Exploring geo-tagged photos to assess spatial patterns of visitors in protected areas: the case of park of Etna (Italy)

Giovanni Signorello\textsuperscript{a,c}, Giovanni Maria Farinella\textsuperscript{b,c}, Lorenzo Di Silvestro\textsuperscript{c}, Alessandro Torrisi\textsuperscript{c}, Giovanni Gallo\textsuperscript{b,c}

\textsuperscript{a}Department of Agriculture, Food and Environment (Di3A)
\textsuperscript{b}Department of Mathematics and Computer Science (DMI)
\textsuperscript{c}Centre for the Conservation and Nature and Agroecosystems (CUTGANA)

University of Catania
Catania, Italy
\{g.signorello,aletorrisi\}@unict.it, \{gfarinella,disilvestro,gallo\}@dmi.unict.it

Abstract—In this paper we use the georeferenced images publicly available on Flickr platform as a source of information to monitor visitors in nature areas. In particular, we propose and evaluate a method composed by three main steps to perform the analysis of social images related to natural park of Mount Etna. At first metadata of the georeferenced images are explored to identify trends, patterns and relationships among the information acquired. Then data mining techniques are used to generate a traveling model. Association rules that highlight locations visited jointly are identified using the Apriori data mining technique. Finally, we consider places that are likely to be jointly visited and analyze the tourist flow from one to another.

Index Terms—Georeferenced Data, Digital Footprints, Flickr, Recreation, Protected Areas, Data Mining, Apriori Algorithm

I. INTRODUCTION

Monitoring affluence, temporal and spatial patterns of visitors’ flow in nature areas is mandatory to improve management of natural capital, to estimate the social and economic value of recreation and to assess the impact of visitors on environment. This information is traditionally acquired by administering questionnaires to visitors in-situ or ex-situ, or by using other counting methods [1]. These data gathering techniques are cost and time consuming, and limited in terms of spatial and temporal coverage [2].

For the aforementioned reasons we need to exploit another source of data. The recent proliferation of web platforms for photo-sharing comes in handy, since photography is one of the most common attributes of tourist behavior [3]. In addition the increasing use of smartphones and digital cameras that natively supports GPS sensors allows tourist to share georeferenced contents producing what Goodchild et al. [4] define as volunteered geographic information (VGI), a special case of the more general web phenomenon of user-generated content (UGC).

Among online platforms that allows users to publicly share their photographs, Flickr [5] is one of the most popular. It shows a growth for both the number of registered users and the amount of publicly available data. Since its launch in 2004, over 75 million of registered users uploaded more than 6 billion photos on the platform [6]. In 2006 Flickr platform enabled users to pin-point the geographical location of their photos on a map. One year after the inclusion of this functionality over 20 million of geotagged photos were hosted on the website [7]. Some years after more than 197 million of photos upload to Flickr are geo-tagged, due to the exponential use of smartphones that georeferences automatically taken photos [8]. Moreover, Flickr include information about the user (photographer), i.e. gender and nationality.

Mining Flickr images gives to the analyst the chance to learn useful clues about the behavior of tourists. In this paper we introduce a simple model to learn tourist behavior by exploiting Flickr data. We process the information on the georeferenced images to infer the tourist flow through a graph representation whose nodes are related to portions of the monitored area. The area is subdivided into a discrete grid and each cell grid is used as a node in our graph representation. All images are associated with a cell grid according to GPS coordinates. Two cells, i.e., two nodes in the graph, are connected by a weighted edge: the weight is proportional to the number of single Flickr users that have taken photos in both cells. Each connection on the graph is hence equivalent to a likely route covered by tourists.

This method allows us to gain information about main routes followed by tourists during their visits. Statistical association rules that model tourists’ movements among neighbors sites are obtained using the Apriori data mining algorithm [9]. Most followed paths are then selected for further analysis. In particular, given a path of interest, we want to understand which route is selected by tourists to move from one node of the graph to another. Find out factors that influence tourists when they choose a specific path is very relevant to infer more clues about their preferences. This information is useful to tourist managers to conduct planning activities regarding the transportation system, tour operators tasks and positioning of accommodations and info-points.

The remainder of the paper is organized as follows: in Section II we present relevant research works that use social media.
data in the context of tourists behavior analysis; in Section III we describe the work-flow of the proposed method; Section IV details the application of the proposed approach on real data. Finally, Section V draws the conclusion and describe some possible future directions for research.

