JOURNAL OF SPATIAL INFORMATION SCIENCE
Number 19 (2019), pp. 29–52

doi:10.5311/JOSIS.2019.19.508

RESEARCH ARTICLE

Identifying the use of a park based on clusters of visitors' movements from mobile phone data

Roberto Pierdicca, Marina Paolanti, Raffaele Vaira, Ernesto Marcheggiani, Eva Savina Malinverni and Emanuele Frontoni

Universitá Politecnica delle Marche, Ancona, Italy

Received: March 10, 2019; returned: April 13, 2019; revised: June 11, 2019; accepted: September 27, 2019.

Abstract: Planning urban parks is a burdensome task, requiring knowledge of countless variables that are impossible to consider all at the same time. One of these variables is the set of people who use the parks. Despite information and communication technologies being a valuable source of data, a standardized method which enables landscape planners to use such information to design urban parks is still broadly missing. The objective of this study is to design an approach that can identify how an urban green park is used by its visitors in order to provide planners and the managing authorities with a standardized method. The investigation was conducted by exploiting tracking data from an existing mobile application developed for Cardeto Park, an urban green area in the heart of the old town of Ancona, Italy. A trajectory clustering algorithm is used to infer the most common trajectories of visitors, exploiting global positioning system and sensor-based tracks. The data used are made publicly available in an open dataset, which is the first one based on real data in this field. On the basis of these user-generated data, the proposed data-driven approach can determine the *mission of the park* by processing visitors' trajectories whilst using a mobile application specifically designed for this purpose. The reliability of the clustering method has also been confirmed by an additional statistical analysis. This investigation reveals other important user behavioral patterns or trends.

Keywords: clustering, human behavior analysis, public open spaces, ICT, landscape planning

1 Introduction

Dealing with green urban areas, understanding their dynamics, and analysing the behavioral patterns of the people who use these areas are difficult tasks. Nevertheless, academics and policy makers cannot be exempted from developing methods and best practices that enable the more efficient study and planning of such areas. In fact, public open spaces (POS) offer humans great benefits in terms of both individual and societal well-being, especially given that urbanization continues unabated at present, representing a threat to all of humankind. POS are a way to link people with nature by reducing stress and encouraging pro-social and sustainable behavior [2]. Providing planners with information and communication technologies (ICT) tools that can facilitate the definition of guidelines or protocols for their investigation should be fundamental to determine the significance of such areas for society. Yet, regardless of the habits and culture of the population, the limited use of POS could be attributed to potential barriers which need to be investigated. Statistics about what users do, how much time they spend, whether they create groups, and which areas they prefer are only some of the simple features that a planner should consider when carrying out the task of planning. In addition, visitors' lengths of stay in a specific area, their preferred paths, the number of points of interest (POIs) and their order of being visited, and the purpose and motivation of the visits to the space, as well as the users' socio-demographic information, can help planners tailor the space to these users' needs. However, such data are difficult to collect. This does not mean that automatic learning tools replace the role of planners; rather, these tools can support them toward improving their capabilities of understanding certain dynamics. Knowledge of a place by an expert still remains the primary element for planning.

The digital footprints [10] that individuals leave whilst performing their daily activities can be used as data for statistical analysis, as well as for extracting metrics about socio-spatial behavior directly from the source. The use of various mobile devices allows the collection of an enormous amount of spatiotemporal data about the movements of POS users. Moreover, the trajectories of mobile phones are information-rich features that reveal the potential of an environment, provide indications of the events that take place, and allow statistical inferences about the interactions between humans and space.

The research presented in this paper exploits the data collected from users' phones to infer useful statistics and provide planners with new hints for the interpretation of POS. A novel information approach to automatically perform user clustering and advanced analytics is described and released in open source. The investigation was conducted by exploiting tracking data from an existing mobile application[1] developed for Cardeto Park, an urban green area in the heart of the old town of Ancona, Italy. The application consists of different elements, including tracks, POIs, entrances (starting points, finishing points, and connections with different parts of the park), bifurcation points, and information (text, pictures and audio), as well as map and navigation tools. Moreover, in the chosen area, we identified three different paths with POIs on each one. The tracks are categorized by the characteristics of their context and depending on the types of information to be shared with visitors, such as naturalistic, historical, archaeological, spiritual, landscape, and urban, to name a few (Figure 1). This specific park was chosen for its historical value and its natural features with a view of the skyline of the old city; it constantly attracts both local and

[1]iOS: https://itunes.apple.com/it/app/cybercardeto/id1219952063?mt=8, Android: https://play.google.com/store/apps/details?id=it.univpm.dii.cardeto&hl=en

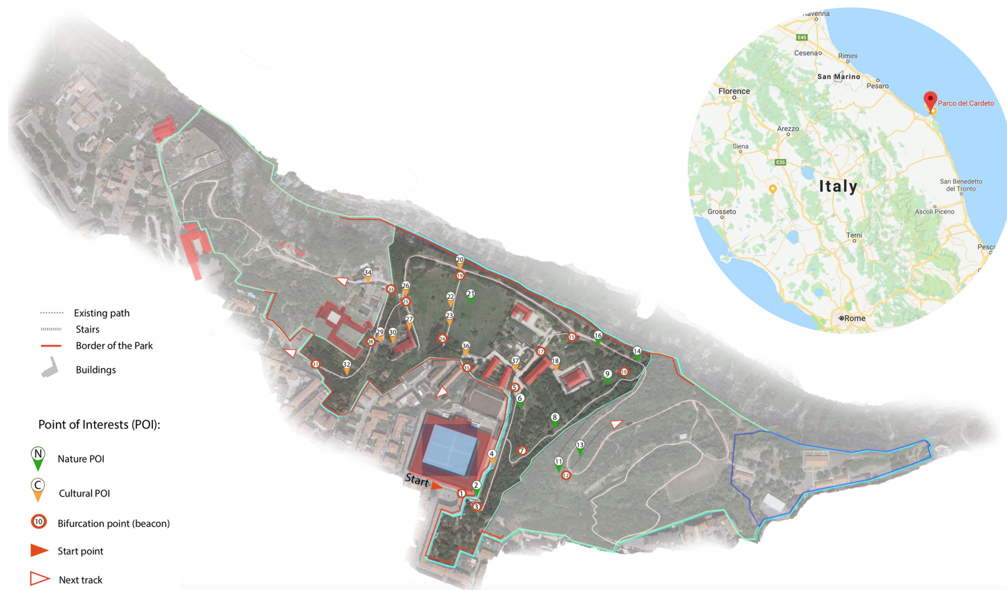


Figure 1: **Overview of Cardeto Park and its location** The general plan depicts the main paths and the POIs which have been valorized by the mobile application developed for the park (background images from https://goo.gl/maps/A6Aw6U9EBFecLssq5).

foreign visitors. On the basis of all these features, Cardeto Park is a good candidate for the present analysis, with a high potential for the examination of its visitors' behavior.

For this purpose, a dataset had been previously recorded from global positioning system (GPS) data and stored in a geodatabase. Our approach also provides the use of several active beacons suitably positioned within a pre-determined area (as shown in Figure 3). Their position was studied according to the characteristics of the park. In fact, Cardeto Park contains both natural and cultural heritage POIs, and the beacons can provide notification to attract users towards them. Other beacons were placed along the track in bifurcation points to allow visitors to choose their preferred path; some other beacons were placed in the different entrances to the park.

This paper is supported by a previous publication in which a tool for pre- and post-data processing was set up to provide data that are appropriate for further processing. Interested readers can find detailed information in [26]. Whereas, in the aforementioned paper, the main novelty was the description of an innovative architectural setup in a real scenario to collect information from users and provide information to them, in this paper, we describe an innovative approach based on trajectory clustering. This was developed to understand the potential *mission of the park* based on pedestrians' movements in order to infer the intention behind a visit. In particular, we are interested in improving the design process of a public green park by clustering the trajectories acquired by GPS data and beacons placed in the park. The trajectories are clustered into sub-trajectories, as considering them as a whole fails to detect when two of them may possess portions that are locally similar, because a trajectory's path may be long and complicated. Thus, even when portions of some trajectories possess common behavioral traits, these might not be detectable at the

Figure 2: Beacons' disposition in the area of interest (background: © OpenStreetMap contributors, CC BY-SA).

level of trajectories when each one is seen as a whole [7, 15]. Discovering common subtrajectories could be very useful in this case and, by this method, the presence of regions of special interest can be noted. This framework, which offers a solution by partitioning a trajectory into a set of line segments and then grouping similar segments together, is called a partition-and-group framework. Two different algorithms, the agglomerative [6] and spectral clustering algorithms [37], are applied, and their performances are compared.

