Whom to Follow? A Comparison of Walking Routes Computed Based on Social Media Photos from Different Types of Contributors

Matan Mor[1] ✉ [0000-0003-4658-2328], Johannes Oehrlein[2] [0000-0003-0478-4298], Jan-Henrik Haunert[2] [0000-0001-8005-943X] and Sagi Dalyot[1] [0000-0002-5639-8009]

[1] Mapping and Geoinformation Engineering, The Technion, Haifa, Israel
[2] Geoinformation Group, University of Bonn, Bonn, Germany
matan.mor@campus.technion.ac.il

Abstract. Since many tourists share the photos they take on social media channels, large collections of tourist attraction photos are easily accessible online. Recent research has dealt with identifying popular places from these photos, as well as computing city tourism routes based on these photo collections. Although current approaches show great potential, many tourism attractions suffer from being overrun by tourists, not least because many tourists are aware of only a few tourism hot spots that are trending. In the worst case, automatic city route recommendations based on social media photos will intensify this issue and disappoint tourists who seek individual experiences. In the best case, however, if individual preferences are appropriately incorporated into the route planning algorithm, more personalized route recommendations will be achieved. In this paper, we suggest distinguishing two different types of photo contributors, namely: first-time visitors who are usually tourists who "follow the crowd" (e.g., to visit the top tourist attractions), and repeated visitors who are usually locals who "don’t follow the crowd" (e.g., to visit photogenic yet less well-known places). This categorization allows the user to decide how to trade the one objective off against the other. We present a novel method based on a classification of photographers into locals and tourists, and show how to incorporate this information into an algorithmic routing framework based on the Orienteering Problem approach. In detailed experiments we analyze how choosing the parameter that models the trade-off between both objectives influences the optimal route found by the algorithm, designed to serve the user’s travel objective and preferences in terms of visited attraction types.

Keywords: Tourism route recommendation · Orienteering problem · Social media · Geotagged photos · User-generated content classification.

AGILE: GIScience Series, 1, 2020.
Full paper Proceedings of the 23rd AGILE Conference on Geographic Information Science, 2020.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras
This contribution underwent peer review based on a full paper submission.
https://doi.org/10.5194/agile-giss-1-16-2020 | © Authors 2020. CC BY 4.0 License.

1 Introduction

Crowdsource geotagged user-generated data and information sources are increasing dramatically; social media websites, such as Twitter and Flickr, are becoming more commonly used to document and share daily activity and experiences. Flickr, for example, allows users to depict and share their everyday activities and events relying on geotagged photos. Mining crowdsource geolocations on urban attractiveness that rely on the photogenic nature of places can be effectively used to recommend attractive routes. Accordingly, trajectories of photographers who share their photos on social media sites, reflect the urban attractivity and can provide informed landmark information and routes that infer current public interest [1].

Travel behaviour of tourists and locals can be very unalike; in particular, while local users usually know the area and collect photos from all over the city, tourists are more oriented towards popular landmarks and famous areas [2]. Furthermore, tourists are more active and travel more over short periods and would prefer to explore new places by different modes of travel - rather than walking - to maximize their travel experience [3]. Moreover, the frequency in which tourists use social media platforms to document famous attractions exceeds that of locals [4].

We hypothesize that computing routes based on social media photos from different types of contributors lead to substantially different solutions. Accordingly, we categorize two main group types of crowdsource contributors from Flickr. One is considered first-time visitors, referred to as tourists, and the other is repeated visitors, referred to as locals. To differentiate between the two groups, we reconstruct the contributing photographers' travel trajectories from their photo locations and interpret specific spatiotemporal activity descriptors associated with their travel trajectories to distinguish both groups. A grid-based clustering method is implemented on the photos to find popular cells traversed by the photographers of the different groups. To compute travel routes, an Orienteering Problem extension is developed using integer linear programming with an objective function reflecting a profit model that maximizes the total popularity profit score. This is done for both groups; thus, different travel routes are calculated. A weight function is used to aggregate the two routing criteria, allowing the analysis of the differences of the routes in terms of popularity measure, places of interest and environmental setting. Experimental results for Manhattan, New-York, are investigated and analysed, validating the results on two levels: areas visited by both groups, and locations (attractions) traversed by both groups. The results raise new ideas and concepts for recommending travel routes that consider both tourism and local perspectives designed to serve the users' travel objective and preferences, supporting the user with a new way to navigate and experience the city.

AGILE: GIScience Series, 1, 2020.
Full paper Proceedings of the 23rd AGILE Conference on Geographic Information Science, 2020.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras
This contribution underwent peer review based on a full paper submission.
https://doi.org/10.5194/agile-giss-1-16-2020 | © Authors 2020. CC BY 4.0 License.

2 Related Work

2.1 Tourism Photography

Sharing tourism information as in ‘digital footprints’ via different social online platforms (e.g., Foursquare, Twitter, and Flickr) is becoming more popular, mainly since nowadays the use of smartphones equipped with embedded GNSS sensors makes it easy to document, upload and share geotagged photos [5,6], whereas social platforms data fusion can be used to study the activity of tourists [7].

Photography is strongly related to tourism experience [8], meaning that most tourists take photos for documentation purposes as proof of consuming the experience of travelling a city [9]. [10] examined the representations and photographic processes of tourists, arguing that even though tourists participate actively in the photographic process and produce photos of very personal significance, the attractiveness of cities implicitly store geographical information in social media photography. This may give a clue as to how visuals affect the construction of destination image and interpretation [11], and the fact that building a destination image by tourist photographers is one of the key concepts of tourism.

2.2 Photographer Groups

Spatial analysis of different contributors of crowdsource data shows that there exists a differentiation in the geospatial distribution of tourists’ photos, which are concentrated around a city’s main sightseeing spots, as opposed to the locals’ photos, which are more dispersed throughout the city [2]. Since tourists mostly have limited time, they usually consume tourism in known popular places, and accordingly will document these places with photos, whereas locals take photos in most city sites. Additionally, according to [8], tourists uploading photos to Flickr vary, originating from diverse social groups involved in tourism consumption, hence it is possible to differentiate these tourism groups according to their unique activity. Different studies have been oriented to classify different groups of visitors using geotagged crowdsourced data, under the assumption that different groups tend to have different travel preferences [12]. [13], for example, used a deep neural network using TensorFlow to classify visitors' behaviour to six groups in Honk-Kong based on geo-located data of the Weibo social check-in platform, receiving an accuracy of close to 90%. In the study of [4], the authors used Weibo data for comparing the activity of locals and tourists in the Shanghai region, showing that the activities of tourists are significantly different from those of locals in terms of their spatiotemporal patterns.

