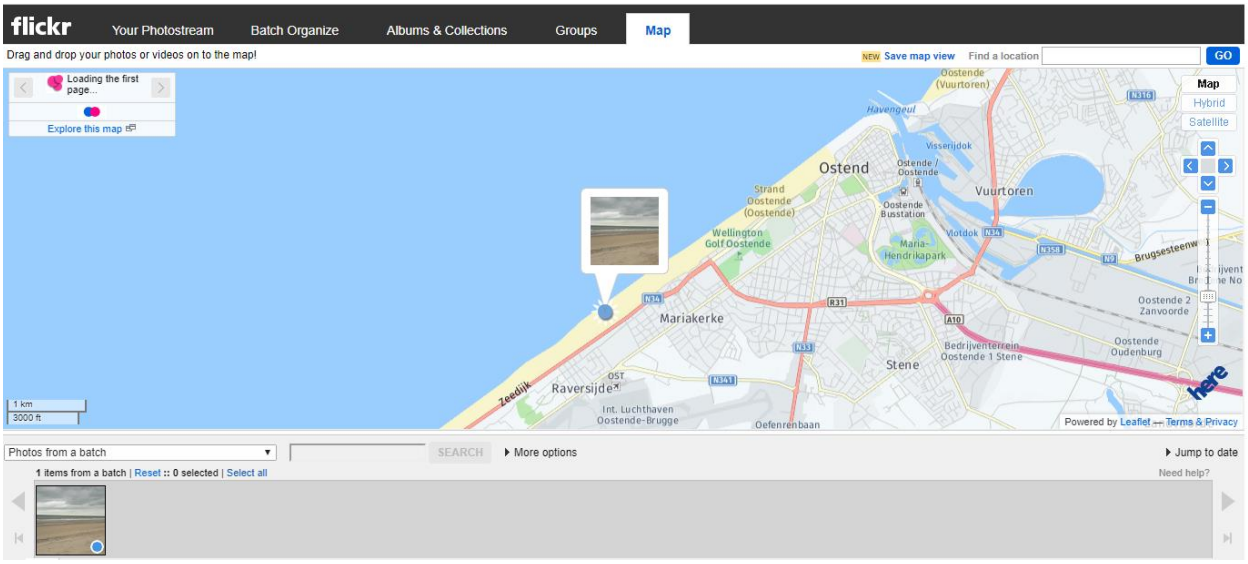


Figure 12 - Uploading process in Flickr website – after the area on the map is zoomed, a users clicks on the position where the photo was taken

[29] https://help.flickr.com/upload-photos-and-videos-to-flickr-BkE9nX01X

[30] https://www.olympiacameraclub.org/files/How-to-Use-Flickr.pdf

3.4. Data from official sources

As it is pointed out, one of our tasks is to compare our findings to the official data thus validate them. Because of the obvious flaws of the collection process, our findings cannot replace, but rather complement and enhance the official tourism figures. The official tourism data we are going to present can be accessed online and is collected and published by the Croatian Bureau of Statistics[31] (Croatian: *Drzavni zavod za statistiku, DZS*) each year in February or March, for the previous year. Another source for the data is Eurostat[32], which provides statistics for Croatia and other European countries.

Among others, the data contains the following figures:

- Tourists per destinations (counties, cities, towns, and municipalities)
- Number of overnight stays
- Trends in the figures (relative and absolute change)
- Sociodemographic profile of tourists (age, country of origin)

While making a sociodemographic profile from data such as ours is possible, and has been done by some researchers, having in mind the complexity of such a task and a rather low importance for our thesis, we will not focus on it.

As it is expected, the data from official sources do not show moving patterns between destinations. This is also difficult, or even impossible, to calculate using the data provided by local authorities, as it would require surveys, polls, interviews, or similar (expensive) techniques. A possible option would be to extract the data from highways (entrances and exits) but they do not necessarily cover tourists and the network of highways is usually not as dense as regular roads. Additionally, there is no data on how many locations tourists visited or how many kilometers they made.

[31] https://www.dzs.hr/default_e.htm
[32] https://ec.europa.eu/eurostat/statistics-explained/index.php/Main_Page

3.4.1. Tourism figures

Dalmatia, much like Croatia itself, has high growth in tourism figures in the observed decade. This goes both for the number of arrivals (tourists stayed at least one night in the area), as well as for the total nights spent, either in private accommodation, hotels, hostels, camps, etc. The growth for Dalmatia mostly follows or slightly exceeds the growth of Croatia, while Split and Dubrovnik show high to very high growth. For example, in 2009, only 1.75% of people visiting Croatia also visited Split, while that number in 2018 was 4.65%. The average of nights spent over this decade is shown only for Dalmatia, as we will compare only those numbers with our figures later on (Table 4).

The growth figures are also supported with a rise in popularity on the Internet. For example, TripAdvisor calculated that Croatia was a country with the highest growth in search and popularity in 2018[33].

Table 4 - Tourism figures for Croatia, Dalmatia, Split, and Dubrovnik

		2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Croatia	Arrivals	10.27	10.64	11.46	11.84	12.43	13.13	14.34	15.59	17.43	18.66
	Nights	55	56.4	60.35	62.74	64.82	66.48	71.61	78.05	86.2	89.65
Change	*Arrivals*	-	3.6	7.7	3.3	5.0	5.6	9.2	8.7	11.8	7.0
	Nights	-	2.5	7.0	4.0	3.3	2.6	7.7	9.0	10.4	4.0
Dalmatia	Arrivals	3.9	4.2	4.5	4.7	5.1	5.5	6	6.5	7.5	8.1
	Nights	22.3	26.9	25.5	26.7	28.3	29.8	32.1	35	39.1	41
	Av. Stay	5,72	6,40	5,67	5,68	5,55	5,42	5,35	5,38	5,21	5,06
Change	*Arrivals*	-	7.69	7.14	4.44	8.51	7.84	9.09	8.33	15.38	8
	Nights	-	20.63	-5.2	4.71	5.99	5.3	7.72	9.03	11.71	4.86
Split	Arrivals	0.18	0.2	0.25	0.27	0.32	0.38	0.49	0.58	0.72	0.87
	Nights	0.43	0.41	0.64	0.68	0.86	1.05	1.34	1.72	2.12	2.49
Change	*Arrivals*	-	11.11	12.50	10.80	11.85	11.88	12.89	11.84	12.41	12.08
	Nights	-	-4.65	15.61	10.63	12.65	12.21	12.76	12.84	12.33	11.75
Dubrovnik	Arrivals	0.52	0.56	0.61	0.66	0.73	0.82	0.89	0.99	1.17	1.26
	Nights	1.86	2.03	2.16	2.37	2.59	2.82	2.98	3.37	3.88	4.06
Change	*Arrivals*	-	10.77	10.89	10.82	11.06	11.23	10.85	11.12	11.82	10.77
	Nights	-	10.91	10.64	10.97	10.93	10.89	10.57	11.31	11.51	10.46

(Note: Figures for arrivals of tourists and nights spent are in millions; Source: DZS)

[33] https://www.croatiaweek.com/croatia-no-1-for-rise-in-popularity-on-tripadvisor-in-2019/

Another statistic we will extract is a figure that represents the annual distribution of tourists. Croatia is among the top countries of Europe when it comes to seasonality. This means that the difference between months with most stays, compared to bottom months, is very large. In the case of Croatia the ratio between August, the peak month, and February, the bottom month, is 55.7 (there is 55.7 times more night spent in August in contrast to February).

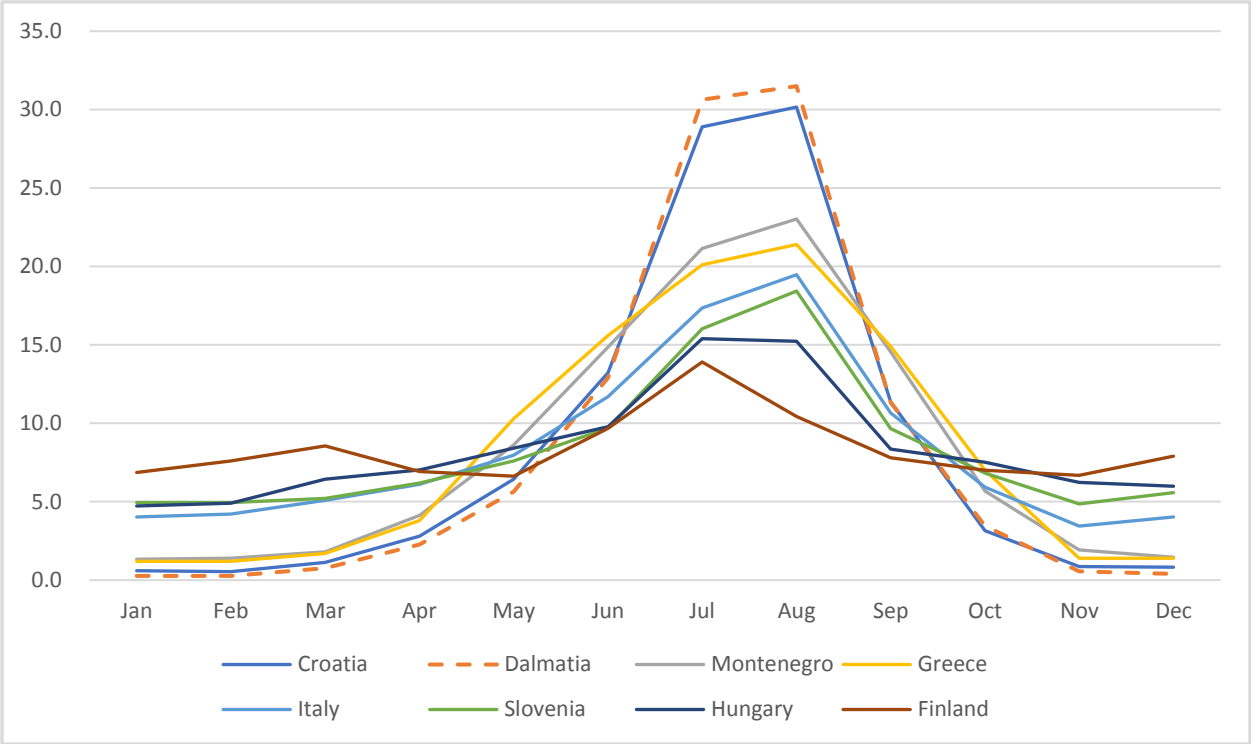


Figure 13 - Relative amounts of overnight stays within selected European countries and Dalmatia. Source of data: Eurostat

We extracted the data for Dalmatia and, for comparison reasons, a few other European countries. The graph on Figure 13 can be compared with graphs on Figure 42, to observe to which point UGC data corresponds to the data from these official sources. Additionally, and according to the official data, the region has an even larger difference between least visited months (January or February at 0,11 million) and most visited August (12.82 million). Such a ratio is 116.6, which again explains strong seasonality of the region's tourism (Table 5).

Table 5 – Night spent figures for Croatia and Dalmatia from the official sources (2018, in millions)

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Croatia	0.51	0.55	0.95	2.51	5.76	11.85	25.92	27.1	10.18	2.82	0.78	0.73
Dalmatia	0.11	0.11	0.31	0.93	2.3	5.25	12.47	12.82	4.61	1.41	0.23	0.16
Ratio (%)*	21,6	20,0	32,6	37,1	39,9	44,3	48,1	47,3	45,3	50,0	29,5	21,9

** Ratio of nights spent in Dalmatia within Croatia. Source: DZS*

3.5. Game of Thrones series / Ultra festival

As we presented in the Background part (sec 2.8), filming of popular tv shows, series or movies can influence the popularity of a destination, as well as increase the number of visits and total income from tourism, and the same can be applied for music/art festivals. Here, we are going to introduce a popular TV series Game of Thrones, which is filmed in Croatia from 2012, as well as electronic music festival Ultra Europe, held in Dalmatia from 2013.

Ultra festival, officially *Ultra Europe*, is an outdoor music festival that hosts some of the most popular electronic music artists and is held in Croatia from 2013. The most prominent locations include the city of Split, including the Poljud stadium, and islands of Brac, Hvar, and Vis. The festival attracts thousands of guests from all over the world. A part of them is accommodated in Official Ultra's campsite, while the rest take private accommodation, hotels or hostels. It is estimated that over 100,000 visitors attend the festival every year[34] [35].

Game of Thrones (GoT) is a US-fantasy/ drama series based on the novels of George R.R. Martin „A Song of Ice and Fire". The series count 8 seasons in total and was filmed in various locations which include Malta, Morroco, Spain, Iceland, Canada, and Croatia, during the period between 2011-2019. It is one of the most influential, popular, and viewed series of all times. It is among the highest-rated series on the popular website IMDb, scoring 9.4 as of August 2019, as well as the series with the highest count of votes, at 1.5 million. The filming in Croatia started in the second season when Dubrovnik replaced Malta for King's Landing, one of the central locations for the story. In later seasons, the series was also filmed in Split, Sibenik, Trogir, and Kastel. For obvious reasons, the locations were visually changed using CGI in post-production[36].

[34] https://ultraeurope.com/previous-lineups

[35] https://www.total-croatia-news.com/lifestyle/26310-ultra-europe-in-numbers-10-facts-about-ultra-2017

[36] https://gameofthrones.fandom.com/wiki/Filming_locations

Chapter 4 –

Methodology

This chapter will firstly introduce important definitions used in the thesis. While we provided some discussion on terms in previous chapters (for example, term „trajectory“ is also discussed in **sec. 2.1.**), here we only give simplified definitions that apply to our work. After this, we will present software used in the thesis, as well as steps and scripting behind the pre-processing. Then, we will explain the extraction of the Points of interest, the process behind grouping photos into relevant clusters which are joined to destinations. We also explain how and why we compare these extracted points with the data from authorities, as well as how the comparison between periods should work. After POIs, the process behind trajectories is presented.

The next step is to present a simple method to show temporal changes in tourism in the last ten years. Since UGC data can be a rich source to check trends in tourism, we decided to see if the seasonality of tourists' visits can be shown here as well. Finally, to see to what extent Dubrovnik's image changed over time, we will observe the tags and titles of the photos in the area. The pre-processing part is explained in the Dataset part of the thesis (sec. 3.2.) while the results of all of the tasks will be presented in the next chapter. Also, a part of the codes for some tasks will be shown in this section, while the full code can be found at the end of the thesis.

4.1. Preliminary definitions

Geotagged photo – also in a text named as a *photo* (photography), *upload* or *image*, can be defined as a multimedia-type (in contrast to a text) of upload on the Internet by social media site users - photographers. Each photo contains metadata which, even when redundant, can be extracted and explored. Photo, in order to be considered as geotagged, needs to have a *geotag*, standardized code that can be inserted into information to note its appropriate geographic location (Goodchild, 2007), set by latitude and longitude. Another metadata content of such a photo includes unique photo ID, photo's temporal context (date and time taken and uploaded); user name or unique ID, and tags (Zheng, 2012 and Memon et al, 2014). Metadata can also consist of name, description, URL, etc.

Tags – also known as *labels*, are keywords added to photos. The process of creating tags is known as tagging (Murugesan, 2007), which can be also explained as „the act of adding human understandable, descriptive keywords to photos" (Spyrou and Mylonas, 2013). For an example of one photo's tag, see Figure 14.

Tags	Example Photo	Comment
id	26851585730	kept
accuracy	16	no use
lng	16.2473	kept
lat	43.51541	kept
owner	42038165@N02	kept
title	Trogir - Kamerlengo	kept
tags	croatia;unescoworldheritagesite;worldheritagesite;napoleon;trogir;hrvatska;dalmatia;dalmacija;saracen;habsburg;coloman;	kept
description	NA	no use
taken	Tue Apr 01 00:00:00 CEST 2008	kept
posted	Fri May 20 09:52:33 CEST 2016	no use
license	0	no use
placeId	4Boo_xpZV7o7r6E	no use
url	https://flickr.com/photos/42038165@N02/26851585730	no use
secret	9dff0ff4b8	no use

Figure 14 - Example of a photo with its metadata

Photo collection – a collection of photos, specifically for this thesis – geotagged photos – that a user or users take and upload in particular space and time. Visualized and connected can ideally present the trajectory and movement of a user.

Photographer – a Flickr-website user which uploads and shares his photos online. In work, we sometimes also use the word *contributor*.

Tourist – to some extent, it can be used along, or instead of, the term *photographer* – it refers to a person who is traveling or visiting a place for pleasure and usually spends at least a night within the place.

Destination – „refers to popular places, such as attractions, sights or landmarks, within a city or a region" (Lu et al, 2010). The minimum and necessary information of a destination is the name, latitude, and longitude.

Points of Interest (POIs) (similar to Regions of Interest, ROIs, the term rarely used in this work) – are „a specific point location that is of interest" and an area within an urban and natural environment which attracts people's attention (Kuo et. al, 2018). The term is also similar to the term „destination", but it is derived from UGC data.

Trajectory – is defined as a sequence of footprints represented by uploaded geotagged photos. We can also call it *route* or *path*.

Additionally, the following definitions are related to the technical part of the thesis:

Cluster – in a broader sense, a group of similar objects positioned closely together. In our work, a cluster represents a group of photos uploaded close to each other, with a defined maximal distance.

Heatmap – a graphical representation of data where different values showing different densities are represented by colors. The density is based on the number of points within the space. In this case, density is based on the number of uploaded photos. Ideally, heatmaps allow identification of popular locations (hotspots).

Buffer - a zone around a map feature measured in units of distance (meters in our work). As an input, we have a point, and as output, a polygon.

4.2. Software, Data extraction and pre-processing

The dataset was provided to us by the staff of the University of Zurich (UZH), Geography Department, Geocomputation, using the code available at the end of the thesis. Because of the changes made by Flickr, the code cannot be used anymore for the extraction. Several steps were made in order to pre-process the data, also by using a few different software. The data then was ready for analysis.

4.2.1. Software

In addition to Microsoft Office software, such as Word and Excel, three other software was used for various steps within the thesis, namely Rstudio (R programming language), PyCharm (Python programming language), and QGIS. Through the thesis, they are used alternately. We are going to present them shorty here, including practical use of them.

RStudio – open-source Integrated development environment (IDE) for R programming language, usually used for statistical calculations and visualization of data. It is available in the desktop version (RStudio Desktop), which we used, and RStudio Server, which can be used with a web browser. It was founded in 2009 and released in 2011 by RStudio, Inc[37]. We used RStudio mostly for data pre-processing, removing unneeded rows, deleting zero values, tag analysis, and similar.

PyCharm – is an IDE for Python programming language. It offers an intuitive graphic design which analysis the code in real-time. The initial release was in 2010, by JetBrains[38]. It was used for most of the programming of the work, including visualizations that are presented in the Results chapter. The calculations for word (tag) ratio were also done in PyCharm.

QGIS – is a Geographic information system (GIS) free, open-source desktop software, a volunteer-driven project[39]. Its development began in 2002 by Gary Sherman and QGIS Development Team. It was initially released in 2009. The primary use is to edit, visualize, and analyze geographic data. It supports raster and vector data. The most common format of the data is shapefile. We primarily used WGS 84 (EPSG: 4326) Coordinate Reference System, as it is standardized and used by most of the layers we downloaded. For some specific calculations, we have also transformed it into WGS 84 / Pseudo-Mercator (EPSG:3857), as it has meters as a unit. For our visualization, we used Google Satelite as a base map, set at XYZ Tiles. There are various Internet and official sources for shapefiles we used. In addition to the visualization and export of maps, some calculations of the data are done in this software.

[37] https://rstudio.com/about/
[38] https://www.jetbrains.com/pycharm/
[39] https://qgis.org/en/site/about/index.html

4.2.2. Data extraction

To extract the data from the Flickr website, we used the code provided by the UZH. The code uses open API of the Flickr website to collect all photos uploaded on the study area, set by coordinates of northwestern- and southeasternmost points (Figure 12).

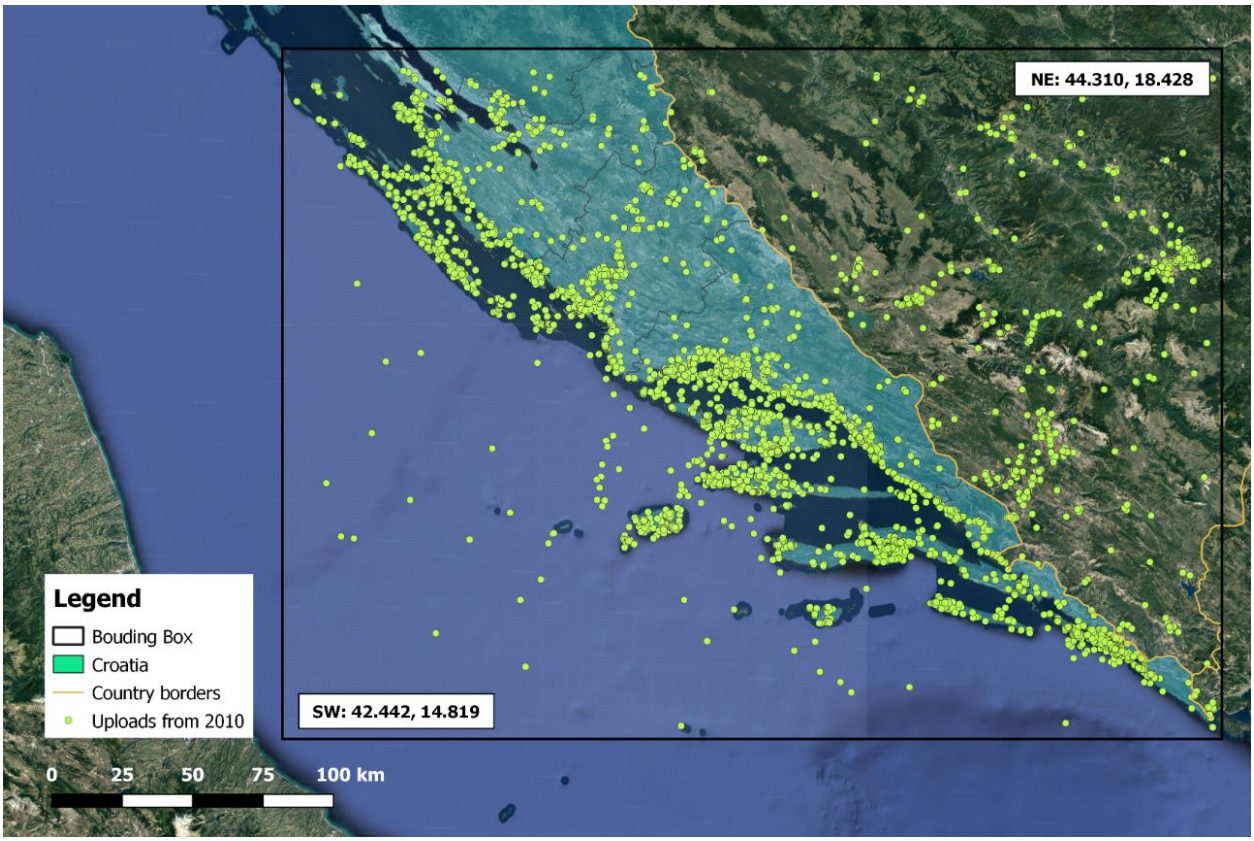


Figure 15 - Bounding box is determined by two points; the image shows data from the year 2010

We ran the code ten times with the date set from January 1st to December 31st for each of the years. The option was also to run the code once for all 10 years and manually separate the data, but this shown to take more time and produced crashes. Each run required an API key, which is gained from the Flickr website after registration and can be reused. The extraction for each dataset took around 10-15 minutes, and the code ran with no major interruption so the data quality is expected to be consistent. By using the Flickr App Garden[40], it is possible to retrieve the data without the code we used.

As we have shown, Flickr removed a part of the uploads from some users on March 12th, 2019. This did not affect our dataset but does affect future researchers which aimed to do a similar task. As we wanted to compare different years and the significance of PoIs and trajectories, so the data consists of 10 .csv files, each representing user uploads of photos in each of the years from the period 2009

[40] https://www.flickr.com/services

to 2018. Each entry in a *.csv* file, representing a photo, consists of the following information: *id, accuracy, lng, lat, owner, title, tags, description, taken, posted, license, placeid, url,* and *secret*.

4.2.3. Data pre-processing

The *.csv* files were used interchangeably in the aforementioned software. For starts, the part of the metadata which is mentioned is of no use for the task, so we only left following information: *lng* (longitude), *lat* (latitude*), owner, title, tags, description, and taken* (time when a photo was taken, or for part of the data, uploaded). This was done in RStudio software by simple row dropping. This increased the data readability and made it easier to process. An example of photo metadata is shown in Figure 14.

