Márcio Martins*, Arlindo Santos

Exploring the potential of Flickr User–Generated Content for Tourism Research: Insights from Portugal

https://doi.org/10.2478/ejthr-2024-0019
received January 25, 2024; accepted April 5, 2024

Abstract: The purpose of this research is to present and discuss the methods of identifying visitors and local population in tourism studies using geotagged photos from the Flickr social network, and differentiating them according to their country of residence. This study focuses on 1,434,268 photos taken in Portugal between 2010 and 2022 and uploaded by 31,286 Flickr users. Different approaches to cleaning the database, to distinguishing tourists from locals, and to identify their country of residence were employed and discussed. After data cleaning, the photos database corresponds to 1,144,981 photos shared by 29,890 users. Using the information provided in each user's profile and the time zone, 12,144 users (41%) were classified as visitors and 2,659 users (9%) as locals. The monthly distribution of the percentage of photos uploaded by users classified as visitors coincides with the high season of tourist activity in Portugal. The distribution of users by country of residence coincides with the main inbound markets in Portugal: Spain, United Kingdom, France and Germany. Building on a country-wide case study, the contribution of this paper is a refined understanding of the use of Flickr user-generated content in tourism studies, proposing a framework to facilitate all researchers to use this data source more frequently.

Keywords: Flickr; Portugal; Tourism; User-Generated Content (UGC); Visitor's activity.

***Corresponding author: Márcio Martins,** Adjunct Professor in Instituto Politécnico de Bragança, CITUR, Portugal,
Email: marcio.martins@ipb.pt
Arlindo Santos, Adjunct Professor in Instituto Politécnico de Bragança

1 Introduction

The use of big data has increased in tourism studies (Mariani & Baggio, 2022). The production of massive quantities of data – i.e., big data – can be collected by different sources such as internet searches, records of mobile phone activity, images recorded with video cameras or social networks, among others, and with these large volumes of data, in particular geolocated photos, scholars in tourism can analyse visitors' activity in destinations from a spatial and temporal perspective (Mariani & Baggio, 2022; Salas-Olmedo et al., 2018). One of the reasons for the increasing interest of researchers in using User-Generated Contents (UGCs) is the easy access to databases, such as Flickr, and also their low cost. Besides that, UGCs from social networks have a huge amount of data and metadata covering a long time period worldwide (Derdouri & Osaragi, 2021). However, some scholars have mentioned that for the generation of relevant knowledge large volumes of data are not enough; analytics is required "to access, store, analyse and interpret data for the identification of meaningful patterns in the data" (Mariani & Baggio, 2022, p.232).

In studies on the spatiotemporal behaviour of tourists, various tracking techniques can be used (Martins & Costa, 2022; Padrón-Ávila & Hernández-Martín, 2020). One of the most popular is the use of GPS technology: the researcher requests the tourists participating in the study to carry a GPS device with them during their visit (Caldeira & Kastenholz, 2018; Shoval et al., 2018). More recently, GPS applications uploaded in tourists' smartphones have been used (Martins et al., 2022; Miyasaka et al., 2018; Yun & Park, 2014). In either case, the use of GPS technology has some disadvantages, such as the resultant small sample, the time-consuming process – and in the case of GPS devices, it can be expensive. GPS technology is also intrusive. It can be difficult to motivate tourists to participate in the research. Some

problems may arise, also, when collecting data, such as the possible blockage of GPS signal due to buildings and denser vegetation, or the limited equipment battery (Martins & Costa, 2022).

When GPS technology is used, it is easy to collect information on participants with the use of a questionnaire. The same cannot be said for the use of the UGC metadata, in which user sociodemographic profiles are often not available. And as visitors and residents have different spatiotemporal behaviours, how can differences between them be established through the photographic metadata? Derdouri and Osaragi (2021) summarise the different approaches existent in literature: the heuristic methods (e.g., based on stay periods), supervised machine learning (ML) algorithms, and Shannon entropy (SHEN).

The authors mention that heuristic approaches cannot be evaluated, and their outcomes cannot be validated. In other words, what is the best way to find the thresholds related to the minimum length of stay to identify visitors? For Derdouri and Osaragi (2021), ML-based methods neglect factors related to the distinction between visitors and locals, including weather, population density, and the content of users' posts. With the SHEN method, the dispersion of values of a given variable is calculated. Tourists and locals can be identified by temporal entropy – tourists remain for a brief time in destinations taking photos – and by spatial entropy – tourists and locals have different travel trajectories with different spatial distributions of images – but it can only differentiate two groups.

Straumann et al. (2014) also mention that time-based approaches have significant weaknesses: the thresholds chosen by scholars are subjective and all residents that don't take or share photos with frequency can be identified as foreigners. They highlight that time-based approaches rely on large amounts of data which need a long processing time.

The purpose of this research is to present and discuss the methods of identifying visitors and local population in tourism studies using geotagged photos from the Flickr social network and distinguishing them according to their country of residence. As also observed, and to our knowledge, there is no conceptual framework that could help scholars to choose the best method for distinguishing locals from visitors using photos from Flickr. This paper intends to address this gap and is structured as follows: after introduction, Section 2 reviews the existing literature on the use of Flickr social network photos on tourism research and presents the different methods of identifying visitors and residents through geotagged photos. Section 3 is dedicated to methodology, and Section 4 analyses and discusses the different methods of data cleaning to distinguish visitors from residents, to identify the visitors' country of residence, and to identify the number of visits made by each visitor during the study period. Section 5 presents the main conclusions.

2 Literature Review

2.1 Flickr social network

Flickr is a popular photo-sharing platform recognised as a source of data for research (Thomee et al., 2016) where users can share and organise their photos. It was developed by Ludicorp, a Vancouver, Canada-based company founded in 2004 by Stewart Butterfield and Caterina Fake and is one of the most popular sites worldwide, ranking top 400 globally, top 290 in USA, and top 2 in its category – photography (Broz, 2022). This social network allows users to add geographical references (longitude and latitude) to each uploaded photo. Users can also associate other data such as photo description, hashtags, location, the capture and publication date, among others. Developers, using the Flickr API interface, access this data for further analysis.

In February 2017, Flickr hosted approximately 13 billion photos from 122 million users that come from 72 countries (USA, 31.03%, UK, 9.83% and Germany, 5.26%), getting up to 60 million visits per month. By 2022, around 25 million photos were uploaded to Flickr every day (Broz 2022). Even if not all social groups or travellers use Flickr, Kádár and Gede (2021) highlight that it is a valid research database for measuring tourism demand, delivering comparable data on visitor flows from different parts of the world. For operational purposes, in this article, a user corresponds to a person who is registered on the Flickr social network to host and share photos.

2.2 Flickr geotagged photos in tourism research

The Flickr social network has stored a huge quantity of georeferenced photographs that have been used by many tourism scholars to study the spatiotemporal behaviour of tourists (Giglio et al., 2020; Höpken et al., 2020; Jing et al., 2020), particularly tourists' decision-making choices (Bettaieb & Wakabayashi, 2021), destination consumption (Solazzo et al., 2022; Spyrou et al., 2017) or the development of predictive factors (Domènech et al., 2020). The classification and quantification of visitors are also

interesting topics explored by several scholars (Kádár, 2014; Wood et al., 2013). Using 16 European historic tourist cities, Kádár (2014) concludes that the correlation coefficient between registered tourist bed nights and both user numbers and number of photographs are very high. Wood et al. (2013) used the locations of photographs in Flickr from 836 recreational sites around the world and found that crowd-sourced information is a reliable proxy for empirical visitation rates.