II. RELATED WORKS

There are, by now, several popular web photo sharing services (Flickr [5], Instagram [10], 500Pixels [11], etc.). Among these we choose to use Flickr data because of the rich information for each picture uploaded on this platform and for the availability of suitable Application Program Interface (API) for data extraction. The most relevant information we look at is of course the geographical coordinates data. Geo-tagging error of a photo taken with a modern mobile phone is less than 10m [12] and in [13] is showed that Flickr data are spatially accurate. Besides location information, the date when each picture has been taken and uploaded together with information about the user (age, nationality, gender, etc.) are also collected. The importance of data analysis techniques in the e-tourism domain is discussed in [14]. This kind of systems store travel plans of users as positive examples, then during a travel planning session the system suggest analogue cases to the one under consideration. The similarity function uses all the available information, including the user’s profile details. This leads to a ranked list of highly recommended results. The authors of [15] present a method to capture travel information from geotagged photos on Flickr. First, they conduct a cluster analysis in order to identify the major areas of interest visited by inbound tourists in Hong Kong. Then, a reconstruction of tourist movements is obtained according to the time information of the pictures extracted on a daily basis. To enrich this information with a model of tourists flow from a location to another, a Markov model is employed. The authors in [16] propose a recommendation system by combining topic models and Markov chains. Their final tourist behavior model provide a set of personalized travel routes that match the user’s current location, user’s interest, and user’s spare time. Another travel route recommendation system is proposed in [17] where the authors describe an approach to build structured models of human travel as a function of spatially-varying latent properties of locations and travel distances, analyzing affinities between locations. Similarly to what is proposed in [15], the authors make the assumption that human travel can be treated as a Markov process. In addition, individuals are grouped into clusters with distinct travel models. Another relevant work in the context of this paper is the one described in [18] where the authors propose to use Flickr data for building a tool to analyze the distribution of tourists with respect to gender, nationality (e.g., domestic vs foreign), also considering the date of the visit (e.g., distribution of pictures for month/year). The work considers also the estimation of a probability density function to better locate most popular sites. The authors of [19] used Flickr geotagged images taken in specific lakes to perform a regression analysis to determine which lake attributes are the best indicators for lake visitation rate and travel costs. They discovered a relationship between visitors’ affluence and lake’s water clarity.

III. PROPOSED METHOD

In this section we detail the method used to extract insights about tourists’ behavior from images. Firstly, we describe how the image dataset has been collected from Flickr, then we describe our visual approach for data exploration. Extracted data are then exploited as source for different data analysis’ task. Taking into account a grid subdivision of the area of interest, we propose to build a graph of the most popular routes followed by tourists during their trips. The model obtained is hence used to reason about tourists’ flow.

A. Data extraction and exploration

Social media platforms generate a huge amount of data every day. An automated procedure to have a constant access to these data for tourism management purposes is therefore very helpful. In addition to the browsing mode, Flickr allows users to download public data through its web based API [20]. The first thing to do in order to extract data from a social platform is the selection of an area of interest. Flickr API’s allow two ways to retrieve geotagged pictures. The first way retrieves pictures located inside a circle specified by latitude and longitude of its center and a radius. The other way is to retrieve the pictures located inside a bounding box specified with the coordinates of the bottom-left and top-right corners. There is however an issue: Flickr server does not provide more than about 4000 distinct records for each query. This led us to adopt a simple recursive subdivision strategy using the bounding box approach. In particular, the initial box that includes the entire area of interest is splitted quadtree-likewise until we have a reasonable confidence that all the pictures in the area have been taken. This first gathering of records is refined and integrated using other Flickr API’s methods. In particular, using flickr.photos.getInfo and flickr.people.getInfo, we complete the records of each picture with more information about time, location and about gender and provenience of the photographer Latitude-longitude pairs provide a representation of spatial distribution of photographers inside the area of interest.