To remove the uploads within the bounding box but outside of the region, we used QGIS. Most of the layers used *EPSG:4326–WGS 84* coordinate system. First, we uploaded layers of Croatia with county borders and of Europe. Then, we visualized all of the photos from *.csv* files by using *lng* and *lat* information. Then, by using the „Select by location" feature, we cropped out photos of unneeded territories. Finally, we exported such *.csv* files again.

As discussed in the previous chapters, by removing outliers (users whose photo collection exceeds the count of 150), as well as those who uploaded under 5 photos, we lower the effect of user bias. The value of thresholds were set after visualizing and observing the data. It was clear that those users with too many uploads can alone make „trends", especially in less-visited destinations. Contrary, users who uploaded under 5 photos usually do not make meaningful or precise trajectories. This task of removing such users was made in RStudio. We related the information of photographers' unique ID ($\mathbf{u_p}$) and the upload count ($\mathbb{P}$). The users whose upload count falls under 5 or exceeds 150 were removed from any further research. Then we tried to identify if the uploads are made by the local residents or by tourists. As it is discussed in the Background section (**sec 2.2.**), different study areas, whether big cities or regions, should have different thresholds for such tasks (Budapest at three, Paris, Amsterdam, Vienna at five, Madrid at seven and similar). Based on the data from official sources, which shows that tourists stay on average 5-7 days, and as high as 10 days for some destinations, we wanted to include as many tourists as possible, so we took the threshold of 20 days. Similar to removing outliers, we made a condition in PyCharm that users whose timestamps between the first and the last photo in the photo collection is more than 20 days are removed. For this, we calculated the day count of each user and transform it into seconds. If the total day count exceeds 20, expressed in seconds, the user was removed. We also tested different thresholds, such as 14 and 30 days, and while 14 days excluded a somewhat significant number of Flickr users when setting the number to 30 days, it almost did not affect the output data. This suggested that many tourists stayed for more than 2 weeks. The code for this is presented at the end of the thesis. This completed the pre-processing part, which is summarized in Figure 16.

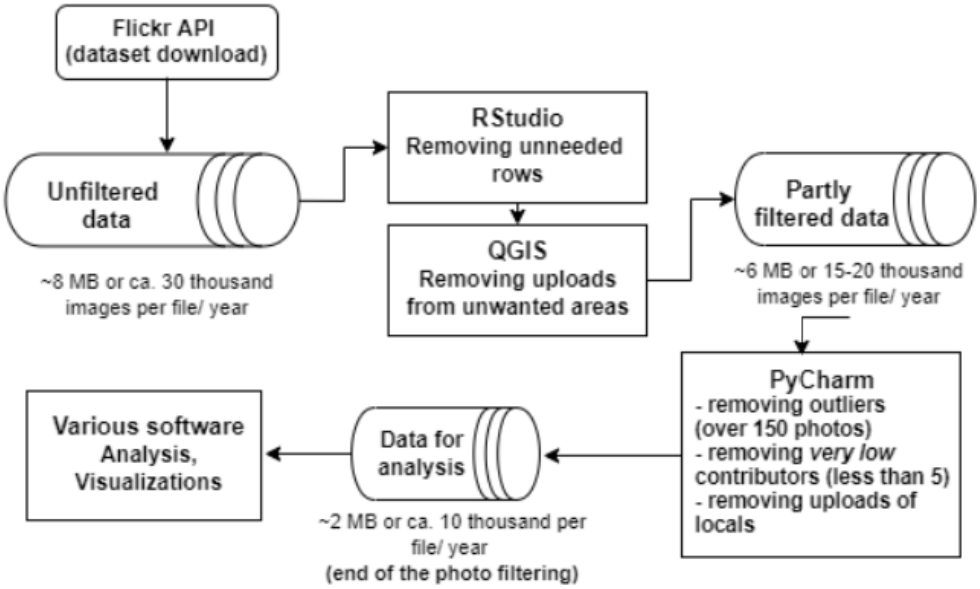


Figure 16 – Steps of filtering (pre-processing) the data

Also, as will be shown in sec. 5.1., such pre-processing steps left us with on average 10 thousand photos a year, with a trend of decreasing, both in terms of the number of users contributed and photos uploaded. Table 6 gives us the average figures for 2009. We split users we will use for further work into two groups, namely „low" and „high" contributors. It can be seen how the users with more than 150 photos have almost 50% of the contribution within the total number of photos for the year. Additionally, all years had a few users who uploaded over 500 or even 1000 photos. In the Result chapter, such data is given for all of the years. We also calculated Average uploads by dividing the total number of photos from selected users and their total number.

Table 6 - Example of the data on user upload for the year 2009

Number of photos per user	Number of users	Total number of photos
1 - 4 photos:	921	1726
Low-contrbutors (5-30):	202	1987
High-contributors (31-150):	260	10346
151 and more photos:	46	10885
Total	**1526**	**24432**
Selected users	**462**	**12333**
Average	**26.7**	

4.3. Points of Interest

Points of interest should represent the most popular destinations within the region of Dalmatia. Different approaches to extracting them are presented in the *Background* part of the thesis (**sec 2.3.**). We will use the one which is also used by the majority of the authors and which proved to work well. Again, to do this, we used both R and QGIS software.

4.3.1. Discovery of POIs

Before clustering the uploads into POIs, we are going to make „heat-map" of the dataset. Heatmaps are useful when we want to find the density of points, as they visualize different density with different colors, making it easy to distinguish less from more „active" areas. There are several options to do heatmaps, and we are using two different techniques, combining them both into one map.

First, we uploaded all the points of a year in QGIS and used Kernel Density Estimation. After we selected the layer we wanted, we needed to chose on a radius that works for the best considering the data. The unit of the radius is in the unit of the layer, so we needed to transform our projection to EPSG:3857 - WGS 84. There are only a few parameters given to change the output, namely selection of the radius (buffer size around each of the points), and the number of rows (number of rows in output raster). For output, we selected radius at 200 meters and a number of rows of 3000. To give an intuitive visualization of different density of photos, seven classes of different values were chosen, by choosing Render type: Single-band pseudocolor.

Another possibility is to visualize points by selecting Symbology: Heatmaps. We selected the radius of 25 millimeters and the maximum value as automatic. The result of both methods can be seen in Figure 15 (example) and in the Results chapter.

However, the issue with the heatmap visualization methods is a noticeable polarization of the distribution of photos. This specifically means that only a very few locations had a very large number of photos (namely Dubrovnik and Split), which made classes for visualizations to vary greatly. This means that light colors, such as pink (see Figure 17) can represent areas of only a few uploads (2-5) while very dark red can have over 1000 uploads. As it would be expected for any region for a period of just one year, most of the area has no uploads.

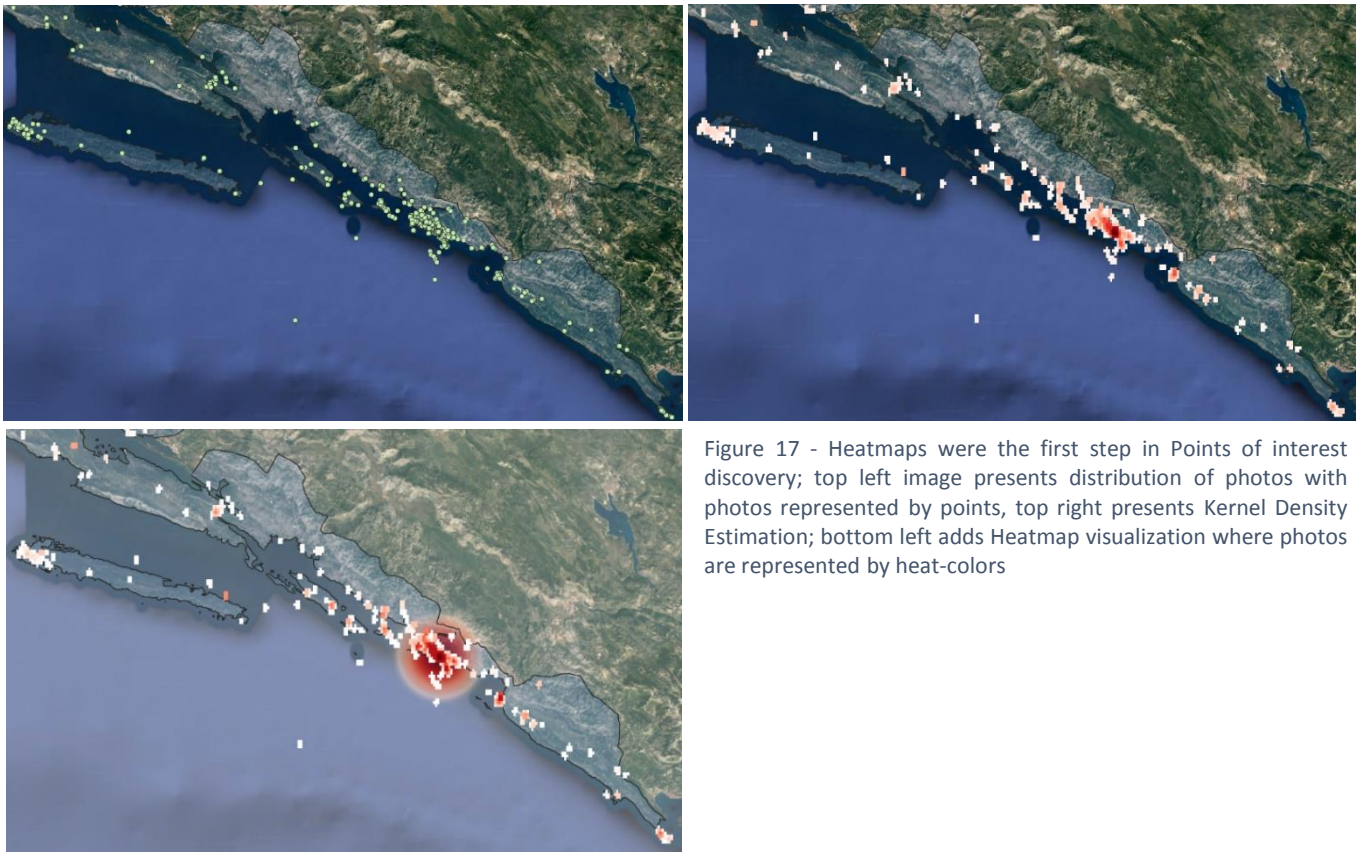


Figure 17 - Heatmaps were the first step in Points of interest discovery; top left image presents distribution of photos with photos represented by points, top right presents Kernel Density Estimation; bottom left adds Heatmap visualization where photos are represented by heat-colors

Then, to identify highly photographed areas and classify them as POIs, we applied the DBSCAN method[41]. For the same goal, this method was used by other researchers, eg. Memon et al., 2014, Huang (2015) or Zeng et al. (2012). To use the method, it is required to set two parameters, namely a minimal number of points (MinPts) which make the cluster, and size of the Optimal Epsilon (Eps), which is a maximal distance between two points to join them into the same cluster. Additionally, the method extracts points not joined to any cluster and categorize it as „noise".

The number of clusters, which correspond to POIs, is variable as it depends on the value of two parameters mentioned. For a different type of area, namely big European cities, Huang (2015) used set MinPts = 100 and Eps = 30 m to find around 20 POI per city, while Zeng (2012) used Eps of 30 and MinPts of 5 for (rather) small area such as Forbidden City in Beijing, China. Since we wanted fewer POIs and we had a different area with different upload properties, we decided for Eps = 45 m [42] and MinPts = 40.

After some test runs, it is decided that in the top 10 POIs each year will be extracted within the region. They are represented by circles whose size depends on the number of points that make the buffer (Figure 18, right). To add names to those circles, we exported their location as a csv file and

[41] Abbervation given in previous chapters
[42] Expressed in degrees, 0.04

imported it into QGIS. Then, the tabular overview of those top buffers over the years will be presented.

The clusters, naturally, change their shape and area they cover. This also means that we will, in a few cases, name them differently. For example, in some years, Sibenik and Krka will be represented by one unique buffer, while in others they will have individual clusters. This could be manually fixed by joining clusters once when exported in the .csv file or in QGIS software, but we decided to show the clusters as they came up after the DBSCAN method. Similarily, Orebic-Korcula is joined in all years, as the center of the buffer is directly half-way between those two places which are anyway nearby. For this reason, we also joined destinations when the official data is represented.

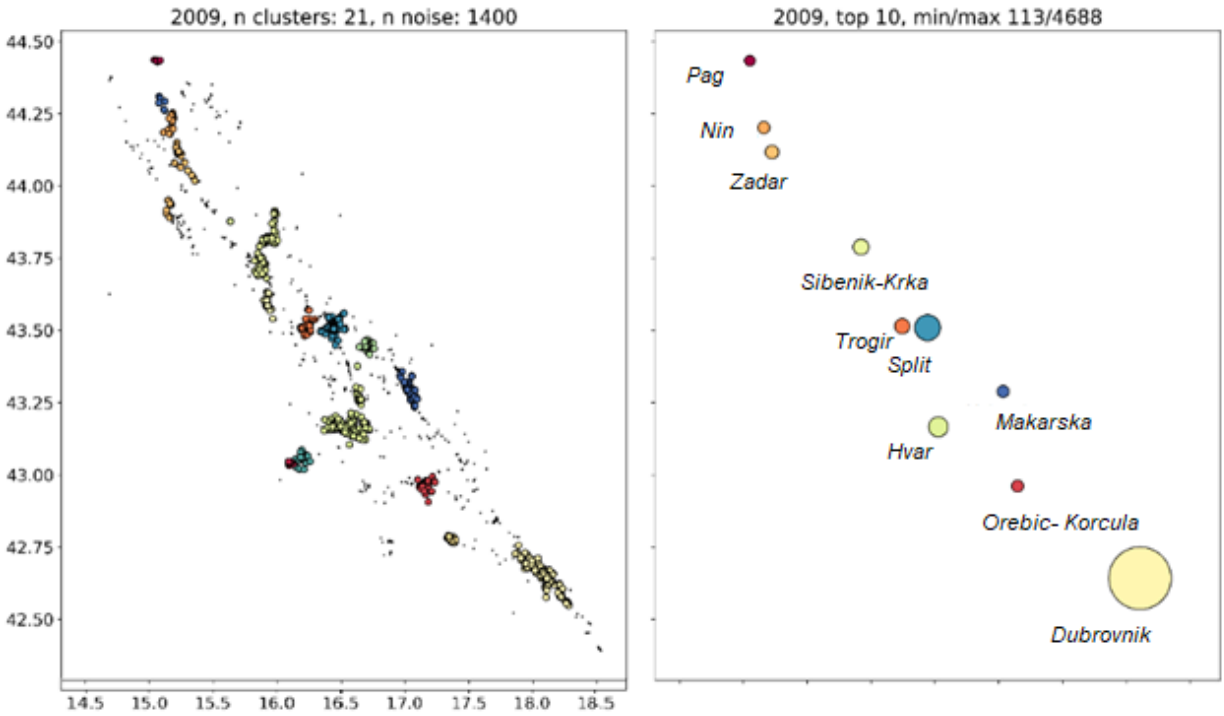


Figure 18 - Example of the clustering to discover POIs; different colors represent different buffers; the size of buffers on the right depends on the number of uploads in the area. The data from 2009.

4.4. Trajectories

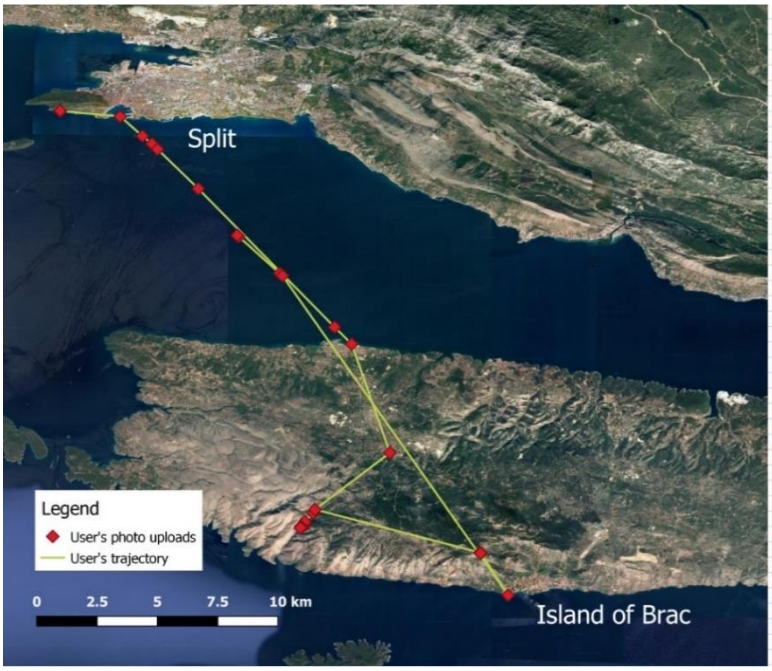
The main ideas behind the extraction of trajectories are to show the movement of tourists within the region and to find meaningful similarities and changes within 10 years by showing patterns within destinations of the area. Using spatial and temporal information from all points of each user, we extracted the distance individuals made, as well as the time they spent, in Dalmatia.

4.4.1. Discovery of trajectories

Trajectories are derived from the final data. As stated before, this means that only the users who uploaded 5 or more, as well as 150 or fewer photos, were taken into focus. We also excluded the local population, as our primary focus are tourists.

The Python code we made grouped all points (photos) uploaded by users (each user individually), taking into account the temporal component (date/time when it was taken). As each point has spatial and temporal information, it was simple to extract trajectories and later on additional information, such as length and time spent.

An example of one user's trajectory and data from which it is derived can be seen in Figure 19. The points (red dots) represent locations where, ideally, the photo was taken, while the yellow line connects dots in a temporal order and gives approximate movement in the space.



16.52899	43.29317	10082046(	Tom	croatia;pustinjablaca;geostat2009	5/9/2009 11:16
16.52854	43.29272	10082046(	Martina	croatia;pustinjablaca;geostat2009	5/9/2009 11:05
16.62041	43.27936	10082046(	Tom	croatia;geostat2009;2kmneofmurvic	5/9/2009 13:00
16.62042	43.27934	10082046(	Shoot!	croatia;geostat2009;2kmneofmurvic	5/9/2009 12:57
16.62041	43.27936	10082046(	The croatian coast	croatia;geostat2009;2kmneofmurvic	5/9/2009 13:00
16.62014	43.27936	10082046(	Photographer cowboy	croatia;geostat2009;2kmneofmurvic	5/9/2009 12:52
16.5357	43.30153	10082046(	White flowers	croatia;pustinjablaca;geostat2009	5/9/2009 11:58
16.61996	43.27955	10082046(	Croatian islands seen from Vidova	croatia;geostat2009;2kmneofmurvic	5/9/2009 12:51
16.62042	43.27934	10082046(	Sarah	croatia;geostat2009;2kmneofmurvic	5/9/2009 12:59
16.62039	43.27931	10082046(	Crocodile Dundee	croatia;geostat2009;2kmneofmurvic	5/9/2009 13:01
16.62038	43.27934	10082046(	On the edge	croatia;geostat2009;2kmneofmurvic	5/9/2009 12:57
16.62041	43.27934	10082046(	Bol	croatia;geostat2009;2kmneofmurvic	5/9/2009 12:58
16.62043	43.27933	10082046(	Bol - a little piece of paradise	croatia;geostat2009;2kmneofmurvic	5/9/2009 12:57
16.40567	43.50465	10082046(	The croatian coast in Split	croatia;split;geostat2009	5/7/2009 16:05
16.40567	43.50465	10082046(	Melita	croatia;split;geostat2009	5/7/2009 16:06
16.49682	43.44057	10082046(	Sunset in Split	croatia;split;geostat2009	5/9/2009 17:53
16.63427	43.2579	10082046(	Preparing lunch	croatia;geostat2009;3kmwofbol	5/9/2009 13:51
16.51772	43.42188	10082046(	Paul	croatia;geostat2009;11kmseofsplit	5/9/2009 17:46
16.40567	43.50465	10082046(	Sarah	croatia;split;geostat2009	5/6/2009 17:12
16.63434	43.2579	10082046(	Lunch by the sea	croatia;geostat2009;3kmwofbol	5/9/2009 14:01
16.40567	43.50465	10082046(	Roger	croatia;split;geostat2009	5/6/2009 17:11
16.61996	43.27955	10082046(	Hvar	croatia;geostat2009;2kmneofmurvic	5/9/2009 12:50
16.62027	43.2793	10082046(	Tom, Catherine, Miguel and Els	croatia;geostat2009;2kmneofmurvic	5/9/2009 12:55
16.51914	43.42064	10082046(	Olaf	croatia;geostat2009;11kmseofsplit	5/9/2009 17:45
16.49598	43.44129	10082046(	Sunset in Split	croatia;split;geostat2009	5/9/2009 17:54
16.40567	43.50465	10082046(	Juan learning R	croatia;split;geostat2009	5/6/2009 22:07
16.52955	43.2933	10082046(	Peter the astronomer	croatia;pustinjablaca;geostat2009	5/9/2009 10:34
16.53067	43.29374	10082046(	Clemens observing	croatia;pustinjablaca;geostat2009	5/9/2009 10:23
16.52899	43.29284	10082046(	Visiting the monastery	croatia;pustinjablaca;geostat2009	5/9/2009 10:05
16.5294	43.29347	10082046(	Campanula	croatia;pustinjablaca;geostat2009	5/9/2009 9:34
16.53484	43.29965	10082046(	Me and Jaime hiking to the Glagolit	croatia;pustinjablaca;geostat2009	5/9/2009 9:18
16.53573	43.30146	10082046(	Olaf and Tom observing some ants	croatia;pustinjablaca;geostat2009	5/9/2009 9:05
16.52844	43.29271	10082046(	Roger	croatia;pustinjablaca;geostat2009	5/9/2009 11:03
16.52996	43.29316	10082046(	GEOSTAT2009 around a piano	croatia;pustinjablaca;geostat2009	5/9/2009 10:44
16.40567	43.50465	10082046(	Paul	croatia;split;geostat2009	5/6/2009 17:07
16.52961	43.29324	10082046(	Aux armes citoyens...	croatia;pustinjablaca;geostat2009	5/9/2009 10:30
16.40567	43.50465	10082046(	Inside the basket	croatia;split;geostat2009	5/6/2009 17:05
16.40567	43.50465	10082046(	It's mine!	croatia;split;geostat2009	5/6/2009 17:03
16.53057	43.29533	10082046(	Mountains, sea and islands seen fr	croatia;pustinjablaca;geostat2009	5/9/2009 9:29
16.52866	43.29276	10082046(	Helena	croatia;pustinjablaca;geostat2009	5/9/2009 11:04
16.52831	43.2927	10082046(	Martina	croatia;pustinjablaca;geostat2009	5/9/2009 11:03
16.52828	43.29222	10082046(	Bells in the monastery	croatia;pustinjablaca;geostat2009	5/9/2009 10:59
16.40567	43.50465	10082046(	Fighting hard	croatia;split;geostat2009	5/6/2009 16:59

Figure 19 - Randomly selected user and their trajectory (left) and part of the data used (right) with geocoordinates of each point, title, tags, and the date and time uploaded

As we said, the code we wrote also calculates the duration of stay and trajectory length for each trajectory, which is later transformed into averages and compared with the official data and calculate the similarity. Figure 20 visualize data for 2009; the rest of the results, comparison, visualization, and analysis will be shown in the next chapter. Figure 20 also displays why the method of visualization of all trajectories would not give meaningful and readable findings, so we used tables to show patterns. In addition to averages, for length we made also boxplot visualization. It should be pointed out that trajectories are line distances between locations, rather than actual, road or sea distances, so the observed users who uploaded their photos with any distance between them made a longer actual path. To transform the line length into road length, we used method from Boscoe et al. (2012) where they concluded how in area such is this adding 20-40% is sufficient. We came with similar figures by calculating the line (280) and actual distance (350 km) from Zadar and Dubrovnik, which adds to ca 25% of the difference.