The main contributions of this research can be summarized as follows. It proposes a framework for clustering users' trajectories inside the park. This enables planners to possibly discover common trajectories, which previous frameworks were not capable of achieving. In addition, our approach allows planners to compare the impacts of different areas of the park, reflecting the preferences of pedestrians, thus making it possible to determine the *mission of the park*. As we mentioned, our work is validated using data gathered from a real environment, making our results more reliable and our experiments repeatable. The data used is made publicly available in an open dataset, making it the first one based on extensive real world data in this field. The method proposed in this paper can generate high-quality clusters on Cardeto datasets. Finally, some statistics about the correlation between the environment and users' behaviors are provided, which are also useful for validating the results of the clustering algorithm.

The remainder of this paper is organized as follows. Section 2 presents the most recent literature in the field of data collection and clustering to give readers a complete overview of the latest research trends. Section 3 presents the data (collection and preparation) and then provides an overview of the case study in which the method has been tested and validated. The trajectory clustering algorithms are given in detail in Section 3.2, which supports the results (Section 4), in which the main findings in terms of both the park's uses and the relations between the users and the park are presented. Our comments on this work, together with some concluding remarks, are provided in Section 5.

2 Related work

Research on the use of mobile phone data for behavioral studies and planning purposes has been widely conducted in recent years [5, 11]. The literature presents a plethora of papers in which UGD and the environment are jointly compared for several purposes. In fact, given the worldwide diffusion of mobile and wireless devices, information about where and when people are present can be obtained, allowing the determination of activity patterns by using counts of connections to the cellular network [17,36]. This process allows identifying and tracking groups, as well as gathering and/or extracting their behavioral patterns.

Nowadays, data collection from users is mainly dependent on mobile phones' location [28], social media [13], wireless sensors networks [9] and GPS tracking from mobile applications [25]. For instance, in [16], an empirical study uses a phone-based floating population to measure spatial accessibility to public transportation. In this regard, [32] shows the influence on individual car usage and CO_2 emissions by using a spatial and urban environment design and by examining individuals' data. In addition to their usefulness for the interpretation of urban dynamics, UGD are also useful for understanding the exposure of humans to a phenomenon, as in [24], in which the authors exploit mobile device-based mobility patterns to quantify the exposure of the population to air pollution. This study was conducted on a worldwide scale, demonstrating the huge potential of using pervasive computing. From the path planning point of view, a significant study conducted in Rome has been reported by [28], in which they delineate the basis of urban planning by using mobile phone data and location-based services. Another interesting application is proposed in [12], in which mobile phone traffic coming from different cities is analysed and some statistical patterns are created for marketing policies and economic growth.

Social media is also worth considering for dealing with digital footprints. Mobile social networking and geo-coding can potentially serve as decision support systems for urban planners. In [30], the Alexplore application was developed with the goal of determining the influence of social networks on tourism and city planning. In the most recent literature, the challenges and opportunities involved in this data collection method can be found in [19].

However, mobile phone networks and social media are not the only means of data collection; recent research trends demonstrate that Bluetooth technology is also increasingly taking hold. Bluetooth is simple and low cost when it comes to analysing spatial behavior and providing users with contextual information. In [31], an efficient method, proximity-based Bluetooth tracking, is considered for the analysis of the complex spatio-temporal dynamics of visitors' movements in mass events. In this aspect, the work of the Senseable City Lab [2] is noteworthy. Among other studies, [34] analyses pedestrians' behavioral patterns in the shopping environment in the historical centre of Barcelona, Spain. The authors

[2]http://senseable.mit.edu

use a Bluetooth detection technique meant for capturing a large-scale dataset of pedestrian behavior over a one-month period. Some other meaningful examples can be found in [20,27,29]. Generally, the collection of media access control (MAC) address data allows the analysis of the spatio-temporal dynamics of humans in terms of their use of shared space, time spent and frequency [1]. However, regardless of the source of UGD, common limitations still exist: first, the amount of data is largely conditioned by ownership of a smartphone and access to an Internet connection. Second, collecting information from large metropolitan areas rather than rural or peri-urban areas is easier. Third and foremost, data can be affected by social, cultural and economic aspects that may influence the results of the analysis. In this scenario and returning to the main novelty introduced by our work, knowledge of positioning data alone is not enough for a fruitful analysis. Introducing the concept of semantic trajectories, referring to the sequences of episodes with geo-located information, is therefore necessary [22,39]. Generally, the increasing use of location-aware devices has led to the increasing availability of trajectory data [4]. Therefore, research efforts in this field have been devoted to developing methods for the analysis of trajectories, including also various data mining methods [33]. Trajectory data mining can encourage the discovery of individuals' behaviors and can facilitate their comprehension by comparing them for different types of recommendations [35]. The use of clustering strategies can be relevant for, among other things, urban planning. Interested readers can find the categorisation made in that paper useful. Other interesting papers that deal with trajectory clustering for applications to traffic are [18,21].

Regardless of their use, users' trajectories are information-rich features that can reveal the structure of a place and give new insights into the events taking place, as well as the kinds of interactions that occur [38]. Within the framework of our research, we have used two different algorithms, optimising them for the purpose of understanding the potential of POS. These are the agglomerative [6] and spectral clustering algorithms [23,37], which will be explained in detail in the following sections. As an unsupervised learning system, our approach overcomes the above-mentioned limitations.

3 Description of the method

This section provides a comprehensive explanation of the present research, which uses phone-based data to cluster different kinds of users during their visits to a public park. We use the trajectory clustering method to transform point data into flow data in order to better understand their relations with the specific characteristics of some parts of the park. We also propose the use of these flows to determine sub-categories of users and identify the mission of the park, referring to the intention behind a pedestrian's visit. For the sake of completeness, Figure 3 presents the general schema of the proposed method.

3.1 Data set and trajectory detection

The raw data stored in the .csv file are in the form of a series of rows, each of which contains four fields that represent an identifier of the device that comes into contact with the beacon, coordinates of the point expressed as latitude and longitude, and a timestamp that represents the temporal instant of the device beacon contact.

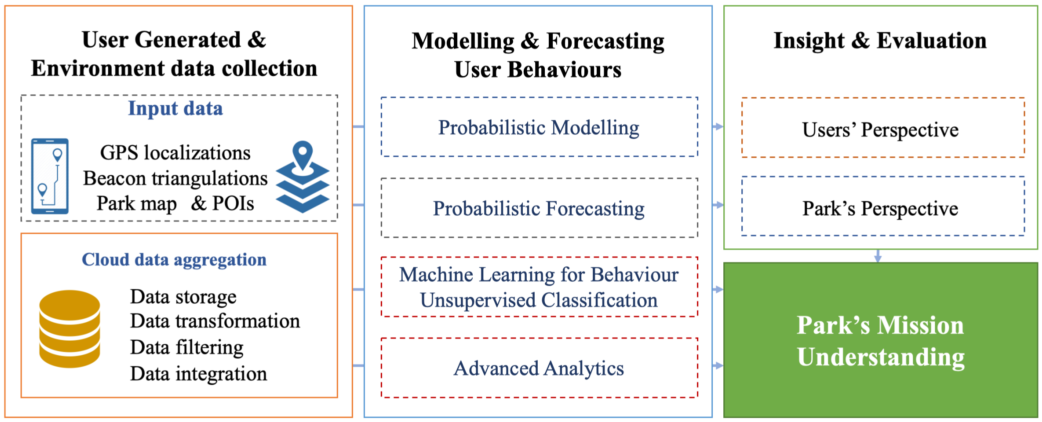


Figure 3: **General overview of the research**. From the collection of the data to their modelling, analysis and interpretation for planning purposes.

For the conversion of the coordinates to the appropriate reference system, the Geopy module was utilized for the planar projection of the GPS data into the X, Y coordinates. We therefore obtained the conversion for both GPS tracks and the positions of the beacons in the WGS84 reference system. Furthermore, to obtain local coordinates, the origin of the system was moved to the upper left part of the map. A specific class was then defined for populating the database (using the PeeWee module); in particular, the table containing the positions of the beacons gives each one a distinct ID. However, for the GPS trajectory, the ID cannot be unique because the same device can be tracked at different times, thus generating different trajectories (the device ID is therefore not a unique key, and for this reason, a pre-processing operation must be performed before one can obtain a unique trajectory). The remaining structure of the database had been fed with time stamps and the sampling times of the tracking, with a threshold of 5 seconds. The tracking process was in operation for a period of 6 months, from March 2017 to September 2017. The total number of devices tracked after cleaning and filtering is about 150 users. To obtain a whole trajectory from the point data, the database is queried to keep all the coordinates of a single device. Given that the GPS data are coarse and that some temporal holes may occur during an individual tracking session of the same device, the threshold used for splitting a trajectory is 80 seconds of absence of transmission. The Cardeto Dataset is publicly available[3] for researchers who aim to perform further clustering tests. Specifically, if the signal of a specific device is lost for a time frame longer than the threshold and then appears in another region, this signal is split into two trajectories. The algorithm, better described in the following subsection, requires a specific definition of the concept of a *trajectory*. For each object, which is composed of the list of points tracked for a specific device, associating a set of parameters, such as distance, total length and duration, is possible. All these features will pave the way for the final clustering of the trajectories, described in the following. Previously, consistent data filtering is required to ascertain that the algorithm is working properly.