Studies analysing and visualizing spatial interactions between tourists and locals using Flickr have determined the location of origin of different users by implementing an analysis of the users' profile location information [12,14]. [15] developed a semi-automatic approach of classification using the users' origin attribute by correction of irregularities in the users’ address description. Still, only 32% of Flickr users are providing information about their origin [16]. Therefore, temporal travel constraints are adopted to distinguish between locals to tourists. For instance, classifying users by using the number of photos taken, and the photographing duration. [2] defined differentiation between locals and tourists by using a threshold of at most 5 consecutive days for considering tourists, under the assumption that longer stays are the result of repeated visits or locals that mostly do not plan to visit only the main tourism sites and attractions. Other approaches, such as the one of [17], set a limit visit duration of 21 days of travel itineraries, while [18] and [19] determined that the temporal criterion of visit duration of tourists will not exceed one month.

Spatiotemporal approaches are also proposed, where [20] searched the country in which a user took the most photos in a period that is greater than six months and classified this user as a local in that country. Another method was introduced by [15], where they used time and speed filters to distinguish movement patterns between two distinct groups: inhabitants of Switzerland and foreigners in Zurich, while [21] used spatiotemporal activity descriptors, such as travel speed and travelled distance, to classify two groups of contributors in Flickr data.

2.3 Route Planning

[22] use three different models for describing the tourist trip design problem (TTDP):

1. the profitable tour problem (PTP), introduced by [23], which means maximizing the collected profit and minimizing the travel cost in one objective function.
2. the prize collecting traveling salesperson problem (PCTSP), introduced by [24], which means minimizing travel cost with defined total tour profit being not smaller than a given value.
3. the orienteering problem (OP) introduced by [25], which means maximizing the total collected profit while maintaining the travel cost under a given value.

All these problems are identified with variants of the Traveling Salesperson Problem, a fundamental problem of combinatorial optimization [26]. The OP is known to be NP-hard [27], i.e. it is unlikely that an efficient algorithm exists that guarantees to find optimal solutions. However, it can be modelled and solved as an integer linear program (ILP), which is a common approach in combinatorial optimization [28].

Since we aim for pedestrian route planning in an urban surrounding, various parameters and perspectives could influence pedestrian navigation. [29] studied the influence of different parameters for computing pleasant routes, such as: safety, quietness, and criteria related to social places along the route. The personal feeling of a pedestrian in a route is analysed in [30], such as: happiness, quietness – and more. The main challenge of a personalized parametrized route computation is to find the compensation between the shortest distance of the route to comprehensively investigated parameters.

New perspectives for routing computation are evaluated by traditional surveys and questionnaires among pedestrians from different groups. [17] used the Amazon Mechanical Turk (AMT) platform with local experts for evaluating their recommended routes. [30] compared results for route recommendations that are based on Flickr data using a survey among local participants in the city of London and Boston. Other approaches are evaluated by comparing their recommended routes to external tourism guides, such as TripAdvisor, Lonely Planet and more ([14, 21]).

3 Methodology

3.1 Data and Software Availability

The methodology builds on data extracted from Flickr API and downloaded data of OSM network. The data is available on: https://github.com/EcslTechnion/Agile2020_SocialMedia. The OP approach that was developed in Java is available on: https://github.com/GeoinfoBonn/OrienteeringProbelm.

3.2 Photographers Classification

We identify Flickr contributors who show tourism activity and descriptors according to a set of spatio-temporal parameters and identifiers that characterize tourism activity; all other contributors are considered locals. Flickr stores explicit photograph metadata information of location (φ,λ) and time-stamp, including the user id. Since tourists mostly show similar tourism consumption for a specific area (e.g., [2,12]), we employ a set of adaptive tourism descriptors calculated for each area by implementing a holistic approach (excluding photos of a user that are taken in accumulated distance of 1 m or in a time interval of 1 second). Subsequently, adaptive descriptors are calculated to retrieve the shared tourism activity related to tourism users' travel trajectories:

1. Travel Time: a single trip time interval, possibly covering multiple days, between the first (t_s^i) and the last (t_e^i) geotagged photograph time-stamp is calculated for each user i. The average visit duration time t_{avg} among n users is defined as:

$$t_{avg} = \frac{\sum_{i=1}^{n}(t_e^i - t_s^i)}{n} \quad (1)$$

2. Number of Photos: a threshold of at least three distant photos per user is defined, also suggested by [17] as the minimum number of points that represent a trajectory.
3. Travel Distance: A maximum threshold of fifty kilometres is used to ensure the retrieval of walking activity.
4. Travel Speed: a single trip travel speed V_u, including multi-day trips, is calculated (Equation 2) according to the accumulated travelled distance D_u divided by the time interval between the last (t_e) and first (t_s) timestamp. Outliers larger than 10 km/hr are excluded to ensure walking activity only.

$$v_u = \frac{D_u}{t_e - t_s} \quad (2)$$

By changing the value of the travel time descriptor, we have computed the descriptors simultaneously of a specific area, analysing the differentiation between iterations for defining the most common shared tourism activity; photographers who do not share these descriptors are defined as locals.

3.3 Grid-based Clustering

We divide the area to equal size cells to retrieve the main popular and frequently traversed locations (places of interest, POIs) visited by the different contributors, thus analysing the activity pattern according to their trajectories. This allows fast and intuitive clustering of photos for a specific area. We use a cell size of 250×250 meters, assuming this size represents the common grid structure of city streets, and the fact that it is rare to find adjacent attractive POIs within the same cell. For each cell, a centroid calculation is implemented on all the photos that fall in the cell's extent to identify the POI location. For each POI location we count the number of contributors in each group (i.e., photographers traversing that cell in their trip), to construct a popularity score matrix (profit); the higher the values, the more popular the cell is for a specific group.