Table 7 – Trajectories - example data for 2009

Total number of users	462
Of those, trajectories > 0 km	380
Total length	58,716 km
Average length	154.5 km
Transformed length	194 km
The total duration of stay	2093 days
The average duration of stay	4.53 days

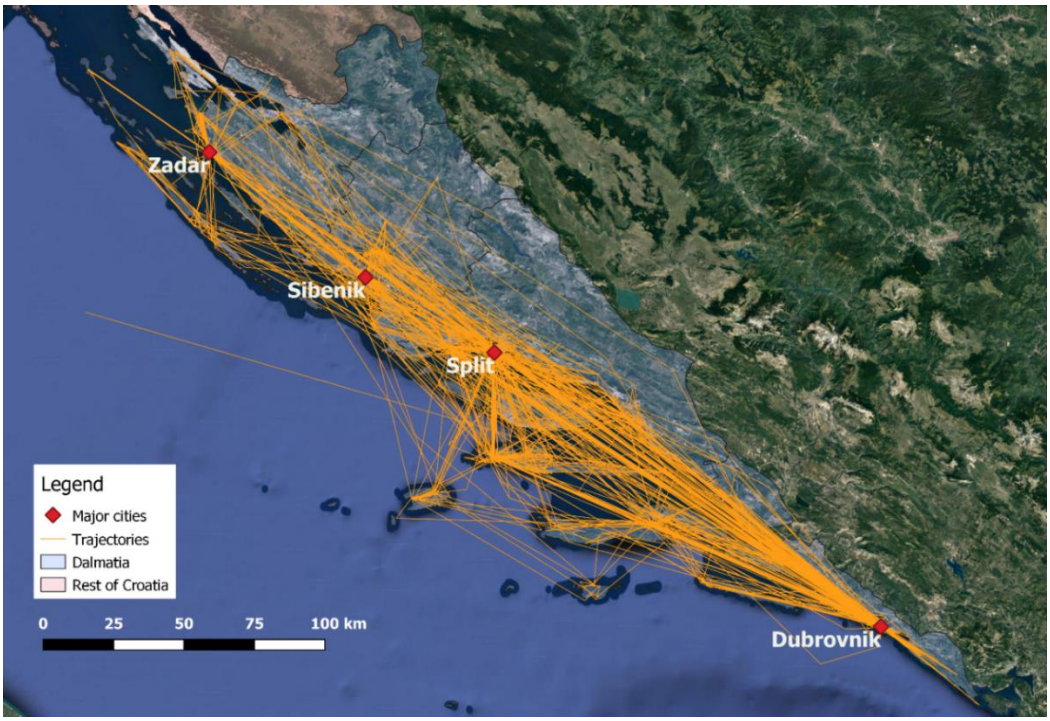


Figure 20 - Trajectories within Dalmatia in 2009

4.4.2. Patterns between destinations

Having the information on tourists' movement, represented by trajectories, also means we can calculate which destinations within the regions are most visited, and what are relations between them. This, in particular, means that we can quantify how many tourists from our dataset visited a selected destination, and out of that number, how many visited other selected destinations. A similar approach was also presented by Zheng et al. and Cao et al. Simply put, we can quantiy and represent visit patterns between selected destinations.

Firstly, we selected 13 locations that vary in size and geographical characteristics (Table 8 and Figure 21). Our choice of destinations is partly based on the points of interest we presented in the previous sub-chapter. However, we wanted to exclude destinations we found to be similar to other destinations. Instead of such places, we included all national parks of the region (Paklenica, Krka, Mljet, and the Kornati islands), all regional capitals and some major towns, some of which are not detected with our POI detection approach.

We used the layer of Croatian towns and municipalities to make polygons of destinations. Because of our code depends on points that make polygon, original municipality, city, and NP borders could not be used because of their complexity, which made the code to crash. Instead, we simplified the original shapes, also adding seaside area to some of them, as many photos were uploaded on it. Because of this approach, the polygon area is, in some instances, much larger than the original area of the place (Table 8).

Table 8 - Selected locations from the north to south. In addition to extraction from POI, "Comment" presents another reason to chose a particular destination

Destination (abbreviation)	Area (km²)	Polygon area (km²)	Comment
Paklenica NP (PA)	170.7	233.8	National Park, popular for climbing
Zadar (ZA)	52	75.6	County Capital, the second largest place of the region, UNESCO site
Kornati NP (KOR)	103	465.8	National Park, Group of tiny islands
Krka NP (KR)	127	135.1	Second-most visited National Park
Sibenik (SIB)	44.1	65.1	County Capital, 2 UNESCO sites
Trogir (TR)	11.5	18.6	UNESCO site
Split (ST)	22	39.5	Regional Capital
Makarska (MA)	26.2	39.5	Picturesque town
Hvar (HV)	28.1	89.8	Most visited island place
Korcula (KORC)	105.8	150.5	Picturesque island destination
Mljet NP (MLJE)	28.4	66.3	National Park, Island
Dubrovnik (DU)	12.1	22.6	Most visited destination of the Region
Konavle (KON)	209.6	297.9	Southernmost destination of Croatia

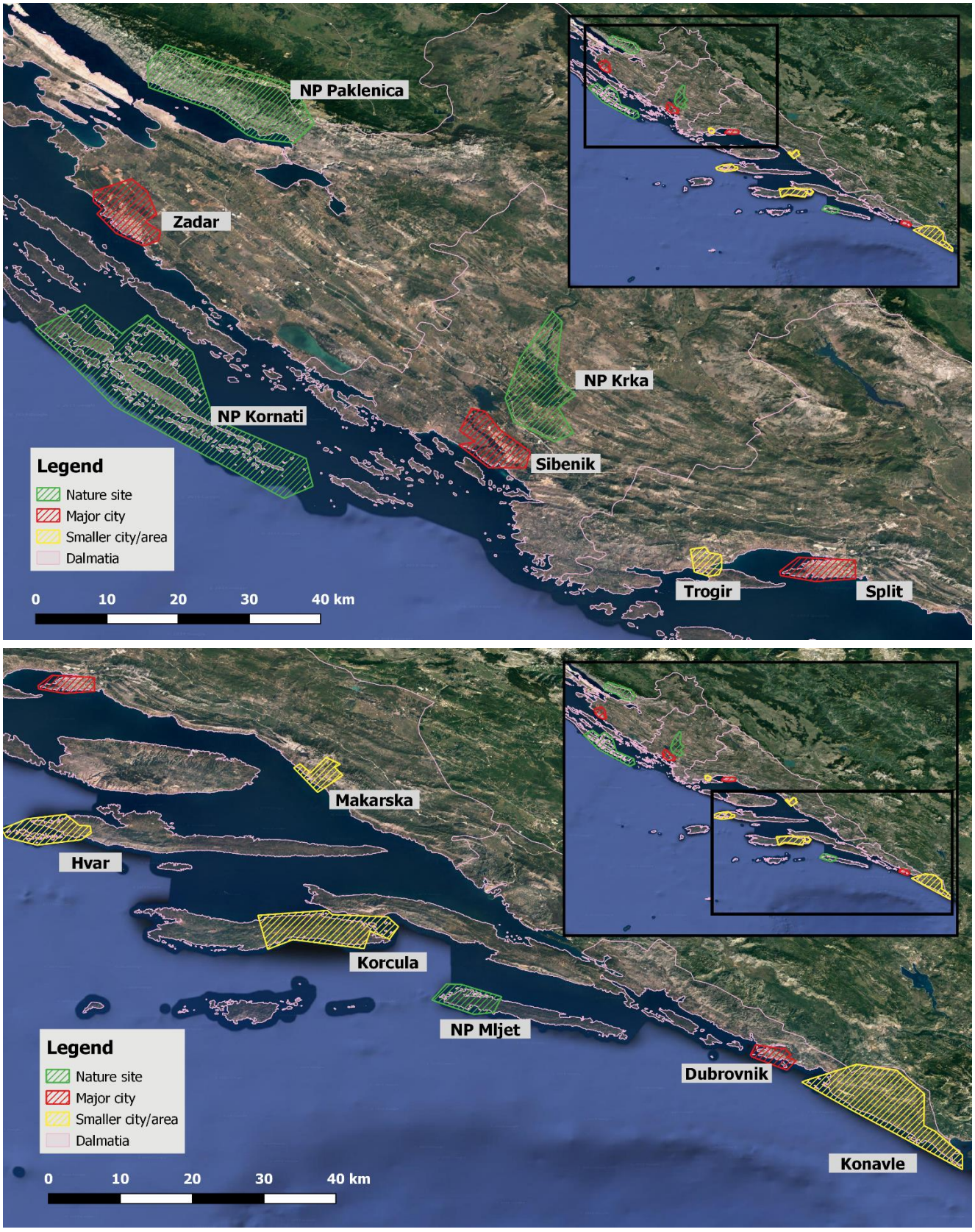


Figure 21 - Selected sites for the analyzis, North and Central Dalmatia on the image up, and South Dalmatia on the image down

Once the destinations and their polygons were extracted, we linked the informaton of coordinates of POIs to the code. The code only detects „a visit" to a destination (polygon) if there is a photo within its borders, which means that if a trajectory touches, go over the polygon but there are no „stops", it is not detected as a visit. It is also important to point out that one user can make only one visit to one destination, so it does not depend on how many photos a user uploaded, or if he or she made a return trip. This is partly represented in Table 9, as it shows absolute figures of visits of the destinations selected (the blue-marked fields). Then, the relative figures were calculated for each destination by relating absolute counts of each destination (total visits) and the absolute figures which show mutual visits. The data was exported to Excel where the Table was edited. In the Results section, we presented absolute figures for destinations visits, and both absolute and relative figures to show relations. Thus the tables in the Result section are not mirrored, unlike Table 9. Instead of showing the results for year by year, we combined two years which still should show trends.

Table 9 - Visits between destinations in 2009, absolute figures

	PAK	ZAD	KOR	KRK	SIB	TRO	SPL	MAK	HVA	KORC	MLJE	DUB	KON
PAK	9	2	1	3	1	2	3	1	2	0	1	1	0
ZAD	2	58	8	8	12	15	18	3	8	4	1	16	0
KOR	1	8	23	7	8	6	7	0	0	1	0	1	0
KRK	3	8	7	40	14	9	23	5	0	0	0	17	1
SIB	1	12	8	14	39	18	21	1	3	4	1	20	3
TRO	2	15	6	9	18	67	36	7	12	9	1	36	2
SPL	3	18	7	23	21	36	145	17	30	12	5	77	4
MAK	1	3	0	5	1	7	17	31	7	5	2	20	2
HVA	2	8	0	0	3	12	30	7	59	14	5	24	2
KORC	0	4	1	0	4	9	12	5	14	37	4	25	1
MLJE	1	1	0	0	1	1	5	2	5	4	11	9	0
DUB	1	16	1	17	20	36	77	20	24	25	9	250	25
KON	0	0	0	1	3	2	4	2	2	1	0	25	30

Note: blue-marked fields represent total numbers of visits to the destination

Table 10 presents an example of absolute figures of such a method, the relation between two places selected for the study area. As we said, relative figures are easier to understand and compare, so we will use such an approach later on.

Table 10 - Example figures for 2009 and relations between Split and Dubrovnik

Total number of trajectories	380
Visited both places	77
Visited Dubrovnik	250
% of those visited Split	24%
Visited Split	145
% of those visited Dubrovnik	50%

4.5. Comparison with Authoritative data

As discussed in the sec. 3.4, one of the goals of the thesis is to both support and enhance the data from official sources. Some figures for the tourism of Croatia and Dalmatia, as well as for Split and Dubrovnik, are already given in sec. 3.1., while additional statistics, also taken from the DZS website, will be presented in the Results chapter, namely section 5.2. We plan to compare the findings for POIs and partly for trajectories, as most of our findings are not tracked by the officials. This is also valid for user activity over time, as well as extraction of user impressions from tags and titles.

When it comes to POIs, we are going to make a list of the top ten most photographed places within the region. An example of this is already presented in Figure 14. Such a task is repeated for each of the years. Then, by using the official data sources, we are going to extract ten most visited places by the number of tourists (in contrast to the number of nights spent within destinations). We can expect that major towns of the region will correspond well, while some places might have more visits but be less covered by photos.

As we said, we will use trajectories to extract the average duration of tourists' stay in days. This will be compared to the figures from officials. This will be done with the formula:

$$PD = \left(\frac{n2 - n1}{n2}\right) x\ 100$$

where PD is percentage difference, n 1 is smaller of the two figures, and n 2 is larger of the two figures. The correctness of our findings might suggest that other figures derivated from trajectories, such as their average length and patterns between destinations, might be correct, or incorrect, as well.

4.6. User activity over time

In addition to the spatial component of users' behavior, UGC data offers a possibility to extract temporal as well. This gives us a chance to observe in which hours of the day the tourists are most active, throughout the whole year. The data is, naturally, available as every entry has the exact time of the upload, and aside to some outliers, we can assume there is fair correctness to it.

We first wanted to observe if tourism trends can be reflected in this data. More specifically, we have already presented how in a few summer months the region has almost all of its tourist arrivals, so we want to present it with our data and seek for any changes and trends. To do this, we again used the information on the upload time of photos in our data, as it gives the exact time when a photo is taken. For visualization, we used PyCharm. We chose to visualize our data in pixels, using darker tones for more uploads, and the light tones for periods with fewer uploads. On the *X*-axis, we showed the time in hours, and every hour is represented by two pixels. *Y-axis* represents months. Additionally, we set the photo-count curve on the right side, with points for easier reading of monthly figures and amounts.

We experimented with visualizations of relative and absolute counts. Figure 22 displays relative counts for pixel visualization, and absolute for the curve. When we reversed the visualization, showing absolute for pixels and relative for curves, it was more intuitive to read the results. Such visualizations are shown for some years in the Results chapter. We are going to offer a qualitative analysis of such results.

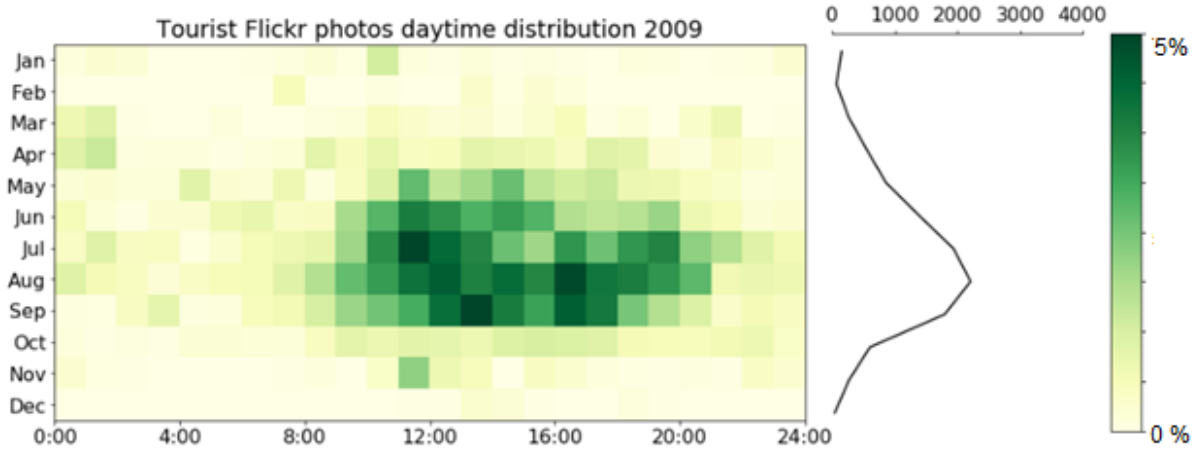


Figure 22 - Example visualization for 2009 – Daytime visualization (relative counts), upload count graph (absolute values)

4.7. Gaining knowledge using the metadata

It is discussed in sec. 2.8 how events, such as concerts or filming of popular series/ movies, can influence the image, popularity and therefore visit count of a destination. We have also presented how the usual method to validate such influence is interviews and polls. Using our data and its metadata, namely tags and titles, we wanted to observe if this applies to Split and its Ultra festival, and Dubrovnik, which was used for the filming of series Game of Thrones.

To assure that the data from the tags is of similar useability within the decade, first, we wanted to check the figures of the data in terms of ratio on title/tags. It is expected that not all of the metadata will have content – many of the photos lack in titles and tags. To check this, we calculated the ratio of the photos without titles or tags in Rstudio by extracting the amount of entries with *NULL* as value.

After this, to observe changes in users' impressions, we are used a method which, to our knowledge, was not used in any works of this kind. Namely, we chose two periods within ten years to observe different trends in 1) upload count and ratio and 2) semantics from user-made tags. For the first period, we chose the first three years of our data (2009-2011) – as it's the period before the festival and filming of GoT in Croatia. The second period includes years between 2013-2015 as the period during the festival/filming.

To achieve those tasks for Split, we again used the polygon which corresponds to city borders on land, adding a small buffer zone on the sea. We separately inserted photos from these two periods. We then selected only the photos within the polygon borders. Next, we extracted the photos from the area where the Ultra festival was held. In this case, it is Poljud Stadium – we made a buffer around it, selecting and exporting only photos within it (Figure 23). The buffer size was self-chosen to include photos that gravitate to the Stadium, as we tried to leave it to be the only attraction of the buffer. Then, we simply export *.csv* files of both periods and areas. The following steps, calculation of the change in relative proportions, and the analysis of tags and titles is explained later in this section.

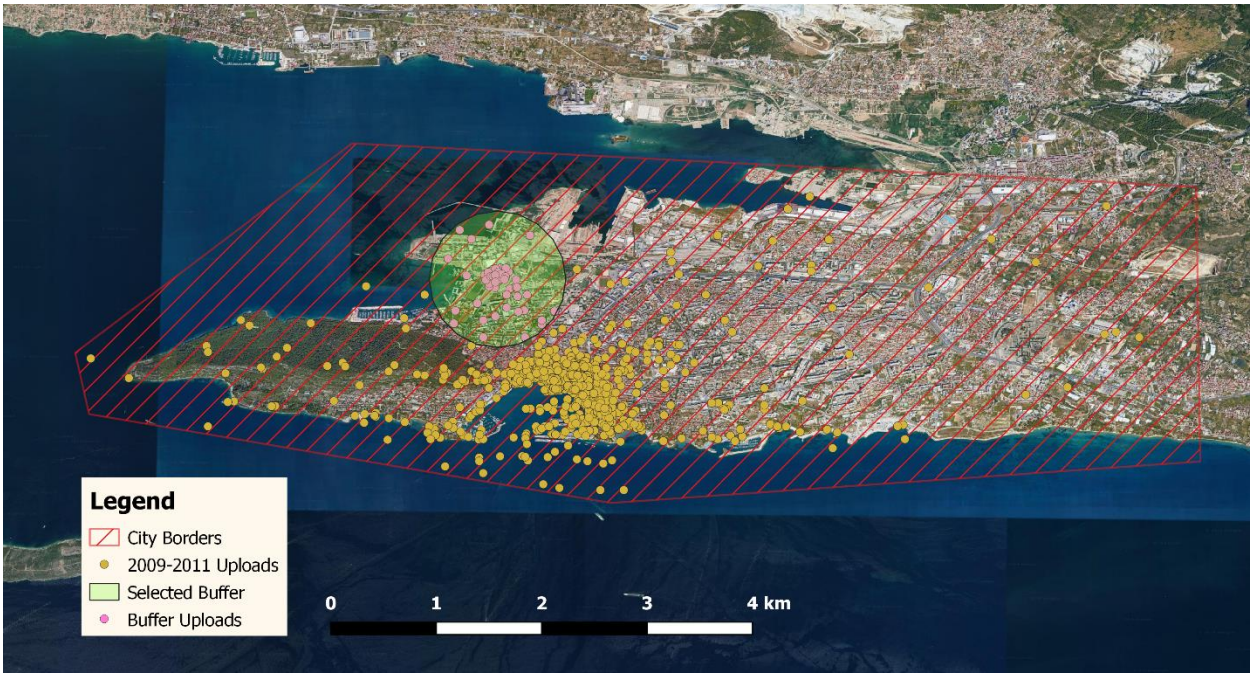


Figure 23 - Split with its borders ; green buffer is around Poljud Stadium

When it comes to Dubrovnik, it has been already shown in the work by Tkalec et. al, 2017 that the filming of the GoT series has affected tourism figures of the place. Moreover, numerous online sources and articles about Dubrovnik, almost without exception, mention the series when presenting the town. Therefore, it is safe to say that its public image has changed, especially among younger travelers, as it went from being a primarily historical, natural, and cultural site to be recognizable for the GoT series.

There are many locations within the Dubrovnik area that were used for filming, most being in the Old town (as presented in Figure 24). The other locations include Arboretum Trsteno, Hotel Beldevere, and the island of Lokrum. We used various Internet sources to make the locations list as *.csv* file with the coordinates, which we uploaded and mapped in QGIS.

Since a majority of tourists and their activity are concentrated within the Old town, which areawise represents a rather small portion of the destination, we approached somewhat differently. In addition to extracting the uploads from the buffer which includes the Old Town, we selected two locations, namely Dubrovnik Port and Srdj Hill, and made buffers around their central point (Figure 25). They are selected because of relatively high upload count and for being attractions independent of the Old Town. The method of the creation of buffers and the selection of photos within them was the same as for Split.

Figure 24 - Filming locations of Game of Thrones series. Some locations are left from the map because they are outside of the Old town (e.g. Arboretum Trsteno, Lokrum Island)

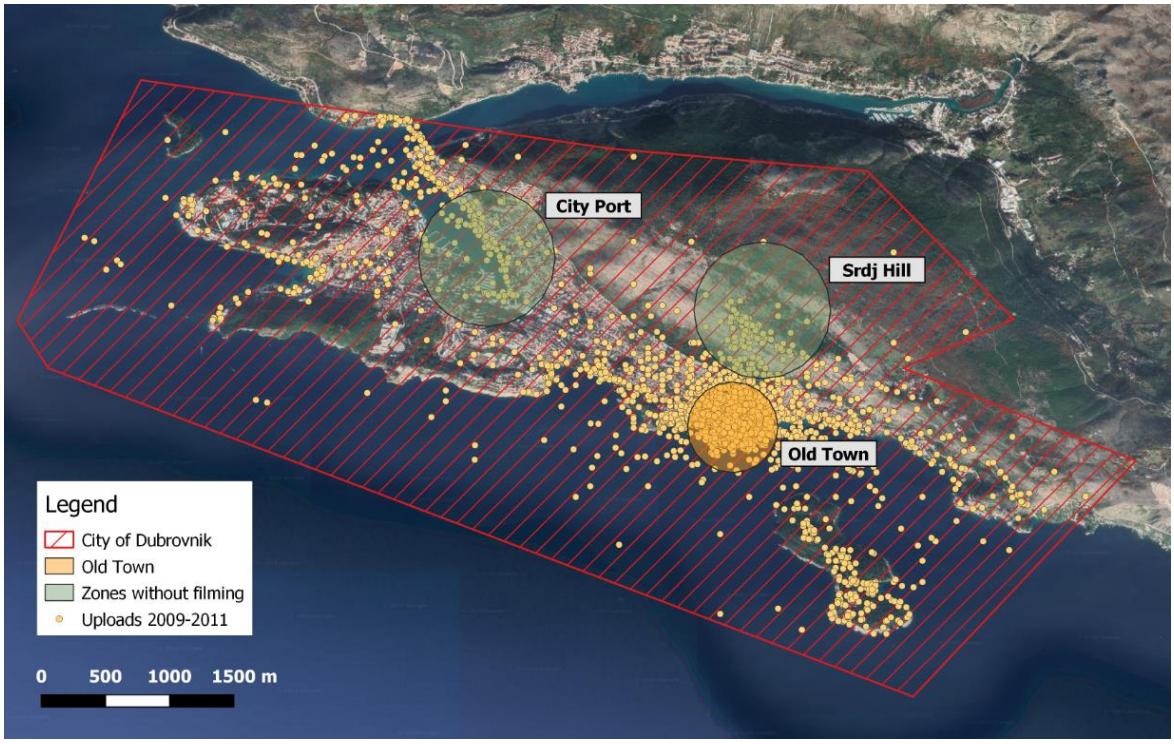


Figure 25 - Dubrovnik in its city borders, with a short zone to include sea surface. Green buffers present popular destinations where there was no filming, while orange buffer includes the Old Town and small area around it

4.7.1. Relative proportions and tag analysis

To compare if the relative proportion of photos uploaded within the festival/filming zones changed and thus to show if the selected area is more relevant for the photographers and visitors, we use and compare the counts from entire polygon and from selected buffers with the following formula:

$$r\ (buffers, selected\ period) = \frac{\text{n (selected photos)}}{\text{n (all photos)}}\ x\ 100$$

We also calculate the most commonly used words for titles and tags, including relative and absolute counts. We have this both for buffers/city borders, as well as for both periods. It is expected that the ratio of the terms related to the festival/series should be significant in the second period (2013-2015), as well as that it will be higher within the buffers where festival/filming happened.