[3]http://vrai.dii.univpm.it/content/cardeto-dataset

3.1.1 Data filtering

Three filtering operations are applied before clustering: one during the process of identifying a trajectory (positions out-of-bounds) and two after the identification (densities of points and Kalman filter).

- *Positions out-of-bounds*: Because of noise in the detection of position by the beacons, some devices may be placed outside the bounds of the map (positions with negative x and/or y coordinates). Pedestrians having positions that are out-of-bounds are skipped and treated like outliers in the trajectory identification process. Outliers [15] refer to objects that are inconsistent with the general behavior or pattern of the data. In this study, the outliers are generated by the inaccuracy of GPS or by the chaotic behavior of pedestrians. Furthermore, a point whose distance form the previous point is unreal (the distance to be covered needs an impracticable speed for a pedestrian) is treated as an outlier and, for this reason, discarded.
- *Densities of points*: The filtering of points whose distance is too low gives us limited information at the trajectory level. In the worst case, points with a null distance result in the failed execution of the algorithm. For this reason, points that are too close are discarded through the use of a density filter.
- *Kalman filter*: A standard Kalman filter [14] is then applied to each trajectory to smoothen it by removing noise and GPS inaccuracies.

Once the dataset was successfully prepared, the clustering approach was then applied. In this research, two different algorithms were used: the agglomerative and spectral clustering algorithms. The first one enables the user to define a priori the number of clusters, whereas the latter calculates the centroid of each single trajectory cluster and performs the segmentation automatically. To make the spectral algorithm work properly, the trajectories were further subdivided into sub-trajectories by using the following criteria: with the active beacons placed at the strategic POIs of the park, the trajectory is split if it passes a control region (areas marked on the map with a blue rectangle, indicating zones of particular interest in the park).

3.2 Trajectory clustering

This section presents an overview of the design of our framework. We assume that not all trajectories in the park begin in the same area. The purpose is to avoid constraining the algorithm from starting the computation from a specific part. Section 3.2.1 formally presents the problem statement. Section 3.2.2 and Section 3.2.3 present the two clustering methods used for the comparison. We write d_{ij} for the distance between the *i*th and *j*th input trajectories (this is used by the agglomerative method), and k_{ij} represents their similarity degree after the selection of an appropriate scale (this is used by the spectral method).

3.2.1 Problem statement

The partition-and-group framework introduced in [15] is used to develop a clustering algorithm. Given a set of trajectories $I = \{T_1, T_2, \ldots, T_n\}$, our algorithm generates a set of clusters $O = \{C_1, C_2, \ldots, C_n\}$, as well as a representative trajectory for each cluster. The following definitions are given for trajectory, representative trajectory and cluster. A trajectory T is a sequence of points $\{p_1, p_2, \ldots, p_i, \ldots, p_k\}$, $(1 \leq i \leq n)$, where p_i, $(1 \leq i \leq k)$ are points. The length of this trajectory is k. Different trajectories can have different lengths. A trajectory $p_{d1}, p_{d2}, \ldots, p_{dt}$, $(1 \leq d_1 < d_2 < d_t \leq k)$ is called a sub-trajectory of T.

A cluster is a set of line segments $p_i p_j$ (i < j), with p_i and p_j being the points chosen from the same trajectory. Line segments belonging to the same cluster are considered to be close to one another. It is important to stress that one trajectory can be a part of several clusters. In fact, a trajectory is partitioned into several line segments, and it is these line segments that are clustered. A representative trajectory (which means a sub-trajectory that occurs frequently) is a sequence of points similar to an ordinary trajectory. It is an imaginary trajectory that represents the behavior of the parts of the trajectory (i.e., the line segments) that are a part of the cluster.

Worth mentioning here are some rules that have been set up for conducting a trajectory clustering that will be suitable for the kind of data available for this study.

- With the agglomerative algorithm, clusters are made up of complete trajectories, whereas with the spectral algorithm, a cluster can be composed of sub-trajectories (hence composed of segments of lines *p-i; p-j*, where the latter belong to the same line. In the case of using the spectral algorithm, this means that a line can belong to more than one cluster.
- Some control regions have been placed, through which most of the trajectories pass and which are therefore considered strategic areas, generally because of the high density of visitors and being located where flow decisions are taken.

3.2.2 Agglomerative clustering

In order to execute the agglomerative clustering algorithm, we have to choose, in the first instance, a measure of similarity between trajectories. In this section, we propose a trajectory similarity measure based on the Hausdorff distance. Considered the universal set Ω, the metric $d : \Omega \times \Omega \rightarrow \mathbb{R}$ (corresponding to the Euclidean distance in our case), the (directed) Hausdorff distance between two sets $P \subseteq \Omega$ and $Q \subseteq \Omega$ is defined as follows:

$$h(P, Q) = \max_{\mathbf{p} \in P} \left\{ \min_{\mathbf{q} \in Q} d(\mathbf{p}, \mathbf{q}) \right\} \tag{1}$$

In order to extend the distance measurement h to a metric, it is defined as follows:

$$H(P, Q) = \max h(P, Q), h(Q, P) \tag{2}$$

where it should be noted that in general, $h(P, Q) \neq h(Q, P)$.

A weakness of such a distance is that a single outlier in one of the two sets can lead to an arbitrarily large distance between these sets even if they are, in fact, identical. In order to overcome this issue, we reject a certain proportion of the worst matches under the maximum obtained in (1), thus narrowing the set under which the minimum is taken:

$$h_\alpha(P, Q) = \overset{\alpha}{\underset{\mathbf{p} \in P}{ord}} \left\{ \min_{\mathbf{q} \in Q} d(\mathbf{p}, \mathbf{q}) \right\} \tag{3}$$

More specifically, the operator $ord^{\alpha}_{s \in S} f(s)$ denotes the value among the image $f(S)$ of the set S that is greater than $\alpha|f(S)|$ of all the values. The meaning of α can be easily understood using some examples:

$$\overset{1}{\underset{s \in S}{ord}} f(s) = \max_{s \in S} f(s) \tag{4}$$

$$\overset{0}{\underset{s \in S}{ord}} f(s) = \min_{s \in S} f(s) \tag{5}$$

$$\overset{1/2}{\underset{s \in S}{ord}} f(s) = \underset{s \in S}{median}\, f(s) \tag{6}$$

The rejection of a certain proportion of large distances in the Hausdorff distance has been suggested in [8], in which it is pointed out that the choice of the maximum between the shortest distances is extremely sensitive to a spurious point (even a single outlier in one set can affect the distance measurement).

A choice of $\alpha < 1$ not only makes the distance robust to a fraction of outliers $1 - \alpha$ but also allows partial trajectories to match and can lead to violations of the triangle inequality, resulting in inconsistent clustering. The overall clustering results are sensitive to the choice of α; running the algorithm with different values of α, increasing the differences in α, leads to different clustering results as more partial comparisons are allowed. Starting from these considerations, the value of α is fixed to 0.88. The value was determined experimentally.

Once the measure of similarity between trajectories has been defined, we execute the agglomerative clustering algorithm by initially assigning each trajectory to a singleton cluster. The two closest clusters are then iteratively merged until all clusters are separated by a distance greater than a specified threshold τ. If I and J denote the sets of indices of the two disjoint clusters' trajectories, the distance between the clusters is as follows:

$$d_{avg} = \frac{1}{|I||J|} \sum_{i \in I} \sum_{j \in J} d_{ij} \tag{7}$$

where d_{ij} is defined as follows:

$$d_{ij} = \frac{1}{h_\alpha(i,j) \cdot h_\alpha(j,i)} \tag{8}$$

with $h_\alpha(i,j)$ representing the modified Hausdorff distance between the trajectory with index i and that with index j. In our tests, we fix the value of threshold τ as 10% of the maximum distance (the distance by which all trajectories are merged into a single cluster). This threshold has been chosen according to the following criteria: if, overcoming a certain value, the algorithm joins two clusters that are very different from one another, the threshold is set to avoid this from occurring. The distance can be deduced by the dendrogram. After the distance values between the clusters and the relative dendrogram are observed (Figure 4), the number of clusters to be obtained is fixed at a value equal to 8.

