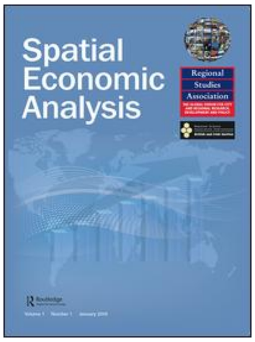

Spatial Economic Analysis

Routledge
Taylor & Francis Group

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rsea20

A pre-processing and network analysis of GPS tracking data

Antonino Abbruzzo , Mauro Ferrante & Stefano De Cantis

To cite this article: Antonino Abbruzzo , Mauro Ferrante & Stefano De Cantis (2020): A pre-processing and network analysis of GPS tracking data, Spatial Economic Analysis, DOI: 10.1080/17421772.2020.1769170

To link to this article: https://doi.org/10.1080/17421772.2020.1769170

Published online: 03 Jul 2020.

SPATIAL ECONOMIC ANALYSIS
https://doi.org/10.1080/17421772.2020.1769170

A pre-processing and network analysis of GPS tracking data

Antonino Abbruzzo [a], Mauro Ferrante [b] and Stefano De Cantis [c]

ABSTRACT

Global Positioning System (GPS) devices afford the opportunity to collect accurate data on unit movements from temporal and spatial perspectives. With a special focus on GPS technology in travel surveys, this paper proposes: (1) two algorithms for the pre-processing of GPS data in order to deal with outlier identification and missing data imputation; (2) a clustering approach to recover the main points of interest from GPS trajectories; and (3) a weighted-directed network, which incorporates the most relevant characteristics of the GPS trajectories at an aggregate level. A simulation study shows the goodness-of-fit of the imputation data algorithm and the robustness of the clustering algorithm. The proposed algorithms are then applied to three cases studies relating to the mobility of cruise passengers in urban contexts.

KEYWORDS

cluster-based method, network analysis, spatio-temporal data, global positioning systems

HISTORY Received 12 July 2019; in revised form 28 April 2020

INTRODUCTION

Global Positioning System (GPS) devices can record unit travel times and the spatial coordinates of locations to a high degree of temporal precision. Nowadays, GPS devices are compact, equipped with significant autonomy and, of the utmost importance, they can memorize the geographical coordinates of a statistical unit at a given moment in time. The widespread availability of GPS data permits the analysis of social behaviour across time and geographical scales (Butz & Torrey, 2006). Given the wide interest in understanding human mobility, applications of GPS tracking technologies in social science research may be found in various disciplines, including urban economics (McCabe et al., 2013; Theo, 2011), urban geography (Ahas et al., 2010; Raanan & Shoval, 2014), population geography (Silm & Ahas, 2010) and environmental health (Elgethun et al., 2003; Zenk et al., 2011), to mention but a few. The availability of high-quality data, in terms of temporal (seconds) and spatial (metres) resolution, demands statistical methods that facilitate the extraction of knowledge from a copious amount of information relating to the various unit trajectories.

CONTACT

[a] antonino.abbruzzo@unipa.it
Department of Economics, Business and Statistics, University of Palermo, Palermo, Italy
[b] **(Corresponding author)** mauro.ferrante@unipa.it
Department of Culture and Society, University of Palermo, Palermo, Italy.
[c] stefano.decantis@unipa.it
Department of Economics, Business and Statistics, University of Palermo, Palermo, Italy

Despite the accuracy of GPS tracking data, they require a pre-processing phase to correct outliers and missing data points. Therefore, after summarizing the main steps required in the analysis of GPS tracking data (and with a special focus on the context of travel surveys), the contribution of this paper to the literature is threefold: first, two algorithms for the pre-processing of GPS data will be proposed in order to deal with outlier identification and missing data imputation; second, a clustering approach based on an existing algorithm (namely the DBSCAN algorithm) (Ester et al., 1996; Palma et al., 2008) will be adapted herein to identify points of interests from GPS trajectories at the individual level; and third, individual-level points of interest will be collated into a weighted directed network to summarize the most relevant characteristics at an aggregate level of the GPS trajectories. The goodness-of-fit and sensitivity of the proposed algorithms will be evaluated through simulation experiments.

From an empirical perspective, the approach proposed in this paper will be applied to three cases studies relating to cruise passengers' trajectories in the cities of Palermo (Sicily, Italy), Dubrovnik (Croatia) and Copenhagen (Denmark). The results obtained may assist in identifying the most visited places, in addition to identifying the network of relationships among the various points of interest, with pertinent implications from a destination management perspective.

From a more general perspective, this paper fits into the ongoing debate as to how GPS tracking data can be processed and analysed, and which type of information and methodology can be extracted from this type of data. In the analysis of this type of data, different approaches regarding various problems and research aims will be presented and methodological proposals presented, according to specific analytical and research aims. These regard the pre-processing of GPS tracking data and the identification and analysis of points of interests in trajectories, in terms of their relationships, interpreted as a network.

The paper is structured as follows. The next section provides a background of the use of GPS technology in social research. The third section describes the statistical methods used to pre-process, locate the points of interest and construct the network from GPS data. The fourth section evaluates the goodness-of-fit and sensitivity of the proposed algorithms through a simulation study. Finally, the fifth section expounds three empirical applications using the proposed methods.