To do this, we exported our *.csv* files with the data consisting of only titles and tags. Using Rstudio, we found and excluded all unnecessary symbols, numbers, or words (such as *the, in, of*), replacing them with empty space. This was then exported as .txt file, which was then analyzed in Python. We searched for 50 most common terms. A large number was, as expected, related to the location (city, county or country name). Additionally, words such as „Croatia“ were written in various languages. Finally, we present 20 most common tags/terms.

Chapter 5 –

Results

This chapter will present all the results we gained from the data using the methods we presented. The results are given in the same order we explained methods in the previous chapter. We first present the data facts (photo contributors and total uploads counts). This will follow the presentation on POI. Next, trajectories derivated from user uploads will be presented, as well as frequency patterns between them. We will continue and conclude with tags analysis and tourists' activity over time results.

5.1. Data Summary

After we finished all data preprocessing steps, we present here upload statistics, year by year. Table 11 shows the following data: the number of Flickr users who uploaded any number of photos within the study area (Total users), the number of users which will be included in further research (Selected users). Then, we for the context purpose, there is the number of photos within the Bounding box (*Bounding b.*), of those within Dalmatia (*Dalmatia*), users with 5 -150 uploads (*Limited upload*), and finally, user-count which exclude locals (*Selected users' u.*), which are our targeted Flickr users. Finally, we calculated the average upload figures of selected users (*Avg*).

Table 11 - Data Statistics

Year	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Total users	1526	2007	1998	1781	1463	1252	1092	958	868	616
Selected users	462	581	591	538	503	476	379	335	329	241
Bounding b.	29283	33419	43876	36317	32450	35065	29147	24371	23681	24260
Dalmatia	24432	27412	34696	29301	26881	29230	24035	20402	20220	19854
Limited upload	14856	18558	19124	17223	16162	13317	10963	10963	11381	8745
Selected users' u.	12333	15301	16272	13729	14129	11253	9593	9381	9684	7655
Avg	26.69	26.34	27.53	25.52	28.09	23.64	25.31	28.00	29.43	31.76

Furthermore, we present graphs of the total number of photos and the total number of users over the years. The graphs demonstrate removed users (those which upload count is under 5 and over 150), but also separate users which are taken into the research into two groups, namely, those who uploaded 5-30 photos (*Low contributors*) and to those who uploaded 31-150 photos (*High contributors*). The classes are self-selected to present how even in selected data there is still some user bias left, as for example 27% of users of the selected pool which are marked as "high contributors" contributed almost 75% of the total number of photos. However, this still lowers the effect of the "90-9-1" rule discussed in the 2.7. section. Particulary, in some years a fraction of users (around 2%) uploaded as many as 60% of the total photos count (see year 2011 in Figure 26).

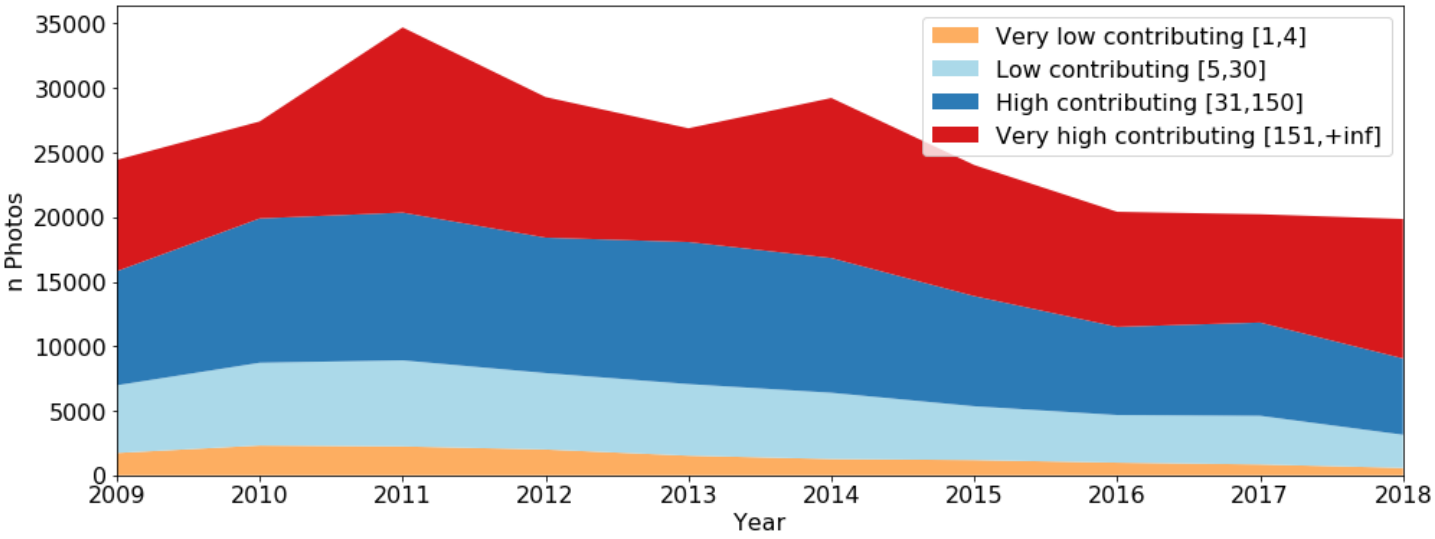


Figure 26 - Number of photos uploaded year by year. We split observed users (5-150 uploads) into two categories

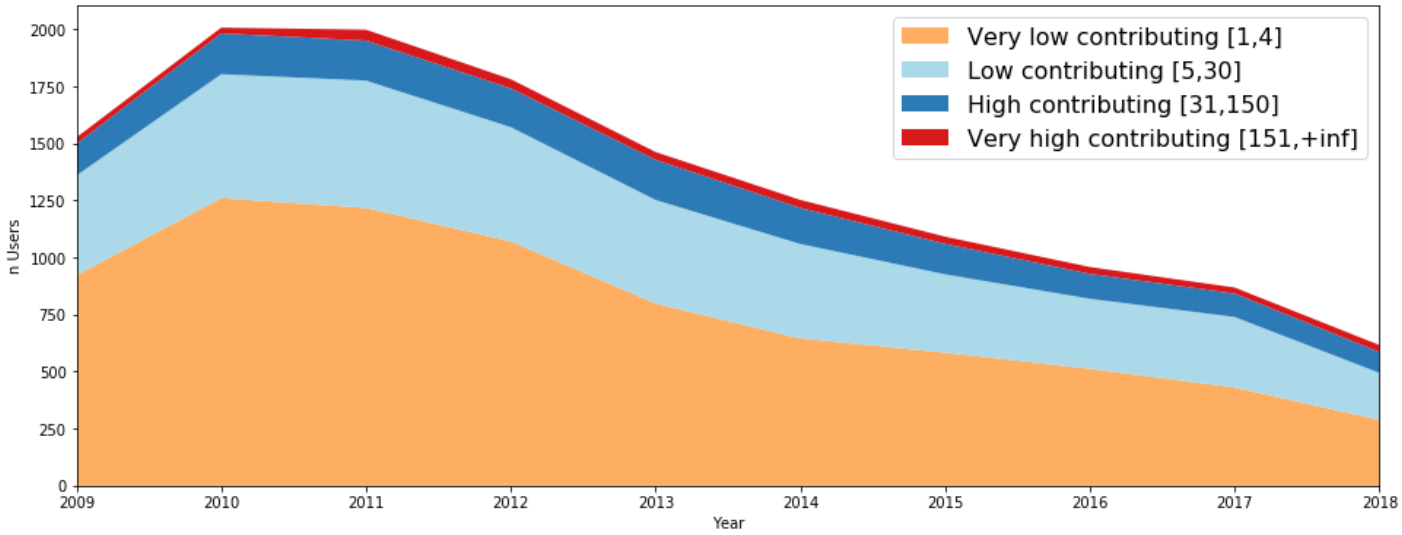


Figure 27 - Number of users which uploaded any number of photos

As discussed, there is a clear drop in the number of users which contributed any amount of photos, selected users, and the number of uploaded photos. The drop in user count started in 2011 and had a sharp drop between 2017 and 2018 (28% drop). However, the average number of uploaded photos grew almost continuously, so the number of photos for the analysis did not drop as sharp. Flickr statistics of uploads, presented in the section 3.3., do not correspond to these numbers entirely, as there the fall in upload count is less obvious. It can be assumed that many travelers migrated to other social media sites, such as Instagram.

Figure 28 also presents how the data was affected by the outliers. We used the data from 2009 which shows that some areas had many photos almost exclusively contributed by the same user. If not removed, this might suggest the conclusion that, in this case, areas of Kastela (an agglomeration of seven coastline villages) or Supetar, a small town on the island of Brac, are among most visited places of the region.

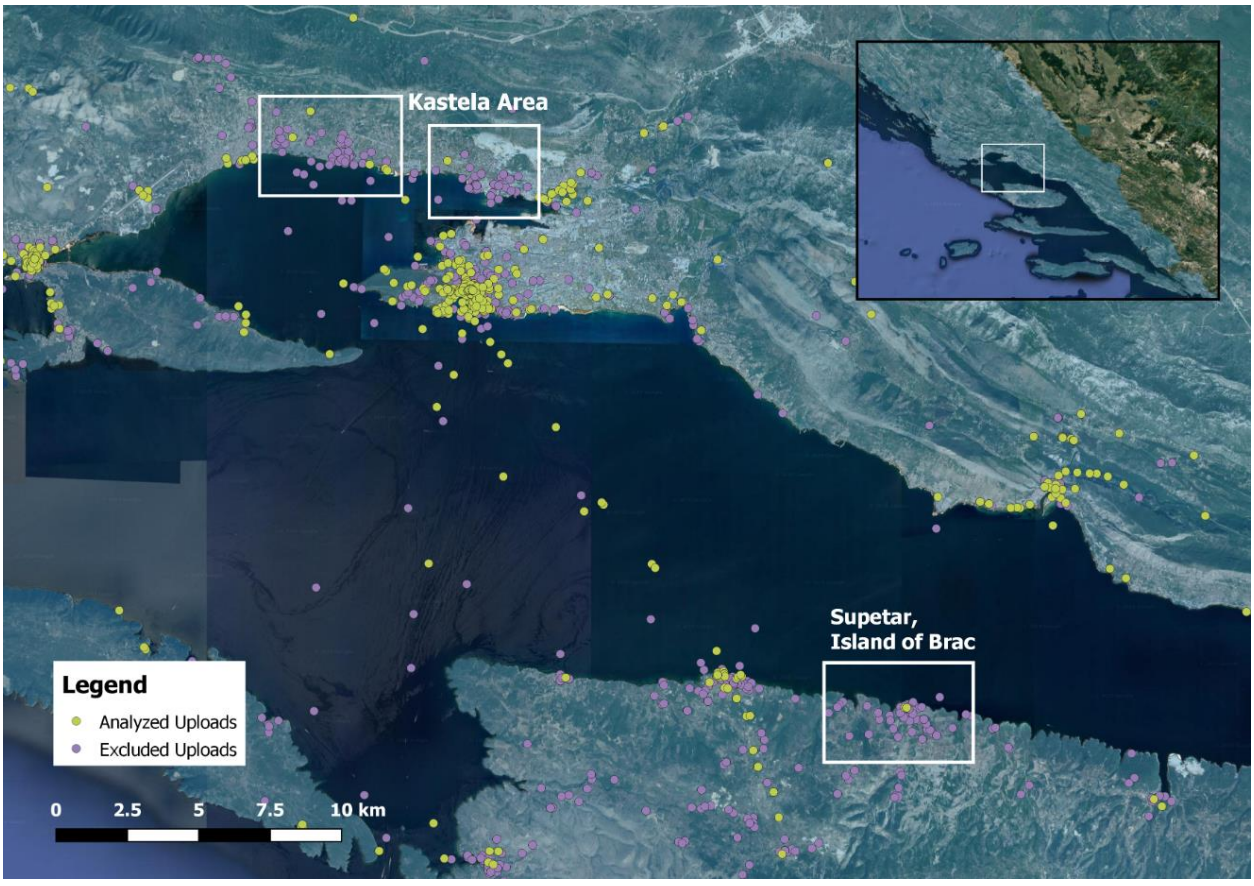


Figure 28 - Area around Split shows how data, without pre-processing, could lead to biased results when analyzed. The purple dots are removed by removing the users which uploaded over 150 photos. Data from 2009

5.2. Points of interest

Firstly, we are presenting the heat map of the entire region. We visualized two techniques explained in the Methods section, namely Kernel Density Estimation with parameters of 1.5 km for radius and 2500 for row number, and Heatmap type of visualization in QGIS, which is mostly noticeable around Dubrovnik and Split. These methods did not show as very effective for the discovery of Points of interest since the area is rather polarized when it comes to user's uploads. This means that there are two major areas of uploads, Dubrovnik and Split (having almost 60% of all photos of the region), and the rest is much less covered by photos. This is visualized in Figure 29.

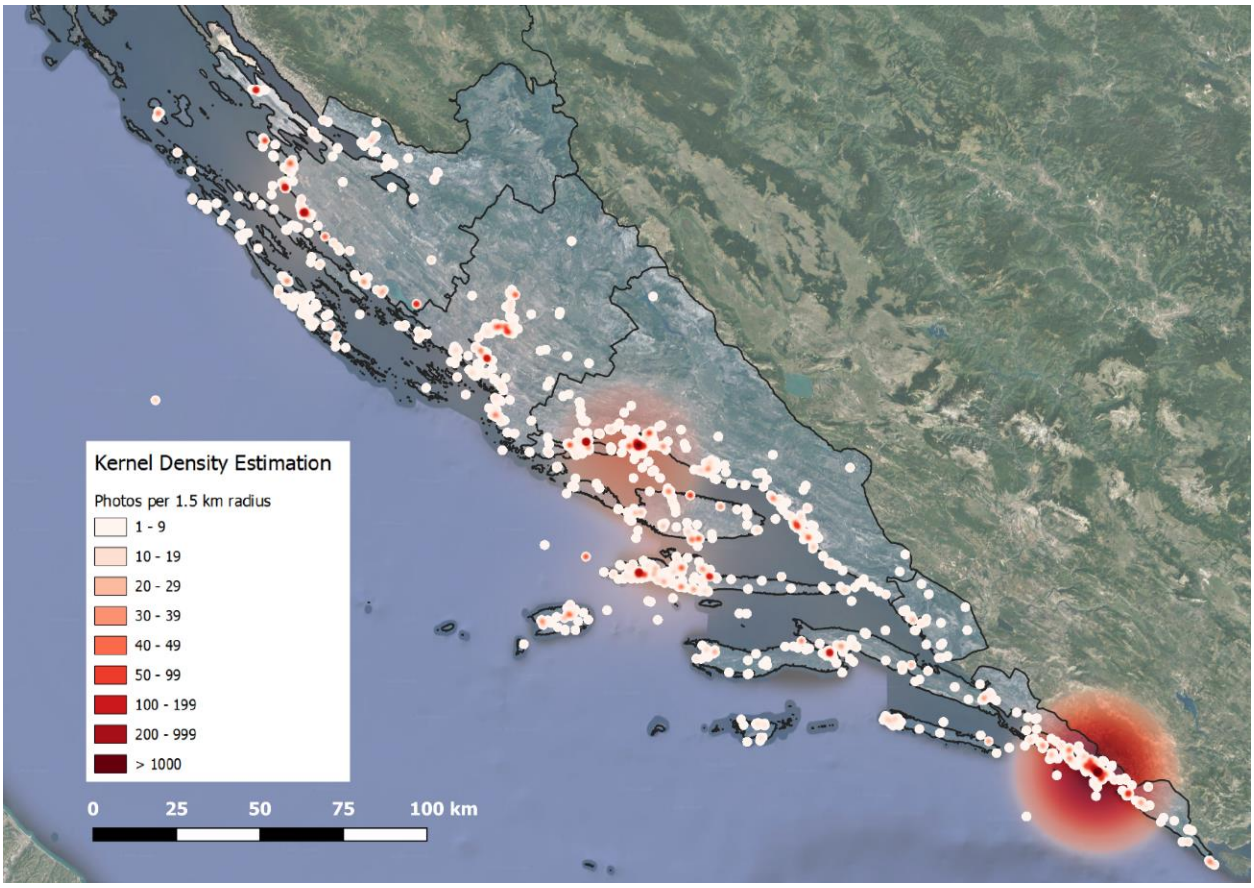


Figure 29 – Kernel density estimation and Heatmap visualization combined; based on uploads from 2009

As we wanted to extract main POIs within the region, we used DBSCAN method. With our parameters set as Eps = 45 meters and MinPts = 40, we extracted 17 to 28 clusters a year, with on average 1300 photos being considered as „noise“ and not joined to any cluster. We extracted the top 10 clusters for each year. We present the output of the code in Figure 30 and in the Appendix. The top ten clusters for each of the year is presented and compared in Table 12.

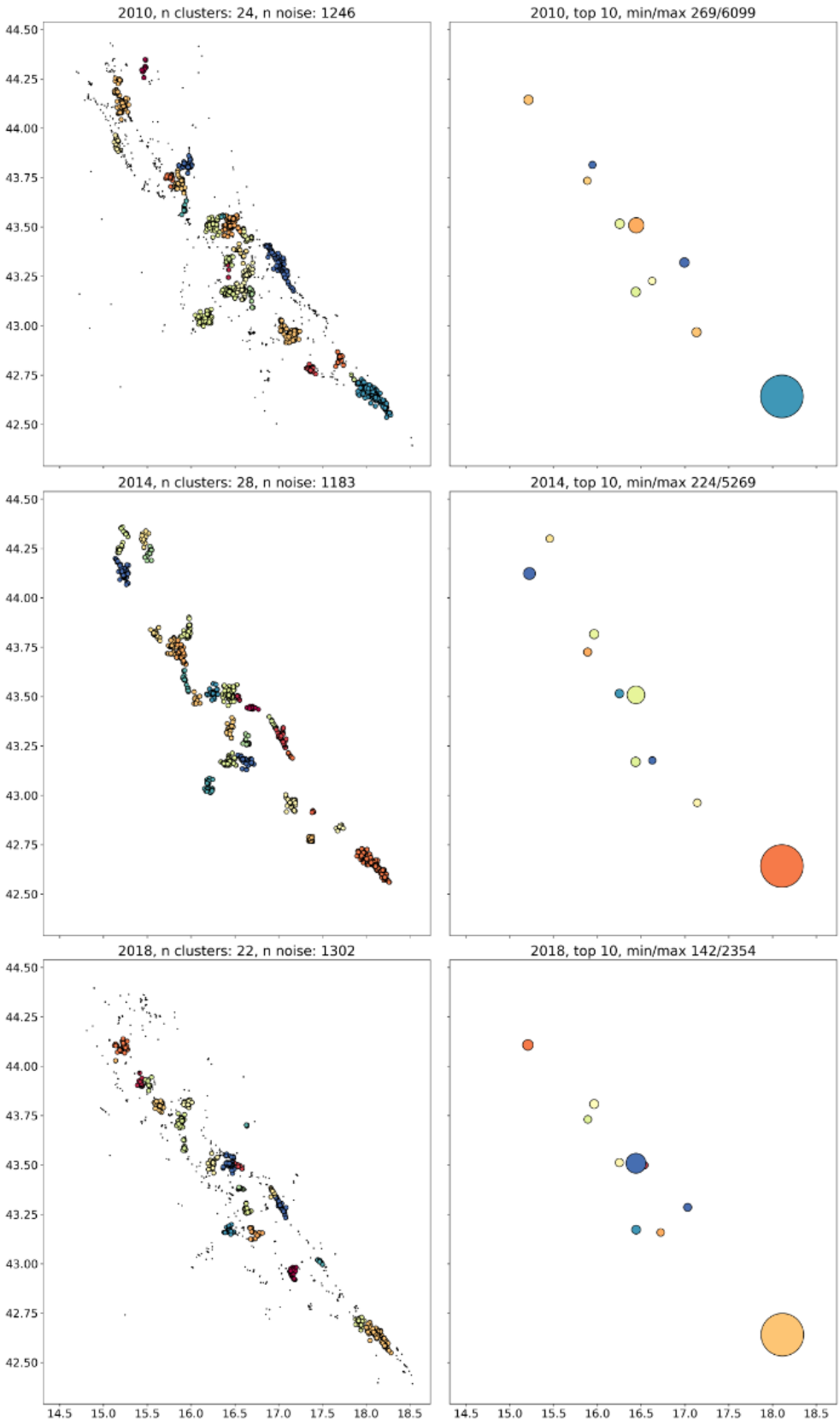


Figure 30 - Clusters with the noise (left) and top 10 of the year (right)

Table 12 - Ranking of the destinations, according to our data

Rank	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
1	Dubrovnik	Dubrovnik	Dubrovnik	Dubrovnik	Dubrovnik	Dubrovnik	Dubrovnik	Dubrovnik	Dubrovnik	Dubrovnik
2	Split	Split	Split	Split	Split	Split	Split	Split	Split	Split
3	Hvar	Makarska	Zadar	Zadar	Zadar	Zadar	Krka	Hvar	Zadar	Zadar
4	Sibenik/Krka	Zadar	Trogir	Sibenik	Hvar	Hvar	Trogir	Trogir	Trogir	Krka
5	Trogir	Trogir	Hvar	Trogir	Orebic/Kor.	Krka	Zadar	Zadar	Krka	Hvar
6	Zadar	Hvar	Krka	Orebic/Kor.	Trogir	Trogir	Sibenik	Krka	Hvar	Trogir
7	Nin	Orebic/Kor.	Orebic/Kor.	Krka	Krka	Sibenik	Hvar	Paklenica	Orebic/Kor.	Makarska
8	Orebic/Kor.	Hvar	Makarska	Hvar	Makarska	Orebic/Kor.	Hvar*	Orebic/Kor.	Sibenik	Sibenik
9	Makarska	Sibenik	Hvar*	Paklenica	Sibenik	Paklenica	Orebic/Kor.	Makarska	Ston	Hvar*
10	Pag	Krka	Sibenik	Hvar*	Paklenica	Hvar	Makarska	Sibenik	Makarska	Stobrec

Table 13 - Most visited destinations according to the official data (tourists, not overnight stays)

Rank	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
1	**Dubrovnik**	**Dubrovnik**	**Dubrovnik**	**Dubrovnik**	**Dubrovnik**	**Dubrovnik**	**Dubrovnik**	**Dubrovnik**	**Dubrovnik**	**Dubrovnik**
2	**Zadar**	**Zadar**	**Zadar**	**Zadar**	**Zadar**	**Split**	**Split**	**Split**	**Split**	**Split**
3	**Sibenik**	**Split**	**Split**	**Split**	**Split**	**Zadar**	**Zadar**	**Zadar**	**Zadar**	**Zadar**
4	**Split**	**Sibenik**	**Sibenik**	**Sibenik**	**Sibenik**	**Sibenik**	**Sibenik**	**Sibenik**	**Sibenik**	**Sibenik**
5	Vodice	Vodice	Vodice	Vodice	Vodice	**Orebic/Kor.**	**Makarska**	**Makarska**	**Makarska**	**Makarska**
6	Biograd	Podgora	**Makarska**	Makarska	**Makarska**	Makarska	**Orebic/Kor.**	Vodice	Vodice	Vodice
7	**Orebic/Kor.**	**Orebic/Kor.**	Biograd	Biograd	Biograd	Vodice	Vodice	**Hvar**	**Hvar**	**Hvar**
8	**Hvar**	**Makarska**	**Orebic/Kor.**	**Hvar**	**Hvar**	Biograd	**Hvar**	**Orebic/Kor.**	**Orebic/Kor.**	Biograd
9	Podgora	Biograd	**Hvar**	**Orebic/Kor.**	Podgora	**Hvar**	Baska Voda	Biograd	Biograd	Orebic/Kor.
10	Seget	**Hvar**	Gradac	Podgora	**Orebic/Kor.**	Gradac	Biograd	**Seget***	**Seget***	**Trogir**

**Seget is located right next to Trogir*

It can be noticed that there are matching between two tables (green marked places in Table 13). As expected, Dubrovnik is both by visits and by our data in the first place every year. Split only became the second most popular destination by visits in 2014, while it always had the second place within our data. Zadar, Hvar, and Sibenik were also correctly detected in all of the years. On the other hand, Biograd or Vodice were among the top most visited places in many years, yet were not detected by our data as the top 10 most photographed places. The main reason might be due to the fact that those destinations are more oriented to mass tourism, rather by being visited because of historical or natural attractions. Another reason could be that photos of those places can be joined to clusters of nearby, more popular places. We gave an example of Orebic and Korcula, and similar could go for Sibenik and Vodice, or Makarska and Podgora and Baska Voda. So, to some extent, our method detected even more places than the Table suggests.