Some researches on itineraries and attraction recommendations (Lim et al., 2018; Sarkar & Majumder, 2021; Sun et al., 2019) and tourism attractions/destinations rankings (Al-Sultany, 2018; Zhou et al., 2015) have used geotagged photos from Flickr. In addition, it is important to highlight that currently one of the most researched topics is sustainability: not only in studies focused on national parks, coastal areas, and other protected areas, but also on climate change (Alieva et al., 2022; Arkema et al., 2021; Barros et al., 2019, 2020; Sottini et al., 2021).

According to Derdouri and Osaragi (2021), researchers are using more photo-based and geotagged social data in tourism research. These user-generated contents (UGC) can be very useful to better understand tourist behaviour at destinations, especially the popular hotspots and coldspots, tourist mobility patterns, and more. But as these authors note, "distinguishing between tourists and locals from this data is problematic since residence information is often not provided" (Derdouri & Osaragi, 2021, p.575).

As mentioned before, scholars of tourism have used different approaches for distinguishing tourists from locals (Derdouri & Osaragi, 2021): the heuristic methods, supervised machine learning (ML) algorithms, and Shannon entropy (SHEN).

Several authors have used the heuristic approaches, choosing a threshold related to the minimum length of stay to identify visitors (Table 1). A time span of 21 days between the first and last photo taken and at least two POIs visited

Table 1: Methods used to distinguish visitors and residents through Flickr photos.

Geographic scale	Authors	Study areas	Method
Urban	(De Choudhury et al., 2010)	Barcelona, Spain; Paris, France; London, UK; San Francisco, New York City, USA.	A time span of 21 days between the first and last photo taken and at least two POIs visited in the same city to be identified as a tourist
	(Straumann et al., 2014)	Zurich	A semi-automated methodology to classify the user location attribute of Flickr user profiles was employed (extract the countries of residence)
	(Kádár & Gede, 2013)	Budapest	Used a threshold of 5 days. If this difference is smaller than 5 days, the user can be considered a visitor; otherwise, he/she is a local.
	(Yan et al., 2017)	Central Philippines islands region	Classification of tourist vs. non-tourist based on user profiles.
Regional	(Kádár & Gede, 2022)	Danube Bend	Locals if they have an interval at least 30 days long or have at least 4 intervals visit; space–time patterns, user data and profile analysis
	(Girardin et al., 2007, 2008)	Province of Florence	Visitor if all photos are taken within a period of 30 days, and inhabitant if photos have an interval greater than 30 days.
	(Chen et al., 2019b)	China	Data screening, text data similarity calculation, geographical location clustering, and time series data modelling.
National	(Önder, 2017)	Austria	A time span of 30 days between the first and last photo taken to identify tourists.
	(Stepchenkova & Zhan, 2013)	Peru	Photos with the tags "Peru" and "travel" considered indicators of images taken by travelers rather than residents
Global	(Wood et al., 2013)	836 sites in 31 countries around the world	User's current location was the origin of each trip and used this information to calculate the proportion of photographers originating from each country.

Source: Adapted from Derdouri & Osaragi (2021).

in the same city identifies the photographer as a tourist, was the range selected by De Choudhury et al. (2010), and a time span of 30 days between the first and last photo taken, was the range chosen to identify tourists by Kádár & Gede (2022), Önder (2017), and Girardin et al. (2007, 2008). Kádár and Gede (2013) considered a threshold of 5 days – i.e., if photos were taken in a period shorter than 5 days, the user is classified as a visitor. These authors state that in an urban context, tourists stay in a destination for a shorter period of time. In the case of Budapest, Kádár and Gede (2013) mention that typical visitors stay on average 3 days, in line with Balińska (2020), who mentions that trips to cities are usually short-term (max 3 nights).

Straumann et al. (2014) employed a semi-automated methodology to classify the user location attribute of Flickr user profiles, extracting the countries of residence. At a global scale, Wood et al. (2013) did the same, considering that the user's current location was the origin of each trip. For Derdouri and Osaragi (2021), ML-based methods neglect factors related to the distinction of visitors and locals, including weather, population density, and the content of users' posts. With the SHEN method, the same authors mention that the dispersal of values of a given variable is calculated. Tourists and locals can be identified by temporal entropy: that is, tourists remain for a brief time in destinations, taking photos for a limited time, and by spatial entropy: that is, tourists and locals have different travel trajectories with different spatial distributions of images. Using these parameters, they can only differentiate two groups.

After presenting the main methods of distinguishing tourists from residents, we can conclude that there is no consensus and optimal method. Time-based approaches present weaknesses such as the different thresholds chosen by researchers or the inclusion of residents that do not share photos with frequency in the group of visitors (Straumann et al., 2014). These authors also highlight that time-based approaches rely on large amounts of data which need a long processing time.

This research aims to present and discuss the methods of identifying visitors and residents in studies using geotagged photos available in Flickr social network.

3 Methodology

This research uses geotagged Flickr photos and their associated user data, from the period between 1 January 2010 and 31 December 2022 in the area of mainland Portugal. For this 12-year period, a total of 1,434,268 photos were identified on Flickr, uploaded by 31,286 users. The "flickr.photos.search" and "flickr.photos.getInfo" methods of the Flickr API were requested by Python script to store data in the MySql database management system: namely, data of geolocalized photographs (P) in mainland Portugal and data regarding the users (U) who submitted those same photographs (Figure 1).

This database will be used in this paper as a basis to describe and discuss the sequence of steps that could be followed by scholars using georeferenced Flickr photos

Figure 1: The flowchart of data collection
Source: Author's construction.

in their tourism investigations. In the following section, the process of cleaning the collected data and the method of identifying visitors and their nationalities will be analysed and described .

4 Results and Discussion

4.1 Data clean

When uploading photos to Flickr, the metadata from the digital camera is also uploaded (Pereira et al., 2011). But users can also place their photos on a map, and Flickr automatically assigns longitude and latitude values. This whole process introduces noise and redundancy that require pre-processing to clean and remove faulty data (Figure 2). For Chen et al. (2019a), Flickr datasets contain erroneous records resulting from malfunctioning or inaccurate hardware, or photos with missing true data. Pre-processing is therefore necessary to clean and remove faulty or irrelevant data. If the purpose of an investigation is to analyse the trajectory of each user, date-related metadata will be a very important factor to consider in order to ensure that the data has a correct timestamp associated with it. For example, if a photo does not contain the EXIF property, Flickr automatically defines when the photo was captured as being equal to the upload date, and this may generate bias in the database. However, as Girardin et al. (2008) highlight, this seems to have little impact in their study.

The different techniques used to perform the data clean are described below, and are summarised in Figure 2. The database of photos taken in mainland Portugal between 2010 and 2022 and available on Flickr, corresponds to 1,144,981 photos shared by 29,890 users (20.2% of photos were deleted).