3.4 Defining the Routing Problem

A routable network containing the origin point $s \in R^2$ and the destination point $t \in R^2$ is modelled as a graph $G = (V, E)$, where V is a set of vertices and E is a set of edges. The vertex set V contains three types of vertices, as shown in Fig. **1**. These types are (1) *terminals*, i.e., s, t as well as vertices representing the centroids of the photo clusters, (2) vertices representing the junctions of the road network with at least three outgoing road sections, and (3) vertices representing the points in the road network nearest to the terminals. The set of terminals is referred to as $T \subseteq V$. Each edge represents a link of the road network or a *feed link* connecting a terminal with the nearest vertex in the road network. A road section with bends between two consecutive junctions is represented with a single edge unless it is subdivided by a vertex of type (3). This way, computations are not slowed down by unnecessary degree-two vertices in the road network. For each edge $e \in E$, the edge weight $w(e) \geq 0$ is defined as the geometric length of the corresponding road section, or, in the case of feed links, simply as the Euclidean distance between the two incident vertices. The reason to choose the geometric length as edge weight is to reflect the general preference for short routes. Since G is undirected, we assume that the same edge weight is counted when traversing the edge in different directions. Especially for pedestrians, the geometric distance travelled can be seen as a good proxy for travel time.

However, since distance and time do not sufficiently reflect the attractiveness of a route, we introduce a popularity score $p(v)$ for each terminal $v \in T$. To summarize, we define an enriched road network graph as follows.

Definition 1 (Enriched Road Network Graph): *An enriched road network graph consists of a graph $G = (V, E)$, a set $T \subseteq V$ of terminals, an edge weighting $w: E \rightarrow R_{\geq 0}$ and a function $p: T \rightarrow R_{\geq 0}$ representing popularity scores.*

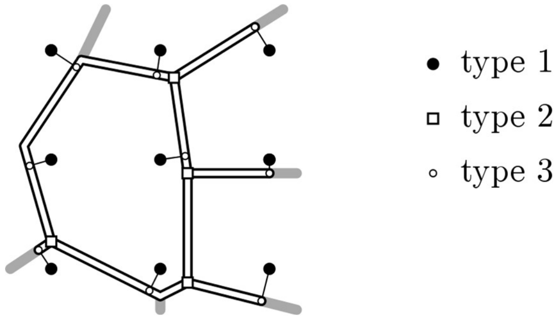


Fig. 1. The enriched road network graph G with its three different types of vertices. The road network should be chosen large enough to contain the shortest path between any two terminals (vertices of type 1). Dead ends not leading to any terminal (gray) do not need to be considered as road links.

In general, we are interested in finding a not necessarily simple s-t path in G that is short with respect to the total edge weight (which we also refer to as travel cost) and accumulates a large total popularity score. Allowing non-simple paths ensures that we can connect a cluster center to the solution by traversing a feed link forth and back. Formally, the two criteria are modelled as follows.

Definition 2 (Travel Cost): *The travel cost $W(\pi)$ of a path $\pi = (e_1, \dots, e_k)$ in G is the sum of weights of the edges in π, i.e., $W(\pi) = \sum_{i=1}^{k} w(e_i)$.*

Definition 3 (Popularity Score): *The popularity score $P(\pi)$ of a path π in G is the sum of vertex popularity scores for the terminals visited by π, i.e., $P(\pi) = \sum_{v \in T \cap V(\pi)} p(v)$, where $V(\pi)$ is the set of vertices visited by π.*

Note that, according to these definitions, traversing an edge e twice (e.g., once in each direction in the case of a feed link) will add $2w(e)$ to the travel cost, yet when visiting a vertex v multiple times its popularity score $p(v)$ is counted only once. This is plausible, since for a tourist it just matters whether or not an interesting sight is connected to his or her route.

To aggregate the two criteria, we impose a strict upper limit W_{max} on the travel cost $W(\pi)$ and maximize the total popularity score $P(\pi)$ of the path π. This model reflects a pedestrian who has a fixed amount of time to visit preferably many and attractive sights. The following formal definition summarizes this model.

Definition 4 (Tourist Routing Problem): *Given an enriched road network graph $G = (V, E)$ with start $s \in V$ and destination $t \in V$ and an upper limit $W_{max} > 0$, find an s-t path $\pi = (e_1, \dots, e_k)$ in G such that $W(\pi) \leq W_{max}$ and $P(\pi)$ is as large as possible.*

Note that by choosing different values for W_{max} and resolving the model one can generate the set of all Pareto optimal solutions to the problem, which could be used to provide the user with multiple route suggestions.

3.5 Solution of the Tourist Routing Problem by Reduction to the Orienteering Problem

According to [31], the OP assumes a set $N = \{1, \dots, |N|\}$ of sites as input, a score $S_i \geq 0$ for each site $i \in N$, a travel time $t_{ij} \geq 0$ for each $i, j \in N$, as well as a time budget $T_{max} \geq 0$. Similar to our tourist routing problem, the OP asks to find a route from site 1 to site $|N|$ that maximizes the score while staying within the time budget. The main differences between the two problems are that in our case only the terminal vertices (i.e., cluster centers) are associated with scores and that for many pairs of vertices $(u, v) \in V^2$ there is no direct link, i.e., $\{u, v\} \notin E$. Despite these differences, the tourism routing problem can be reduced to the OP in such a way that any algorithm for the OP can be used for our purpose. In essence, this reduction is based on the following definition.

Definition 5 (Terminal Graph): *Given an enriched road network graph, i.e., (V, E), T, w, and p, the terminal graph $G' = (T, E')$ contains a vertex for each terminal in T and an edge $e = \{u, v\}$ for each $u, v \in T$ with $u \neq v$. The weight of $e = \{u, v\}$ is defined as the total weight of the shortest u-v path in G.*

Fig. **2** shows how the terminal graph is constructed (Step 1) and the solution to the OP is computed (Step 2) and transferred back to the original graph (Step 3). To accomplish these three steps, the following details need to be considered:

Step 1: Shortest paths in $G = (V, E)$ need to be computed between every two-terminal vertices. This step may take relatively long if G and the set T of terminals are large, i.e., it requires $|T|$ runs of Dijkstra's algorithm and, thus, $O(|T||V|\log|V| + |T||E|)$ time in the worst case. However, the bottleneck in terms of the running time is Step 2, i.e., solving the OP, which is NP-hard [27]. For our instances, the first step was easily accomplished with standard software, i.e., the ArcGIS network analyst based on the OpenStreetMap (OSM)[1] road network.

Step 2: For solving the OP we implemented an exact approach based on mathematical programming (published on GitHub[2]). For this, we used an integer programming formulation by [28], and the solver Gurobi[3] that uses an algorithm based on branch-and-cut. With this, we were able to solve almost all instances with proof of optimality. For some instances, the solver yielded a solution, but we terminated the experiment due to long computation time without knowing whether the solution was optimal. This shows that the running time is crucial, and that one should consider implementing fast heuristics if the aim is to deploy the method in practice, e.g., as a location-based service. It is worth mentioning that based on the travel cost limit W_{max} (the time budget T_{max} in the OP) one can rule out many terminal vertices without losing the optimality guarantee of the method. More precisely, a site i with $t_{1i} + t_{i|N|} > T_{max}$ in the OP cannot be visited in a feasible solution. We used this fact to reduce the size of the terminal graph before solving the OP.