GPS TRACKING DATA ANALYSIS IN TRAVEL SURVEY: A REVIEW

The widespread availability of GPS tracking devices have facilitated the collection of spatio-temporal information on human movements, providing new challenges and opportunities in social and economic research. GPS technology was initially in travel surveys used in the late 1990s and this use has developed rapidly over the past decade. Shoval and Ahas (2016) present a summary and evaluation of the development and progress in this emerging field of scholarship within the field of tourism research. Before the introduction of the GPS technology, conventional travel surveys were generally conducted by paper-and-pencil interviews (PAPI), computer-assisted telephone interviews (CATI) or computer-assisted self-interviews (CASI). However, these traditional survey methods may have been burdensome to respondents since it was necessary to recall detailed information (including trip timing, travel modes and trip purposes) for each trip during the survey period (Wolf et al., 2003). GPS was used to record the location of the participants and their arrival and departure times so that the obtaining of the boarding information, the routes of the trips and transit trip times would be more precise. Owing to the fact that GPS devices are very accurate at recording time and positional characteristics of a given journey, GPS surveys can improve the accuracy and extent of travel survey data and correct any trip-misreporting issues caused by respondents (Murakami & Wagner, 1999; Wolf et al., 2001). As a result, these technologies have become a critical data collection tool in human mobility studies (Shoval, 2008).

It is widely acknowledged that GPS surveys can report accurate data, which is unbiased by users' perceptions (Blanchard et al., 2010; Wolf et al., 2003). Given that GPS devices have the potential to provide very large quantities of data, it is necessary to develop specific algorithms to deal with the main issues that may arise in the analysis of these data. However, the literature on the analysis of GPS tracking data are rather fragmented among various disciplinary fields. The latter ranges from computational geometry and data-mining to specific applications, such as animal tracking, road network, vehicle movements monitoring and hurricane formation, to mention but a few (Bermingham & Lee, 2017). Moreover, research typically deals with a particular problem and related solutions arising from the analysis of this type of data. On the other hand, comparative approaches regarding existing and proposed algorithms (for the same purpose) and a general framework for the analysis of space–time data (derived from GPS tracking devices) are lacking (for a survey of problems and methods in spatio-temporal data, see Atluri et al., 2018).

In the context of travel surveys, steps in the processing of GPS tracking data and the main topics of interest in their analysis can be summarized as follows:

- Pre-processing of the information:
  - Outlier detection.
  - Missing data imputation.
  - Smoothing of the series.
- Extraction and synthesis of relevant information, such as:
  - Activity identification and change detection.
  - Cluster identification:
    - at the individual level;
    - at the aggregate level.
  - Network relationships:
    - at the individual level;
    - at the aggregate level.
  - Analysis of trajectories.

Moreover, spatio-temporal information – when available – may be related to other types of information derived from questionnaire-based surveys or other source of information, such as map layers containing destination-related attributes, socioeconomic or environmental data (Van der Spek et al., 2009). This will facilitate the analysis of the relationships between space–time behaviour and other individual and contextual-level information.

Before conducting any meaningful data analysis, a set of pre-processing operations, aimed at reducing the magnitude of data problems, must be implemented. Stopher et al. (2008) describe various reformatting steps required to derive relevant information from the basic set of data, which have been collected from the GPS tracking device. These include (Stopher et al., 2008), for example, a calculation of local time and date, latitude–longitude correction, distance computation from the previous observation, time interval from the previous observation, and cumulative time intervals from the beginning of recording.

A crucial stage in pre-processing GPS data involves the identification of anomalies in data (also referred to as an *outlier*), and missing data imputation due to signal loss and outlier observations. Outlier observations may occur for several reasons, such as a cold or a warm start. These usually occur at the beginning of tracking operations (i.e., a cold start) or when the GPS device switches from *sleep mode* to *working mode* after a person stops for one or two hours (i.e., a warm start) (Shen & Stopher, 2014). Signal loss is generally determined by closed areas, urban canyons, etc. The latter are formed by roads cutting through blocks of tall buildings, which leads to a loss in signal (Stopher et al., 2008). They can have an impact on GPS signal reception and cause GPS data loss. Signal problems can result in missing or partial trips, and those generating spurious trips (a sequence of points generated by a stationary GPS device, incorrectly identified as *a trip*).

In the context of GPS tracking data, the joint presence of spatial and temporal components needs to be considered in the identification of outlier observations (Atluri et al., 2018), and several approaches have been proposed in the literature (e.g., Erenoglu & Hekimoglu, 2010; Třasák & Štroner, 2014). Atluri et al. (2018) propose a review of the main contributions dealing with anomaly detection in the context of spatio-temporal data-mining, according to the different type of data considered (point, trajectory, raster).

Once outlier observations have been identified and eliminated, it is necessary to implement imputation techniques relating to missing observations. Although several studies have discussed issues relating to signal loss and outlier observations (Gong et al., 2012; Tsui, 2005), only a few authors (Chen et al., 2010) have suggested how to deal with errors in GPS tracking data. The latter is still, therefore, a considerable challenge in GPS studies (Shen & Stopher, 2014).

The final step in pre-processing of GPS data may involve reducing the noise level by means of smoothing techniques (Jun et al., 2006). Smoothing splines (Castro et al., 2006), Kernel smoothers (Schuler et al., 2014) and Kalman filters (Jun et al., 2006) are often used in the pre-processing stage of GPS data, and other approaches have also been proposed (Chazal et al., 2011). Jun et al. (2006) compare various smoothing techniques for vehicle GPS data and evaluate the performance of the considered algorithms in terms of their capability at minimizing the impact of random errors on the estimation of speed, acceleration and distance.