3.2.3 Spectral clustering

As mentioned above, we use the clustering algorithm proposed in [3]. The number of clusters is chosen automatically using the distortion metric for selecting the optimal global scale [23]. The input to the spectral clustering method is the affinity matrix K, containing

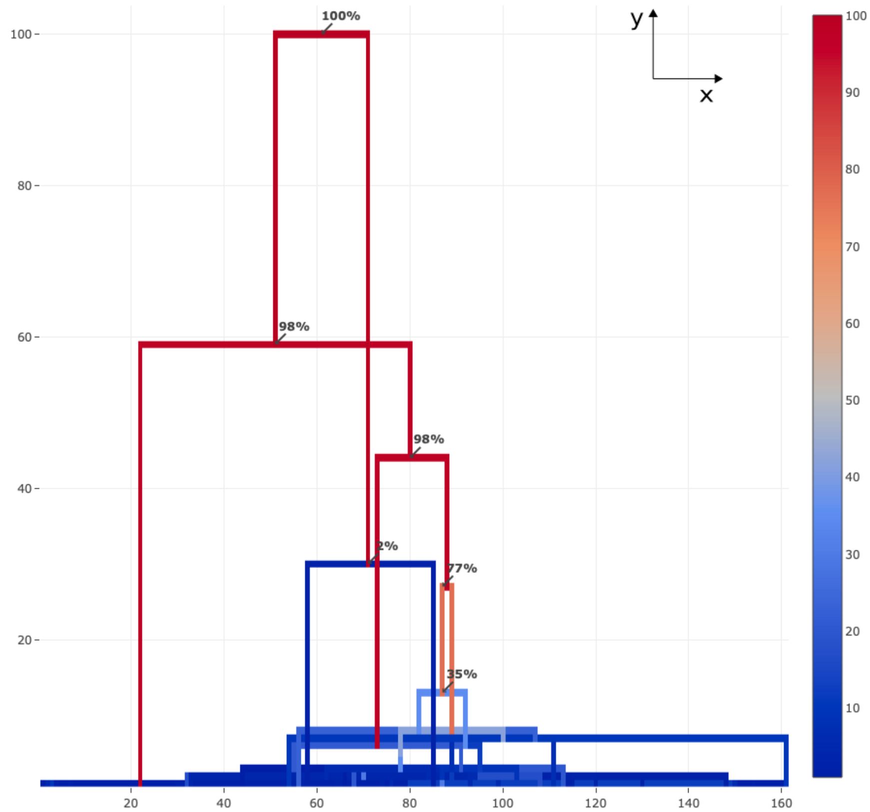


Figure 4: Clusters' Dendrogram related to the Agglomerative Algorithm. On the y-axis, we have the values, in percentage, of the distances between clusters referred to the maximum ones ; on the x-axis, there are the indexes of the 161 trajectories found, starting from the dataset in our possession.

the elements k_{ij} for $1 \leq i, j \leq n$. The process begins with the normalization of the rows and columns of K, leading to a normalized affinity L:

$$L = W^{-1/2} K W^{-1/2} \tag{9}$$

where the *i*the lement of the diagonal matrix W is defined by

$$w_{ij} = \sum_{1 \leq j \leq n} k_{ij} \tag{10}$$

The matrix L is block diagonal and has g eigenvalues equal to 1 (one for each cluster found) and $n - g$ eigenvalues equal to zero. Counting the number of eigenvalues with a value of 1 is not useful for understanding the number of clusters [37]. For this reason, the range for the number of clusters is found using the spectrum of L. Assuming that $1 \geq \lambda_1 \geq \lambda_2 \cdots \geq \lambda_n \geq 0$ are the eigenvalues of L, the minimum number of clusters g_{min} in the data can be determined by counting the number of eigenvalues greater than 0.99; alternatively, gmax can be estimated by counting the number of eigenvalues greater than 0.8. The thresholds mentioned here were selected experimentally. In the case of some experiments, the possible range is very narrow, whereas for others, it is large enough. After the search range for the number of clusters is selected, the eigenvectors of L corresponding to the top g_{max} eigenvalues are computed. It is assumed that v_i represents these vectors for $1 \leq i \leq g_{max}$. In the case in which an eigenvalue repeats, the corresponding eigenvectors are required to be mutually orthogonal. The clustering algorithm proceeds by applying the following steps, which are repeated for each value of g that is an integer, ranging inclusively from g_{min} to g_{max}. With the purpose of making the algorithm consistent (i.e., making it converge to a final solution or making it produce a reasonable number of clusters), the trajectories have been split into sub-trajectories; in case one of them crosses a control region, it is split.

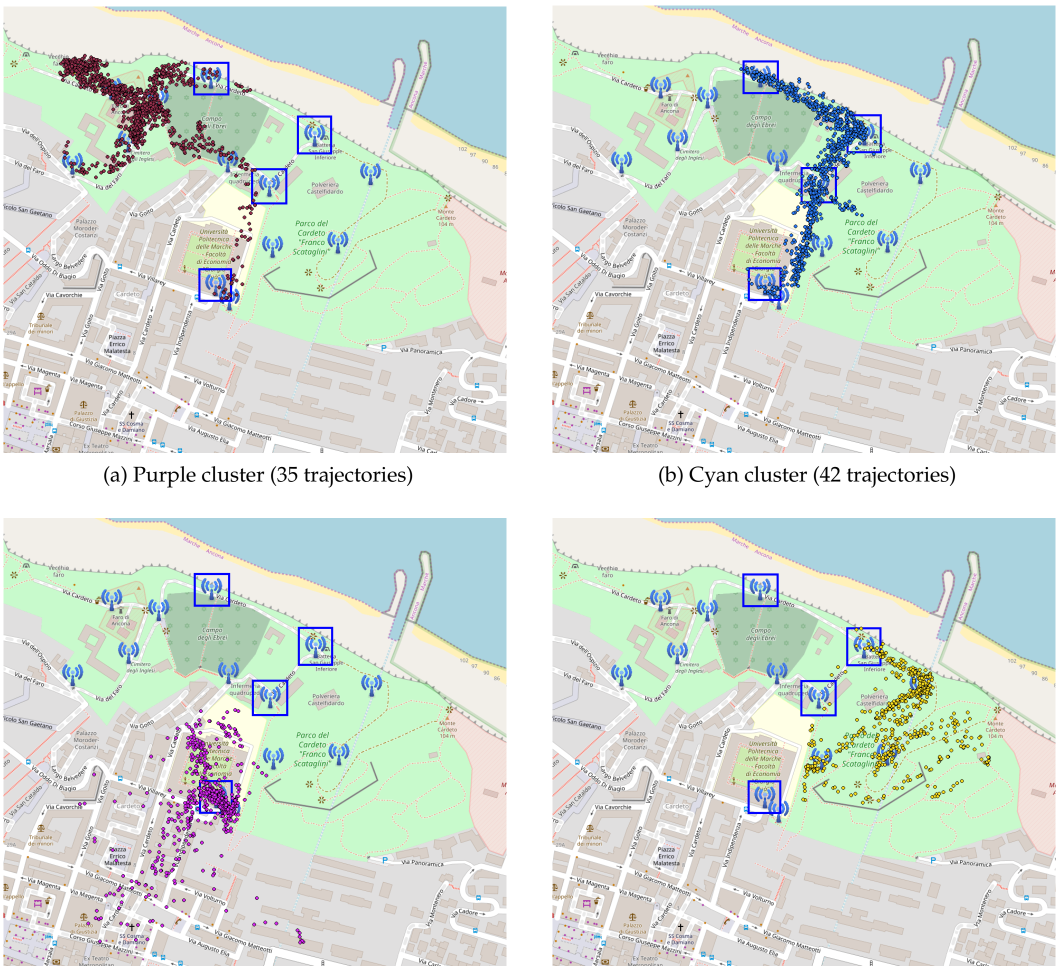
- Assume that g is an integer between g_{min} and g_{max}. Then, the $n \times g$ matrix $V = [v_1, \ldots, v_q]$ is formed.
- Each row of V is normalized to have unit length $R = SV$, where S is diagonal with elements $s_i = (\sum_{j=1}^{c} V_{ij})^{-1/2}$.
- The rows of R have to be considered as g-dimensional data points and clustered using k-means. Let $\mu_1, \mu_2, \ldots, \mu_q$ represent the centres of the g clusters (as row vectors), and let $c(i)$ represent the cluster that corresponds to the ith row $r_{(i)}$.
- The within-class scatter is $W = \sum_{i=1}^{n} ||r_{(i)} - \mu_{c_{(i)}}||_2^2$ and the total scatter is $T = \sum_{i=1}^{n} \sum_{j=1}^{k} ||r_{(i)} - \mu_j||_2^2$. The value of g that produces the least distortion p_g is the number of clusters determined automatically, and $c(i)$, which is obtained with the number of clusters, is the class indicator function for the input trajectories.