[1] https://www.geofabrik.de/data/download.html

[2] https://github.com/GeoinfoBonn/OrienteeringProblem

[3] https://www.gurobi.com/

https://doi.org/10.5194/agile-giss-1-16-2020 | © Authors 2020. CC BY 4.0 License.

Step 3: Since each edge in G' corresponds to the shortest path connecting two terminals in G, the optimal solution to the OP can be easily converted into an optimal solution to the tourist routing problem (TRP) by concatenating the paths corresponding to the selected edges in G'. For reconstructing these paths, one could again use Dijkstra's algorithm. Alternatively, assuming that the shortest paths computed in Step 1 have been memorized and associated with the edges of G, one can simply reconstruct the solution based on the paths associated with the selected edges. In our implementation, we used the first approach since we use ArcGIS tools for Step 1 and 3, but an external application (Gurobi) for Step 2.

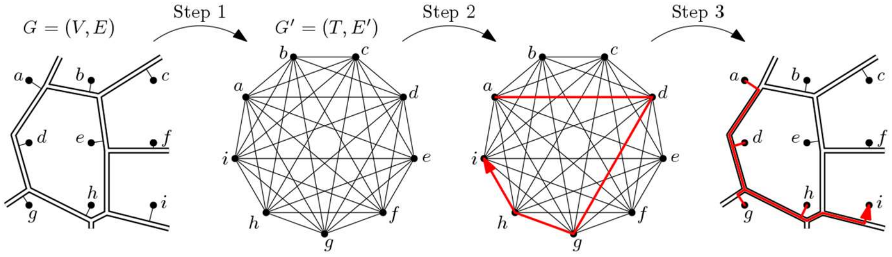


Fig. 2. The solution of the TRP via the OP. The instance of the OP is generated in Step 1 and solved in Step 2. In Step 3, the solution is transferred to the original graph.

3.6 Weighted Contributor Routing

For each group of contributors, i.e., tourists and locals, we can define a corresponding score function, i.e., $p_{tourists}$ and p_{locals}, respectively. Both are defined based on the corresponding number of accumulative users in the grid cell. Accordingly, setting the popularity score in our problem definition $p = p_{tourists}$ or $p = p_{locals}$ may yield different solutions. To generate even more alternative routes, we define a weighted version of the popularity score.

Definition 6 (Weighted Contributor Routing): *Weighted Contributor Routing is the special case of the TRP with*

$$p = \alpha \cdot p_{tourists} + (1 - \alpha) \cdot p_{locals} \quad (3)$$

where $\alpha \in [0,1]$ is referred to as the trade-off parameter needed to be set by the user. Using this model, a user can aggregate the two routing criteria, i.e., trade-off visiting popular tourist sights against experiencing activities and places known to locals, thus not visiting overcrowded tourism areas.

https://doi.org/10.5194/agile-giss-1-16-2020 |

4 Experimental Results

4.1 Contributor Classification

Two main areas were analysed in Manhattan, New York. Close to 500,000 Flickr geotagged photos were downloaded using Flickr API[4] and shared on GitHub[5]. The first area is in Midtown Manhattan; this area is very famous and includes cultural, entertainment and leisure attractions, such as: shopping areas, Times Square, The New York Public Library and more. This area is visited by tourists and locals alike, where the existing landmarks and attractions are dense and usually crowded with visitors. The second area chosen is in Central Park, which is one of the most famous parks in the United States, visited by locals and tourists alike for various purposes, including: scenery, culture, sports and more. It is an area of interest that includes many attractions that are geographically distributed in the park, including landmarks, museums, architecture and more. The park serves as a major recreation area for the citizens of Manhattan, for resting, strolling and jogging. Both selected areas are popular and visited by locals and tourists alike; nevertheless, their travel objective and hence activity - the explored sites and areas - is different.

Implementing our contributor classification, we have found that the average tourism travel time in Manhattan is 5 days, a value also validated in [2]. Implementing the other travel indices, i.e., speed and distance thresholds, excluded users traveling long distances and/or over a short time, resembling a travel mode other than walking. Differentiating both groups allows us to analyse the activity differences between them, where 1,400 contributors are first-time visitors (i.e., tourists), and 22,000 contributors are repeating visitors (i.e., locals). Fig **3** depicts the number of contributors versus the number of visit days, showing that the value roughly remains the same even after 5 days.

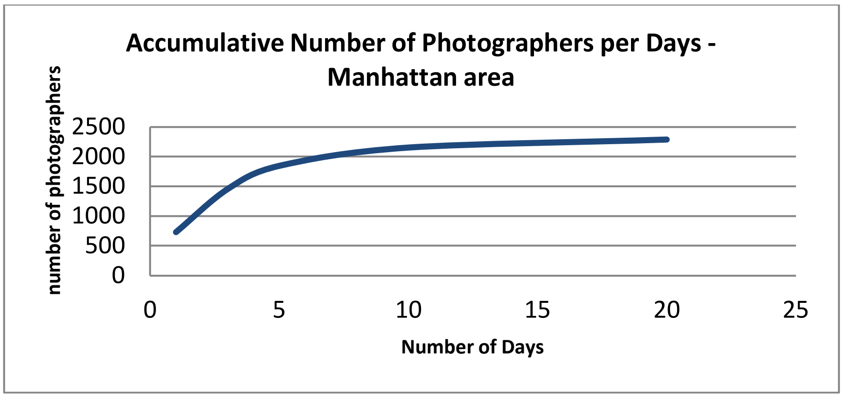


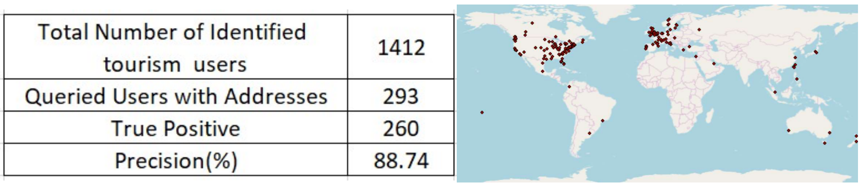
Fig. 3. The number of contributors vs. travelled number of days for the Manhattan areas.