After pre-processing the information, the extraction and synthesis of relevant information from the large quantity of data provided by GPS is required. Invaluable information regarding individual movements is related to activity detection, which is in turn mainly related to the identification of the different travel modes by which movements can occur (Siła-Nowicka et al., 2016). The most commonly variables to consider regarding travel mode detection are average speed, average absolute acceleration and travel distance. This step is generally based on the segmentation of trajectories, which requires the identification of segments characterized by different behaviour types in terms of mobility variables. Several approaches have been proposed for travel mode identification, such as, for example, machine-learning algorithms (Torrens et al., 2011), neural networks (Hu et al., 2004), Bayesian networks (Xiao et al., 2015). However, the availability of contextual information regarding the location of movements may also be effectively used for travel mode identification (Sester et al., 2012; Siła-Nowicka et al., 2016).

Having described trajectory features, the identification of stop locations is a relevant step in the extraction and synthesis of relevant information from tracking data, since they may indicate relevant locations or points of interest for the subject. Gong et al. (2015) review the main research results on stop location identification and propose a classification of methods according to five groups, namely centroid-based methods, speed-based methods, duration-based methods, density-based methods and hybrid methods. For a review of this topic, see Gong et al. (2018). Nonetheless, in evaluating any method for identifying stop locations, the characteristics of the data set used, as well as the research aims, study context and related assumptions, must be taken into consideration.

By aggregating individual-level stops, attractive locations visited by many users in the region being analysed may be identified. In the context of tourism, this step may reveal popular places, such as tourist attractions, restaurant locations or shopping centres. This approach has been undertaken to classify and rank places (Tiwari & Kaushik, 2013), evaluate the shared characteristics of users (Andrade & Gama, 2019), and even for the determination of optimal meeting points for a set of users (Khetarpaul et al., 2012).

Having obtained the main points of interest in a given location, a further step in the extraction of relevant information may consider each point of interest as the vertices (or nodes) of a graph, with the aim of analysing the network relationships among these nodes (Gilad, 2001). At the individual level, individuals' mobility can be described as a graph, where each node indicates various points of interests for each unit; the edges of the graph may be characterized by features, such as the duration and frequencies of transitions (Lin & Hsu, 2014; Zheng et al., 2009). Similarly, at the aggregate level, the nodes on the graph are represented by relevant locations for a given number of units, and the edges of the graph may be characterized in terms of the number of units passing from one node to another. Applications of this approach can be found in an analysis of the travel route decision process (Knapen et al., 2020; Knapen et al., 2016). For example, Lin and Hsu (2014) contain summaries of studies regarding graph-based trajectory-mining using GPS tracking data.

Finally, a wide range of applications of GPS tracking data dealt with the analysis of GPS trajectories. These trajectories are represented as a set of stops and moves (Spaccapietra et al., 2008). A stop is characterized by consecutive GPS data points within a predefined time and distance threshold. Several studies have focused on models for efficiently analysing moving objects (Brakatsoulas et al., 2004; Du Mouza & Rigaux, 2005; Güting et al., 2006; Wolfson et al., 1998). The main focus of enquiry in Wolfson et al. (1998) relies on the geometric properties of trajectories, while Brakatsoulas et al. (2004) and Du Mouza and Rigaux (2005) consider semantics and background geographical information. Moving patterns have been extracted from data by Du Mouza and Rigaux (2005), who define the patterns a priori. For instance, they suggest locating all the trajectories that move from zone A to zone B, crossing zone Z. They state that moving patterns are those trajectories that follow a given pattern.

Alvares et al. (2007) extract moving patterns from data that are unknown a priori but whose minimal number of trajectories are frequent. An algorithm named *SMoT* (Stops and Moves of Trajectories) in Alvares et al. has been provided to extract stops and moves from trajectory sample points. This work also demonstrates how simple trajectory data analysis develops when this semantic model is used. Moving patterns have been extracted from stops and moves by Campello et al. (2013), and the resulting patterns modelled in the geographical conceptual schema in order to visualize trajectory patterns in geographical space. The user specifies the places of interest in all the previous approaches, given that stops and moves are defined from an application point of view. The main drawback of this assumption is that important places, which may lead to the discovery of interesting patterns, can be missed if they are unknown to the user.

STATISTICAL METHODS

Trajectories have been generally considered as a path, which is followed by an object moving in space and time (Güting et al., 2006; Wolfson et al., 1998). Each point on this path represents one position in space at a given instant of time. Typically, trajectory data are obtained from GPS devices that capture the position of an object at specific time intervals. An analysis of GPS trajectories usually involves several steps (Van der Spek et al., 2009). This section describes the pre-processing of data, which relates to the set of procedures implemented for outlier detection and missing data imputation from raw GPS data. In a second step, similar to Palma et al. (2008), a statistical approach based on a density-based cluster algorithm, the DBSCAN (Ester et al., 1996), is proposed in order to recover the main points of interest of a trajectory. Finally, a directed network approach is defined to summarize the density-based clusters for every unit, and to identify the most significant characteristics of their movements. At the aggregate level and in order to summarize the behaviour of all the units, a weighted directed network is derived by means of a second implementation of the DBSCAN algorithm with reference to the set of all the individual points of interest.

Before describing the statistical methods used, let us formally define what is a trajectory and the quantities that can be derived from the trajectory.