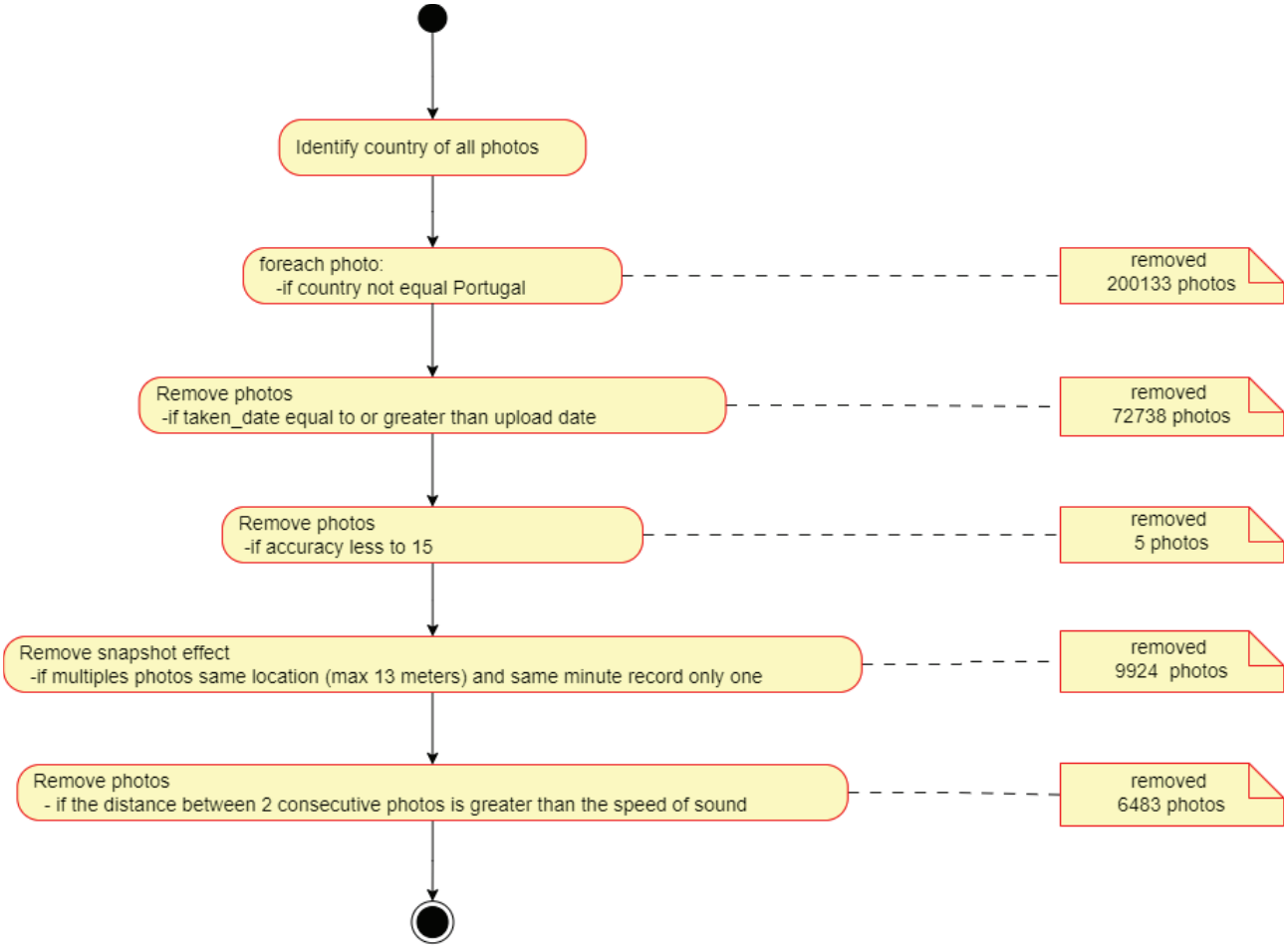

Figure 2: The flowchart of data cleaning
Source: Author's construction.

The following steps were considered:

1. The API flickr.photos.search returns the photos associated with a rectangular space that includes the entire territory of mainland Portugal and some Spanish territory. Considering this fact, we chose to detect the country associated with each geographic point, and then remove the photos associated with Spain. This technique implied the elimination of 200,133 photographs (14.1%).
2. In the database corresponding to photos uploaded to Flickr between 2010 and 2022 in mainland Portugal, the number of photos with the record of the date of capture equal to or greater than the date of upload is 72,738 (5.1%).
3. The Flickr social network also provides information on accuracy attributes, that is, the zoom level of the map used to disclose the location of the photo (Girardin et al. 2007). The accuracy level of the geographic coordinates of each photograph ranges from 1 to 16: world level (1), national level (2-3), regional level (4-6), city level (7-11), and street level (12-16). In order to ensure a higher accuracy of the georeferenced photos, only the accuracy value of 15 and 16 was considered in this work. Therefore, only 5 photos were eliminated.
4. There is also the 'snapshot effect', defined as the bias that can be introduced by 'active users', that is, people who upload a large number of photos (Hollenstein & Purves, 2010). This bias can lead to the emergence of a study area dominated by the behaviours of active users (Hu et al., 2015). To minimise this effect, only one record is retained if multiple images are taken in the same minute. And if the user captures photos at the same time (i.e., in the same minute) in different geographic locations separated by at least 13 metres (Merry & Bettinger, 2019), only the first record is considered. These decisions also result from the fact that this study does not focus on the photographs themselves, but on the photographs as objects that record the user's check-in. Therefore, 9,924 (0.7%) records were deleted.
5. An additional noise factor, mentioned by Cai et al. (2014), corresponds to spatial outliers, that is, extreme longitude and latitude values. In this research, the value of 1,235 km/h was considered as the maximum speed that tourists can humanly reach and which corresponds to the speed of sound. This decision implied the elimination of 6,487 records from our database (0.5%).

4.2 Separating visitors from residents

Separating visitors from residents of the study area is another important step of pre-processing the database. There is no method by which all users can be rigorously identified as visitors or locals, therefore there will always be users who may be misclassified as visitors or locals. However, the existence of a large amount of data always results in a generous sample.

In this research, the authors have decided to use the information provided in each user's profile, as done by Yan et al. (2017), considering that the Flickr user describes the country of residence and the time zone information (see Table 2). Regarding the time zone, it was considered that if the values are different from Portugal (Madeira and Azores), the Flickr user is classified as a visitor. Portugal has two different time zones. In mainland Portugal and the Madeira archipelago, Western European Time (GMT +0) applies. In the Azores archipelago, located west in the Atlantic Ocean, the time zone is GMT-1.

It was also pertinent to analyse the activity of users who uploaded photos of Portugal to their profiles. As we can see in Table 3, a significant percentage of users (26%) uploaded only one photo taken in Portugal, with a reduction in the percentage of users as the number of photos increases. Only 1% of visitors uploaded 12 photos in their profiles. These figures deserve further reflection in future work to understand whether the extreme

Table 2: Algorithm for Flick users' classification

User type	Nr.	%	Rules
Local/resident	2659	9	if(location(user) equal Portugal) then type = Local

Visitor	12144	41	if(location(user) not equal Portugal) then type = Visitor
			OR
			if(timezone(user) not equal Portugal) then type = Visitor

Unknown	15087	50	-
Total	29890	100	-

Source: Author's construction

Table 3: Flickr users' activity (2010-2022)

Number Photos per User	total of Users (%)	total of Visitors (%)	total of Locals (%)
1	26	26	20
2	11	12	10
3	7	7	6
4	5	6	4
5	4	4	3
6	3	3	3
7	3	3	2
8	2	2	2
9	2	2	2
10	2	2	2
11	2	2	1
12	1	1	1

Source: Author's construction

values, that is, users with many photos and users with only 1 or 2 photos, have an influence on the quality of the collected data.

Although the number of users with nationality is already quite large, there are still 50% of users without identification of the area of residence. This can be improved by applying other complementary methods, such as the option followed by Girardin et al. (2008) or Önder (2017): that is, the presence in the area over time as the discriminating factor. These authors explain that this strict threshold was selected in order to capture the real one-time tourists. Both options are correct and if used together could contribute to reducing errors in visitor selection, improving samples quantitatively.