National parks are not covered with the statistics the same way as towns and municipalities are. However, the visit counts are still accessible online. By far, the most visited National park in the region is NP Krka, with over 1.45 million visitors in 2018, which is almost double from 750 thousand in 2008. The growth in popularity can be noticed in our data, as the relative amounts of photos for Krka increased. Better overview of Krka's increase in visit count is presented by trajectories in the next section. NP Paklenica, Kornati, and Mljet have a lower growth of figures, so Paklenica being among the top 10 in some years is more likely due to more uploads of photos of a small number of users, than a high upgrade in the popularity.

There are no significant changes over time when it comes to our data and the statistics from the official sources. Our approach correctly detected 6-8 most visited places in each year.

5.3. Trajectories

5.3.1. General statistics

In the following tables and graphs, we present the statistics calculated from our data and related to users' trajectories. Namely, we calculated the total and average length of trajectories, as well as the total and average number of days spent within the region. Firstly, we wanted to check if the users are uploading their photos from the photo collection with different locations. We found out that there is a clear drop of users who upload all of their photos with the exact same location, as the figures went from around 17% in 2009 to less than 3% in 2018 (Table 14).

We calculated the total and average length of trajectories. There is no significant drop in the average length. To have a somewhat better image of how much average tourist travels, we, as explained in the Method section, added 25% to the numbers extracted from our data, as that possibly better represents driving distances compared to straight-line distances. According to our data, the average distance in the period would be 205 kilometers within the destination (see green row in Table 14). We also split the data on trajectory length to compare the differences between seasons. As we explained, we adjusted the seasons to the following dates: Winter – 01.12. to 28/29.02., Spring 01.03. to 31.05., Summer 01.06. to 31.08., and Autumn 01.09. to 30.10. This also means that a part of the trajectories was split between the seasons. The biggest oscillations are, as expected, during the winter period. This is because of a very small number of uploads suitable for our work. On the other hand, summer had the most consistent length, as the total length in the summer months is the highest in all years. If we can make any conclusions from the data, it would seem that during the winter the trips visitors are making are the shortest (except in 2011, again due to the data quality), while there is a trend of trips during autumn and spring months being longer during this decade. This is also presented by averages (Figure 31) and boxplots (Figures 33-36).

Additionally, we present the statistics on the time users spent in the destination, year by year (Table 15). We excluded the users with timestamp under one day, as they are not counted as tourists by the definitions from the officials[43]. We can notice the calculations from our data in terms of average stays come very close to the official statistics, as the biggest difference is 13%, while on average, correctness is close to 4.6% (see green marked rows). This might suggest that our other results which we cannot be tested by the official data are also correct to the satisfying level. We also calculated the average stay per season (Figure 32). Again, we the dates of seasons to Northern Meteorological Seasons. As expected, summer months are having longer stays. On another hand, winter months have a much lower number of days stayed, except for in some years, but this is again due to a low amount of data.

[43] According to the definition, a person is a tourist only if she or he spents at least one night within the destination

Table 14 – Length results from trajectories

	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Total number of tourists	462	581	591	538	503	476	379	335	329	241
Of those, > 0 km	380	500	511	475	430	443	351	313	312	235
(%)	82.3	86.1	86.4	88.3	85.5	93.0	92.7	93.4	94.8	97.6
Total length (km)	66,032	78,553	90,713	71,518	68,593	75,225	61,024	48,995	49,077	39,869
Average length (km)	173.7	157.0	177.7	150.5	159.5	169.9	173.7	156.6	157.4	169.5
Adjusted av. length (km)*	217.1	196.3	222.1	188.2	199.4	212.4	217.1	195.7	196.7	211.9

**As explained in the Method section*

Table 15 - Duration of stay, year by year

	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Total number of tourists	462	581	591	538	503	476	379	335	329	241
Timestamp > 1 hour	458	531	550	510	467	453	363	319	312	233
Timestamp > 1 day	356	433	455	417	382	376	311	274	254	204
% of the total users	77.06	74.53	76.99	77.51	75.94	78.99	82.06	81.79	77.20	84.65
Total duration (days)	2070	2605	2530	2297	2051	1963	1835	1475	1366	1171
Average duration (days)	5.81	6.02	5.56	5.51	5.37	5.22	5.90	5.38	5.38	5.74
Av. Stay (official data)	5.72	6.40	5.67	5.68	5.55	5.42	5.35	5.38	5.21	5.06
Comparison (per. diff.)*	1.57	5.94	1.94	2.99	3.24	3.69	10.28	0.00	3.26	13.44

**the percentage difference between two figures in order to determine how close they are, relative to the larger value*

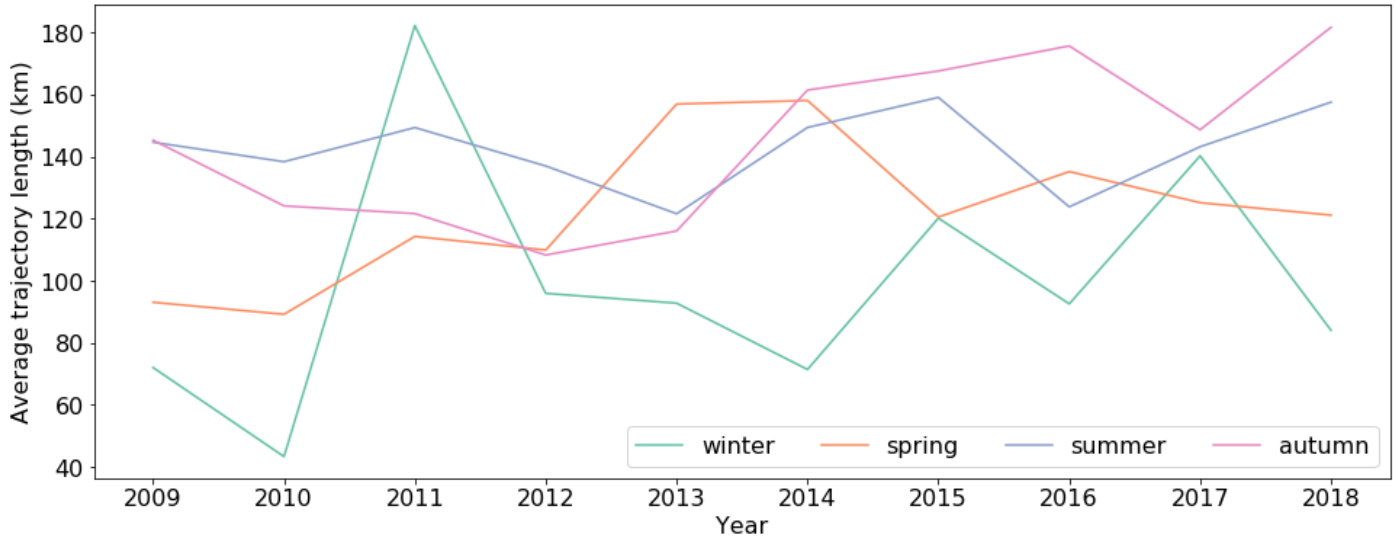


Figure 31 - Trajectory trajectory length per season

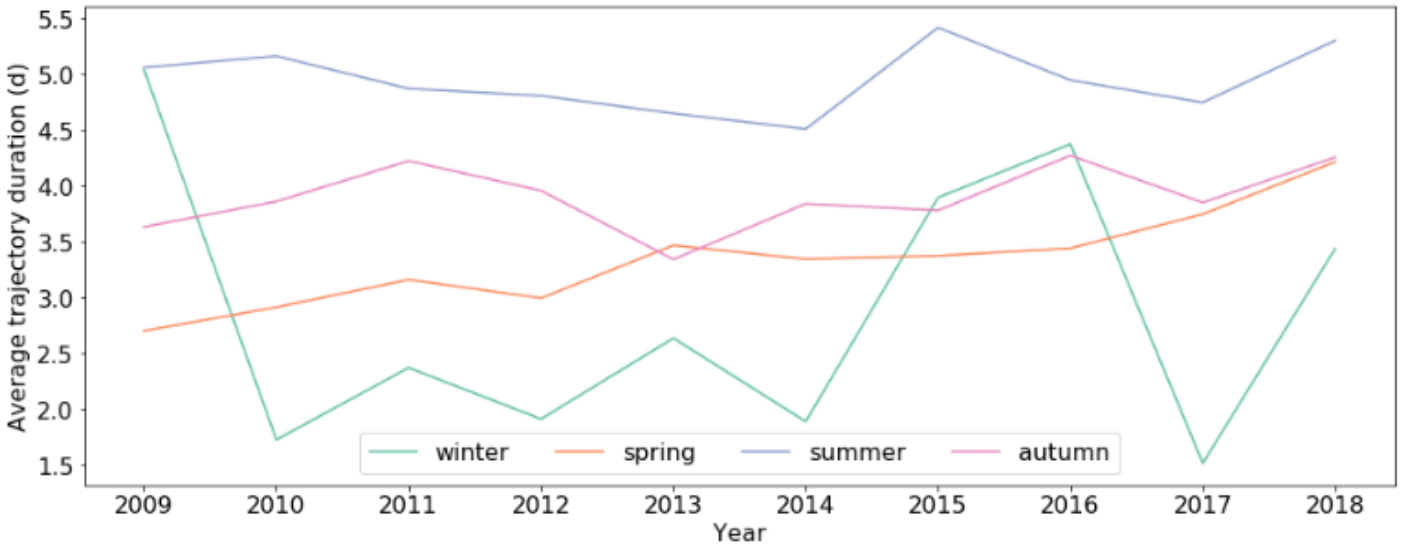


Figure 32 - Average number of days spent per visitor for each season

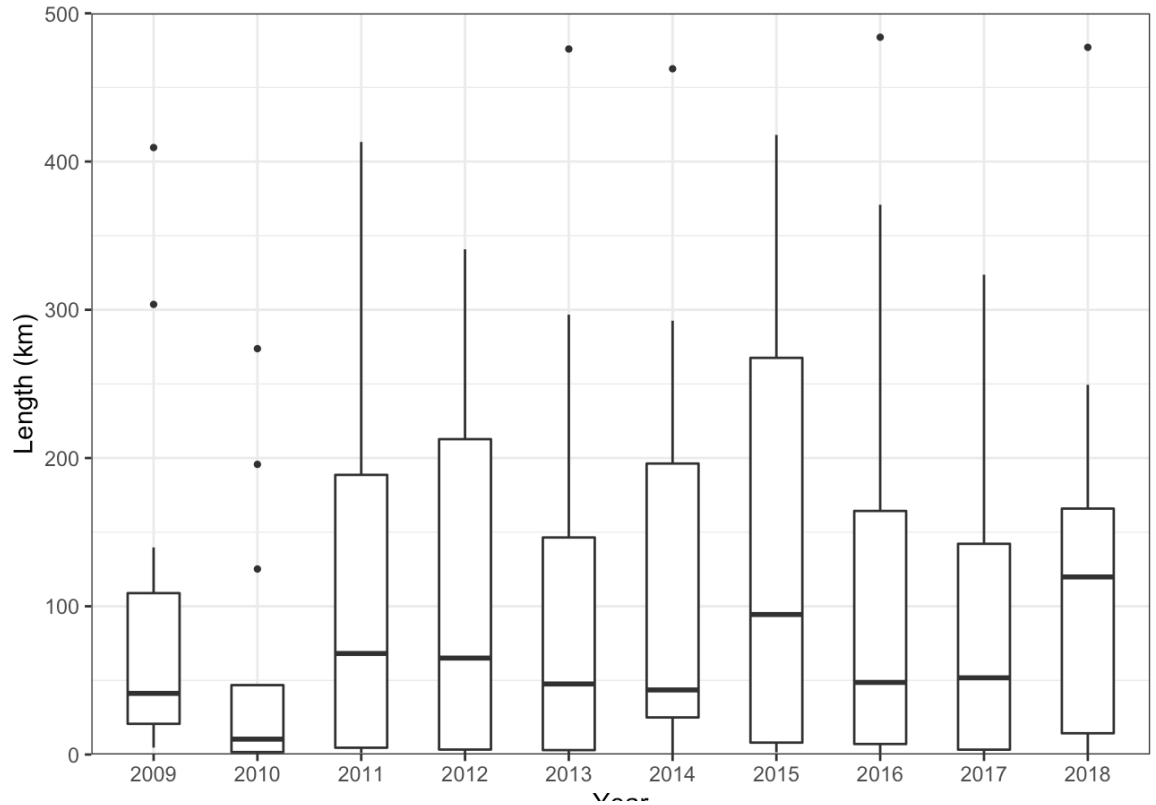


Figure 33 - Winter trajectories length year by year

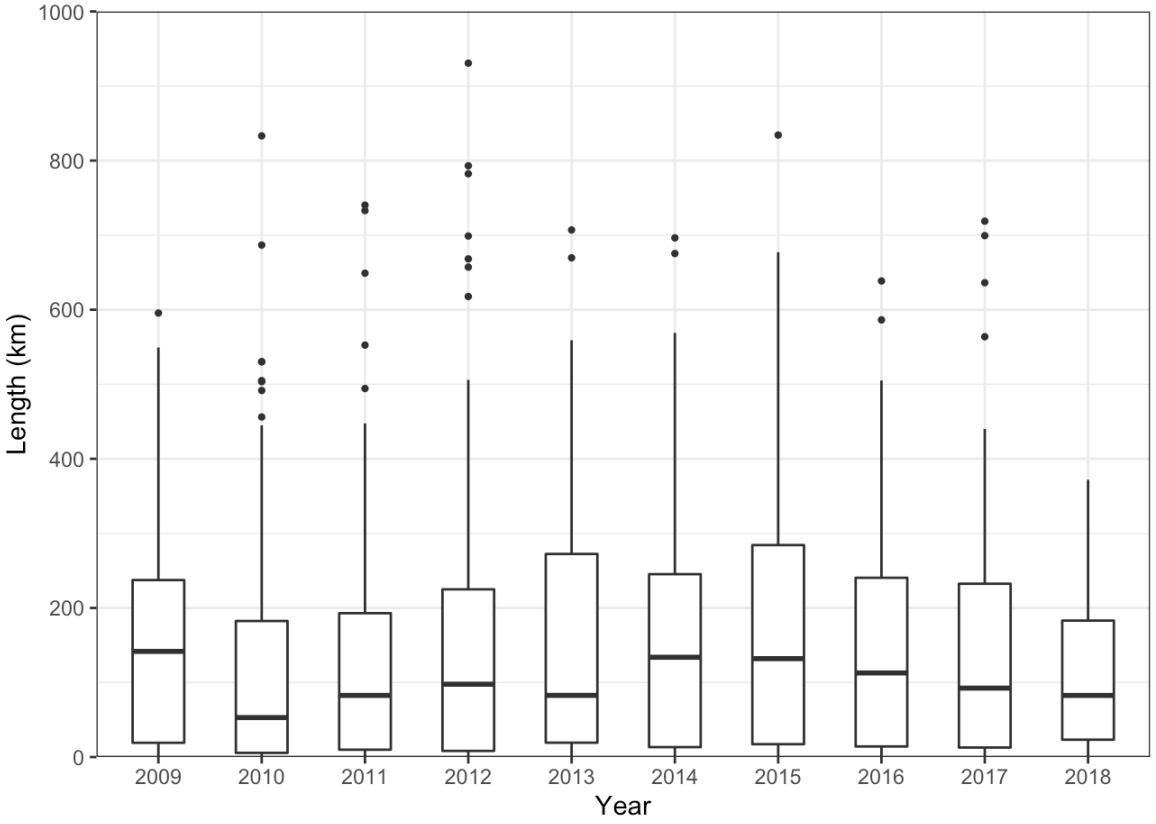


Figure 34 - Spring trajectories length year by year

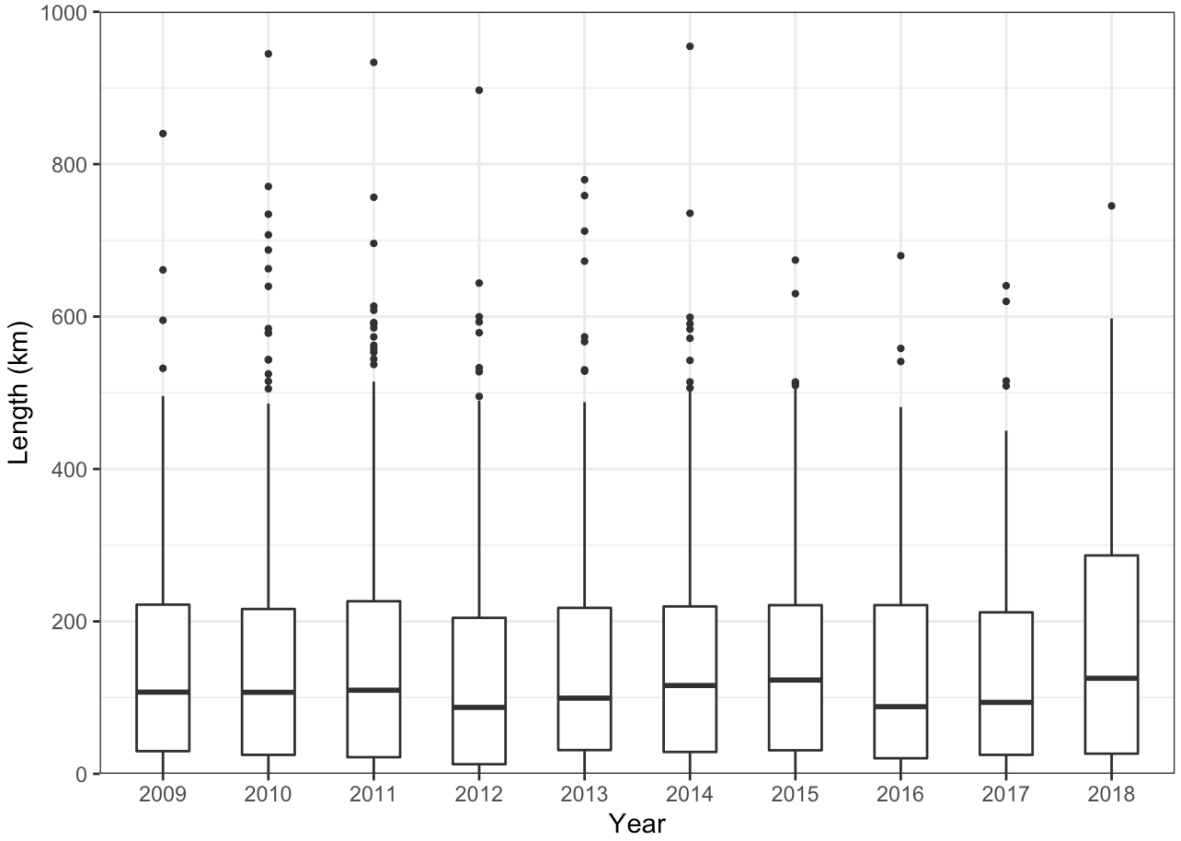


Figure 35 - Summer trajectories length year by year

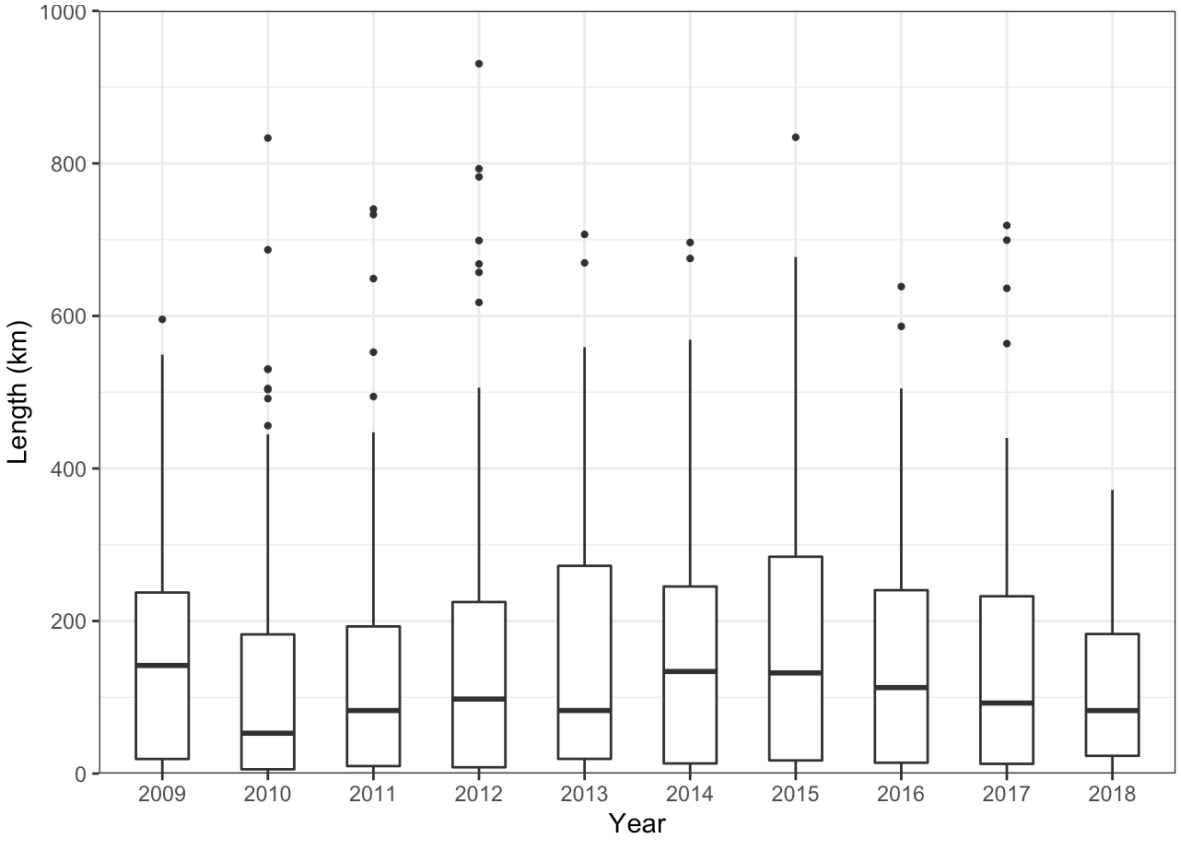


Figure 36 - Autumn trajectories length year by year

5.3.2. Frequency patterns between destinations

As we selected 13 destinations within the region, including Split and Dubrovnik, we wanted to see the frequency of visits between them.

First, we present here visit patterns between Split and Dubrovnik and other destinations (Figure 37). The lines represent relative values or percentages of people that visited both Split/Dubrovnik and other destinations over the years. Tables that follow after are representing also relative figures between all destinations. The blue fields are absolute counts of visits. On the top right field of the tables, the total number of tourists observed is also given (*n* count). Next to each destination, the percentage of visits is displayed. The tables present figures for two continuous years.