Definition 3.1:. Let $D_j^{(i)} = (x_j^{(i)}, y_j^{(i)}, t_j^{(i)})$, where x is the longitude, y is the latitude, and t represents the time, be a sample point from a GPS tracker for unit i at position j. The set

$D^{(i)} = \{D_j^{(i)}, j = 1, \ldots, n_i\}$, where $t_1^{(i)} < \ldots < t_{n_i}^{(i)}$, is the temporal–spatial movement or trajectory of the i-th unit. Let $d_j^{(i)}$ denotes the Euclidean distance between $(x_j^{(i)}, y_j^{(i)})$, $(x_{j+1}^{(i)}, y_{j+1}^{(i)})$, $\Delta t_j^{(i)} = t_{j+1}^{(i)} - t_j^{(i)}$ the time between two consecutive observed spatial points, $T_j^{(i)} = t_{j+1}^{(i)} - t_1^{(i)}$ the cumulative time, and $v_j^{(i)} = d_j^{(i)}/\Delta t_j^{(i)}$ the velocity to pass from the spatial point j to the point $j + 1$.

Figure 1 shows an example of quantities, which can be derived from a trajectory $D^{(i)}$. Specifically, the distance time series is shown on the left, and the velocity time series can be seen on the right.

Pre-processing of GPS data

According to the framework presented in the second section, the first step in GPS data analysis regards pre-processing, which involves outlier detection and the imputation of missing observations from the trajectories (Stopher et al., 2008). Figure 2 shows missing data points (on the left) and an outlier for a statistical unit in a hypothetical trajectory. Algorithm 1 is proposed to detect the outliers, together with Algorithm 2 to impute the missing spatial–temporal points.

Outlier detection

An outlier is a point whose coordinate exceeds a certain distance with respect to the true coordinate. As described in the second section, several approaches can be adopted to detect outlier observations (Shen & Stopher, 2014). In this paper we will propose the application of changepoint

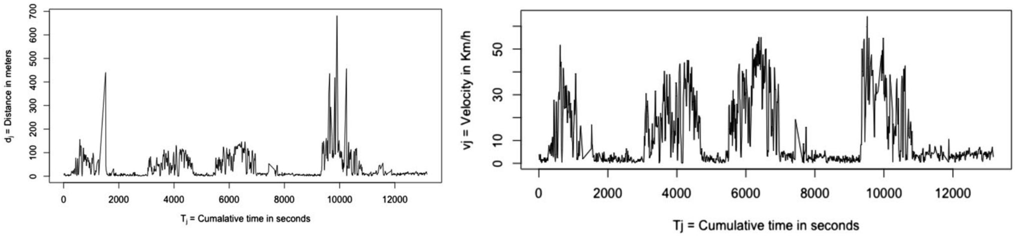

Figure 1. Example of quantities derived from a trajectory. The distance time series is on the left; and the velocity time series is on the right.

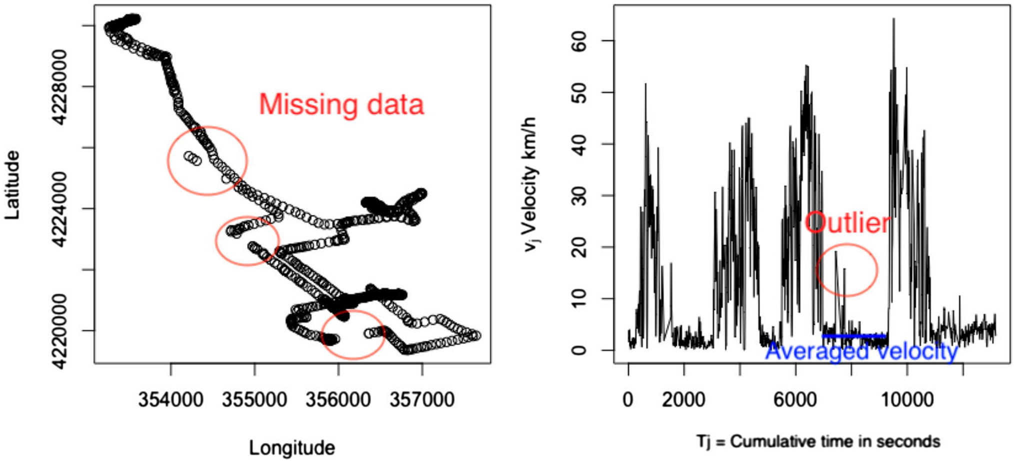

Figure 2. Example of missing data and outlier from a raw hypothetical trajectory.

analysis – generally used in time series to discover multiple segments (Haynes et al., 2017a, 2017b) – to the time series of the velocities. The lower and upper limits for each segment were constructed in order to identify outlier observations. A changepoint will be detected if there is a change in the mean or in the velocity variance at some point $T_j^{(i)}$. Algorithm 1 describes the steps for identifying and dealing with outliers.

Algorithm 1: Outlier detection

(1) Use changepoint analysis for a given time series to discover multiple segments $s = \{s_1, \ldots, s_k\}$ in the velocity time series.
(2) Compute the quantities for each segment s_l of the series $lower\_limit(s_l) = Q_1(s_l) - \lambda * IQR(s_l)$ and $upper\_limit(s_l) = Q_3(s_l) + \lambda * IQR(s_l)$, where $Q_1(s_l)$ is the first quartile, $Q_3(s_l)$ is the third quartile and $IQR(s_l) = Q_3(s_l) - Q_1(s_l)$ is the interquartile range, $\lambda > 0$ is a tuning parameter.
(3) Remove the coordinate $D_j^{(i)}(s_l)$ if $v_j^{(i)}(s_l) > upper\_limit(s_l)$ or $v_j^{(i)}(s_l) < lower\_limit(s_l)$.

A tuning parameter $\lambda > 0$ was used in step 2 of Algorithm 1 in order to detect outliers. According to Tukey (1977), $\lambda = 1.5$ indicates an ‘outlier’ and $\lambda = 3$ indicates that the data are ‘far out’. Figure 3 shows an example of the application of Algorithm 1. The circled points indicate detected ‘far out’ points, and the horizontal lines indicate the mean of each segment.