4 Results

This section presents the results of the experiments conducted on our dataset. In addition to the performance of the agglomerative and spectral clustering algorithms, a statistical data analysis is also presented. The trajectories taken into account have been computed and clustered following the steps previously mentioned. For the agglomerative clustering algorithm, the trajectories are grouped into eight different clusters; as stated above, this number has been chosen on the basis of the observations made on the distance between clusters and the relative dendrogram. However, for the spectral clustering algorithm, the number is computed automatically at run-time.

Figure 5 shows the most significant clusters (i.e., those composed of a meaningful number of trajectories) among the eight clusters calculated by the agglomerative algorithm. Their distribution on the map clearly denotes the different behaviors within the monitored area.

Figure 6a shows the entirety of the clusters calculated by the spectral algorithm, whereas Figure 6b presents a sample cluster chosen among the 37 clusters computed by the spectral algorithm. Thanks to the division into sub-trajectories (a trajectory that crosses a control region is cut into two sub-trajectories at the point where the control region, denoted by the blue squares in the map, is crossed), obtaining the convergence of the algorithm that will autonomously find a number of clusters equal to 37 is possible; furthermore, the division into sub-trajectories is also useful to better understand users' behavior. In fact, with this method, only the clusters containing trajectories that are covered with high frequency



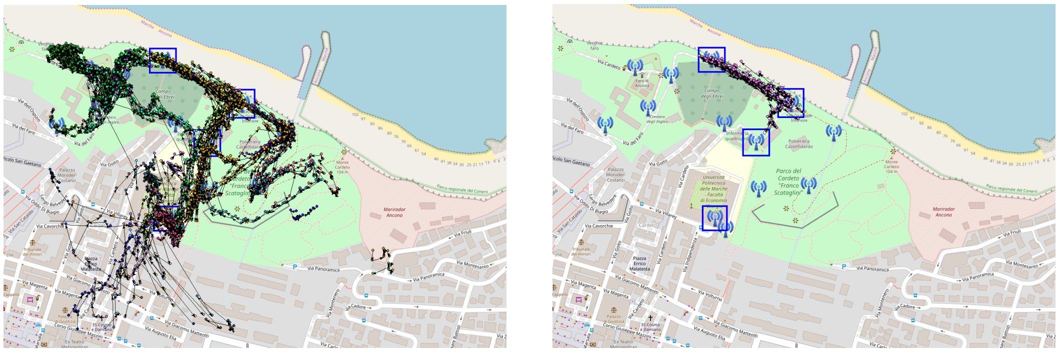
(a) Purple cluster (35 trajectories)

(b) Cyan cluster (42 trajectories)

(c) Pink cluster (48 trajectories)

(d) Yellow cluster (18 trajectories)

Figure 5: Meaningful clusters computed by the agglomerative algorithm. Each cluster is identified by a particular color (background: © OpenStreetMap contributors, CC BY-SA).



(a) (b)

Figure 6: Results obtained through the execution of the spectral algorithm. In Figure 6a, the totality of calculated clusters is shown, whereas in Figure 6b, the details of one of the most significant cluster is shown.(background: © OpenStreetMap contributors, CC BY-SA).

and with a high degree of similarity remain evident, whereas all those that are traveled once are discarded (forming clusters with a low density).

It is fair to say that the tests of the two clustering approaches (spectral vs. agglomerative) have given similar outputs in terms of the categorisation of park uses. Nevertheless, the agglomerative algorithm seems to better fit the kind of data available, as it takes into account a larger amount of data, as demonstrated by the results reported in Table 4; for this reason, we will consider in the subsequent analysis the results obtained from the latter.

Name	Clustering Results (%)
pink	32.43%
cyan	28.38%
purple	23.65%
yellow	12.16%
other	3.38

Table 1: Results of agglomerative trajectory clustering and correspondence between the number of clusters and the percentage of data processed.

Multiple comparisons of the clusters led to the identification of four significantly different regions. This reflects the structure of the park and is in agreement with the four regions that characterize it: the cultural area, the natural area, the walking area and the panoramic area (Figure 7). The identification of such regions within the park is only possible after the clustering of trajectories, which then allows the aggregation of the latter into groups. The visualisation of the trajectories in Figure 6 is therefore to be considered as a heatmap that gives information on the most frequented areas of the park by dividing the entire set of trajectories into sub-groups (clusters).

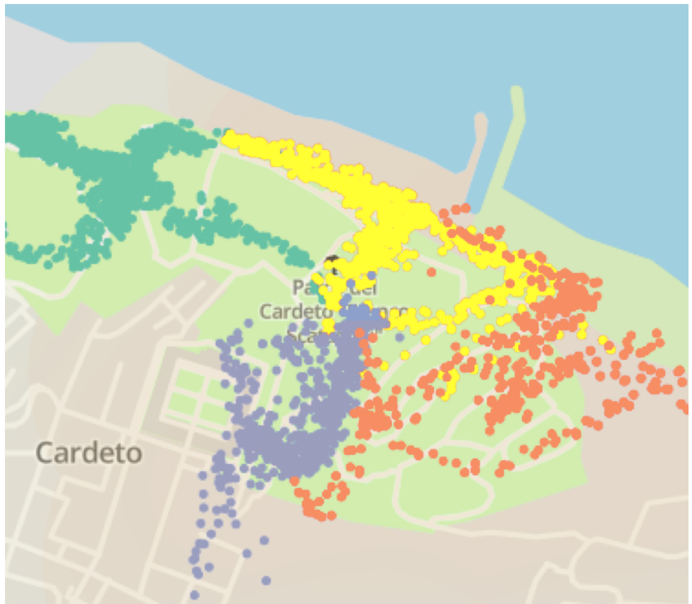


Figure 7: **Regions of the park from the clustering results.** Green indicates the panoramic area, orange indicates the walking area, pink indicates the natural area and blue indicates the cultural area.

4.1 Data analytic layer

The data analytic layer is based on R Studio[4] and provides access to the AWS Redshift data layer for the raw data, the trajectories and the clustered data. The latter, once processed, can be managed within the information system, allowing users to extract statistical information. In particular, the developed information system is designed to merge the geographic information of the park together with users' data (i.e., clustering) to perform advanced analytics. Additional relevant pieces of information are, among others, the i) spatial distribution of users or clusters, ii) sub-region-based statistics, iii) daily and hourly distributions per cluster and region, iv) region-based inbound and outbound analysis (people coming from or going to a certain region) and v) spatial weekly distribution. All these analytic insights can be filtered following different criteria, such as period, cluster and region of the park. To show the significance of the cluster variation in the park and the benefits of the statistical analysis, some examples of possible analysis are reported in the following. We calculated the spatial distribution of visitors in the areas identified with the clustering algorithm. Figure 8 depicts the macro regions into which the trajectory data can be divided after their clustering.

Figure 9a shows, in percentage, how many times regions are visited. In particular, more than 50% visit every region, 21% visit only one region, 11% visit two regions and 14% visit three regions. These data can give insights into whether the park is used in its entirety by visitors. The average of ID identifications in each region is shown in Figure 9b. These data

[4]https://www.rstudio.com/

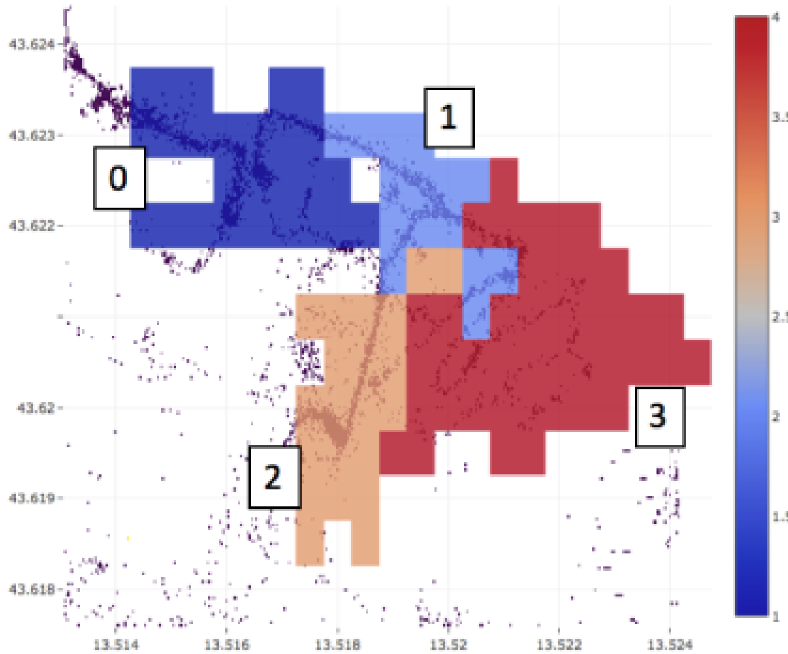


Figure 8: Macro regions that subdivide the data. The values on the horizontal axis and vertical axis represent the longitude and latitude intervals, respectively, for the considered geographical area.

give the planner an overview of the average number of visitors who visit each region of the park.