It is important to note that in this second method, the choice of an appropriate threshold will depend on the research objectives and the size of the geographical area under study. However, there is still no consensus in the existing literature (Önder et al., 2016). It is usual, at a city scale, to use a 5-day threshold to consider the user to be a visitor. As explained by Kádár and Gede (2013), this short period correlates better with the stays of typical visitors in urban destinations. In studies on a national scale, it is advisable to choose a longer time span to include international visitors who made a multi-destination visit to the national territory. And in the case of Portugal, there are also large numbers of Portuguese emigrants who usually return to Portugal during the summer months for their holidays, usually staying for up to 30 days. As this is an experimental work, the authors chose not to apply this second method to distinguish resident and visiting users in this research.

For operationalisation purposes, a visitor is defined as "any person travelling to a place other than that of his/her usual environment for less than 12 months and whose main purpose of visit is other than the exercise of an activity remunerated from within the place visited" (UNWTO, 2007, p.444).

Through users' profiles, from a population of 29,890 users (photographers), 12,144 visitors were identified (Table 2).

The graph in Figure 3 shows the monthly distribution of the percentage of photos uploaded by users classified as visitors between 2010 and 2022. It can be observed that the concentration of photos in the summer months (June to September) coincides with the high season of tourist activity in Portugal. The average seasonality rate in Portugal for the period 2016–2022 is 39.49% (TravelBI, 2023). This value measures tourism demand in the three months of highest demand (July, August and September) and as can be seen from Figure 3, the highest percentage of photographs (33%) is also taken during these three months.

4.3 Users' country of residence

Previous studies have used various methods to detect users' place of residence, such as spatial, spatio-temporal and content-based methods (Heikinheimo et al., 2022). The spatial approach uses the maximum number of posts per user (max posts) as the place of residence, as done by Hawelka et al. (2014) with data from Twitter. According to Heikinheimo et al. (2022), centrality measures based on the spatial distribution of geotagged posts are also used in several researches to estimate the

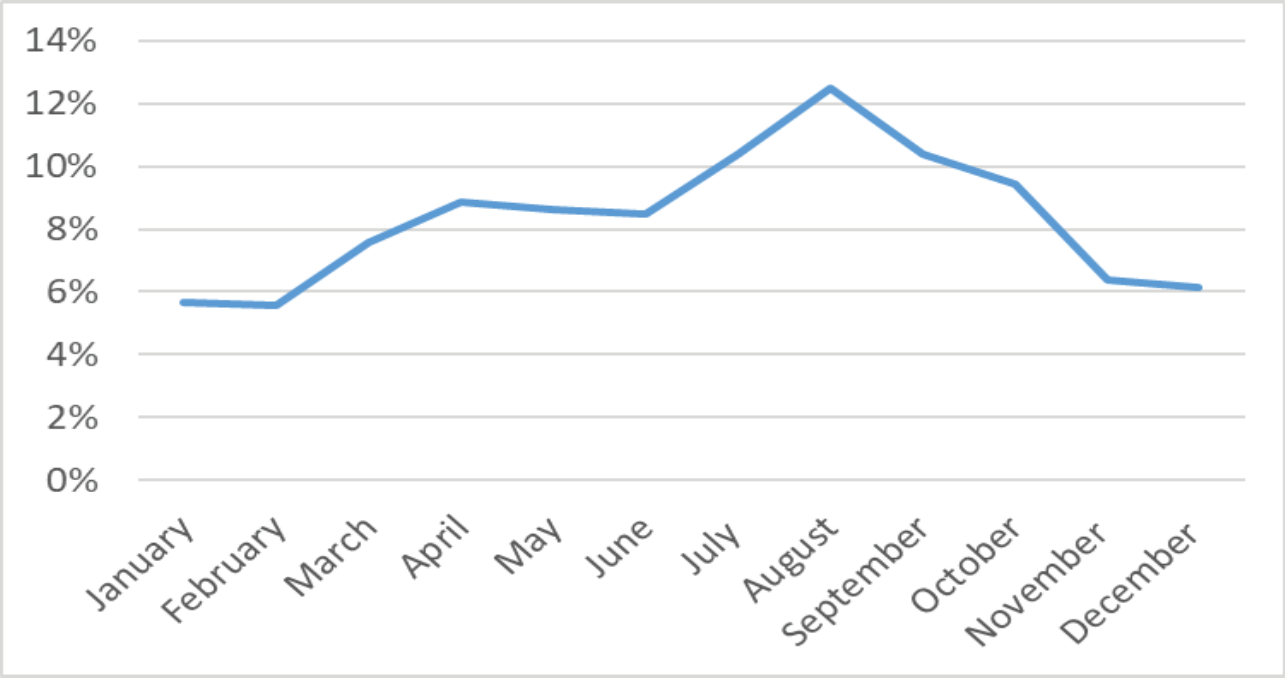

Figure 3: Users' activity by month (2010-2022).
Source: Author's construction

country of residence. In this case, the mean centre or centre of mass correspond to the average (or a weighted average) of the latitude and longitude. If the median centre is selected it minimises the distance to all points and smooths the effect of outlying points compared to the mean (Heikinheimo et al., 2022). Other studies used clustering methods such as the density-based spatial clustering of applications with noise algorithm (DBSCAN) to detect the activity centres of users, as done by Luo et al. (2016) in Chicago.

Through a spatio-temporal approach, scholars can use the Max timedelta method: the place with the longest difference between the date of first and last post; the Max period (a selected period of social media activity differentiates between locals and visitors in the region) as done by Girardin et al. (2008), among others. The content approach uses the self-reported home locations in user profile; or in images or text content.

In this article, the content approach was selected, allowing identification of 7,574 users who specifically indicate the country or city of residence. The time zone displayed by users was also used to detect users' place of residence. This new method, which can be classified as a spatial approach, allows the identification of more than 4,253 user countries. The following table shows the number and percentage of users, by country, who shared photos in Portugal in the timeframe of the study.

According to Table 4, the most representative nationalities of the users who visited Portugal, on average, in the period between 2010 and 2022, are: Spain (21%), United Kingdom (15%), France, United States and Italy with 8% respectively, and Germany (7%). In order to verify the existence of a relationship between the average proportion of users by nationality (2010–2022) and the proportion of average guests for the period 2018–2022, a linear regression was calculated (Figure 4) obtaining a coefficient of determination of 0.9128 [P(F<=f)=0,0135]. If the average proportion of users by nationality for the period 2018–2022 is selected, the linear regression has a coefficient of determination of 0.9024 [P(F<=f)=0,006836].

Although there is no information available on the website of the National Institute of Statistics of Portugal on the number of guests by nationality between 2010 and 2017, it is interesting to note that the main foreign markets according to the number of guests correspond to the most important nationalities in terms of number of foreign Flickr users who visited Portugal. This means that despite the number of Flickr users decreasing worldwide, it is still an excellent and valid research database for measuring tourism demand (Kádár & Gede, 2021).