In the segment in which two ‘far out’ points have been discovered, it seems that the unit was walking at a constant speed and, at some point, the velocity increased sharply twice at around 15 km/h. A similar argument can be used in the last segment where another ‘far out’ point was detected.

Missing data imputation

The second step in pre-processing of GPS data regards missing data imputation. Missing data derives from outlier removal and signal loss. The latter becomes particularly important in the identification of the stops in a trajectory. A loss of signal can be easily detected since GPS devices are generally set to record spatial locations at every c seconds. Algorithm 2 reconstructs a trajectory $D^{(i)}$ where $\Delta_j^{(i)} = t_{j+1}^{(i)} - t_j^{(i)}$ is equal to a constant c. The rationale of this algorithm is to impute missing points by assigning several points proportional to the time lost at the last point, which had been observed before losing the signal. Some of these points will then be used to reconstruct a linear path from the point of signal loss to the next available signal point. The new reconstructed path will be a regularized trajectory in the sense that $\Delta_j^{(i)}$ will be constant for all j. This idea is illustrated in Figure 4. The raw data can be observed on the left and the pre-processed data on the right. A total of 25 signal points were lost in this example: 11 were used to reconstruct the

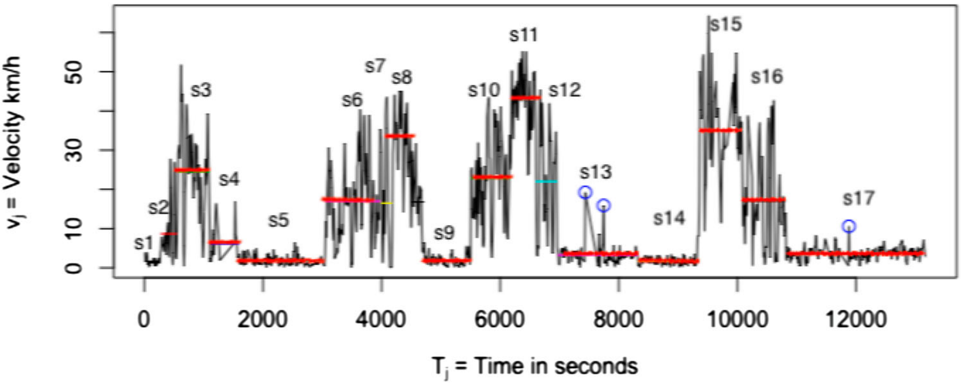

Figure 3. Example of outlier detection.

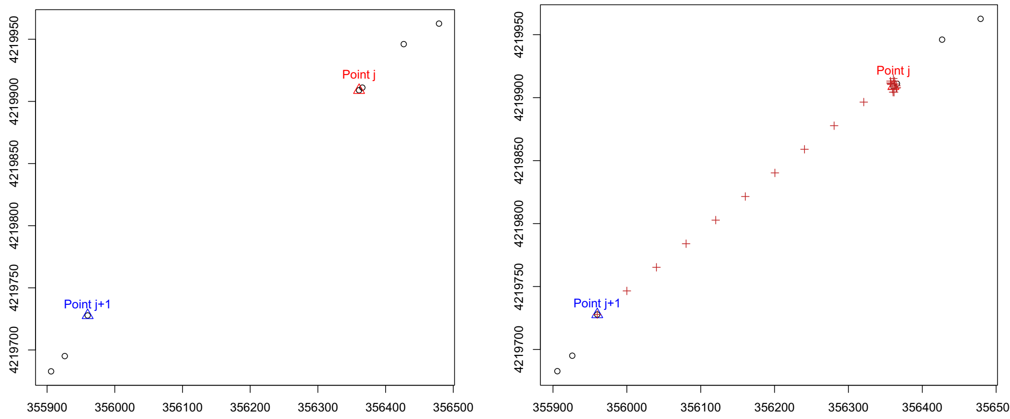

Figure 4. Example of the trajectory of a pre-processed unit. The raw data are on the left; and the pre-processed data are on the right. A total of 25 signal points were lost in this example: 11 were used to reconstruct the linear path between j and $j+1$; and 14 were imputed randomly close to the spatial point j. Points j and $j+1$ are indicated by triangles; and random and linear imputed points are indicated by pluses.

linear path between j and $j+1$; and 14 were imputed randomly close to the spatial point j where the signal had been lost.

Algorithm 2: Missing data imputation for trajectory regularization.
Algorithm 2a: Find the number of points to be imputed

```
1. Given a unit $i$, denote with $D_*^{(i)}$ the regularized trajectory;
2. FOR $j$ in $1{:}n_i$
   • Compute $n_j^{(imp\_points)} = \lfloor +; \frac{\Delta t_j^{(i)}}{c} \rceil$, where $\lfloor +; x \rceil$ is the nearest integer function of real number $x$, that indicates the number of lost points between $j$ and $j+1$;
     o IF $n_j^{(imp\_points)} = 1$ then $D_{*j}^{(i)} = D_j^{(i)}$
     o ELSE $D_{*j}^{(i)} = [D_{1j}^{(i)} \vdots D_{2j}^{(i)}]$ where $D_{1j}^{(i)}$ and $D_{2j}^{(i)}$ are recovered in Algorithm 2b
   END FOR
3. Repeat the steps 1 and 2 for $i = 1, \ldots N$.
```