[4] https://www.flickr.com/services/api/

[5] https://github.com/EcslTechnion/Agile2020_SocialMedia

https://doi.org/10.5194/agile-giss-1-16-2020 | © Authors 2020. CC BY 4.0 License.

To further asses our results, we analyse the contributors' country of origin (geolocation home address); this value can be inserted by Flickr users as text, and is stored as metadata, where we convert the text via geocoding to a geolocation value. We use OSM's API (Nominatim[6]) for geocoding, automatically converting the users' textual location to geographic coordinates. For the analysed data, most users did not insert this information, or the information is only partial, such that we could not retrieve this information for all contributors. Fig **4** depicts the results of this analysis, where for 20% of the contributors who were classified according to our classification algorithm as tourists (close to 300) we could generate their home address location: for close to 90% their home address was defined outside of New-York, validating our classification method.



Total Number of Identified tourism users	1412
Queried Users with Addresses	293
True Positive	260
Precision(%)	88.74

Fig. 4. Tourism photographers in Manhattan. Distribution map of the users' origin and comparison of our classification method to reference data.

A comparative differentiation of contributor groups is implemented according to the Dissimilarity Index [14], that for our purpose measures the uneven distribution of the two groups – tourists and locals. V_i is the unique number of tourists in a cell, while V is the total number of unique tourists in the area. Accordingly, L_i is the unique number of locals in a cell, while L is the total number of unique locals in the area. Index R in Equation 4 represents the relative activity between the two groups. A positive R value indicates a higher tourism activity in the cell, and a negative R value indicates a higher activity of locals.

$$R = \frac{V_i}{V} - \frac{L_i}{L} \tag{4}$$

Implementing the index, a heat map (matrix) is generated, depicted in Fig **5**, showing the relative activity of the two groups. Positive values (where the maximum value of 1 is coloured in red) resemble tourism activity, and negative values (where the minimum value of -1 is coloured in green) resemble the activity of locals. The clustered red cells are primarily visible near main tourism attractions and landmarks mainly in upper Manhattan, such as: Times Square, the Empire-State Building, the Rockefeller Center and the Hudson River. The activity of locals is more heterogeneously dispersed in the area, where popular areas are: the Grand Central Terminal (public transportation area), Bryant Park and Lower Manhattan (business area). Accordingly, this matrix is valuable for determining the relative popularity for a specific area among the two groups, where values are used as input in the OP routing algorithm.

[6] https://nominatim.org/release-docs/develop/api/Overview/

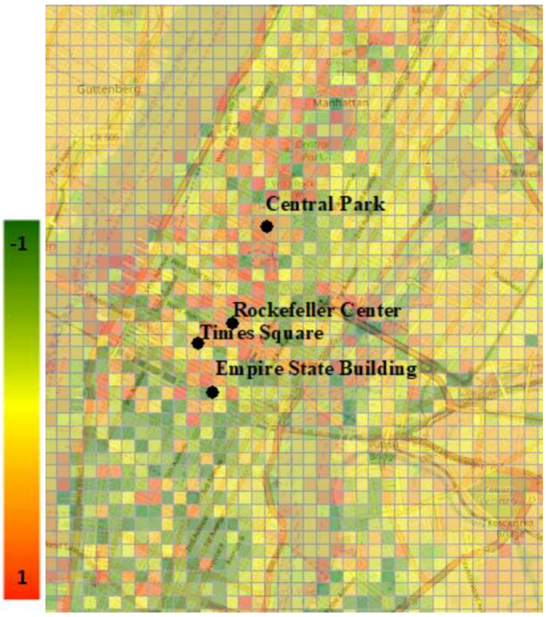


Fig. 5. Dissimilarity Index results for the area of Manhattan.

4.2 Comparison of Routes

We use the Dissimilarity Index values as input for the OP route planning. We compute a tourism route that relies on the tourists' popularity index, a locals' route that relies on the locals' popularity index, and an aggregated route that uses a varying α value according to Definition 6.

Midtown Manhattan. The origin point is defined in Madison Square Garden, and the destination point is defined in Times Square. This is relatively a very popular route commonly visited by numerous tourists. Since many users use nowadays crowdsource tourism applications to travel unknown areas [32], we compare and evaluate the generated routes to routes from GPSMYCITY[7], which is a sightseeing application guide that recommends walking routes generated by local tourism guides. Fig **6** depicts the tourism route of GPSMYCITY between the above-mentioned origin and destination points, with a length of 6.6 km traversing many attractions in the area (depicted with consecutive numbering).

[7] https://www.gpsmycity.com/

https://doi.org/10.5194/agile-giss-1-16-2020 | © Authors 2020. CC BY 4.0 License.

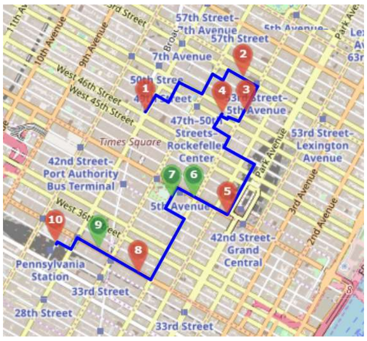


Fig. 6. GPSMYCITY recommended tourism route for Midtown Manhattan: red attractions are considered must-see according to GPSMYCITY, green attractions are recommended.

Using the OP algorithm with a maximum distance of 7 km, we generate a tourism route, i.e., defining $\alpha = 1$. Examining the route, depicted in Fig **7** (left), we discover that it traverses many famous and popular attractions and landmarks in the area of Midtown Manhattan. More than 90% of the attractions this route traverses appear in the GPSMYCITY route, validating the premise that the route validates the tourism objective. Although the Museum of Modern Art (MoMA) has a relatively high tourism popularity index, it was not included by the OP algorithm due to the used distance threshold. The Majestic Theatre, on the other hand, was not found to be an attractive place according to the Flickr tourism photographers, as opposed to the Chrysler Building, which does not exist in the GPSMYCITY route, but is considered as a place that attracts many tourism photographers.

Using the OP algorithm with the same distance threshold of 7 km, we generate a locals' route, i.e., defining $\alpha = 0$. Observing the route, depicted in Fig **7** (right), we discover that it traverses different locations, which are considered less touristic (roughly 70% match when compared to GPSMYCITY route), but do serve as local popular attractions, such as: Grand Central Terminal and Bryant Park, which is a park liked by many families in the area. The route does not traverse the Rockefeller Center, which is mostly overcrowded by many tourists, most likely avoided by locals. Table **1** summarizes the traversed attractions and locations according to the three routes.