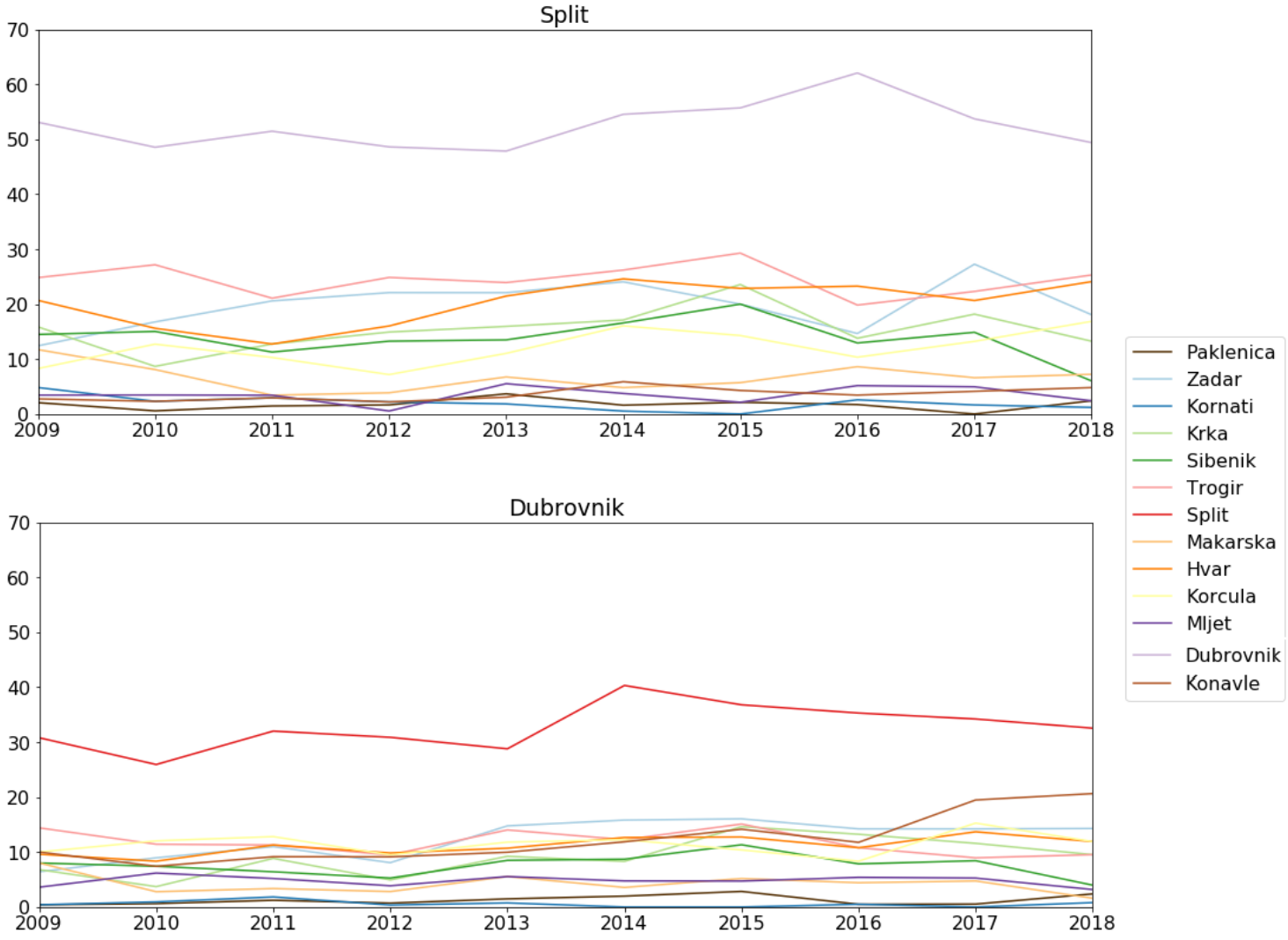


Figure 37 - Visits between Split and Dubrovnik and other destinations

Table 16 - Relative relations between destinations, 2009 and 2010

n = 1074	PAK	ZAD	KOR	KRK	SIB	TRO	SPL	MAK	HVA	KORC	MLJE	DUB	KON
PAK (1.8)	**19**	4	9	6	1	3	1	2	2	0	3	1	2
ZAD (12.1)	26	**130**	32	21	31	21	15	9	11	9	12	8	3
KOR (4.4)	21	12	**47**	12	11	6	3	0	1	4	0	1	2
KRK (7.5)	26	12	19	**77**	33	16	12	14	2	3	0	5	3
SIB (8.6)	5	21	21	38	**88**	27	15	5	7	9	3	8	8
TRO (13.1)	21	23	17	29	43	**141**	26	23	20	17	9	13	8
SPL (29.6)	21	36	23	49	53	59	**318**	54	47	33	33	28	13
MAK (5.3)	5	4	0	10	3	9	10	**57**	9	10	9	5	3
HVA (11.4)	11	11	2	3	10	17	18	19	**122**	29	30	9	5
KORC (9.9)	0	7	9	4	10	12	11	18	25	**102**	42	11	3
MLJE (3.1)	5	3	0	0	1	2	3	5	8	14	**33**	5	3
DUB (53.4)	16	35	9	38	50	52	51	51	42	63	88	**574**	82
KON (5.6)	5	2	2	3	6	4	3	4	2	2	6	9	**60**

Table 17 - Relative relations between destinations, 2011 and 2012

n = 1173	PAK	ZAD	KOR	KRK	SIB	TRO	SPL	MAK	HVA	KORC	MLJE	DUB	KON
PAK (2.4)	**28**	3	4	5	6	2	2	7	0	2	0	1	3
ZAD (15)	21	**176**	49	29	40	25	21	10	15	11	9	10	5
KOR (4)	7	13	**47**	10	7	6	3	0	3	0	3	1	2
KRK (10.7)	21	20	28	**125**	38	24	14	24	13	7	9	7	2
SIB (7.5)	18	20	13	27	**89**	22	12	5	7	5	3	6	3
TRO (12.1)	11	20	19	27	35	**142**	23	21	18	14	14	10	10
SPL (32.8)	21	47	21	42	53	62	**385**	33	47	36	23	31	16
MAK (3.6)	11	2	0	8	2	6	4	**42**	3	3	6	3	0
HVA (10.1)	0	10	9	12	9	15	14	10	**118**	32	31	11	6
KORC (8)	7	6	0	6	6	9	9	7	25	**94**	49	11	22
MLJE (3)	0	2	2	2	1	4	2	5	9	18	**35**	5	10
DUB (52.3)	21	34	15	34	40	45	50	45	55	73	80	**613**	89
KON (5.4)	7	2	2	1	2	4	3	0	3	15	17	9	**63**

Table 18 - Relative relations between destinations, 2013 and 2014

n = 1020	PAK	ZAD	KOR	KRK	SIB	TRO	SPL	MAK	HVA	KORC	MLJE	DUB	KON
PAK (2.6)	**27**	7	3	10	4	2	3	4	2	2	10	2	3
ZAD (18)	48	**183**	34	36	36	27	23	10	19	19	33	15	12
KOR (2.8)	4	5	**29**	10	9	3	1	2	2	1	3	0	2
KRK (12.3)	48	25	45	**125**	43	20	17	23	15	13	17	9	6
SIB (9.5)	15	19	31	34	**97**	32	15	10	13	9	7	9	6
TRO (12.7)	11	19	14	21	42	**130**	25	23	23	23	30	13	3
SPL (34.3)	33	44	14	46	55	68	**350**	42	61	56	53	34	24
MAK (4.7)	7	3	3	9	5	8	6	**48**	5	8	20	5	5
HVA (13)	7	14	10	16	18	23	23	15	**132**	36	43	12	8
KORC (8.4)	7	9	3	9	8	15	14	15	23	**86**	43	12	9
MLJE (2.9)	11	5	3	4	2	7	5	12	10	15	**30**	5	3
DUB (51.3)	33	44	7	37	46	53	51	50	46	73	90	**524**	86
KON (6.5)	7	4	3	3	4	2	5	6	4	7	7	11	**66**

Table 19 - Relative relations between destinations, 2015 and 2016

n = 743	PAK	ZAD	KOR	KRK	SIB	TRO	SPL	MAK	HVA	KORC	MLJE	DUB	KON
PAK (2.4)	**18**	7	0	6	1	3	2	5	0	2	0	2	2
ZAD (15.6)	44	**116**	17	30	40	28	18	15	14	17	9	15	11
KOR (3.1)	0	3	**23**	3	5	2	1	0	0	4	0	0	0
KRK (16.3)	39	31	17	**121**	57	36	19	28	15	19	9	14	7
SIB (11.3)	6	29	17	40	**84**	35	17	15	19	21	9	10	8
TRO (13.7)	17	25	9	31	43	**102**	25	15	30	25	13	13	11
SPL (34.5)	28	39	13	40	51	63	**256**	45	69	62	39	36	16
MAK (5.4)	11	5	0	9	7	6	7	**40**	7	4	4	5	3
HVA (11.6)	0	10	0	11	19	25	23	15	**86**	33	30	12	7
KORC (7)	6	8	9	8	13	13	12	5	20	**52**	57	9	7
MLJE (3.1)	0	2	0	2	2	3	4	2	8	25	**23**	5	7
DUB (56)	39	54	4	48	48	53	59	50	57	75	91	**416**	89
KON (8.2)	6	6	0	3	6	7	4	5	5	8	17	13	**61**

Table 20 - Relative relations between destinations, 2017 and 2018

n = 603	PAK	ZAD	KOR	KRK	SIB	TRO	SPL	MAK	HVA	KORC	MLJE	DUB	KON
PAK (2.5)	**15**	6	8	2	0	0	1	3	0	0	0	1	2
ZAD (16.7)	40	**101**	50	30	38	27	24	9	10	13	0	14	9
KOR (2)	7	6	**12**	2	2	3	1	0	1	3	0	0	0
KRK (14.3)	13	26	17	**86**	50	27	16	9	15	15	17	11	6
SIB (8.3)	0	19	8	29	**50**	26	11	6	13	10	11	7	3
TRO (12.8)	0	21	17	24	40	**77**	24	25	20	12	6	9	8
SPL (33.8)	13	48	25	38	46	62	**204**	44	63	50	44	34	14
MAK (5.3)	7	3	0	3	4	10	7	**32**	7	8	11	3	3
HVA (11.8)	0	7	8	13	18	18	22	16	**71**	37	28	13	5
KORC (10)	0	8	17	10	12	9	15	16	31	**60**	56	14	12
MLJE (3)	0	0	0	3	4	1	4	6	7	17	**18**	4	3
DUB (52.4)	27	45	8	40	42	38	52	34	58	73	78	**316**	97
KON (10.8)	7	6	0	5	4	6	4	6	4	13	11	20	**65**

In the five tables above, the relations between destinations are presented. As we mention, the numbers in tables are not *mirrored*. This means that to check the visits *from* a particular destination, the columns should be observed. For example, the data for Dubrovnik in the years 2017 and 2018 (Table 20) shows that there were 316 visits of the place, which is 52 % out of 603 visits of the region, concerning our data. Out of that figure for Dubrovnik, and in no particular direction, 34% also visited Split, 20% also visited Konavle, and so on. If we want to see how many visitors visited Dubrovnik from the total visit count of other places, we read the rows. So, according to the last table, 97% of visitors of Konavle also visited Dubrovnik, as well as 78% visitors of Mljet, 58% of Hvar, etc.

A few conclusions can be taken from the tables:

- The closer the destinations are, the more likely it is it will share the same visitors. This, naturally, does not come as a surprise, as it supports Tobler's First Law of Geography[44]. The most obvious examples within the region are, in addition to Dubrovnik and Konavle, Split and Trogir, or Hvar and Split. This is mostly due to daily trips tourists make by using agencies or traveling by themselves.
- Low-visited places do not offer enough data to observe trends. For example, Paklenica usually has only ca 2% of the total visits of the region, which in the case of our data means only 15-30 total visitors in each two-year period. Trajectories of such a low number do not provide meaningful figures to compare within the years.
- National Parks Mljet, Kornati, and Paklenica are not well-visited. This is partly because of them being placed off the busy routes and being island destinations (Kornati and Mljet). The exception is NP Krka. This is supported by the official figures, as Krka has over a million visitors per year, while other NPs have not more than 150 thousand. The fact that in this period Krka doubled number of visitors can be read from our data (Table 16: 7.5% of visitors of the area, Table 20: 14.3%).
- Aside of strong visits between bigger places, there are no noticed patterns when it comes to different types of places. In Figure 18 we classified places according to their significance and type (mayor places, smaller towns, nature sites). Probably because of low visits of primarily nature places, we cannot make conclusions out of it. However, this approach, presented by Arase et al. (2010), with better data could be used to classify tourists.

[44] "Everything is related to everything else, but near things are more related than distant things."

5.4. Figures of Users' temporal activity

We present here the visualization of the upload activity during the year. As explained in the Methods section, the visualization shows how the activity is distributed depending on months and hours during the day. Each *pixel* in the square represents the total amount of photos uploaded on that hour within that month, thus there are 24 pixels for each month. The pixelated square represents absolute, while the curve represents relative amounts, as we found this to better represent our data. We present graphs for every second year of the decade observed, while the rest can be found in the Appendix.

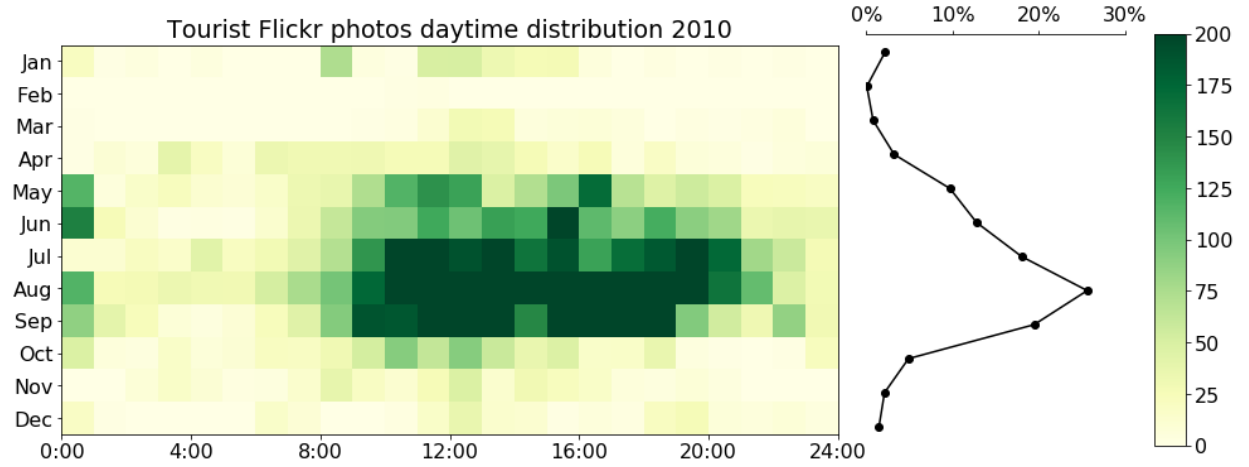


Figure 38 - Distribution of uploads, 2010

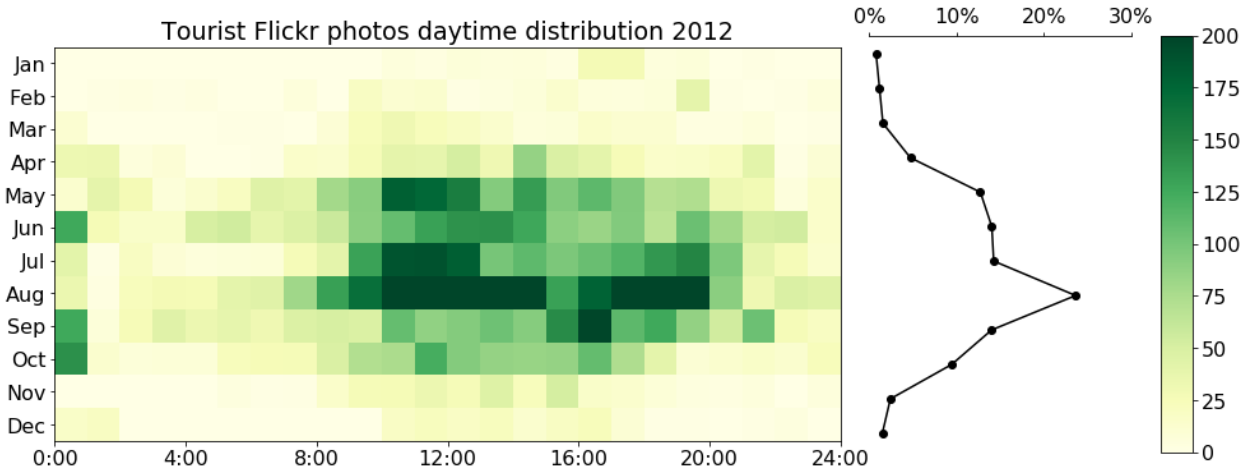


Figure 39 - Distribution of uploads, 2012

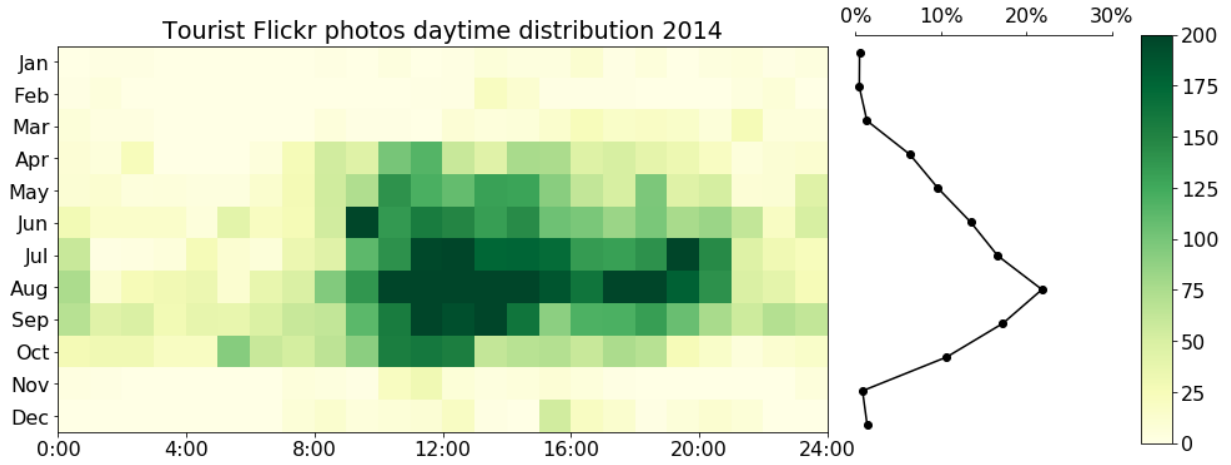


Figure 40 - Distribution of uploads, 2014

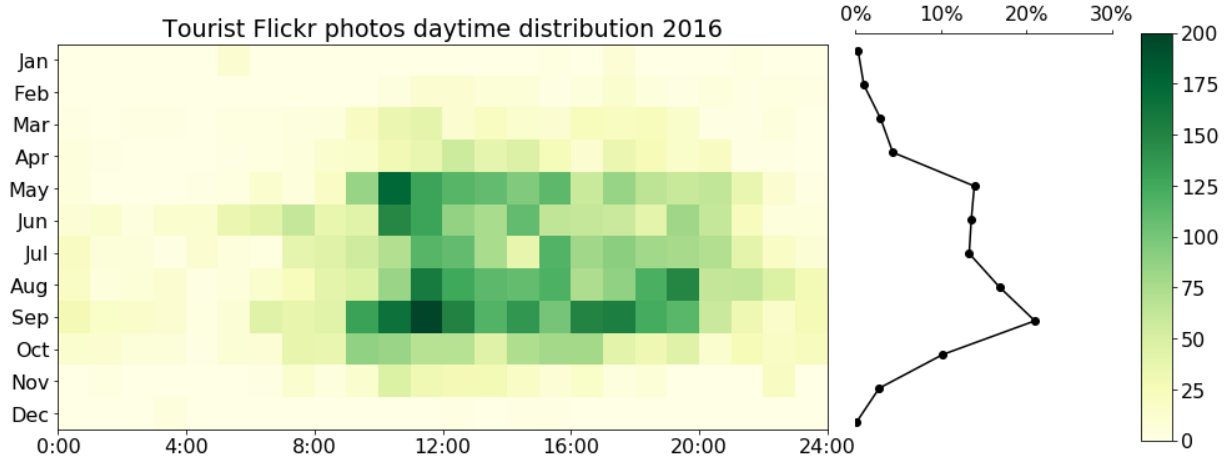


Figure 41 - Distribution of uploads, 2016

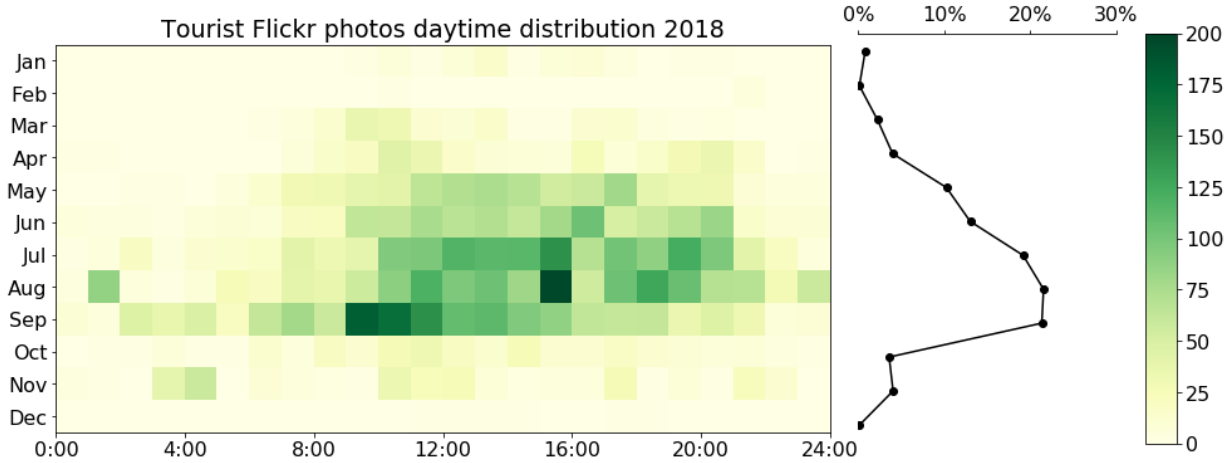


Figure 42 - Distribution of uploads, 2018

We did not detect unexpected patterns in the graphs. Users' activities are concentrated from May to September, which some activity noticeable in early Spring. Winter months are almost without tourist activity, which correlates with the data from authorities, and as shown in Figure 13, sec 3.4. Additionally, the visualization offers expected results that hours from 9 or 11 in the morning till 9 hours in the evening are the hours with the most uploads. The unexpected uploads out of these hours, in some months, are probably related to the uploads „all at once“ when at home, and confirms that our data still is not bias-free.

5.5. Analysis of tags and titles

By using the tags and titles we also wanted to show how impressions during different periods can potentially differ and can be affected by events that occurred. We extracted the most commonly used terms from tags and titles. We set our code to extract the top 50 terms, excluding any numbers or symbols.

First, we wanted to explore if our data is consistent over the ten years period. This simply means that we wanted to see if users are adding titles and tags to their photos. As Table 21 shows, and unlike with the trajectories where there is a trend of more desirable uploading, the ratio of users which add photos without tags increased, while adding the title stayed on a similar level.

Table 21 - Proportion of photos with tags and titles

	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Total count	10096	12248	11984	10875	11200	10993	8759	7537	6873	5872
With tags	7660	9660	7895	7389	7496	7266	5381	4854	3958	3113
No tags	2436	2588	4089	3486	3704	3727	3378	2683	2915	2759
(%)	24.1	21.1	34.1	32.1	33.1	33.9	38.6	35.6	42.4	47
With Title	9550	11594	11023	10181	10528	10491	8126	6967	6328	5458
No title	546	654	961	694	672	502	633	570	545	414
(%)	5.41	5.34	8.02	6.38	6.00	4.57	7.23	7.56	7.93	7.05

After this, we wanted to check the upload count (absolute), upload change (relative figures), on the selected areas. Namely, we selected areas within Split and Dubrovnik, as well as two different time periods (2009-2011 and 2013-2015). We wanted to show if the selected areas within the towns have changed their relevance, based on the upload count. When it comes to the analysis of the tags and titles, we also wanted to compare the two periods and different buffers we selected. For Split, we selected the buffer around Poljud stadium, wanting to see if in the second period there are words related to the Ultra festival. In Dubrovnik, three buffers were selected, one within the old town, where the filming locations are placed, and two unrelated to the filming, but rather attractions on their own (the city harbor and Srdj hill).

Split

We are presenting here the most common words extracted from tags and titles of the metadata within the city of Split. We exclude words such as „the“, „and“, „in“, and similar. The words in bold are not related to the geographical names.