Algorithm 2b: Create set of linear and random points between two points of a spatial–temporal trajectory

```
1. Given $n_j^{(imp\_points)} > 1$ missing points, create two sets of points of length $n_j^{(linear)}$ and $n_j^{(random)}$ as follows:
   • Estimate a regression linear model where $y = (y_j, y_{j+1})$ and $x = (x_j, x_{j+1})$
   • Compute $\bar{v}_j = \frac{\sum_{i=j-4}^{j+4} v_{i \neq j}}{8}$ where $v_i$ indicates the velocity, then $\bar{d}_j = \bar{v}_j * \sum_{i=j-4}^{j+4} \Delta t_{i \neq j}$ is the averaged distance of the unit in a neighbourhood of $j$.
   • Set $n_j^{(linear)} = \lfloor d_j / \bar{d}_j \rceil$ and $n_j^{(random)} = n_j^{(imp\_points)} - n_j^{(linear)}$
   • Create a set of times of length $n_j^{(imp\_points)}$ from $t_j$ to $t_{j+1}$ and denote with $t_1$ the first $n_j^{(random)}$ time points, and with $t_2$ the remaining
2. Create an equidistant sequence $x_{imp}$ from $x_j$ to $x_{j+1}$ of length $n_j^{(linear)}$ and obtain the predicted values $\hat{y}_{imp}$ from the linear regression. Then $D_{2j}^{(i)} = (x_{imp}, \hat{y}_{imp}, t_2)$
```

3. Create two sets u_{imp_x} and u_{imp_y} of length $n_j^{(random)}$ from $U \sim Unif(-\alpha, \alpha)$ then $D_{1j}^{(i)} = (x + u_{imp_x}, y + u_{imp_y}, t_1)$

Density-based cluster algorithm for GPS data

Let $D^{(i)}$ be the pre-processed trajectory for unit i. A cluster can be defined as a minimum set of temporal–spatial points, which are sufficiently close (regarding distance) to form a cluster. For example, in a tourism application, the cluster may indicate some points of interest, such as: a monument, a museum, a hotel or restaurant. Note that these places may have any geometrical shape.

The DBSCAN is a density-based algorithm (Ester et al., 1996), which is designed to discover arbitrary-shaped clusters, where the clusters are sets of spatial points which fall within a certain distance. Concurrently, the algorithm can identify the *noise* points, which are spatial points not belonging to any cluster. Each cluster in this study was interpreted as a point of interest, that is, a virtual place where the unit spent a certain amount of time.

Let $p = (x, y)$ be a point in the trajectory of a generic unit. The ϵ*-neighbourhood* of a point p is defined by $ne_\epsilon(p) = \{q \in D{:}||p, q|| \leq \epsilon \in \mathbb{R}^+\}$, where $||p, q||$ is a distance function. If the cardinality of an ϵ-neighbourhood of a point p, that is, $|ne_\epsilon(p)|$, is at least greater than a minimum number ($minpts \in \mathbb{R}^+$) then p is a *core point*.

A point p is *directly density-reachable* from the object q with respect to ϵ and $minpts$ if $p \in ne_\epsilon(q)$ and $|ne_\epsilon(q)| \geq minpts$.

A point p is *density-reachable* from the object q with respect to ϵ and $minpts$ if there is a chain $p_1, \ldots, p_l, p_1 = q, p_l = p$ such that p_{i+1} is directly density-reachable from p_i.

An object p is *density-connected* to object q, with respect to ϵ and $minpts$, if there is an object o such that both p and q are density-reachable from o with respect to ϵ and $minpts$.

Definition 3.2:. A *cluster* C is a non-empty subset of D satisfying the following requirements:

- $\forall \quad p, q$: if $p \in C$ and q is density-reachable from p with respect to ϵ and $minpts$, then $q \in C$;
- $\forall \quad p, q \in C$: p is density-connected to q with respect to ϵ and $minpts$.

Let $C_1, \ldots, C_k$ be the clusters of D with respect to ϵ and $minpts$, then $p \in D$ is a *noise* point if it does not belong to any cluster C_i. The algorithm starts with the first point p in the database D, and it retrieves all the neighbours of a point p with respect to ϵ and $minpts$. If p is a core point, this procedure will yield a cluster concerning ϵ and $minpts$. If p is not a core point, no points will be density-reachable from p, and the DBSCAN algorithm will proceed to consider the next point of the database. The DBSCAN has been deployed in the DBSCAN R package (Hahsler et al., 2018).

Example: from a trajectory to clusters

Set $minpts = 30$ (5 min) and $\epsilon = 40$ (metres). The clusters for the trajectory D of a hypothetical unit are illustrated in Figure 5 and Table 1.

The implementation of this algorithm may provide important information. In this example we observe that the number of cluster for the considered unit is five, with a tour length of 219.66 min (3.65 h), and an average time per cluster of 13.2 min.

Choice of the tuning parameter ϵ and minpts

The DBSCAN algorithm requires two parameters, ϵ and $minpts$, to be selected. The easier-to-set parameter of DBSCAN is the $minpts$ parameter, whose choice is also related to the minimum amount of type required to characterize a cluster as meaningful. On the other hand, and as suggested by Schubert et al. (2017), the parameter ϵ is often harder to set, and the authors