In addition, to detect the country of residence, it was decided to test two more methods included in the Spatial approach: the Mean centre – the place where the geographic mean of user posts is located; and the Median

Table 4: Flick users by nationality and Guests in Portugal

Country of residence	Number (#) of Flickr users (2010-22)	Users (2010-22) %	Number (#) of Flickr users (2018-22)	Users (2018-22) %	Guests (2018-2022) %
Spain	2001	21	349	20	10,5
United Kingdom	1477	15	306	18	7,9
France	772	8	161	9	6,9
United States	808	8	167	10	5
Italy	771	8	89	5	3
Germany	654	7	147	9	5,4
Brazil	541	6	70	4	4,8
Netherlands	455	5	103	6	2,6
Canada	252	3	41	2	1,4
Belgium	201	2	32	2	1,4
Switzerland	174	2	36	2	1,3
Russia	125	1	18	1	0,6
Ireland	115	1	21	1	1,8
Sweden	103	1	18	1	0,6
Australia	108	1	19	1	0,5
Austria	93	1	16	1	0,5
Poland	77	1	16	1	1
Norway	74	1	15	1	0,3
Denmark	70	1	9	1	0,5
Finland	73	1	15	1	0,3
China	51	1	9	1	1,1
Argentina	46	0	9	1	0,3

Source: Author's construction

centre – the place where the geographic median of user posts is located (Heikinheimo et al., 2022). All photographs with public access were extracted from each user (more than 33 million photos) and through the geographical coordinates provided in the photographs, centrographic measures mean centre and median centre were calculated. Subsequently, the validation of the information obtained was carried out, assuming that the users entered their location correctly. Through the mean and median centre, the degree of correctness of the country of residence was also evaluated for 10 countries, including Portugal (Table 5).

It is concluded that the median method obtains better results**,** in line with the results obtained by Heikinheimo et al. (2022), where the hierarchical median centres method corresponded the most to the official visitor statistics. It can also be observed that in general, the greater the distance between Portugal and the user's country of origin, the greater the error. Spain, UK, France or Italy register the smallest errors and distant countries, such as the USA, Brazil or Canada register lowest accuracy. This is a topic that will deserve further analysis in future studies.

For users classified as visitors who do not have their country of residence registered on Flickr, it is possible to use the median method together with the time zone data available in the user's profile. After the algorithm calculates the median of the geographic coordinates (latitude and longitude), the data obtained is validated: that is, for each user, it is confirmed whether the country resulting from the median coincides with the time zone information entered by the user. If so, the country of residence is considered valid. Subsequently, the name of the country that resulted from the median calculation is compared with the user self-reported home location (Content approach). In this way, the degree of accuracy of the algorithm is validated. These results are presented in Table 6.

In fact, the combination of methods considerably improves the accurate identification of the users' country of origin, with very significant levels of accuracy for countries such as Portugal, Brazil, United States or Spain.

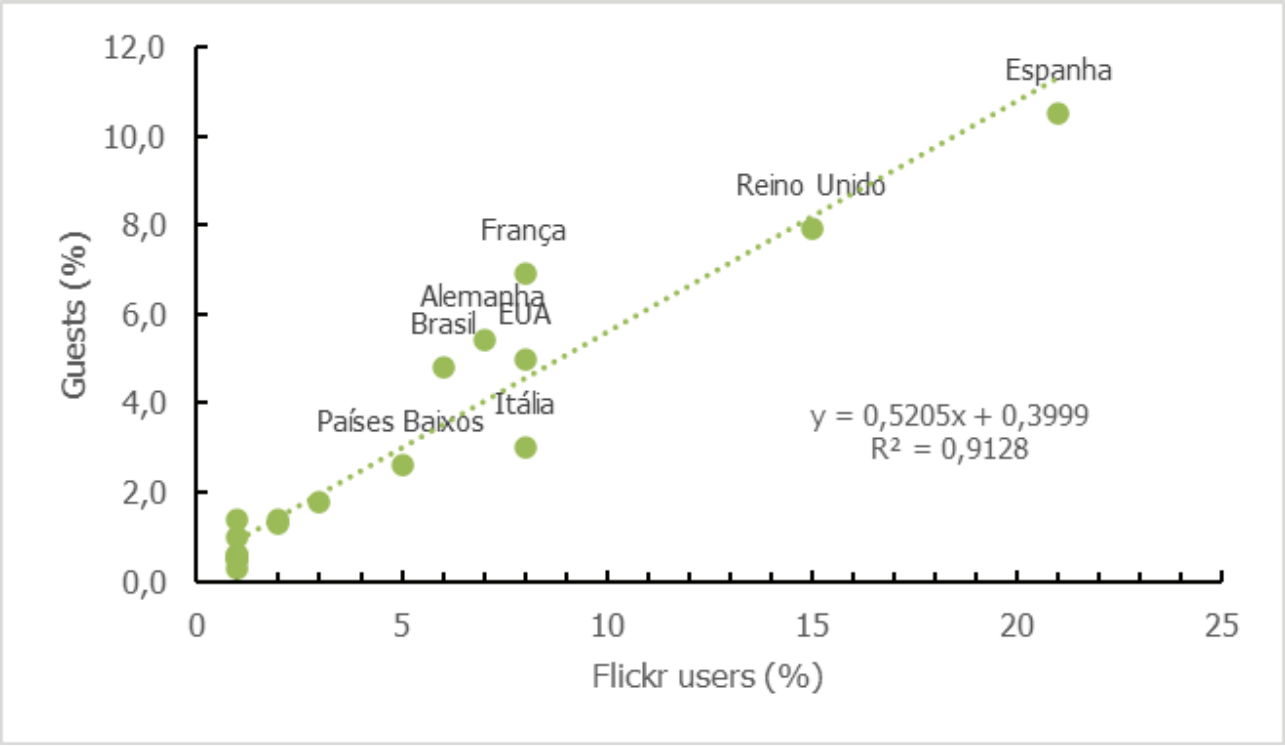

Figure 4: Guests (2018-2021) versus Users (2010-2022) in Portugal
Source: Author's construction

Table 5: Users' country of residence using centrographic measures (mean and median centre)

Countries	Total	Median - Accuracy (%)	Media - Accuracy (%)
Portugal	2659	88%	63%
Spain	2001	83%	56%
United Kingdom	1477	65%	32%
Italy	771	64%	29%
France	772	66%	36%
United States	808	53%	62%
Germany	654	56%	21%
Brazil	541	55%	53%
Netherlands	455	48%	24%
Canada	252	45%	58%

Source: Author's construction

4.4 Group of visitors

It is also possible to analyse the number of visits made by each user/visitor, identifying the first-time and repeat visitors. For this purpose, a spatio-temporal approach was selected using the Max period, that is, a selected period of social media activity differentiates between locals and visitors in the region. Taking into account that Max days corresponds to the maximum time each user stays in Portugal, a threshold of 30 days was selected. Then the number of visits of each user was counted.

Figure 5 shows the number of visits made by each visitor between 2010 and 2022, and it can be seen that 79% of users have visited Portugal only once, which means that for the period under analysis, they are considered first-time visitors.

Table 6: Users' country of residence accuracy combining crossing median centre, time zone, and self-reported home location methods.

Country	Accuracy (%)
Portugal	95%
Spain	94%
United Kingdom	85%
United States	91%
France	84%
Italy	83%
Germany	70%
Brazil	96%
Netherlands	62%
Canada	79%

Source: Author's construction

As can be observed in Table 7, the number of repeaters is relatively low (21%). A more detailed analysis of the number of visits according to the country of residence shows that the Spanish, the main inbound market, are the ones who most repeat their visits to Portugal (31%), which may be justified by geographic proximity.