Given N total visitors per hour, M total number of visitors per area per hour, P total visitors per area per day of the week, Q total visitors per day, R total number of visitors per area per week day interval and S total visitors per interval of days, more information about the hourly distribution can be extracted. This is represented in Figure 10a, where $Count = N$ and the different colors represent in which area they were.

Figure 10b is the hourly normalized distribution, where *Count* is the hourly distribution of people in different areas, i.e., $Count = M/N$. In particular, how often some areas are frequented at different times can be observed. For example, there is a peak in region 0 at lunch time. Furthermore, in the morning, people visit regions 3 and 1.

Further spatial distribution information can be deduced by analysing user behavior during week days (Figure 11a, where $Count = P/Q$) and during weekends (Figure 11b, where $Count = R/S$. 0 is from Monday to Sunday, and 1 represents Saturday and Sunday). In particular, Figure 11b shows how some areas are frequented during the weekend or during the rest of the week.

We repeated these calculations for both weekly and hourly spatial distributions. In the hourly case, we defined time intervals, and the color bar ranges from 0 to 1, with the color grey indicating few people and the color red indicating many people. The spatial weekly distribution is shown in Figure 12; Figure 13 shows instead the spatial hourly distribution

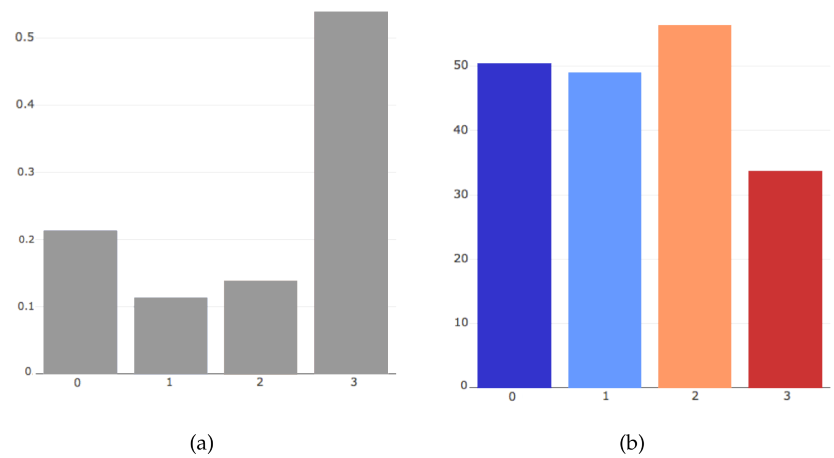


Figure 9: Statistical Results. Figure 9a is the percentage of visitors who visit the areas of the park (where the horizontal axis is the progressive number of areas and the vertical axis is the percentage). Figure 9b is the average of ID identifications in each region (where the horizontal axis is the specific area and the vertical axis is the average number).

defined by four time slots: 8 a.m. to 11 a.m., 12 a.m. to 2 p.m., 3 p.m. to 7 p.m. and 8 p.m. to 10 p.m. In other words, the color changes as a function of visitor density in

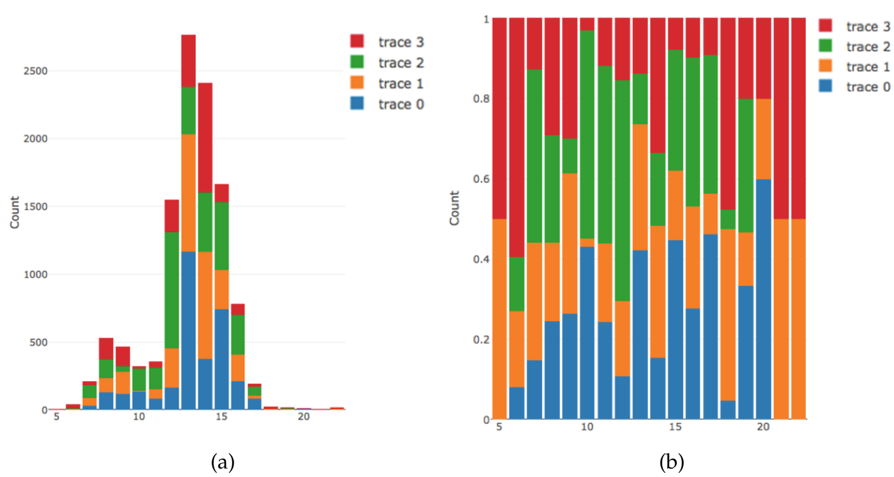


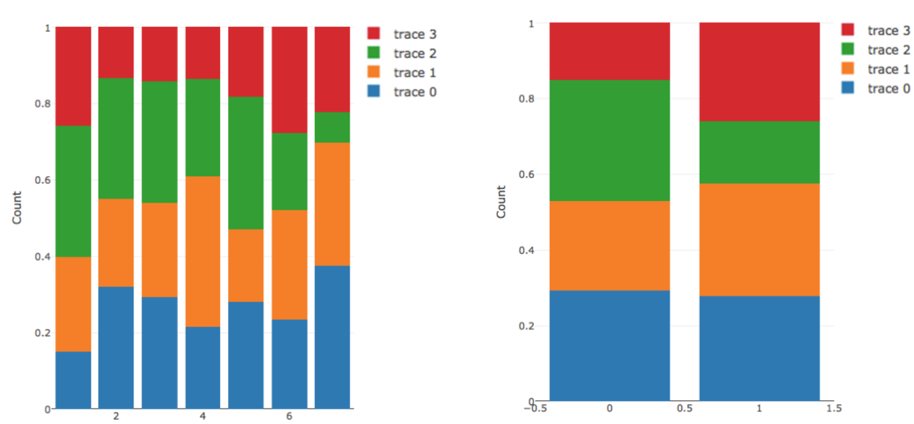
Figure 10: Hourly distribution.Figure 10a is the number (count) of visitors for each region divided per hour (only the opening hours of the park are considered). Figure 10b represents the normalized distribution. Each region of the park is identified by a different trace, as shown in the legend.

one specific region with respect to the total number of visitors in a specific time slot of a week day. When the area is red, it means that almost all visitors went through that region; consequently, if more red cells appear, it means that all visitors explored more regions. From our findings, in the first two time slots, people visit only peripheral areas of the park, while in the afternoon can be observed a higher density of visitors in all the areas of the park. These pictures also denote a clear change in visitors' path depending on the week day; for instance, the weekend changes the path preferred by users.

5 Discussion and conclusion

This paper has presented a novel method for tracking and analysing the spatio-temporal dynamics of humans in terms of spatial behavior. To achieve this purpose, real data collected from an urban park over a period of six months have been used. The data analytic layer, thanks to the dashboard, enables one to visualize data processed by the clustering algorithm, which proved to be reliable. It allows computing point data and transforming them into trajectories that enable the identification of specific patterns of behavior, thus inferring the main uses of the park, i.e., its mission. From the analysis, we found four main categories: green, cultural, panoramic and walking. This new method of processing GPS data for the purposes of urban planning sheds new light on the way in which this information can be provided to planners, allowing the discovery of aspects of pedestrian behavior, such as the number of places visited and the order of their being visited, and users' length of stay in the park. This makes an important contribution to the study of POS.

The reliability of the clustering method has also been confirmed with an additional statistical analysis. In fact, a comparison between the clustering results and the macro-regions discovered (as shown in Figure 10) shows that they lead to analogous results. The clus-



(a) (b)

Figure 11: Week days and weekend analysis. Figure 11a is the weekly user distribution in the four regions from Monday to Sunday, and Figure 11b is the weekend user distribution in the four regions.