Table 22 - List of most common words within the city of Split

2009-2011		2013-2015	
Split	Poljud Buffer	Split	Poljud Buffer
split	split	split	**pictures**
croatia	croatia	croatia	**party***
travel	poljud	**travel**	**above**
europe	**wallpaper**	**palace**	**house***
conference	**side**	croatie	europe
palace	**art**	**sea**	**ultra***
city	**history**	europe	**records***
holiday	**city**	diocletian's	croatia
architecture	**west**	spalato	**stadium**
cruise	**summer**	**summer**	**art***
spalato	**east**	**annual**	split
croacia	**people**	eurashe	**skyline**
lemeridien	**hajduk**	**conference**	**umf***
diocletian's	**torcida**	**street**	**hajduk**
sea	**sky**	**cathedral**	poljud
street	**life**	splitcroatia	**stadion**
people	**street**	**city**	**travel**
vacation	**roman**	**roman**	**festival***
hrvatska	**museu**	**old**	**music***
adriatic	**north**	**pictures**	hrvatski

** related to the Ultra festival*

We also wanted to compare if the relative popularity of the buffer increased. With the formula presented in the Methods section, we came to the following results:

Table 23 - Absolute and relative proportions of photos within Split

	2009-2011	2013-2015
Total:	4846	4362
Poljud Buffer:	111	158
%	*2.3*	*3.6*

Dubrovnik

Similarly, here we present the most common words extracted from tags and titles of the metadata within the city of Dubrovnik. Same as the results for Split, we pointed out words which are not representing the name of a geographic location.

Table 24 - List of most common words within the city of Dubrovnik, Old Town and buffers without filming locations

2009-2011			2013-2015		
Dubrovnik	Old Town	No filming buf.	Dubrovnik	Old Town	No filming buf.
dubrovnik	dubrovnik	dubrovnik	dubrovnik	dubrovnik	dubrovnik
croatia	croatia	croatia	croatia	croatia	croatia
old	europe	croacia	**old**	**old**	**area**
city	**city**	europe	**city**	**city**	**sunset**
europe	**old**	**sea**	**town**	**town**	**sea**
travel	**travel**	from	europe	europe	**cruise**
sea	hrvatska	**sunset**	**travel**	**travel**	**city**
cruise	**town**	hrvatska	**sea**	hrvatska	**travel**
holiday	**cruise**	**old**	balkans	balkans	**cable**
town	**holiday**	croatie	**sunset**	**oldtown**	**car**
croatie	croatie	**city**	croatie	**street**	**old**
hrvatska	**oldtown**	**view**	**street**	**walls**	mljet
walls	croacia	**harbour**	hrvatska	**sea**	**town**
oldtown	**sea**	**town**	**walls**	croatie	adriatic
vacation	**walls**	**ship**	croacia	croacia	croacia
croazia	balkans	adriatic	from	dalmatia	europe
sunset	**vacation**	**travel**	dalmatia	**church**	**got***
trip	**trip**	europa	lokrum	**wall**	dalmatia
lokrum	**night**	**above**	**oldtown**	ragusa	**night**

** related to Game of Thrones series*

Table 25 – Absolute and relative proportions of photos within Dubrovnik

	2009-2011	2013-2015
Total:	14665	12981
Old Town:	10635	9510
%	72.5	73.3
No filming buffers:	803	928
%	5.5	7.1

In both areas, as expected, there are many words extracted strictly related to the geographic position where the photo was taken. Many of them give the name of the broader (Dubrovnik, Split, Adriatic sea, Croatia in various languages) or precise geographic location (Poljud, Diocletian's Palace, Le Meridian Hotel). Some, on another hand, are rather descriptive (harbor, old town, walls, church). In the case of Dubrovnik, and contrary to our expectations, there was only one *Game of Thrones* – related tag. This still might be due to the large number of tags and the fact that tourists (still) primarily come because of the landscape and cultural heritage. Additionally, while the increase in the number of tourists is partly due to the GoT series, it is expected that such an audience use other social networks, rather than Flickr. Split's Poljud, on another hand, shows many more results related to the festival in the buffer created around it. This also might be due to the data quantity, but also due to the fact that, by tourists, this is usually not a visited point within the city.

We can also notice, on example of Dubrovnik, how different buffers we selected show relations to main motifs of the buffers. The Old town buffer has terms such as „old town“, „walls“ „church“, while buffers over harbour and Srdj hill has terms „harbour“, „ship“, or „cable“ (relating to cable car to the Srdj Hill peak). Also, among the top 50 most common terms, there was no any which would be marked as negative. This might suggest tourists' high satisfaction of the location.

We also calculated the relative amounts of photos within the buffers where the festival and the filming took place. Both buffers have slightly increased in importance. The Poljud buffer went from 2.3 to 3.6% of the total number of photos within the city of Split. In Dubrovnik, the Old town has increased from 72.5 to 73.3% of the total number of uploads.

Chapter 6 –

Discussion

The goal of this chapter is to further discuss the results we gained. We discuss the meaning, importance, and relevance of the results, relating them to some relevant literature we presented. We will follow the structure of the thesis, meaning that we first present the data, discussing its quantity and quality, and pre-processing steps. Then, we discuss to which extent is our method of POIs detection and trajectory extraction successful, and how does it correlate to the official data. We discuss also our approach to the tags/titles analysis and temporal analysis. We include possible future work at the end of each section, giving also an overview of what could have been done better.

6.1. The data

The lower figures of the later years in the dataset should be correlated with the visit count, rather reflect the decrease of the user pool and the popularity of Flickr. As the number of Internet users and social networks increased, this additionally suggests that users migrated to other platforms, such as Instagram, as there are at the moment 14 million results for #croatia query. This is in particularly bad for researchers, as Flickr is one of rare platforms with open API. This also shows the problem of the platform bias we discussed.

In Table 10 an overview of how different steps in preprocessing excluded different amounts of photos is given. For example, despite the fact that the region of Dalmatia took less than half of the territory of the bounding box, it had almost 80% of photos. However, once we removed users which uploaded over 150 photos, the photo count dropped as much as 50% in some years. This reduced the effect of so-called „90-9-1" rule and demonstrated the importance of the exclusion of the outliers, which Grossenbacher (2014) and Nielsen (2006) discussed. We can speculate if a better approach to data pre-processing would increase data quantity or quality. For example, leaving users with any upload count, but relating the upload count to trajectory length and time spent within the destination. We did not discuss in detail the problem or upload locations, but it can be assumed that a large proportion of photos are to some extent incorrectly uploaded. This is almost impossible to detect with large amounts of data, as it would take a manual approach.

Lastly, by using the threshold of 20 days, we excluded locals, relatively small proportions of photos. As we discussed, different researchers suggested different thresholds for the exclusion of locals. Kadar and Gede set only 3 days for Budapest, Huang set 5 days for 6 major European cities. Knowing the difference between city-break tourism and destination for longer vacation, we can say that our threshold is well-chosen and can be suggested for similar destinations and researchers, such as other Mediterranean regions.

6.2. Points of interest, trajectories, and temporal change

Discovery of Points of interest, and comparison with the official data, was one of the main tasks of the thesis. We used heatmap and Kernel Density Estimation, which offer visual, rather than quantitative results. Additionally, the polarisation of uploads within the region cannot be visualized properly, as very small areas, old towns of only a few places, are disproportionally covered by photos, as also concluded by Cai et al („city centers are hubs for photographers"). Similar to many other works with related topics, we used DBSCAN method which group photos into clusters, removes noise and gives measurable, quantitative, and comparable results. Once we compared the top ten of our clusters and the top ten most visited destinations according to the official data, we had each year 6-8 matchings. Because of the nature of the DBSCAN method, and as we demonstrated with the data, the clusters which represent most photographed areas are not consistent over time, no matter which parameters are set. Additionally, it needs to be pointed out that the methods above can detect most photographed places, and not the most visited ones, or the ones with most nights spent. There are several locations that the method did not detect as prime points of interest, yet they constantly have over half a million nights spent by tourists. Because of this, the method can complement the data from the officials, and by no means can it replace it. We have also used the method for the region and did not try to detect most photographed places within a specific city or town. This is because all of the places of the region are by area too small for such an approach. Cities like Paris, Vienna, Amsterdam, as demonstrater by Huang (2016), or San Francisco, Los Angeles, or New York, as demonstrated by Zheng et al. (2012), could have meaningful results from the method. Also, because of the data quantity, the DBSCAN method is not ideal for the discovery of less-visited yet relevant POIs, thus for destinations with lower uploads count.

Trajectory extraction was not covered as much as POIs by the researchers, either within countries, regions, or cities. At first, we wanted to only extract trajectories and statistical data from them. We demonstrated a method of calculating an average trajectory length, and by adding 25%, we speculated what could be the actual travel distance of the tourists within the region. When compared results year-by-year, there is no significant change, and as there are no official data covering such figures, we can only guess how close the numbers are to the reality. On another hand, by knowing the time of the first and the last uploaded photo, we calculated an average time tourists spent within the destination. When compared to the official data, our figures were relatively close to even completely overlapped. Despite such results, this again cannot replace traditional approaches but tells that the results from VGI/UGC data can be taken as relevant for other tasks. For both lengths of trajectories and nights spent, we presented the data into seasons, finding how summer months offer, by far, the most precise results, while winter is heavily influenced by outliers as the user pool is very low.

Trajectories provided other interesting findings. Once we selected destinations of interest, we wanted to show how tourists move between them. For starts, our findings correspond to those reported from Cai et al. (2014), where they noticed how closer places are more likely to be connected by trajectories, or how short trips, or nearby movement, are more common than distant movement.

The exception being the most visited places of the regions, such as Zadar, Split, and Dubrovnik, as they are interchangeably rather well- visited, despite being far away from each other (Zadar and Split are 160, while Split and Dubrovnik are 230 kilometers away). Despite this being somewhat expected, this is an interesting finding not presented by other researchers. Those destinations are more and more mutually visited over the years. Same can be said for some near-movement trips. For example, more tourists every year are visiting Konavle from Dubrovnik (or, to less extent, vice-versa), or Trogir from Split. The biggest issue with the approach is that some destinations have a small sample size, thus create inconsistent results where a few users present *trends* that might not be based in reality. The approach presented is partly manual, namely destination borders are self-drawn. Another approach that can be used is adding grids which would be combined with borders of towns or municipalities.

As it was discussed, it is clear that trajectories are mostly imprecise, as many are made out of only a few points yet represent relatively long paths. It also lacks information on speed, but contrary to what some researchers claim, it provides the direction of the movement. Also, a possible alternative of using GPS data has its own disadvantages, such as problems with sample size, devices used, and issue of privacy. This has been demonstrated by Khairi and Ismaili (2015), where their sample size was less than 20. Our data provided a minimal sample size od 250 in 2019, up to 600 in some years. Our method gives us some additional possibilities. We could, naturally, visit also patterns between destinations of different countries; we could go to more micro (urban level), observing how many tourists are visiting museums, restaurants, or other attractions. A posibility, presented by Arase et al. (2010), is to relate the different type of destinations (natural, cultural, urban, a specific type of attractions, and similar) and try to extract patterns between them, categorizing tourists according to their visits. For this sample size needs to be larger and some of the pre-processing steps more critical in excluding some Flickr users. As our method mines popular paths, it can be part of the recommendation system. In particulary, it can be used for smaller areas, giving an overview of popular day trips, as also discussed by Lu et al. (2010).

The official data naturally offers figures on visits within the year, meaning that it can be found how much is a destination, either a county or a city, visited each month. However, it is rather a general overview which does not offer an in-depth temporal component of a tourists' behavior. The method we presented showed not only precise annual distribution but the daily distribution of tourists' activity. We developed a visually appealing, easy to read visualization of such activity. As we have no official data to compare it and thus show correctness, we can only speculate how our method is close to reality. When it comes to annual distribution, our data relatively correctly presented a strong polarization between months and seasons in Dalmatia.

6.3. Analysis of tags and titles

We analyzed two towns of Dalmatia, Split and Dubrovnik, to observe if impressions of tourists can be read from the tags and titles of the photos. We focused on possible effect the two specific events occurring there, Ultra festival in Split and filming of the popular series Game of Thrones in Dubrovnik. We took two periods of three years, namely the period of 2009-2011 and 2013-2015. We expected that the second period, as it had a peak of popularity of both festivals and series, will reflect this in our data. As for Poljud, the stadium where Ultra festival takes place showed that many tags and words from titles are related to this event. However, within the whole Split, we did not detect words related to the festival. Similar goes for Dubrovnik, as we found only one term related to the series, and it was outside of the buffer where filming took place. When it comes to most used terms within the places, a majority is related to naming the geographical location of the place, primarily towns' names, then country names, both in various languages (English, Spanish, Italian, and others). We also found terms such as Adriatic (sea), Europe, and Dalmatia. The terms not related to the name of the geographic location did not change too much over the time, as they were related to vacation (holiday, vacation, trip, summer) or describing attractions (walls, harbor, old town).

It was concluded that buffers with different motifs also offer different most common terms. As we presented, Old town will have „oldtown“ as one of top 10 terms, the harbour will have „harbour“, while hill has „cable“ (cable car which transports people to the hill's peak) as one of the most occured terms. It was also detect that relative proportions of uploads increased both in the Poljud buffer and Dubrovnik's Old town (1.3 and 0.8%, respectively). Despite somewhat low figures and low amount of terms related to the events, this might suggest more interest in such locations, and the part of the reason might be in aforementioned events.

Our approach could have been extended to more destinations, and we could also visualize our findings with word clouds, circle packing, cartograms, or similar. However, since it did not detect as meaningful results as we expected, we only showed and commented the most common tags and relative changes. Additionally, as Spyrou and Milonas (2016) discussed, Flickr is a social network which has a high interaction between users, mostly when it comes to groups and commenting. Mining comments and impressions from them could be done with the approach presented as well.

Chapter 7 –

Conclusion

This thesis aimed to use freely available UGC data to check trends in tourism on the area of the Croatian region of Dalmatia thus to see to which extent such data can complement the data from the official sources. Additionally, the thesis took the time span of ten years within Dalmatian tourism almost doubled, assuming that this will be reflected in tourists' behavior, their temporal and spatial activity and impressions.

Even after we did necessary preprocessing steps with our data, it has still shown some flaws, such as less quantity in the later years, problem of over contributors, or incorrect upload location. Despite that, our findings proved to correlate well with the data from officials when such comparison was possible to be made, which would suggest that even the findings not provided by the authorities are relevant and correct to some extent. For example, by using the DBSCAN method, we successfully detected 6-8 out of 10 most visited places in each year of the decade 2009-2018. This was expected, as such and similar approaches have been already tested by other researchers, mostly in big cities over the world.

Additionally, we extracted trajectories in order to seek for patterns between destinations. A similar approach was presented by some researchers, however, many did not present such comprehensive quantitative results as we did. Our approach showed to some extent expected, yet interesting findings on how destinations are mutually visited. This showed strong patterns between major places of the region, even when the distance between them is relatively long. It also showed how close destinations are mutually visited, which would suggest that many tourists make daily trips while in a popular destination. It is safe to assume that similar approach on the area such as Paris, Amsterdam, New York, or any other well-visited western city with more UGC data could use this approach for gaining the knowledge on tourism and tourists' interests. Having this in mind, this can be used in tourism planning and recommendation systems. Our findings on trajectories expanded to a length of trajectories, suggesting that a tourist, on average, make up to 200 kilometers, which is also not covered by the official data. As we successfully detected how many days on average tourists spent, we can assume that traveling distance tourists make is not too far from reality.

Official data also does not track tourists' activity over the day and year. We have visualized most popular months within the year, as well as hours over the day, finding that tourists' activity is concentrated mostly from 9 in the morning to 21 in the evening, naturally during late spring and summer. While this is expected for such region, this approach could be used for micro-regions, particular destinations, and relate to impressions from tags, finding if tourists are more or less hapy during particular hours and if this is related to crowds, jam and similar.

Our thesis also wanted to detect if changes in tourism, namely popular events which can affect the destination's image, can be read from the impressions of tourists. For this, we used a large corpus of words from tags and titles uploaded with the photos. Depending on a destination, our data showed that less visited areas reflect popular events more easily. However, many tags showed to be rather generic and related to the location's name, than on descriptive words which would detect the impression and satisfaction of a tourist. There is also a possibility that users will in general leave only positive and neutral impressions when on vacation, so we evaluate this approach as not ideal to read objective user impressions.

Our research questions, therefore, can be answered as follows:

RQ 1: Are changes in the behavior and impressions of tourists in Croatia (Dalmatia) reflected in the properties of UGC data?

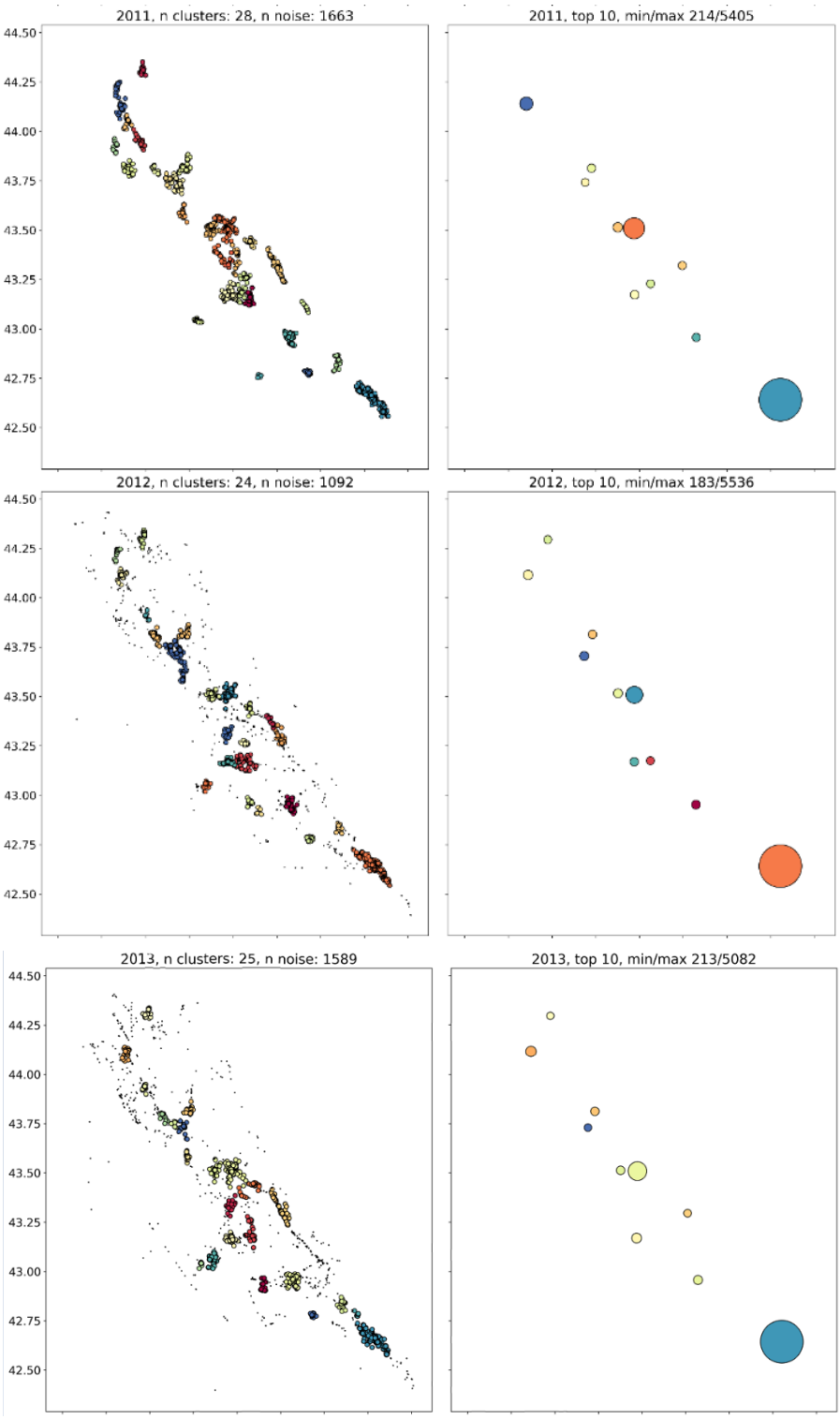
We defined tourists' behavior as a temporal and spatial activity within a destination, and as such, we wanted to not only visualize it, as because of a relatively large amount of trajectories this does not provide findings, but also quantify it. As we already stated, there are clear patterns that suggest how some destinations are more and more visited from other destinations. We gave the example of NP Krka, which increased a number of visits from all major towns, or how daily trips, such as those from Dubrovnik to Konavle or Split to Trogir, are more likely to happen. We also wanted to observe if users changed their temporal activity but found that such changes seem to be hard to detect and impossible to validate by the data from officials. Additionally, we proved that users do leave tags related to special events within the space. This, however, was to only a small extent, and depends on the total amount of the data. Particularly, many tags from Poljud stadium were related to a music festival which was held there, while very few related to the popular GoT series in Dubrovnik.

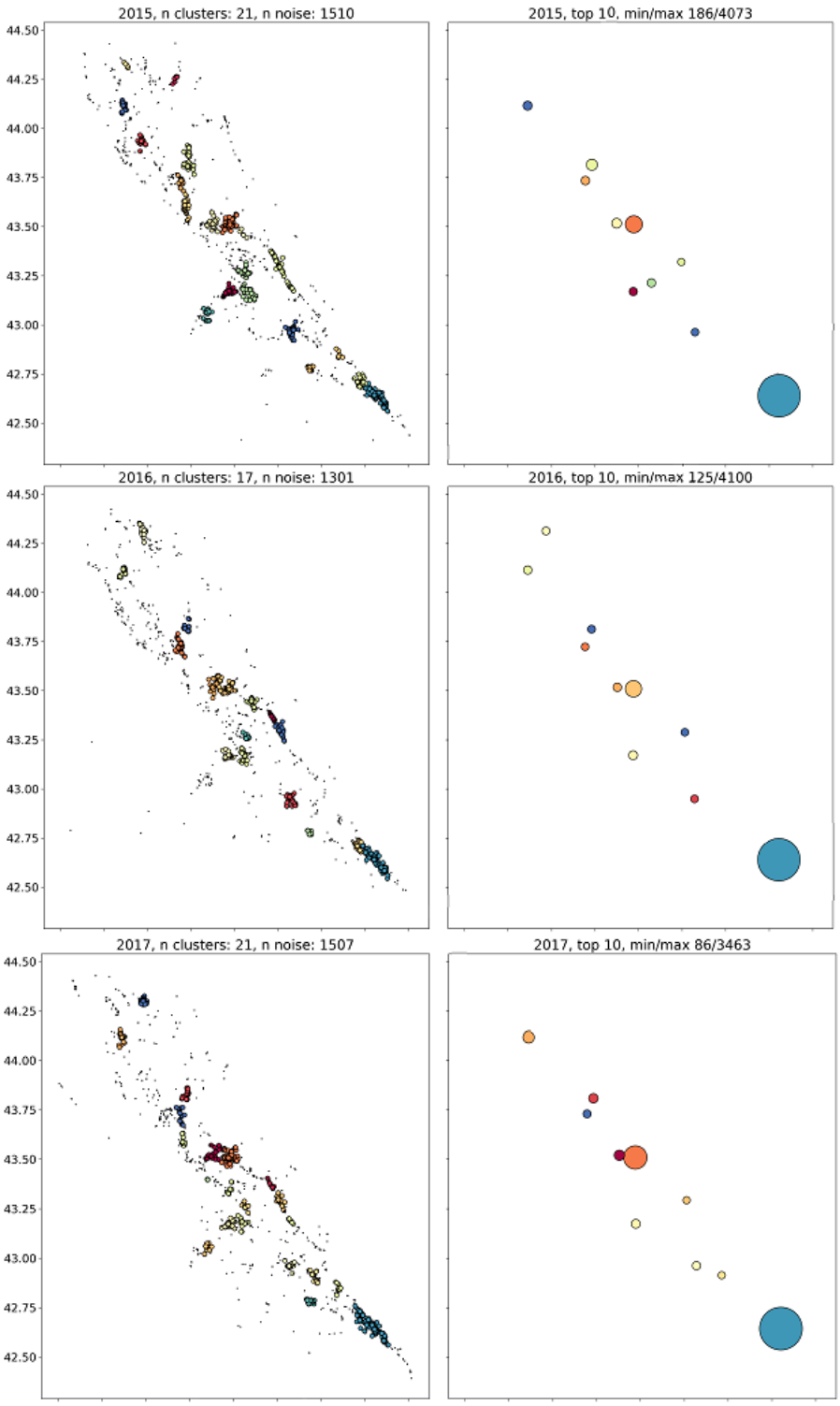
RQ 2: What dimension of these changes could be extracted from UGC?