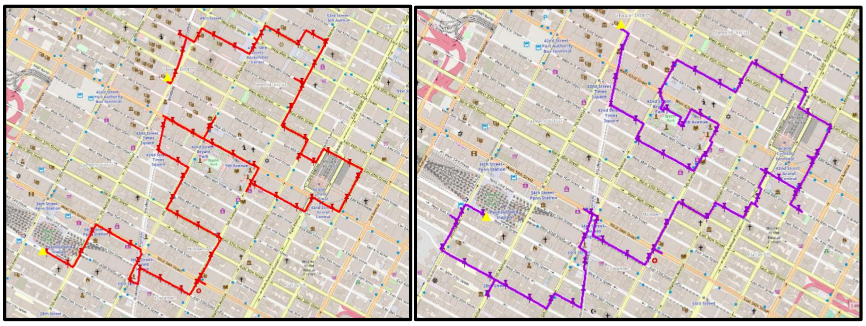
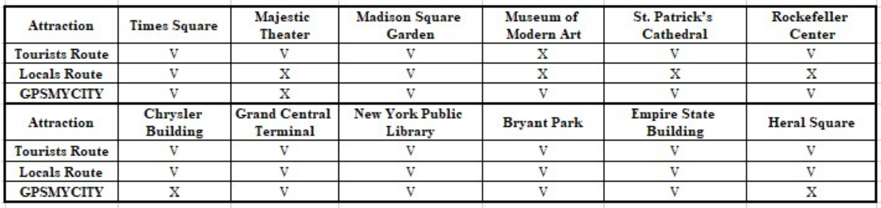


Fig. 7. Route comparison: tourism (left) and local (right) photo contributors.

Table 1. Traversed attractions and locations comparison for Midtown Manhattan: V marks a traversed location, and X marks a location not traversed.



Attraction	Times Square	Majestic Theater	Madison Square Garden	Museum of Modern Art	St. Patrick's Cathedral	Rockefeller Center
Tourists Route	V	V	V	X	V	V
Locals Route	V	X	V	X	X	X
GPSMYCITY	V	X	V	V	V	V
Attraction	**Chrysler Building**	**Grand Central Terminal**	**New York Public Library**	**Bryant Park**	**Empire State Building**	**Heral Square**
Tourists Route	V	V	V	V	V	V
Locals Route	V	V	V	V	V	V
GPSMYCITY	X	V	V	V	V	X

Using the OP algorithm with the same distance threshold of 7 km and $\alpha = 0.5$, we generate an aggregated route, depicted in Fig **8**, that relies on both popularity indices of tourists and locals alike. While the aggregated route traverses popular attractions and landmarks, it occasionally deviates, passing through streets that are parallel to the tourism route, most likely to avoid overcrowded and dense tourism streets, such as the Fifth Avenue. The aggregated route traverses through the Rockefeller Center (marked as B), having a large tourism popularity index, but visits Bryant Park (marked as C) and Grand Central Terminal (marked as A), both of which have high locals' popularity index. The latter, for example, is a very popular destination among locals who take many photos in that area. The aggregated route shows an interesting perspective of both tourism and locals' activities, traversing main attractions frequently visited by tourists, but also places visited and liked by locals.

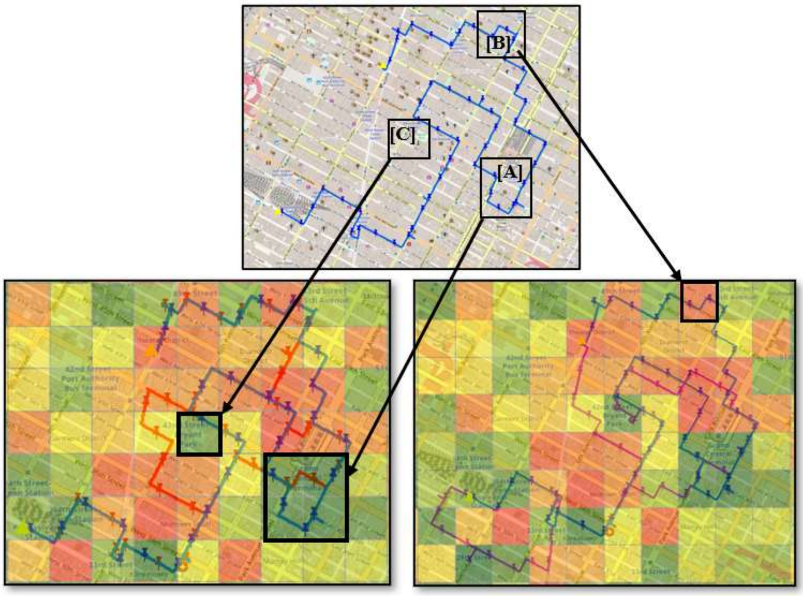


Fig. 8. Aggregated route (top), in comparison to the locals (bottom-left) and the tourism (bottom right) routes, superimposed on the dissimilarity index heatmap.

https://doi.org/10.5194/agile-giss-1-16-2020 |

Central Park. The origin point is defined in Columbus Circle, and the destination point is defined in the Art Museum. Fig **9** depicts the tourism route recommended by GPSMYCITY, with a length of 7.3 km. The route is recommended for a 3-hours walking, visiting many attractions and landmarks distributed along it. The OP is implemented between points 1 and 11 of the GPSMYCITY route.

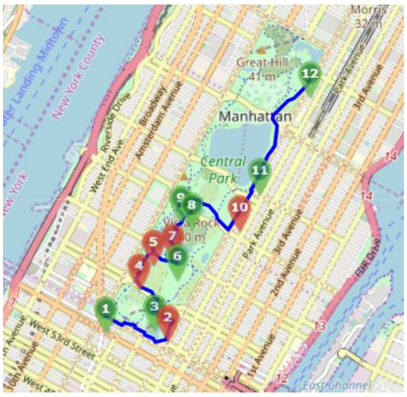
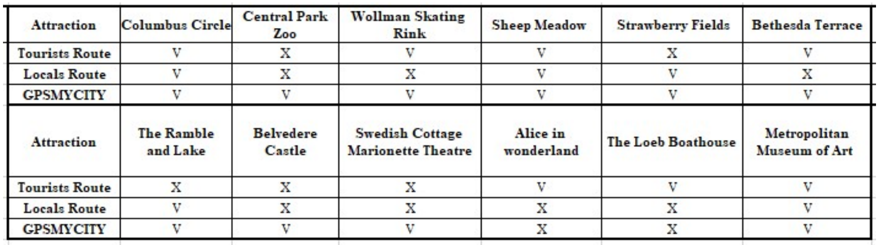


Fig. 9. GPSMYCITY recommended tourism route for Central Park: red attractions are considered must-see according to GPSMYCITY, green attractions are recommended.