5 Conclusions

This article presented and discussed the methods of identifying visitors and local population in tourism studies using geotagged photos from the Flickr social network, and distinguishing them according to their country of residence. As far as the authors are aware, this is the first analysis of its kind to use photos taken in Portugal and uploaded by Flickr users to their profiles. One of the main challenges in using social media data in research and decision-making is the question of representativeness (Heikinheimo et al., 2022). However, the analysis carried out on the data collected allows us to conclude that the geotagged photos taken in Portugal and posted on the Flickr social network constitute an interesting, valuable and representative source of information for tourism studies.

After data cleaning, the photos database corresponds to 1,144,981 photos shared by 29,890 users. Using the information provided in each user's profile and the time zone, 12,144 users (41%) were classified as visitors and 2,659 users (9%) as locals. The monthly distribution of the percentage of photos uploaded by users classified as

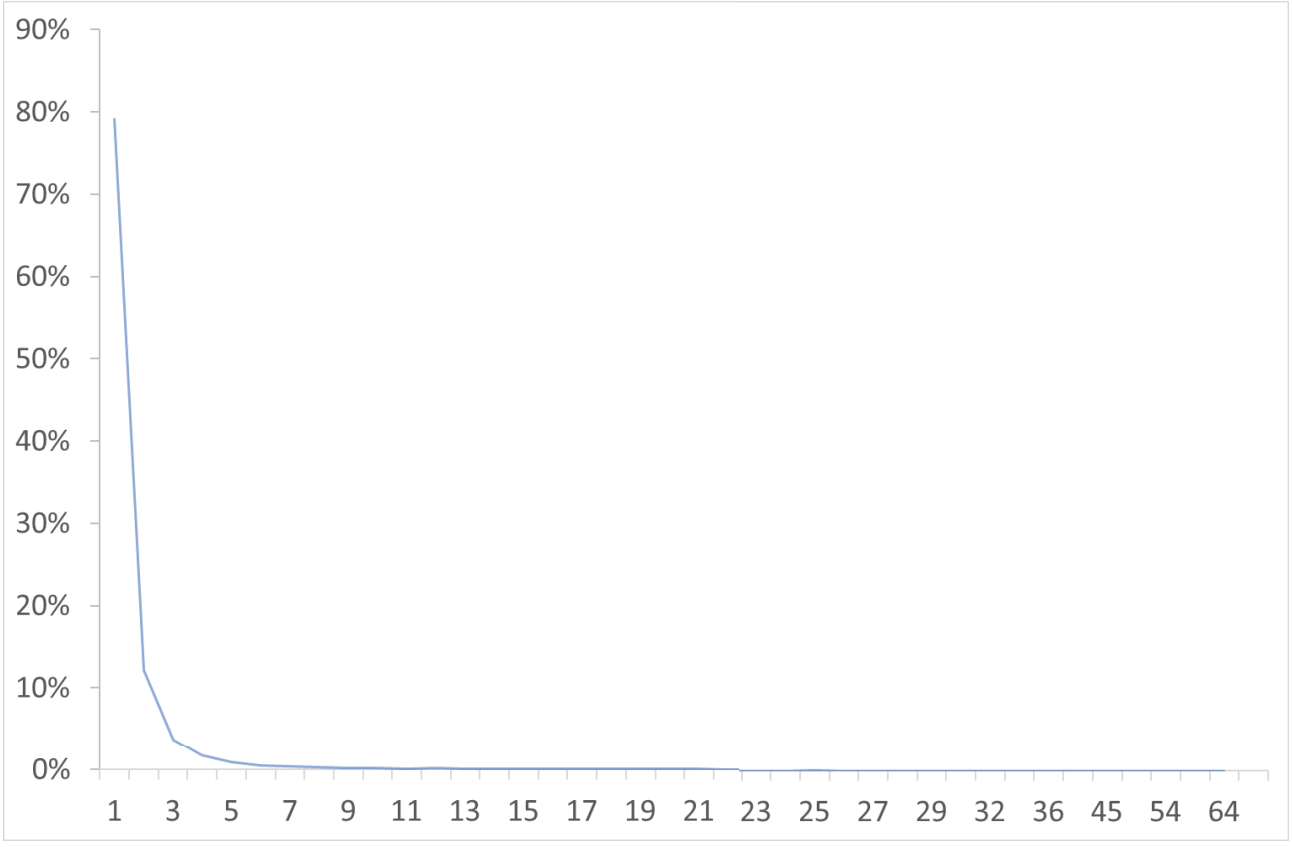

Figure 5: Number of visits by user in Portugal (2010-2022)
Source: Author's construction.

Table 7: Number of visits by Flickr users' country of residence (2010–2022)

Nr. Visits	Total	Spain	United Kingdom	France	United States	Italy
1	79%	69%	77%	81%	83%	87%
2	12%	17%	13%	13%	9%	9%
3	4%	6%	3%	2%	4%	3%
4	2%	3%	2%	2%	2%	1%
5	1%	1%	1%	1%	1%	1%

Source: Author's construction

visitors coincides with the high season of tourist activity in Portugal. Through a content approach, the countries of residence of 7,574 users were also identified. Once again, the data obtained are in line with official statistics on tourism, since the distribution of users by country of residence coincides with the main inbound markets in Portugal: Spain, United Kingdom, France, Germany, Brazil or the United States.

The flowcharts of data collection and data cleaning that were produced throughout this work are a relevant contribution for all scholars who want to use geotagged photos provided by Flickr or any other social network. After data cleaning, the identification of visiting users and resident/local users was carried out. First, the information entered by the users (location and time zone) was used. If this information does not exist, the identification of the country of residence can be obtained by calculating the median of the locations of all the photos published by each user. As shown above, we also considered the time zone information associated with Portugal (Portugal, Azores and Madeira). We can also choose to consider a timeline lower than 30 days/year to identify the number of visits made by each user during the selected period (2010–2022).

Once noise-causing information has been removed and visitors have been identified, scholars can use metainformation from georeferenced photographs to analyse visitors' space-time behaviour at different geographic scales in the Portuguese territory in order to understand who they are, where they go, and how long they stay, among others. The knowledge generated through the use of this important data source also has practical implications for destination management organisations (DMOs), namely for Turismo de Portugal, providing information, for example, on tourist flows in the national territory, both of international and national visitors.

In future, it will be important to explore the possible removal of extreme values, such as the data referring to the number of photos published by each user. The relationship between the number of photos published in Portugal and the total number of photos of the users also deserves to be better explored.

Funding

This work was supported by the National Funds through FCT - Foundation for Science and Technology, I.P., under project Refª UIDB/04470/2020. Thanks to the Centre for Research, Development and Innovation in Tourism (CiTUR).

Conflicts of Interest

The authors of the article "Exploring the Potential of Flickr User-Generated Content for Tourism Research: Insights from Portugal" declare no conflict of interest.