B. Data mining techniques to analyze visitors’ spatial patterns

We take into account the location where pictures have been acquired as well as time information of collected images. In this way we are able to reconstruct the tourist trip considering the sites they have visited and the duration of their trip (e.g., daily or weekly excursion). For this kind of analysis the density distribution described in the previous sub-section is not very helpful. What is needed here is some kind of “discretization” of the locations inside the area of interest. We hence build a list of “sites” that in this context are simply clusters of the locations where the pictures have been taken. These clusters are obtained with a naive, but
We employ the Apriori algorithm [9]. In our settings, for a given rule \( s^1_{(lon,lat)} \rightarrow s^2_{(lon,lat)} \), confidence is proportional to the likelihood that the site \( s^1_{(lon,lat)} \) is visited during the same trip of a photographer who has visited the site \( s^2_{(lon,lat)} \). The last step of our analysis takes into account one at each time the strongest association rules generated by the Apriori algorithm. The aim is to get more details on the tourist flow generated between nodes (sites) of the obtained association rules. In other words, it would be useful to have a comprehensive knowledge of the routes covered by tourists to move from one site of interest to another one along these “most travelled paths”. They could follow the main path, others may choose auxiliary routes. For example, a tourist could get away from the main route and choose a customized way as it presents particular naturalistic details that he wants to see and photograph. The easiest and quickest way to get this type of information involves the use of buffers in ArcGIS [21]. From the geometric point of view, the buffer is a polygon whose perimeter identifies a territorial area which is located at a distance with respect to the path of interest, between a minimum and a maximum value. Our analysis includes the creation of buffers of increasing size from the main path related to association rules to locate different bands of territory and see the evolving of the tourist flow in these areas.

IV. CASE STUDY

A. Dataset

The case study considered here is Mount Etna, the highest active volcano in Europe. The volcano is a natural park of approximately 59,000 ha and it is included in the World Heritage List of UNESCO [22]. The overall natural environment is very attractive for visitors from all the world for naturalistic aspects, volcanic phenomena and outdoor recreational activities. The bounding box for data extraction from Flickr includes the entire protected area of Etna. Flickr allows to add a search filter based on the date on which the photos have been taken. For our study we took into account photos posted between 2006 and 2016. Another parameter that could be used in the search is “accuracy”: it is a measure of how accurate are the GPS coordinates of the image with respect to the point where the image was taken. We have conducted different experiments with different levels of accuracy. It can take any value between 1 (World level) and 16 (Street level). Filtering images using the accuracy parameter has the only effect of reducing, with higher levels, the number of retrieved images. It does not affect the quality of the visual content in the images as we have observed in few experiments. Since for the present application to have large datasets is relevant we choose the minimum accuracy level. The final dataset is composed by 30,692 images taken by 2932 different photographers.

B. Analysis of tourist movements

The next step of our proposal consists in analyzing the tourist traffic within the considered area. As described in

Fig. 1. Reconstruction of tourists trips. Yellow points represent the images taken by a photographer inside the area of interest. Each image is associated to a cell of the grid taking into account its GPS coordinates. The trip of the photographer is then reconstructed using the temporal sequence of the shot images.

effective, method: the area of interest is partitioned into square cells of uniform dimension. The pictures belonging to a “site” are those whose coordinates fall within one of these cells. We keep track of the trips taking into account the temporal sequence of images of each photographer (Figure 1). For all the trips that have been carried out on the area of interest in a specific time interval by the same photographer, we obtain a weighted diagram of the routes followed by the photographers. The proposed method is formally modeled by a directed graph.

We define a graph \( G = (N, E) \) composed by \( N \) different nodes and \( E \) edges. An oriented edge from \( i \) to \( j \) indicates tourist flow from the site \( i \) to the site \( j \). Two main parameters are involved to model the grid graph:

- \( M \): it represents the granularity of the grid. The higher the value of \( M \), the greater the number of nodes in the grid graph (while the area size of the sites of interest decreases). In the example in Figure 1, \( M \) is equal to 5.
- \( \Delta t \): it indicates the duration of the trip. In this way we can choose to model trips that have a daily or weekly basis.

To understand data an analyst may look to the graph to have a comprehensive view of all the paths followed by tourists inside the selected area. A limit in this kind of visualization is that it erases the “sequentiality” of each photographer visit to the cells. Some tourists could stop at a site, others might continue for one or more sites. Understanding how much a particular path (composed by 2 or more nodes) is taken by travelers is a useful information to tourism managers to conduct planning and tourism management of the territory. Because of this we need to extract association rules between sets of nodes that have a certain probability of being jointly visited during a trip. To learn association rules among tourist sites we employ the Apriori algorithm [9]. In our settings, for
Section III-B, the area of interest is subdivided into a $M \times M$ cells grid. The value of $M$ should be chosen carefully. To this aim, experts of the area of interest may consulted. We tried three granularity levels: $M = 3, 5, 7$. The best results have been obtained with $M = 5$. Experts have confirmed that a $3 \times 3$ grid has too large cells and loses important details. On the other hand a $7 \times 7$ grid allows too many empty cells and makes further analysis unnecessarily complex.

The second parameter we need to model the trajectories followed by tourists takes into account the duration of the trip measured in days. The longer the trip, the higher is the probability that completely unconnected areas are jointly visited. Different choices have been considered and we experimentally selected only trips of maximum 7 days. Figure 2 shows the graph obtained considering the method of Section III-B.