A comparison of the tourism route generated by the OP algorithm to the GPSMYCITY route reveals a match of 60% in terms of visited landmarks and attractions, depicted in Table **2**. Comparing the tourism route and the locals' route, depicted in Fig **10** (top), the tourism route (in red) passes the park in the middle, traversing most existing attractions, whereas the locals' route (in purple) passes through the west side of the park. This implies that locals try and avoid overcrowded areas; for example, they travel the west side of the northern lake in the park that is still considered scenery, also visiting the Strawberry Fields, liked by many residents as a meeting place. Locals also visit The Lincoln Center of Performing Arts (marked as A), a local art center - less frequently visited by tourists who visit Central Park - but liked by many locals.

The aggregated route computed by the OP algorithm ($\alpha = 0.5$, in blue), depicted in Fig **10** (top), is very similar to the tourism route, mainly since Central Park is a very popular area among first-time visitors to New York. One distinct difference is the Pilgrim Hill, considered as a picnic and relaxing area, mostly visited by local families (marked B). This comes to show that although the Dissimilarity Index of Central Park is more tourism-oriented, thus the aggregated route is more similar to the tourism route, still, some local areas are extracted that can enrich the experience of visitors that look for a different less overcrowded experience.

Table 2. Traversed attractions and locations comparison for Central Park: V marks a traversed location, and X marks a location not traversed.



Attraction	Columbus Circle	Central Park Zoo	Wollman Skating Rink	Sheep Meadow	Strawberry Fields	Bethesda Terrace
Tourists Route	V	X	V	V	X	V
Locals Route	V	X	X	V	V	X
GPSMYCITY	V	V	V	V	V	V

Attraction	The Ramble and Lake	Belvedere Castle	Swedish Cottage Marionette Theatre	Alice in wonderland	The Loeb Boathouse	Metropolitan Museum of Art
Tourists Route	X	X	X	V	V	V
Locals Route	V	X	X	X	X	V
GPSMYCITY	V	V	V	X	X	V

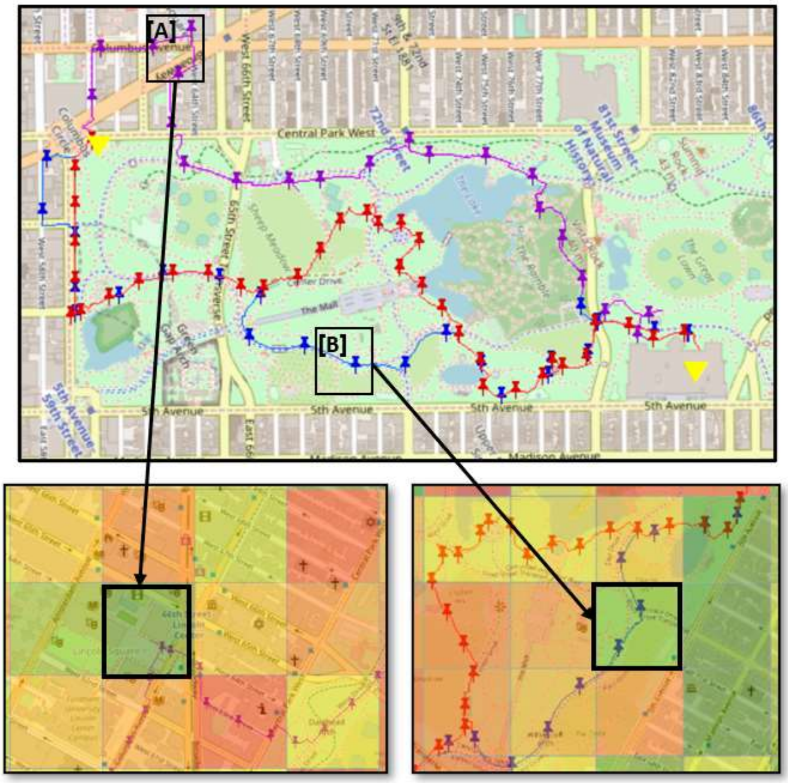


Fig. 10. Route comparison: tourists (red), aggregated (blue), and locals (purple) (top), superimposed on the Dissimilarity Index (bottom).

5 Conclusions & Future Work

This research presents a methodology that enables the investigation and comparison of walking routes computed by mining social media user-generated crowdsource photos.

More specifically, we compared two groups of photo contributors: tourists, considered first-time visitors to the area, and locals, considered repeating visitors. By implementing tailored spatio-temporal travel indices, we were able to classify photographers that show tourism activity patterns, where statistical analysis proved this classification reliable. Accordingly, popularity matrices were generated according to the number of contributors from each group traversing the space. A Dissimilarity Index showed that correlation exists between the locations frequently visited by the two groups and their context. A customized Orienteering Problem algorithm was developed, which uses the popularity score (profit) matrix, to generate routes that maximize the tourism experience or the locals' experience. By using different weights, we produced routes that take into consideration both tourism and locals' context, thus producing routes that can enrich user experience.

Comparing the results for several areas in Manhattan, New York, showed that the tourism routes produced by the customized Orienteering Problem algorithm maximized the tourism experience, traversing most popular tourism attractions and landmarks, also frequently recommended in tourist guides. Locals’ routes, on the other hand, revealed other unique and specific locations, mostly open, cultural and recreation areas, frequently visited by locals, also avoiding overcrowded streets. The aggregated routes traversed popular locations and attractions visited by both groups, enriching the experience of users who aspire to travel less crowded streets and more off-beat local attractions.

Future work will entail the improvement of the Orienteering Problem route computation. For example, in its current version, the popularity weight matrix considers the popularity score of each vertex (cell centroid), whereas adding a score that is associated with the traversed edge between different vertices (i.e., counting the number of users traveling between these locations) can generate more tuned and attractive routes. Improving the clustering approach of the geotagged photos by introducing a density-based method can improve the computation of attraction locations. Understating the urban morphology by analysing its spatial configuration, e.g., integrating visibility or space syntax analysis, could also be used to better understand the travel context about the routes and the different users’ preferences.

This research presents the use of online crowdsource social media for developing context-aware walking routes attuned to the user’s interest and criteria. The categorization between tourists and locals allows the user to decide how to trade the one-off against the other, generating routes that correspond to her/his preferences. We believe that this implementation can be used in location-based services, such as travel guides, enriching user experience when traveling to new areas.