Bionotes

Márcio Ribeiro Martins (Murça, 1979) has a degree in Geography (FLUP, 2002), a Master in Natural Hazards Management (FLUP, 2005) and a PhD in Tourism at the University of Aveiro (DEGEIT, 2020). Since 2021 he is Adjunct Professor at Instituto Politécnico de Bragança (Escola Superior de Administração, Comunicação e Turismo). At the moment, backpacker tourism, tourist space-time behaviour and social media geottagged data are his main research topics. ORCID ID: 0000-0003-3343-3155

Arlindo Santos has been a computer science professor for 25 years at Instituto Politécnico de Bragança, Portugal. Currently a PhD student in computer science at Universidade de Trás-os-Montes e Alto Douro, researching context-aware systems that apply to smart factories. ORCID ID: 0000-0002-7531-9070

References

[1] Alieva, D., Holgado, D., de Juan, S., Ruiz-Frau, A., Villasante, S., & Maya-Jariego, I. (2022). Assessing landscape features and ecosystem services of marine protected areas through photographs on social media: comparison of two archipelagos in Spain. *Environment, Development and Sustainability, 24*(7), 9623–9641. https://doi.org/10.1007/s10668-021-01841-y

[2] Al-Sultany, G. A. (2018). Semantic based geotagged photos similarities for location's ranking purposes. *Journal of Engineering and Applied Sciences, 13*(18), 7716–7720. https://doi.org/10.3923/jeasci.2018.7716.7720

[3] Arkema, K. K., Fisher, D. M., Wyatt, K., Wood, S. A., & Payne, H. J. (2021). Advancing sustainable development and protected area management with social media-based tourism data. *Sustainability, 13*(5), 1–19. https://doi.org/10.3390/su13052427

[4] Balińska, A. (2020). City break jako forma turystyki miejskiej [City break as a form of urban tourism]. *Zeszyty Naukowe Małopolskiej Wyższej Szkoły Ekonomicznej w Tarnowie, 46*(2), 85–95. https://doi.org/10.25944/znmwse.2020.02.8595

[5] Barros, C, Moya-Gómez, B., & García-Palomares, J. C. (2019). Identifying temporal patterns of visitors to national parks through geotagged photographs. *Sustainability, 11*(24), 1-16. https://doi.org/10.3390/su11246983

[6] Barros, C., Moya-Gómez, B., & Gutiérrez, J. (2020). Using geotagged photographs and GPS tracks from social networks to analyse visitor behaviour in national parks. *Current Issues in Tourism, 23*(10), 1291–1310. https://doi.org/10.1080/13683500.2019.1619674

[7] Bettaieb, B., & Wakabayashi, Y. (2021). Comparison of the areas of interest in Central Tokyo among visitors by rountry of residence using geotagged photographs. *Geographical Review of Japan Series B, 93*(2), 66–75. https://doi.org/10.4157/GEOGREVJAPANB.93.66

[8] Broz, M. (2022). Flickr Statistics, User Count, & Facts (September 2022). *Photutorial*. https://photutorial.com/flickr-statistics/

[9] Cai, G., Hio, C., Bermingham, L., Lee, K., & Lee, I. (2014, January 6-9). *Mining Frequent Trajectory Patterns and Regions-of-Interest from Flickr Photos* [Conference Paper]. 2014 47th Hawaii International Conference on System Sciences, Waikoloa, United States of America. https://doi.org/10.1109/HICSS.2014.188

[10] Caldeira, A. M., & Kastenholz, E. (2018). Tourists' spatial behaviour in urban destinations: The effect of prior destination experience. *Journal of Vacation Marketing, 24*(3), 247–260. https://doi.org/10.1177/1356766717706102

[11] Chen, M., Arribas-Bel, D., & Singleton, A. (2019a). Understanding the dynamics of urban areas of interest through volunteered geographic information. *Journal of Geographical Systems, 21*(1), 89–109. https://doi.org/10.1007/s10109-018-0284-3

[12] Chen, W., Xu, Z., Zheng, X., & Luo, Y. (2019b). Geotagged photo metadata processing method for Beijing inbound tourism flow. *ISPRS International Journal of Geo-Information, 8*(12), 1-16. https://doi.org/10.3390/ijgi8120556

[13] De Choudhury, M., Feldman, M., Amer-Yahia, S., Golbandi, N., Lempel, R., & Yu, C. (2010, June 13-16). *Automatic construction of travel itineraries using social breadcrumbs* [Conference Paper]. HT'10 - 21st ACM Conference on Hypertext and Hypermedia, Toronto, Canada. https://doi.org/10.1145/1810617.1810626

[14] Derdouri, A., & Osaragi, T. (2021). A machine learning-based approach for classifying tourists and locals using geotagged photos: the case of Tokyo. *Information Technology and Tourism, 23*(4), 575–609. https://doi.org/10.1007/s40558-021-00208-3

[15] Domènech, A., Mohino, I., & Moya-Gómez, B. (2020). Using Flickr geotagged photos to estimate visitor trajectories in world heritage cities. *ISPRS International Journal of Geo-Information, 9*(11), 1-28. https://doi.org/10.3390/ijgi9110646

[16] Giglio, S., Bertacchini, F., Bilotta, E., & Pantano, P. (2020). Machine learning and points of interest: typical tourist Italian cities. *Current Issues in Tourism, 23*(13), 1646–1658. https://doi.org/10.1080/13683500.2019.1637827

[17] Girardin, F., Dal Fiore, F., Blat, J., & Ratti, C. (2007, November 8-10). *Understanding of tourist dynamics from explicitly disclosed location information* [Conference Paper]. 4th International Symposium on LBS & TeleCartography, Hong Kong. https://www.researchgate.net/publication/228787929_Understanding_of_tourist_dynamics_from_explicitly_disclosed_location_information

[18] Girardin, F., Dal Fiore, F., Ratti, C., & Blat, J. (2008). Leveraging explicitly disclosed location information to understand tourist dynamics: a case study. *Journal of Location Based Services, 2*(1), 41–56. https://doi.org/10.1080/17489720802261138 Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., & Ratti, C. (2014). Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science, 41*(3), 260–271. https://doi.org/10.1080/15230406.2014.890072

[19] Heikinheimo, V., Järv, O., Tenkanen, H., Hiippala, T., & Toivonen, T. (2022). Detecting country of residence from social media data: a comparison of methods. International *Journal of Geographical Information Science, 36*(10), 1931–1952. https://doi.org/10.1080/13658816.2022.2044484

[20] Hollenstein, L., & Purves, R. S. (2010). Exploring place through user-generated content: Using Flickr tags to describe city cores. *Journal of Spatial Information Science, 1*(2010), 21–48. https://doi.org/10.5311/JOSIS.2010.1.3

[21] Höpken, W., Müller, M., Fuchs, M., & Lexhagen, M. (2020). Flickr data for analysing tourists' spatial behaviour and movement patterns: A comparison of clustering techniques. *Journal of Hospitality and Tourism Technology, 11*(1), 69–82. https://doi.org/10.1108/JHTT-08-2017-0059

[22] Hu, Y., Gao, S., Janowicz, K., Yu, B., Li, W., & Prasad, S. (2015). Extracting and understanding urban areas of interest using geotagged photos. *Computers, Environment and Urban Systems, 54*(2015), 240–254. https://doi.org/10.1016/j.compenvurbsys.2015.09.001

[23] Jing, C., Dong, M., Du, M., Zhu, Y., & Fu, J. (2020). Fine-grained spatiotemporal dynamics of inbound tourists based on geotagged photos: A case study in Beijing, China. *IEEE Access, 8*(2020), 28735–28745. https://doi.org/10.1109/ACCESS.2020.2972309

[24] Kádár, B. (2014). Measuring tourist activities in cities using geotagged photography. *Tourism Geographies, 16*(1), 88–104. https://doi.org/10.1080/14616688.2013.868029

[25] Kádár, B., & Gede, M. (2013). Where do tourists go? Visualizing and analysing the spatial distribution of geotagged photography. *Cartographica: The International Journal for Geographic Information and Geovisualization, 48*(2), 78–88. https://doi.org/10.3138/carto.48.2.1839

[26] Kádár, B., & Gede, M. (2021). Tourism flows in large-scale destination systems. *Annals of Tourism Research, 87*(2021), 1-16. https://doi.org/10.1016/j.annals.2020.103113