As we presented, there is a big volume of information which can be extracted from UGC data. Some of them can be supported by the official data, while some can only be gained with expensive polls, surveys, or GPS tracking. The changes are mostly reflected when it comes to tourists' movement, as we showed how patterns between destinations have changed. Conclusively, this approach can be especially useful for areas without well-developed tourism statistics.

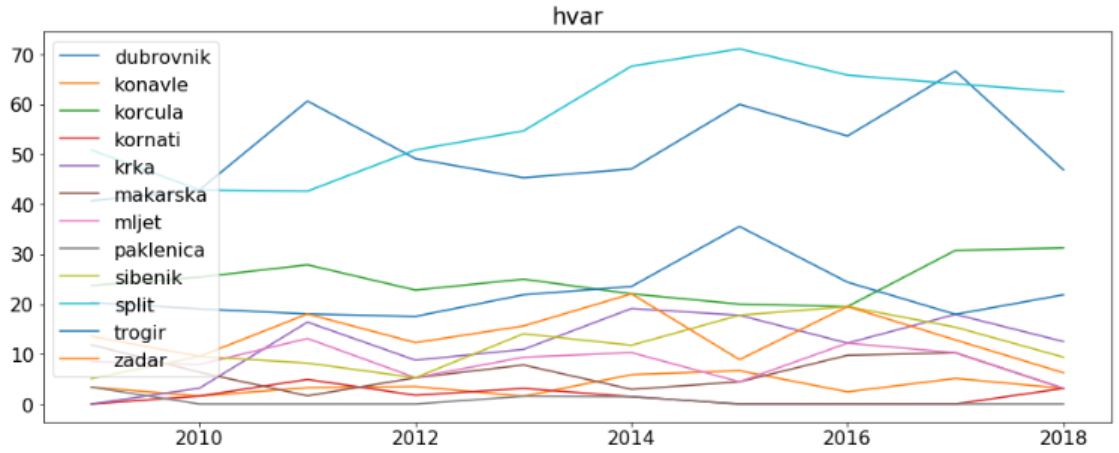
Appendix

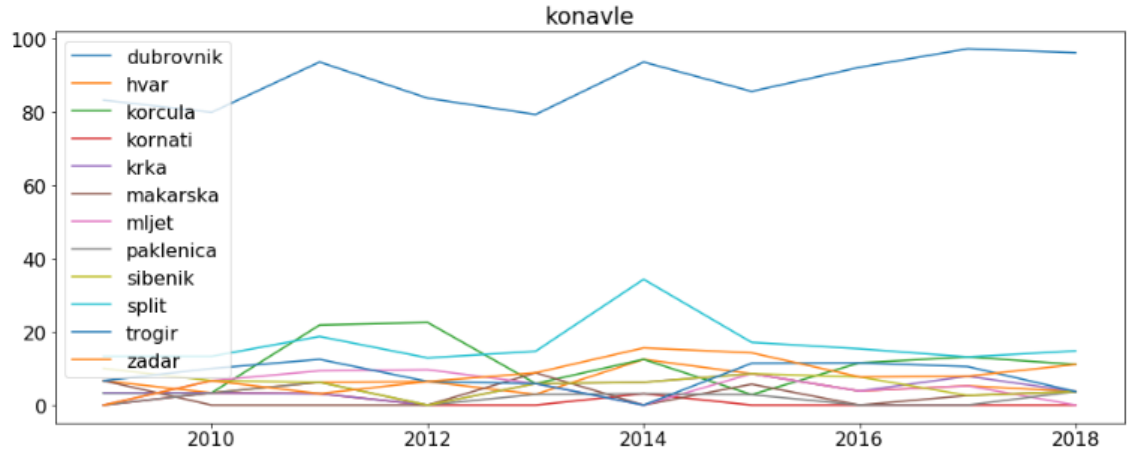
We offer here additional visualizations of our data. The codes which resulted in such visualization are available at GitHub: https://github.com/InspectorTime/Dalmatia2020ThesisCode

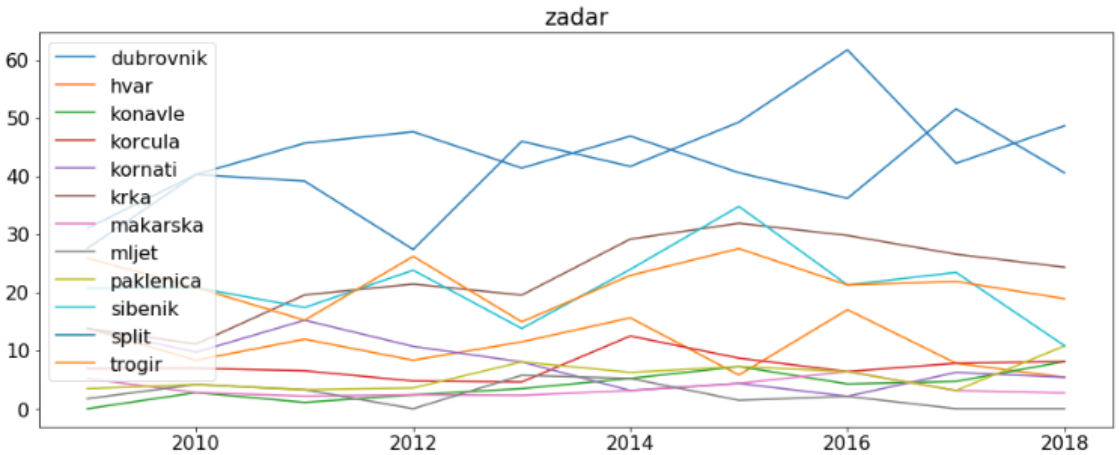




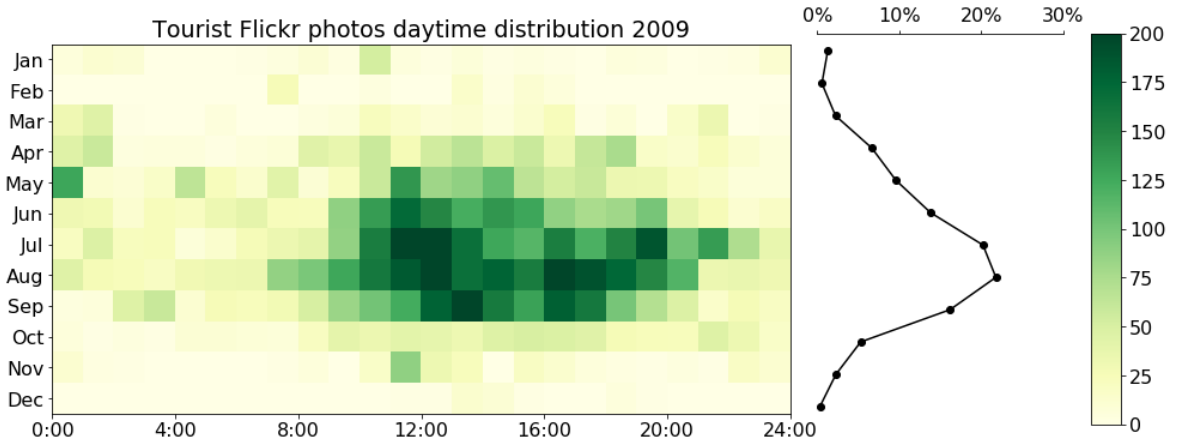
These images present clusters of the most popular destinations within the region of Dalmatia. On the right side, we transformed the top 10 clusters into circles that represent the amount of the uploads of photos, as their size is relative to photo count.

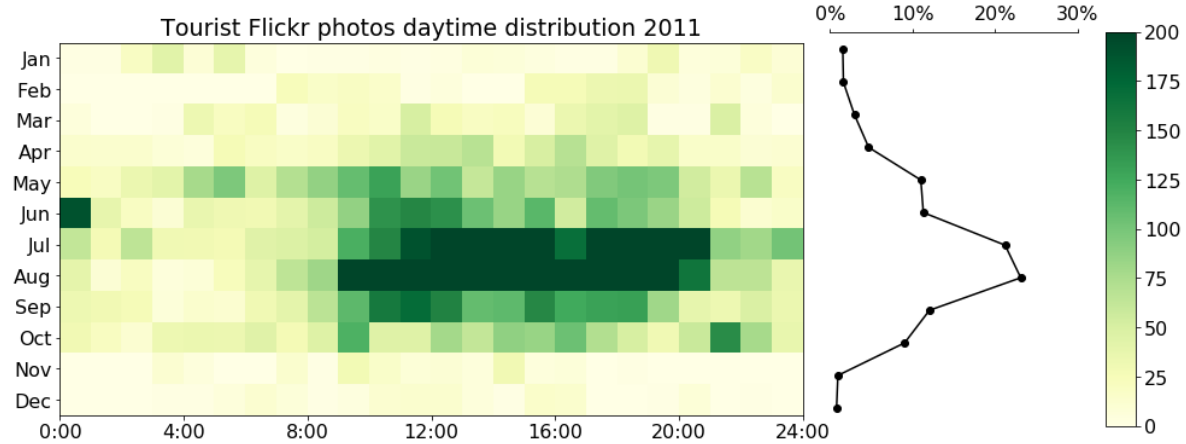


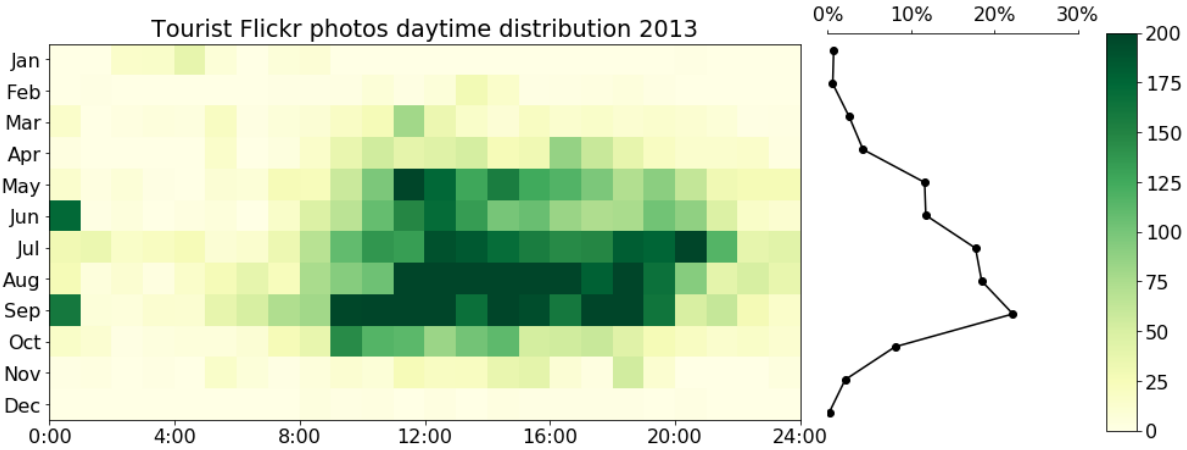


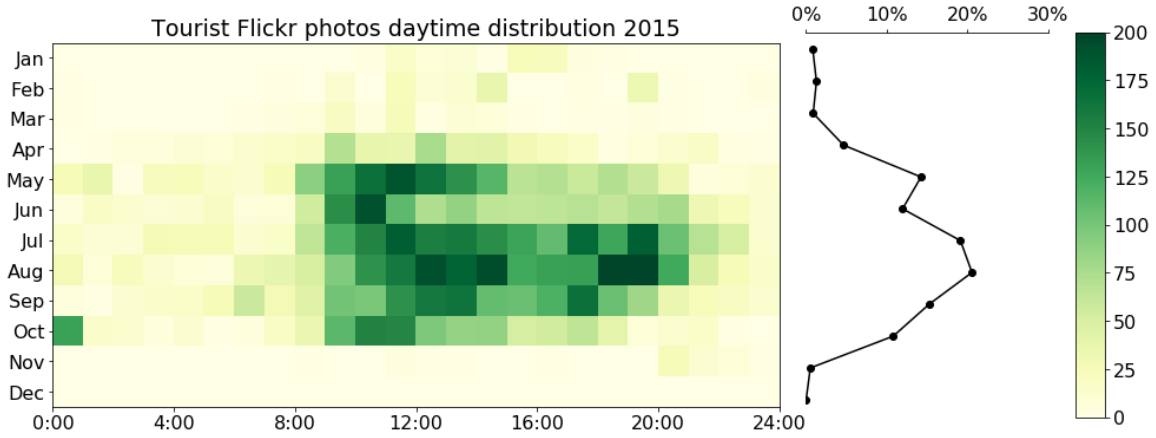


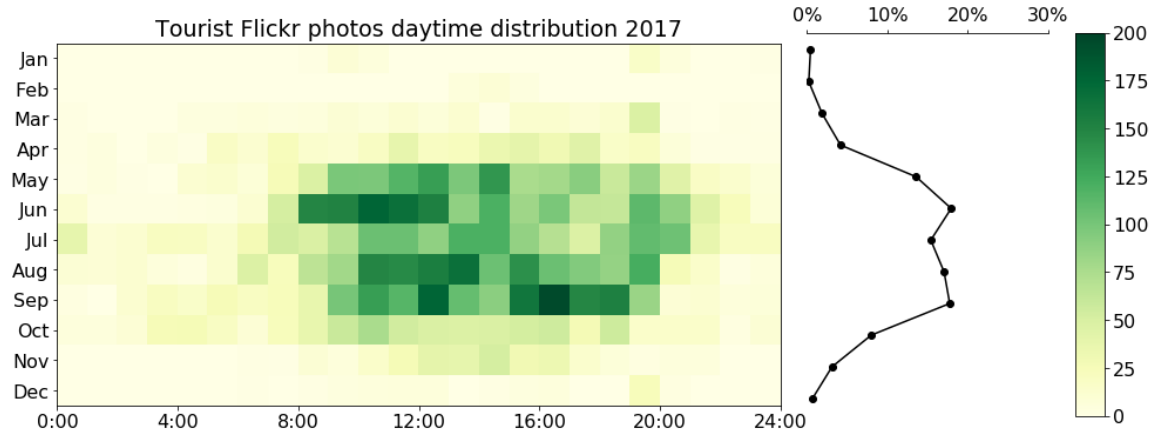
In the work, we only presented these graphs for Dubrovnik and Split, as the data for them is most representative. We add here additional graphs for Hvar, Konavle, and Zadar. Other destinations do not give meaningful visualization and can be read from Tables 16 – 20.


Tourist Flickr photos daytime distribution 2009
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
0:00
4:00
8:00
12:00
16:00
20:00
24:00
0%
10%
20%
30%
200
175
150
125
100
75
50
25
0


Tourist Flickr photos daytime distribution 2011
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
0:00
4:00
8:00
12:00
16:00
20:00
24:00
0%
10%
20%
30%
200
175
150
125
100
75
50
25
0


Tourist Flickr photos daytime distribution 2013
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
0:00
4:00
8:00
12:00
16:00
20:00
24:00
0%
10%
20%
30%
200
175
150
125
100
75
50
25
0





We present here the rest of the matrices which represent the activity of the Flickr users, thus tourists, within our study region.

Literature

Akram, W. and Kumar, A. (2017). A Study on Positive and Negative Effects of Social Media on Society. International Journal of Computer Sciences and Engineering. Volume 5, Issue 10. E-ISSN: 2347-2693

Arase, Y.; Xie, X.; Hara, T.; Nishio, S. Mining People's Trips from Large Scale Geo-Tagged Photos; ACM Multimedia: Mountain View, CA, USA, 2010; pp. 133–142.

Boscoe, F.P., Henry, K.A. and Zdeb M. S. (2013). A Nationwide Comparison of Driving Distance Versus Straight-Line Distance to Hospitals. The Professional Geographer 64(2).

Buchin, M. and Purves, R. (2014). Computing Similarity of Coarse and Irregular Trajectories Using Space-Time Prisms. International Journal of Built Environment and Sustainability SIGSPATIAL'13: Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information SystemsNovember 2013 Pages 456–459

Cai, G., Bermingham, L., Lee, K., Hio, C., and Lee, I. (2014). Mining Frequent Trajectory Patterns and Regions-of-Interest from Flickr Photos. 47th Hawaii International Conference on System Science

Cao, L., Luo, J., and Gallagher, A. (2010). A worldwide tourism recommendation system based on Geotagged web photos. Acoustics, Speech, and Signal Processing, 1988. ICASSP-88.

Cohen, S., Prayag, G., and Moital, M. (2014). Consumer behaviour in tourism: Concepts, influences and opportunities. Current Issues in Tourism.

Croy, W.G. (2004). The Lord of the Rings, New Zealand and Tourism: Image Building with Film. Project: Tourism and Media (especially film tourism)

Dhiratara, A., Yang, J., Bozzon, A. and Houben. G.-J. (2016). Social media data analytics for tourism, a preliminary study. In Proceedings of the KDWeb 2016 content. Pervasive Computing no. 7 (4):36-43

Dodge, S., Weibel, R., and Lautenschütz, A. K. (2008). Towards a taxonomy of movement patterns.

Duarte, P. and Folgado-Fernardez, J.A. (2018). Measurement of the Impact of Music Festivals on Destination Image: The Case of a Womad Festival. Event Management.

Fan, J., Fang H., and Han L.. 2014. "Challenges of Big Data Analysis". National Science Review, 1(2), 293-314.

Fisher, D., Wood, S.A., Roh, Y., and Kim, C. (2019). The Geographic Spread and Preferences of Tourists Revealed by User-Generated Information on Jeju Island, South Korea. Land 2019, 8, 73

Fisher, P. F., Laube, M., and Imfeld, S. (2006). Finding REMO — Detecting Relative Motion Patterns in Geospatial Lifelines.

Girardin, F., Blat, J., Calabrese, F., Dal Fiore, F. and Ratti, C. (2008). Digital footprinting: Uncovering tourists with user-generated

Goodchild, M. (2007). Citizens as Sensors: The World of Volunteered Geography. August 2007 GeoJournal 69(4):211-221

Golder, S. and Huberman, B. (2006). Usage Patterns of Collaborative Tagging Systems. Journal of Information Science 32(2):198-208.

Grossenbacher, T. Studying Human Mobility Through Geotagged Social Media Content. Master's thesis, University of Zurich (2014).

Gudmundsson, J., Laube, P. and Wolle, T. (2012). Computational movement analysis.

Harris, R., Sleight, P., & Webber, R. (2005). Geodemographics, GIS and neighbourhood targeting. Wiley.

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

Huang, H. (2016). Context-aware location recommendation using geotagged photos in social media. ISPRS International Journal of Geo-Information, 5(12):195.

Hudson, S. and Ritchie, J. (2006). Promoting destinations via film tourism: An empirical identification of supporting marketing initiatives. Journal of Travel Research. 2006, 44, 387-396

Joshi, N. (2015). A Quantitative Study of the Impact of Social Media Reviews on Brand Perception. Master thesis.

Juvan, E., Omerzel, D. and Maravić, M. (2017). Tourist Behaviour: An Overview of Models to Date.

Kádár, B.; Gede, M. Where do tourists go: Visualizing and analysing the spatial distribution of geotagged photography. Cartogr. Int. J. Geogr. Inf. Geovis. 2013, 48, 78–88.

Khairi, N. D. and Ismail, H. N. (2015). Acknowledging the Tourist Spatial Behavior for Space Management in Urban Heritage Destination. International Journal of Built Environment and Sustainability

Kuo, C., Chan, T., Fan, I. and Zipf, A. (2018). Efficient Method for POI/ROI Discovery Using Flickr Geotagged Photos. ISPRS International Journal of Geo-Information.

Liu, Y., Wu, L., Yu, L., and Chaogui K. (2017). Quantifying Tourist Behavior Patterns by Travel Motifs and Geo-Tagged Photos from Flickr

Lu, X., Wang, C., Yang, J.M, Zhang, L., and Pang, Y. (2010). Photo2Trip: Generating Travel Routes from Geo-Tagged Photos for Trip Planning. ACM Multimedia 2010 International Conference. January 2010

McKercher B, Shoval N, Ng E, et al. (2012) First and Repeat Visitor Behaviour: GPS Tracking and GIS Analysis in Hong Kong. Tourism Geographies 14(1): 147–161.

Memon, I., Cheng, L., Majid, A., Lv, M., Hussain, I., and Chen, G. (2014). Travel reccomentadion Using Geo-tagged Photos in Social Media for Tourists. Wireless Personal Communications.

Memon, I., Cheng, L., Majid, A., Ly, M., Hussein, I., and Chen, G. (2015). Travel Recommendation Using Geo-tagged Photos in Social Media for Tourist. Wireless Personal Communications. 2015, 80, 1347-1362

Mukhina, K., Visheratin, A. and Nasonov, D. (2018). Building City-Scale Walking Itineraries Using Large Geospatial Datasets. Conference paper.

Murugesan, S. (2007). Understanding Web 2.0. 1520-9202/07/$25.00 2007 IEEE

Nagy, A. and Nagy, H. (2013). The Importance of Festival Tourism in the Economic Development of Hungary. Visegrad Journal on Bioeconomy and Sustainable Development.

Nielsen, J. (2006). Participation inequality: Encouraging more users to contribute [Nielsen Norman Group]. Link: https://www.nngroup.com/articles/participation-inequality/

Ponomarev, A. (2016). Recommending Tourist Locations Based on Data from Photo Sharing Service: Method and Algorithm. 18th Conference of Open Innovations Association and Seminar on Information Security and Protection of Information Technology.

Quaiser, S. and Ali, R. (2018). Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. International Journal of Computer Applications 181(1).

Rorissa, A. (2010). A Comparative Study of Flickr Tags and Index Terms in a General Image Collection. Journal Of The American Society For Information Science And Technology, 61(11):2230–2242, 2010

Seely-Gant, K. and Frehill, L.M. (2015). Exploring Bias and Error in Big Data Research. Journal. Washington Academy of Sciences, Washington, D. C.

Sharma, S. and Godiyal, S. (2016). A Study on the Social Networking Sites Usage by Undergraduate Students. Online International Interdisciplinary Research Journal, ISSN 2249-9598, Volume-VI, Issue-III, May-June 2016 Issue

Simundic, A. (2013). Bias in research. Biochemia Medica. 2013, 23, 12-15

Spyrou, E. and Mylonas, P. (2016). A survey on Flickr multimedia research challenges. Engineering Applications of Artificial Intelligence, February 2016.

Tkalec, M., Zilic, I., and Rechner, V. (2017). The effect of film industry on tourism: Game of Thrones and Dubrovnik. International Journal of Tourism Research. 2017, 19, 705-714

Van Vuuren, C. and Slabbert, E. (2012). Travel Motivations And Behaviour Of Tourists To A South African Resort. Book Of Proceedings Vol. I – International Conference On Tourism & Management Studies – Algarve 2011

Wider, T., Palacio, D., and Purves, R. (2013). Georeferencing images using tags: application with Flickr. Conference: AGILE'13: Proceedings of the 16th AGILE International Conference on Geographic Information ScienceAt: Leuven, Belgium

Wilson, D. W., Lin, X., and Longstreet, P. (2011). Web 2.0: A Definition, Literature Review, and Directions for Future Research. AMCIS 2011 Proceedings, November 2014

Xia, J. (2007). Modelling the Spatial-Temporal Movement of Tourists. School of Mathematical and Geospatial Sciences

Yoon, Yooshik & Uysal, Muzaffer. (2005). An Examination of the Effects of Motivation and Satisfaction on Destination Loyalty: A Structural Model. Tourism Management. 26. 45-56. 10.1016/j.tourman.2003.08.016.

Zeng, B. and Gerritsen, R. (2014). What do we know about social media in tourism? A review. Tourism Management Perspectives 10 (2014) 27–36 Contents

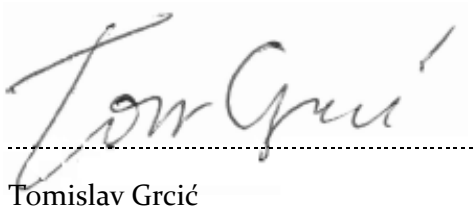
Zeng, Z., Zhang, R., Liu, X., Guo, X., and Sun, H. (2012). Generating Tourism Paths from Trajectories and Geo-Photos. WISE 2012, LNCS 7651, pp. 199-212

Zheng, Y.T.; Zha, Z.J.; Chua, T.S. Mining Travel Patterns from Geotagged Photos. ACM Trans. Intell. Syst. Technol. 2012, 3, 1–18.

Personal declaration

I hereby declare that the submitted thesis is the result of my own, independent work.

All external sources are explicitly acknowledged in the thesis.



Tomislav Grcić

in Zurich, 31.01.2020.