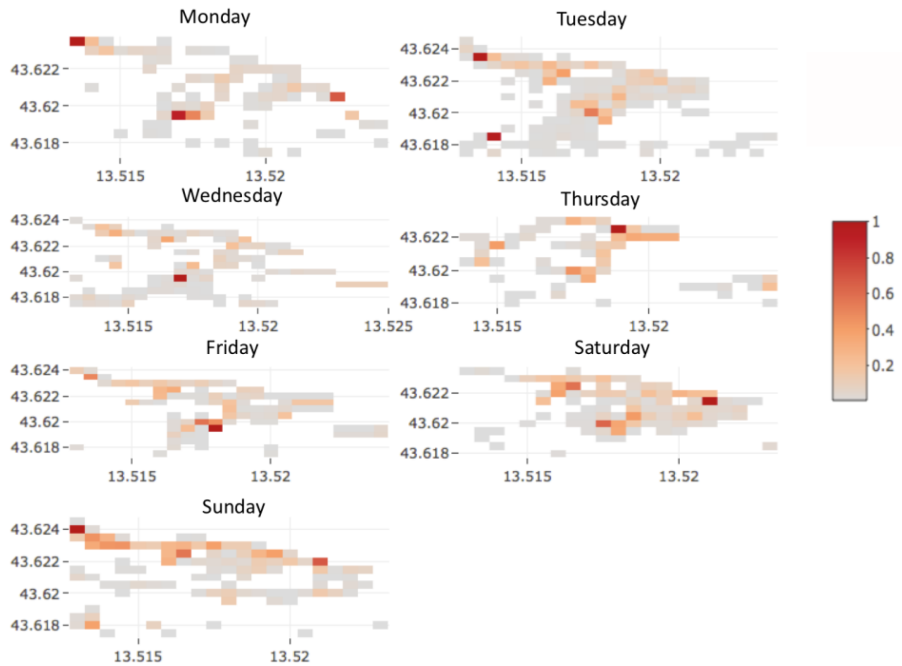


Figure 12: Spatial weekly distribution. For each figure, we have the latitude and the longitude of the study area on the horizontal axis and vertical axis, respectively. The colored sidebar is used to visually represent the number of visitors for each area of the grid. This number is normalized with respect to a maximum value represented in red on the map.

tering algorithm gives suitable information and ensures the removal of outliers, providing more accurate details to planners.

Some interesting findings in terms of how POS are used have also been obtained from the results. Visitors prefer to use the park during the early afternoon. At noon, the park is not visited quite as much, and at certain hours, it is not visited at all. This may be attributed to the lack of security in the park, especially its green areas. Furthermore, there is variability in the use of the park for different purposes. For instance, the panoramic areas are the preferred path, meaning that the park is mainly used for relaxation. Various areas are exploited in different ways because of the environmental variables involved; these require further investigation by park managers. For instance, one variable can be the low attractiveness (lack of points of interest) and/or presence of obstacles that prevent the visit, absence of routes and so on. As this space is very important in terms of cultural heritage, the policies implemented by the municipality to advertise it are perhaps insufficient to make the citizens fully aware of this heritage location. Indeed, the results of the analysis should be deepened by expert urban planners in order to optimize the use of the park, thus exploiting the hidden potential of the park itself. The experiment proves that the system highlights both the potential of and the issues with the park, but their interpretation is entrusted on

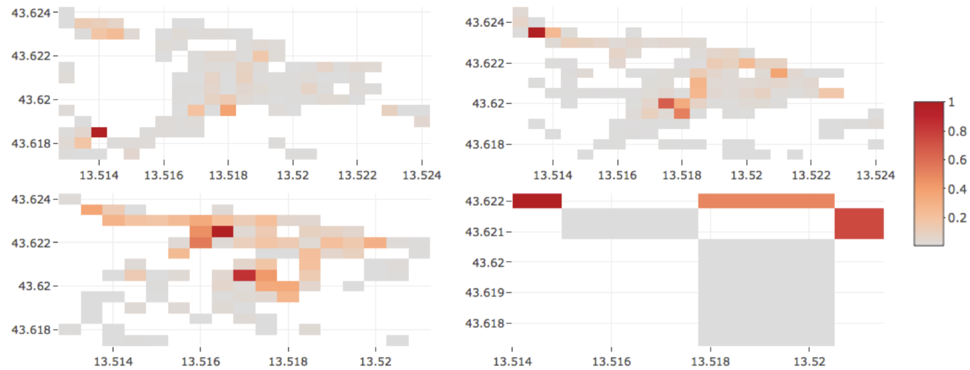


Figure 13: Spatial hourly distribution. For each figure, we have the latitude and the longitude of the study area on the horizontal axis and vertical axes, respectively. Figure 13a shows the spatial distribution for the time interval ranging from 8 a.m. to 11 a.m. The remaining distributions for the subsequent time intervals (i.e., 12 a.m. to 2 p.m., 3 p.m. to 7 p.m. and 8 p.m. to 10 p.m.) are shown in Figures 13b, 13c, 13d. The colored sidebar is used to visually represent the number of visitors for each area of the grid. This number is normalized with respect to a maximum value represented in red on the map.

planners and on knowledge of the park itself. Nowadays, planners, decision makers and public administrators conduct their design process with a *pen-and-pencil* strategy based on visual observation or direct interviews, leading to a subjective interpretation of the data. In the era of big data, a data-driven approach should instead be chosen in order to develop a real *measure to design* policy. When compared with other observational and subjective reporting methods used previously in this research field, the method presented here offers a number of benefits. First, with the use of a mobile application, users' information can be gathered whilst they are interacting directly with the space. This means that the information obtained can be easily applied and standardized for other parks in order to expand the research. Second, previous studies have been conducted with a large scale of observation (i.e., using social media footprints), whereas the method used in the current study is efficient in examining small areas in a very precise way, contributing to the implementation of planning strategies for specific areas of the urban fabric. Despite the above contributions, our method also has some shortcomings, resulting in the limitations of our research. These limitations involve both technical and human factors. The process of collecting the data is application based, meaning that we can collect information only from those visitors who are using the app. Furthermore, the sample is not very consistent, and we might need to collect more data in order to verify whether the monitored trend is maintained over time or for the different seasons of the year. Finally, statistical inferences considering demographic aspects have not been performed, since such data was unavailable; it would be interesting to consider social information to expand the results. Speacking about privacy, users' information is stored using the MAC address, which uniquely identifies every device and thus makes it impossible to link to any personal information. Moreover, users were aware of being tracked, as they were asked to accept terms and conditions before installing the app from the stores. With the advent of the new General Data Protection Regulation, MAC addresses can be considered linkable information. Considering this legal obligation in future implementations would be desirable. To conclude, the suggested approach to the behavioral segmentation and characterisation of visitors may support planning decisions that can have a significant impact on users' level of satisfaction. This approach can help planners in setting some guidelines for a new redesigned park layout. Widening the vision for future perspectives, we can move from a feature-based layout to a mission-based one that can help visitors exploit the real potential of green areas.

Acknowledgment

This work was supported by Cost Action TU1306, which is called CYBERPARKS: Fostering knowledge about the relation between Information and Communication Technologies and Public Spaces supported by strategies to improve their use and attractiveness. The authors would also like to thank Grottini Lab and Dr. Claudia Norscini for their support in processing the data.

References

[1] ABEDI, N., BHASKAR, A., AND CHUNG, E. Tracking spatio-temporal movement of human in terms of space utilization using media-access-control address data. *Applied Geography 51* (2014), 72–81. doi:10.1016/j.apgeog.2014.04.001.

[2] ALBERT, C., AND VON HAAREN, C. Implications of applying the green infrastructure concept in landscape planning for ecosystem services in peri-urban areas: An expert survey and case study. *Planning Practice & Research 32*, 3 (2017), 227–242. doi:10.1080/02697459.2014.973683.

[3] ATEV, S., MILLER, G., AND PAPANIKOLOPOULOS, N. P. Clustering of vehicle trajectories. *IEEE Trans. on Intelligent Transportation Systems 11*, 3 (2010), 647–657. doi:10.1109/tits.2010.2048101.

[4] BAO, J., ZHENG, Y., AND MOKBEL, M. F. Location-based and preference-aware recommendation using sparse geo-social networking data. In *Proc. of the 20th international conference on advances in geographic information systems* (2012), ACM, pp. 199–208. doi:10.1145/2424321.2424348.

[5] BOJIC, I., MASSARO, E., BELYI, A., SOBOLEVSKY, S., AND RATTI, C. Choosing the right home location definition method for the given dataset. In *International Conference on Social Informatics* (2015), Springer, pp. 194–208. doi:10.1007/978-3-319-27433-1_14 .

[6] BUZAN, D., SCLAROFF, S., AND KOLLIOS, G. Extraction and clustering of motion trajectories in video. In *Proc. of the 17th International Conference on Pattern Recognition (ICPR 2004)* (2004), vol. 2, IEEE, pp. 521–524. doi:10.1109/icpr.2004.1334287.

[7] DODGE, S., LAUBE, P., AND WEIBEL, R. Movement similarity assessment using symbolic representation of trajectories. *International Journal of Geographical Information Science 26*, 9 (2012), 1563–1588. doi:10.1080/13658816.2011.630003.

[8] DUBUISSON, M. P., AND JAIN, A. K. A modified hausdorff distance for object matching. In *Proc. of 12th International Conference on Pattern Recognition* (1994), pp. 566–568. doi:10.1109/icpr.1994.576361.

[9] FRONTONI, E., MANCINI, A., PIERDICCA, R., STURARI, M., AND ZINGARETTI, P. Analysing human movements at mass events: A novel mobile-based management system based on active beacons and AVM. In *24th Mediterranean Conference on Control and Automation (MED)* (2016), IEEE, pp. 605–610. doi:10.1109/med.2016.7536047.