[27] Kádár, B., & Gede, M. (2022). The measurable predominance of weekend trips in established tourism regions—The case of visitors from Budapest at waterside destinations. *Sustainability, 14*(6), 1-16. https://doi.org/10.3390/su14063293

[28] Lim, K. H., Chan, J., Leckie, C., & Karunasekera, S. (2018). Personalized trip recommendation for tourists based on user interests, points of interest visit durations and visit recency. *Knowledge and Information Systems, 54*(2), 375–406. https://doi.org/10.1007/s10115-017-1056-y

[29] Luo, F., Cao, G., Mulligan, K., & Li, X. (2016). Explore spatiotemporal and demographic characteristics of human mobility via Twitter: A case study of Chicago. *Applied Geography, 70*(2016), 11–25. https://doi.org/10.1016/j.apgeog.2016.03.001

[30] Mariani, M., & Baggio, R. (2022). Big data and analytics in hospitality and tourism: a systematic literature review. *International Journal of Contemporary Hospitality Management, 34*(1), 231–278. https://doi.org/10.1108/IJCHM-03-2021-0301

[31] Martins, M. R., da Costa, R. A., & Moreira, A. C. (2022). Backpackers' space–time behavior in an urban destination: The impact of travel information sources. *International Journal of Tourism Research, 24*(3), 456–471. https://doi.org/10.1002/jtr.2514

[32] Martins, M., & Costa, R. (2022). Tracking technologies in tourism: A bibliometric and content review. In J. V. de Carvalho, P. Liberato & A. Peña (Eds.), *Advances in Tourism, Technology and Systems* (pp. 215–230). Springer Nature Singapore.

[33] Merry, K., & Bettinger, P. (2019). Smartphone GPS accuracy study in an urban environment. *PLoS ONE, 14*(7), 1–19. https://doi.org/10.1371/journal.pone.0219890

[34] Miyasaka, T., Oba, A., Akasaka, M., & Tsuchiya, T. (2018). Sampling limitations in using tourists' mobile phones for GPS-based visitor monitoring. *Journal of Leisure Research, 49*(3–5), 298–310. https://doi.org/10.1080/00222216.2018.1542526

[35] Önder, I. (2017). Classifying multi-destination trips in Austria with big data. *Tourism Management Perspectives, 21*(2017), 54–58. https://doi.org/10.1016/j.tmp.2016.11.002

[36] Önder, I., Koerbitz, W., & Hubmann-Haidvogel, A. (2016). Tracing tourists by their digital footprints: The case of Austria. *Journal of Travel Research, 55*(5), 566–573. https://doi.org/10.1177/0047287514563985

[37] Padrón-Ávila, H., & Hernández-Martín, R. (2020). How can researchers track tourists? A bibliometric content analysis of tourist tracking techniques. *European Journal of Tourism Research, 26*(2020), 1–30. https://doi.org/10.54055/ejtr.v26i.1932

[38] Pereira, F. C., Vaccari, A., Giardin, F., Chiu, C., & Ratti, C. (2011). Crowdsensing in the Web: Analyzing the citizen experience in the urban space. In M. Foth, L. Forlano, C. Satchell, & M. Gibbs (Eds.), *From Social Butterfly to Engaged Citizen: Urban Informatics, Social Media, Ubiquitous Computing, and Mobile Technology to Support Citizen Engagement*. The MIT Press. https://doi.org/10.7551/mitpress/8744.003.0029

[39] Salas-Olmedo, M. H., Moya-Gómez, B., García-Palomares, J. C., & Gutiérrez, J. (2018). Tourists' digital footprint in cities: Comparing Big Data sources. *Tourism Management, 66*(2018), 13–25. https://doi.org/10.1016/j.tourman.2017.11.001

[40] Sarkar, J. L., & Majumder, A. (2021). A new point-of-interest approach based on multi-itinerary recommendation engine. *Expert Systems with Applications, 181*(2021), 115026. https://doi.org/10.1016/j.eswa.2021.115026

[41] Shoval, N., Schvimer, Y., & Tamir, M. (2018). Tracking technologies and urban analysis: Adding the emotional dimension. *Cities, 72* (Part A), 34–42. https://doi.org/10.1016/j.cities.2017.08.005

[42] Solazzo, G., Maruccia, Y., Lorenzo, G., Ndou, V., Del Vecchio, P., & Elia, G. (2022). Extracting insights from big social data for smarter tourism destination management. *Measuring Business Excellence, 26*(1), 122–140. https://doi.org/10.1108/MBE-11-2020-0156

[43] Sottini, V. A., Barbierato, E., Bernetti, I., & Capecchi, I. (2021). Impact of climate change on wine tourism: An approach through social media data. *Sustainability, 13*(13), 1-18. https://doi.org/10.3390/su13137489

[44] Spyrou, E., Korakakis, M., Charalampidis, V., Psallas, A., & Mylonas, P. (2017). A geo-clustering approach for the detection of areas-of-interest and their underlying semantics. *Algorithms, 10*(1), 1-22. https://doi.org/10.3390/a10010035

[45] Stepchenkova, S., & Zhan, F. (2013). Visual destination images of Peru: Comparative content analysis of DMO and user-generated photography. *Tourism Management, 36*(2013), 590–601. https://doi.org/10.1016/j.tourman.2012.08.006

[46] Straumann, R. K., Çöltekin, A., & Andrienko, G. (2014). Towards (re)constructing narratives from georeferenced photographs through visual analytics. *Cartographic Journal, 51*(2), 152–165. https://doi.org/10.1179/1743277414Y.0000000079

[47] Sun, X., Huang, Z., Peng, X., Chen, Y., & Liu, Y. (2019). Building a model-based personalised recommendation approach for tourist attractions from geotagged social media data. *International Journal of Digital Earth, 12*(6), 661–678. https://doi.org/10.1080/17538947.2018.1471104

[48] Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., & Li, L.-J. (2016). YFCC100M: the new data in multimedia research. *Communications of the ACM, 59*(2), 64–73. https://doi.org/10.1145/2812802

[49] UNWTO. (2007). *UNWTO Metadata project: Common Glossary*. http://statistics.unwto.org/sites/all/files/docpdf/glossary.pdf TravelBI. (2023). *Taxa de Sazonalidade [Seasonality Rate]*. https://travelbi.turismodeportugal.pt/sustentabilidade/taxa-de-sazonalidade/

[50] Wood, S. A., Guerry, A. D., Silver, J. M., & Lacayo, M. (2013). Using social media to quantify nature-based tourism and recreation. *Scientific Reports, 3*(2013), 1-7. https://doi.org/10.1038/srep02976

[51] Yan, Y., Eckle, M., Kuo, C.-L., Herfort, B., Fan, H., & Zipf, A. (2017). Monitoring and assessing post-disaster tourism recovery using geotagged social media data. *ISPRS International Journal of Geo-Information, 6*(5), 1-17. https://doi.org/10.3390/ijgi6050144

[52] Yun, H. J., & Park, M. H. (2014). Time–space movement of festival visitors in rural areas using a smart phone application. *Asia Pacific Journal of Tourism Research*, *20*(11), 1–20. https://doi.org/10.1080/10941665.2014.976581

[53] Zhou, X., Xu, C., & Kimmons, B. (2015). Detecting tourism destinations using scalable geospatial analysis based on cloud computing platform. *Computers, Environment and Urban Systems, 54*(2015), 144–153. https://doi.org/10.1016/j.compenvurbsys.2015.07.006