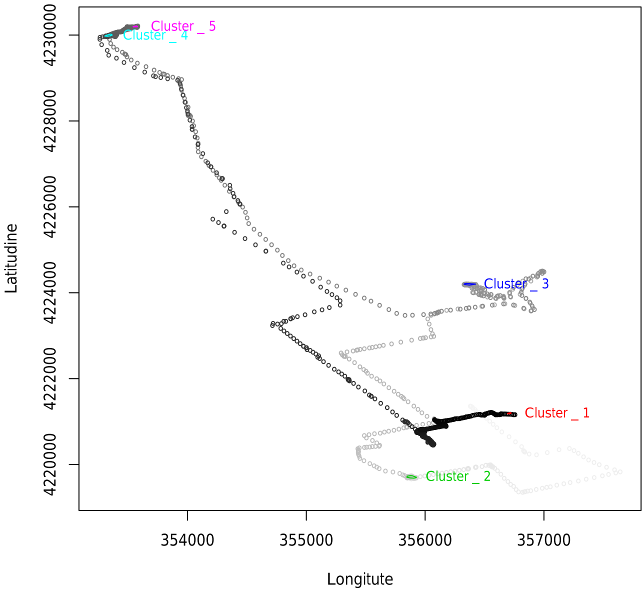

Figure 5. Example of clusters relating to the trajectory of a unit: Geographical location, and direction of the trajectory represented by the scale ofgrey from white (starting point) to black (end-point).

Table 1. Example of clusters relating to the trajectory of the unit in Figure 5: time spentat each cluster.

Cluster	Time (min)	% Time
Noise points	153.66	69.98%
1	5.00	2.28%
2	25.33	11.54%
3	13.66	6.22%
4	16.33	7.44%
5	5.60	2.55%
Total	219.58	100.00%

have made various recommendations in this regard. Several internal validity cluster measures can be used to select the tuning parameter ϵ (Desgraupes, 2013). Such measures quantify the quality of a clustering relying only on properties intrinsic to the data. Examples of internal measures include the silhouette, Davies-Boldin and Calinski–Harabasz measures. However, none of these indices has been proved to be appropriate for selecting the tuning parameter in the context of GPS data analysis (Van Craenendonck & Blockeel, 2015). Consequently, the problem of selecting the tuning parameter ϵ is still an open-ended issue, worthy of a separate paper.

In an ideal scenario, there exists domain knowledge for selecting this parameter, which is based on the application domain. In the scenario described in this paper, that is, in clustering GPS locations, a domain expert may decide that objects within q metres (ϵ) should be neighbours if there are at least a minimum number of points p ($minpts$). In the application outlined in this paper, the $minpts$ was chosen such that the time spent at cluster j was at least x min. For example, if points were collected at 10-s intervals, then $minpts = 30$ guaranteed that the time spent in each cluster was at least $x = 5$ min. Given the parameter *minpts*, we suggest to let the parameter ϵ depend on the distance function $\{d_j\}, j = 1, \ldots, n$, which denotes the Euclidean distance between the coordinates (x_j, y_j). The simulation study described in the fourth section indicates that a choice of ϵ between 10 and 40 provides robust results.

Networks of DBSCAN objects

The application of the DBSCAN algorithm to a unit trajectory $D^{(i)}$ produces a set of clusters $C_1^{(i)}, \ldots, C_{k_i}^{(i)}$ and a set of noise points. These noise points are coordinates that can be removed from the analysis. The centroid of each cluster can be seen as its core location for a given unit. In order to synthesize each trajectory, it is proposed to construct a network that includes the concepts of clusters and moves. Specifically, a directed network can be defined as follows.

Definition 3.3:. Let $\bar{C}_1^{(i)}, \ldots, \bar{C}_{k_i}^{(i)}$ be the centroids of the clusters of a trajectory $D^{(i)}$. A directed network for a generic unit i is the set $G = (V, E)$, where V is a set of nodes with cardinality $|V| = k_i$; and E is a set of links (moves between clusters for the unit), where $e_{kl} = 1$ if the unit goes from the cluster k to the cluster l, and 0 otherwise.

Note that each cluster originally contained a set of spatial points but in Definition 3.3 a centroid was associated with each cluster. The cluster centroids can be used as coordinates to represent the network nodes graphically. Moreover, the cardinality of each cluster provides information regarding the time spent by the unit at each cluster; it also permits the recovery of the direction of the link between two clusters. The latter information is obtained by virtue of the temporal ordering of the spatial movements.

Figure 6 shows an example of application of the DBSCAN to a hypothetical trajectory, displayed on the left, with the recovered directed network displayed on the right. In this example, the implementation of DBSCAN algorithm led to an estimate of $k = 5$ clusters. Consequently, the directed network has five vertices $V = \{1, \ldots, 5\}$ and the directed links show the path of the unit. Moreover, it can be seen that unit one starts from node 1, selecting the following points of interest $1- > 2- > 3- > 4- > 5- > 4- > 1$. Note that the positions of each node in the network correspond to the centroid of each cluster. According to this approach, the information