References

1. Basiri, A., Amirian, P., Winstanley, A., Moore, T.: Making tourist guidance systems more intelligent, adaptive and personalised using crowd sourced movement data. Journal of Ambient Intelligence and Humanized 9(3), 413–427 (2018).
2. Kádár, B. Measuring tourist activities in cities using geotagged photography. Tourism Geographies 16(1), 88-10 (2014).

3. Hall, C. M., Ram, Y. Measuring the relationship between tourism and walkability? Walk Score and English tourist attractions. Journal of Sustainable Tourism 27(2), 223-240 (2019).
4. Khan, N. U., Wan, W., Yu, S. Spatiotemporal Analysis of Tourists and Residents in Shanghai Based on Location-Based Social Network's Data from Weibo. ISPRS International Journal of Geo-Information 9(2), 70. (2020).
5. Farzanyar, X., Cercone, N.: Trip pattern mining using large scale geo-tagged photos. In Proceedings of the International Conference on Computer and Information Science and Technology, p. 113. Ontario (2015).
6. Lim, K. H.: Recommending tours and places-of-interest based on user interests from geo-tagged photos. In Proceedings of the 2015 ACM SIGMOD on PhD Symposium, pp. 33-38. ACM, Melbourne (2015).
7. Salas-Olmedo, M. H., Moya-Gómez, B., García-Palomares, J. C., Gutiérrez, J.: Tourists' digital footprint in cities: Comparing Big Data sources. Tourism Management 66, 13-25. (2018).
8. Kádár, B., Gede, M. 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 (2013).
9. Urry, J.: The tourist gaze – leisure and travel in contemporary societies. SAGE publications, London (1990).
10. Stylianou-Lambert, T.: Tourists with cameras: Reproducing or Producing? Annals of Tourism Research 39(4), 1817-1838 (2012).
11. MacKay, K. J., Fesenmaier, D. R.: Pictorial element of destination in image formation. Annals of tourism research 24(3), 537-565 (1997).
12. Liu, Q., Wang, Z., Ye, X.: Comparing mobility patterns between residents and visitors using geo-tagged social media data. Transactions in GIS 22(6), 1372-1389 (2018).
13. Han, S., Ren, F., Wu, C., Chen, Y., Du, Q., Ye, X. Using the tensorflow deep neural network to classify mainland china visitor behaviours in Hong Kong from check-in data. ISPRS International Journal of Geo-Information 7(4), 158 (2018).
14. Li, D., Zhou, X., Wang, M.: Analyzing and visualizing the spatial interactions between tourists and locals: A Flickr study in ten US cities. Cities 74, 249-258 (2018).
15. Straumann, R. K., Cöltekin, A., Andrienko, G. Towards (re) constructing arratives from georeferenced photographs through visual analytics. The Cartographic Journal 51(2), 152-165 (2014).
16. Milleville, K., Ali, D., Porras-Bernardez, F., Verstockt, S., Van de Weghe, N., Gartner, G.: WordCrowd–A Location-Based Application to Explore the City based on Geo-Social Media and Semantics. In LBS2019, the 15th International Conference on Location Based Services, pp. 231-236. Ghent University, Vienna (2019).
17. De Choudhury, M., Feldman, M., Amer-Yahia, S., Golbandi, N., Lempel, R., Yu, C.: Automatic construction of travel itineraries using social breadcrumbs. In Proceedings of the 21st ACM conference on Hypertext and hypermedia, pp. 35-44. ACM, Toronto (2010).
18. Girardin, F., Fiore, F. D., Ratti, C., Blat, J.: Leveraging explicitly disclosed location information to understand tourist dynamics: a case study. Journal of Location Based Services 2(1), 41-56 (2008).
19. García-Palomares, J. C., Gutiérrez, J., Mínguez, C.: Identification of tourist hot spots based on social networks: A comparative analysis of European metropolises using photo-sharing services and GIS. Applied Geography 63, 408-417 (2015).
20. Bojic, I., Massaro, E., Belyi, A., Sobolevsky, S., Ratti, C.: Choosing the right home location definition method for the given dataset. In International Conference on Social Informatics, pp. 194-208. Springer, Cham, Beijing (2015).

21. Mor, M., Dalyot, S.: Computing touristic walking routes using geotagged photographs from Flickr. In Adjunct Proceedings of the 14th International Conference on Location Based Services, pp. 63-68. ETH Zurich, Zurich (2018).
22. Gavalas, D., Konstantopoulos, C., Mastakas, K., Pantziou, G.: A survey on algorithmic approaches for solving tourist trip design problems. Journal of Heuristics 20(3), 291-328 (2014).
23. Dell'Amico, M., Maffioli, F., Värbrand, P.: On prize-collecting tours and the asymmetric travelling salesman problem. International Transactions in Operational Research 2(3), 297-308 (1995).
24. Balas, E.: The prize collecting traveling salesman problem. Networks, 19(6), 621-636 (1989).
25. Tsiligirides, T.: Heuristic methods applied to orienteering. Journal of the Operational Research Society 35(9), 797-809 (1984).
26. Papadimitriou, C. H., Steiglitz, K.: Combinatorial optimization. Englewood Cliffs: Prentice Hall, NJ, United States (1982).
27. Golden B, Levy L., Vohra R.: The orienteering problem. Naval Research Logistics 34, 307-318 (1987).
28. Vansteenwegen, P., Souffriau, W., Van Oudheusden, D.: The orienteering problem: A survey. European Journal of Operational Research 209(1), 1-10 (2011).
29. Novack, T., Wang, Z., Zipf, A.: A system for generating customized pleasant pedestrian routes based on OpenStreetMap data. Sensors 18(11), 3794 (2018).
30. Quercia, D., Schifanella, R., Aiello, L. M.: The shortest path to happiness: Recommending beautiful, quiet, and happy routes in the city. In Proceedings of the 25th ACM conference on Hypertext and social media, pp. 116-125. ACM, Santiago (2014).
31. Gunawan, A., Lau, H. C., Vansteenwegen, P. Orienteering problem: A survey of recent variants, solution approaches and applications. European Journal of Operational Research 255(2), 315-332 (2016).
32. Cavallaro, F., Galati, O. I., Nocera, S.: Policy strategies for the mitigation of GHG emissions caused by the mass-tourism mobility in coastal areas. Transportation Research Procedia 27,317-324 (2017).