The weight of each edge indicates how many photographers have covered the route. To make the graph more readable, we reported only arcs that have a weight $\geq 10$. As can be seen from the graph in Figure 2, tourist traffic is focused on the central part of the map which correspond to the path that brings tourists to the top of the volcano.

With the Apriori algorithm we obtained two association rules which confirm that the edge between the red nodes in Figure 2 is the most travelled one. In particular the two symmetric rules that characterize the path between the two red nodes have $Support = 8\%, Confidence = 22\%$ in one case and $Support = 8\%, Confidence = 25\%$ in the other case.

In the final step of our study we zoom on the edge between the red nodes. The idea is to analyze which paths are chosen during the crossing between these two sites. The Etna paths are developed on recent and historical lava flows, in wooded areas or in areas without tree vegetation. These often present a variety of slopes resulting from the changing morphology of the volcano. The paths that we focus on are those that connect “Rifugio Sapienza” (node $(4, 3)$ in the graph) with “torre del filosofo” (node $(3, 3)$ in the graph). More specifically, “Rifugio Sapienza” at about 2000mt above the sea level is an area easily reached with cars or buses and local guides offer many trekking opportunities toward the higher areas starting from there. “Torre del filosofo” (english: Philosopher tower) at about 2900mt above the sea level is the mytical place where Heraclitus killed himself jumping into the lava flow. More realistically is a hill from where the main crater at 3000mt above the sea level can be easily and safely observed. To investigate tourist behaviors in this area we have identified what is the main route taken by the guides using ArcGIS. We have considered four different buffers around the main path (Figure 3(a)). We have considered a maximum radius of 200 meters from the center of the path, with incremental steps of 50 meters. Since the GPS coordinates of acquisition devices may differ by several meters from the actual value, the first buffer covers a distance of 50 meters around the main path. Figure 3 shows the details about the areas belonging to each buffer. The box in Figure 3(a) 7259 images. Most of the pictures are located outside the path of interest and the four buffers. This means that tourists move largely away from the main road to explore the surroundings. The photos within the four buffers are only 2333. The first and the second buffers contain both the 32% of the whole set of images and then this percentage decreases to 19% in the third buffer and to 17% in the last one.
C. Discussion

The data taken into account in the case study reveal, through the suggested tool, some interesting insights. Our analysis confirms that the tourist population on Mount Etna is mainly distributed on the eastern and southern versants of the volcano. Although Etna area has been interested by a continuous tourism development, its versants have different degrees of anthropization. The west side of the mountain is strongly linked to traditional agricultural activities; it represents an area of territory less developed from the tourism point of view although it is the most suitable to practice a tourism regarding naturalistic aspects. The southern and eastern sides, on the contrary, offer greater receptive hospitality structures. The presence of nature trails and ski facilities further contributes to making this area the most popular choice among tourists. The seasonal distribution of visits helps to understand when the highest presence of tourists is realized. Our analysis confirms the seasonal nature of the Etna territory. It highlights that the winter tourism (from November to March) is practiced almost exclusively by winter sports enthusiasts. They usually take less photographs than a tourist who prefers a hiking tourism feasible mostly from April to October. The second part of our methodology produces a travel model of the tourism movements within the area of interest.

With the buffer analysis we have noticed there is considerable amount of pictures taken outside the main path of interest. This indicates that many tourists customize their way to go to the north part of the volcano. Analysis of the visual content of their images may help to see which naturalistic aspect influence them to follow a particular route. This will be object of future Computer Vision research.

V. Conclusion

In this paper we have proposed a framework that include different methodologies for data analysis that is useful to explore tourists’ flow in an area of interest and extracts useful clues that helps the tourism managers to conduct a reliable tourism planning activity. In particular, we used for our analysis data available on social media platforms. Indeed, the images collected by social media contain metadata that can provide details about the number of tourists present in an area of interest, along with additional information on their gender, nationality, etc. By using these information, we have built a tool that analyzes the spatial distribution of the tourists. Using data mining techniques we have also proposed a method to infer tourist flow within the area of interest. The method has been applied to monitor recreation in park of Etna. Findings support and confirm the experts’ knowledge about this region. We believe that further investigation and potential applications of our tools can be replicated in other case studies. Future work will consider social media data to build new and automatic tools for the analysis of tourism. The study could be enriched with the addition of data from other types of social platforms. We are also interested in considering the automatic understanding of the visual content of images using Computer Vision algorithms. This may improve our study with details about different types of tourists detected within the area of interest which can be clustered according to which natural features they prefer to see and photograph.

REFERENCES