[10] GIRARDIN, F., CALABRESE, F., DAL FIORE, F., RATTI, C., AND BLAT, J. Digital footprinting: Uncovering tourists with user-generated content. *IEEE Pervasive Computing 7*, 4 (2008). doi:10.1109/mprv.2008.71.

[11] GOLDER, S. A., AND MACY, M. W. Digital footprints: Opportunities and challenges for online social research. *Sociology 40*, 1 (2014), 129. doi:10.1146/annurev-soc-071913-043145.

[12] GRAUWIN, S., SOBOLEVSKY, S., MORITZ, S., GÓDOR, I., AND RATTI, C. Towards a comparative science of cities: Using mobile traffic records in New York, London, and Hong Kong. In *Computational approaches for urban environments*. Springer, 2015, pp. 363–387. doi:10.1007/978-3-319-11469-9_15.

[13] HEIKINHEIMO, V., MININ, E. D., TENKANEN, H., HAUSMANN, A., ERKKONEN, J., AND TOIVONEN, T. User-generated geographic information for visitor monitoring in a national park: A comparison of social media data and visitor survey. *ISPRS International Journal of Geo-Information 6*, 3 (2017), 85. doi:10.3390/ijgi6030085.

[14] JULIER, S. J., AND UHLMANN, J. K. A new extension of the Kalman filter to nonlinear systems. In *International Symposium of Aerospace/Defense Sensing, Simulation and Controls* (1997), vol. 3, Orlando, FL, pp. 182–193. doi:10.1117/12.280797.

[15] LEE, J.-G., HAN, J., AND WHANG, K.-Y. Trajectory clustering: A partition-and-group framework. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data* (2007), ACM, pp. 593–604. doi:10.1145/1247480.1247546.

[16] LEE, W. K., SOHN, S. Y., AND HEO, J. Utilizing mobile phone-based floating population data to measure the spatial accessibility to public transit. *Applied Geography 92* (2018), 123–130. doi:10.1016/j.apgeog.2018.02.003.

[17] LIEBIG, T., XU, Z., AND MAY, M. Incorporating mobility patterns in pedestrian quantity estimation and sensor placement. In *Citizen in Sensor Networks*. Springer, 2013, pp. 67–80. doi:10.1007/978-3-642-36074-9_7.

[18] LIU, Y., WANG, F., XIAO, Y., AND GAO, S. Urban land uses and traffic 'source-sink areas': Evidence from GPS-enabled taxi data in Shanghai. *Landscape and Urban Planning 106*, 1 (2012), 73–87. doi:10.1016/j.landurbplan.2012.02.012.

[19] MARTÍ, P., SERRANO-ESTRADA, L., AND NOLASCO-CIRUGEDA, A. Social media data: Challenges, opportunities and limitations in urban studies. *Computers, Environment and Urban Systems 74* (2019), 161–174. doi:10.1016/j.compenvurbsys.2018.11.001.

[20] MASIERO, A., GUARNIERI, A., PIROTTI, F., AND VETTORE, A. A particle filter for smartphone-based indoor pedestrian navigation. *Micromachines 5*, 4 (2014), 1012–1033. doi:10.3390/mi5041012.

[21] MAZIMPAKA, J. D., AND TIMPF, S. Exploring the potential of combining taxi GPS and Flickr data for discovering functional regions. In *AGILE 2015*. Springer, 2015, pp. 3–18. doi:10.1007/978-3-319-16787-9_1.

[22] MONAJJEMI, P. P. Z. *A Clustering-Based Approach for Enriching Trajectories with Semantic Information Using VGI Sources*. PhD thesis, M. Sc. Thesis, ITC, 2013.

[23] NG, A. Y., JORDAN, M. I., AND WEISS, Y. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems* (2002), pp. 849–856.

[24] NYHAN, M., GRAUWIN, S., BRITTER, R., MISSTEAR, B., MCNABOLA, A., LADEN, F., BARRETT, S. R., AND RATTI, C. "Exposure track"—the impact of mobile-device-based mobility patterns on quantifying population exposure to air pollution. *Environmental science & technology 50*, 17 (2016), 9671–9681. doi:10.1021/acs.est.6b02385.

[25] OSABA, E., PIERDICCA, R., DUARTE, T., BAHILLO, A., AND MATEUS, D. Using ICTs for the improvement of public open spaces: The opportunity offered by CyberParks digital tools. In *CyberParks—The Interface Between People, Places and Technology*. Springer, 2019, pp. 278–293. doi:10.1007/978-3-030-13417-4_22.

[26] OSABA, E., PIERDICCA, R., MALINVERNI, E. S., KHROMOVA, A., ÃLVAREZ, F. J., AND BAHILLO, A. A smartphone-based system for outdoor data gathering using a wireless beacon network and GPS data: From cyber spaces to senseable spaces. *ISPRS International Journal of Geo-Information 7*, 5 (2018). 10.3390/ijgi7050190.

[27] PIERDICCA, R., LICIOTTI, D., CONTIGIANI, M., FRONTONI, E., MANCINI, A., AND ZINGARETTI, P. Low cost embedded system for increasing retail environment intelligence. In *IEEE International Conf. on Multimedia & Expo Workshops (ICMEW)* (2015), IEEE, pp. 1–6. doi:10.1109/icmew.2015.7169771.

[28] RATTI, C., FRENCHMAN, D., PULSELLI, R. M., AND WILLIAMS, S. Mobile landscapes: using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design 33*, 5 (2006), 727–748. doi:10.1068/b32047.

[29] STURARI, M., LICIOTTI, D., PIERDICCA, R., FRONTONI, E., MANCINI, A., CONTIGIANI, M., AND ZINGARETTI, P. Robust and affordable retail customer profiling by vision and radio beacon sensor fusion. *Pattern Recognition Letters 81* (2016), 30–40. doi:10.1016/j.patrec.2016.02.010.

[30] TAHA, D. S. The influence of social networks in visiting, planning and living in cities. Alexplore: A pilot project in Alexandria. *Alexandria Engineering Journal 52*, 3 (2013), 479–488. doi:10.1016/j.aej.2013.04.006.

[31] VERSICHELE, M., NEUTENS, T., DELAFONTAINE, M., AND VAN DE WEGHE, N. The use of Bluetooth for analysing spatiotemporal dynamics of human movement at mass events: A case study of the Ghent Festivities. *Applied Geography 32*, 2 (2012), 208–220. doi:10.1016/j.apgeog.2011.05.011.

[32] YANG, D., AND TIMMERMANS, H. Effects of urban spatial form on individuals' footprints: Empirical study based on personal GPS panel data from Rotterdam and Eindhoven area. *Procedia Environmental Sciences 22* (2014), 169–177. doi:10.1016/j.proenv.2014.11.017.

[33] YOON, H., ZHENG, Y., XIE, X., AND WOO, W. Social itinerary recommendation from user-generated digital trails. *Personal and Ubiquitous Computing 16*, 5 (2012), 469–484. doi:10.1007/s00779-011-0419-8.

[34] YOSHIMURA, Y., AMINI, A., SOBOLEVSKY, S., BLAT, J., AND RATTI, C. Analysis of pedestrian behaviors through non-invasive Bluetooth monitoring. *Applied geography 81* (2017), 43–51. doi:10.1016/j.apgeog.2017.02.002.

[35] YUAN, N. J., ZHANG, F., LIAN, D., ZHENG, K., YU, S., AND XIE, X. We know how you live: exploring the spectrum of urban lifestyles. In *Proc. of the first ACM conference on Online social networks* (2013), ACM, pp. 3–14. doi:10.1145/2512938.2512945.

[36] ZEITZ, K. M., TAN, H. M., GRIEF, M., COUNS, P., AND ZEITZ, C. J. Crowd behavior at mass gatherings: a literature review. *Prehospital and disaster medicine 24*, 1 (2009), 32–38. doi:10.1017/s1049023x00006518.

[37] ZELNIK-MANOR, L., AND PERONA, P. Self-tuning spectral clustering. In *Advances in neural information processing systems* (2005), pp. 1601–1608.

[38] ZHANG, L., AND VAN DE WEGHE, N. Attribute trajectory analysis: a framework to analyse attribute changes using trajectory analysis techniques. *International Journal of Geographical Information Science 32*, 5 (2018), 1043–1059. doi:10.1080/13658816.2018.1435885.

[39] ZHAO, P., QIN, K., YE, X., WANG, Y., AND CHEN, Y. A trajectory clustering approach based on decision graph and data field for detecting hotspots. *International Journal of Geographical Information Science 31*, 6 (2017), 1101–1127. doi:10.1080/13658816.2016.1213845.