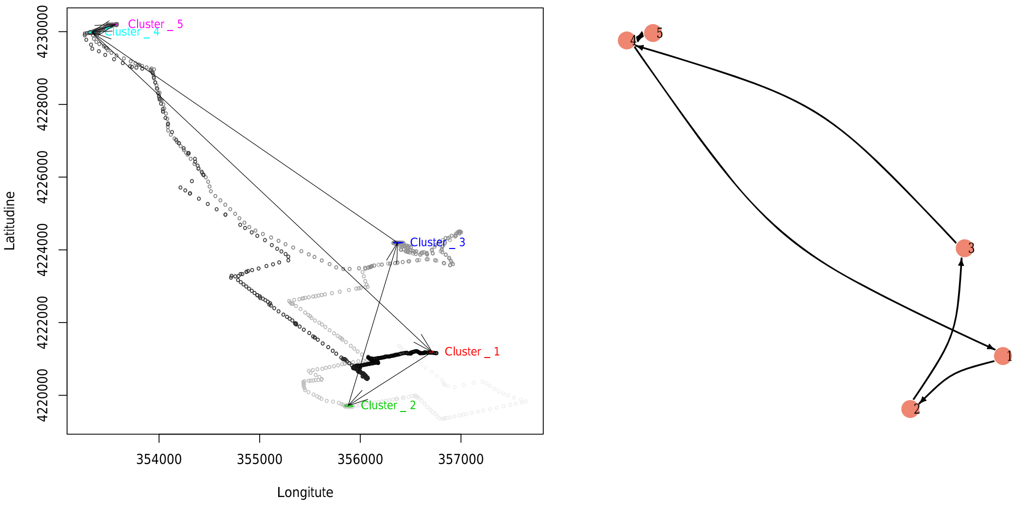

Figure 6. Raw GPS data and clusters obtained from the DBSCAN algorithm (left). Nodes and links (right) represent points of interest (cluster) and moves between clusters, respectively, for the considered unit. The nodes of the network are placed in the centroid of each cluster.

relating to the noise points (i.e., the streets which the units followed during their moves) has been lost, not being of interest in this step of the analysis.

A weighted directed network of clustered trajectories

Thus far, an approach for deriving a set of clusters and directed networks from the temporal–spatial GPS data $D^{(i)}$ for each unit i, $i = 1, \ldots, N$ has been described. This analysis has produced the set of clusters centroids $\bar{C}^{(i)} = \{\bar{C}_1^{(i)}, \ldots, \bar{C}_{k_i}^{(i)}\}$. Let $\bar{C} = \{\bar{C}^{(1)}, \ldots, \bar{C}^{(N)}\}$ be the set of centroid clusters obtained by applying the DBSCAN. By reapplying the DBSCAN algorithm to the set $\bar{C}$, a new set of clusters is obtained. The latter can be interpreted as the points of interest for the units sampled, and a weighted directed network can be reconstructed as follows.

Definition 3.4:. A weighted directed network is a triplet $G = (V, E, W)$, where the set V represents the points of interest of the collected sample units, which have been obtained by applying the DBSCAN algorithm to $\bar{C}$. The set of edges E represents the transitions from one point of interest to another. An edge $e_{lk} = 1$ if a unit goes from cluster l to cluster k and 0 otherwise. The weighted matrix W represents the number of units transferring from one node to another. Specifically, w_{lk} indicates the number of units that have been transferred from node l to node k; and the W_{ll} diagonal represents the number of units that spent a certain amount of time at cluster l.

It is proposed to indicate the procedures described in the third section as the *ClusNet* algorithm.

SIMULATION STUDY

A simulation study was performed in order to: (1) validate the goodness-of-fit of the pre-processing Algorithm 2; and (2) asses the sensitivity of the DBSCAN algorithm with respect to the choice of the tuning parameter ϵ. The simulations were conducted making use of the R statistical software (R Development Core Team, 2008).

Goodness-of-fit of the pre-processing Algorithm 2

In order to evaluate the performance of Algorithm 2, we used 50 collected trajectories for which all data points were observed. To evaluate the goodness-of-fit of Algorithm 2 in missing data imputation, a set of points from the collected trajectories was removed. Considering that these points may be lost both in a given segment of the trajectory, and between a segment of the trajectory and a cluster in empirical applications, a two-factor scheme was developed. The first factor was related to the number of clusters in the trajectory, which provides information regarding the positions in which several points is removed. The second was given by the number of points to be removed in each of these positions.

In order to determine the positions from which a constant number of points is removed, n_{rm}, the number of clusters $C = \{C_1, \ldots, C_K\}$ was estimated for each trajectory by applying the DBSCAN algorithm (with $minpts = 30$ and $\epsilon = 30$). This step permitted the definition of $D = \{D_1, C_1, D_2, C_2, \ldots, D_K, C_K, D_{k+1}\}$. The second factor $n_{rm} = 10, 20$ indicates the number of data points to be removed in each of the K positions of every trajectory. Specifically, if $|D_j| > n_{rm}/2 + 1$, where $|\cdot|$ denotes the cardinality of the set, the last $n_{rm}/2$ data points are removed from D_j and the first $n_{rm}/2$ points from C_j. Note that $|C_j| \geq 30$ since the authors set $minpts = 30$ in the DBSCAN algorithm. Otherwise, only the first n_{rm} points are removed only from the cluster C_j.

The set observations having been removed according to the above described criteria, Algorithm 2 described in the third section was implemented to impute the missing data. The imputed trajectory is denoted by $D_{imp}^{(i)}$. To evaluate the goodness-of-fit of the missing value imputation, pre-processing algorithm, for any considered i-trajectory, a measure of distance between the true and the imputed trajectory, was considered via the dynamic time warping (DTW) algorithm (Rabiner & Juang, 1999), as follows:

$$d_\phi(D, D_{imp}) = \sum_{T}^{t=1} d(\phi_D(t), \phi_{D_{imp}}(t)) m_\phi(k)/M_\phi, \tag{1}$$

where the warping curve $\phi(t)$, $t = 1, \ldots, T$, and $\phi_D(t) \in \{1, \ldots, N\}$ and $\phi_{D_{imp}} \in \{1, \ldots, M\}$ are wrapper functions remapping the time indices of D and D_{imp}, $m_\phi(t)$ is a per step weighting coefficient; and M_ϕ is the corresponding normalization constant, which ensures that the accumulated distortions are comparable along different paths. At the core of equation (1) lies a wrapping curve technique. See Giorgino (2009) for further details on the application of DTW for comparing GPS trajectories.

The results of the simulation study regarding the missing data imputation algorithm are reported in Figure 7 for a value of $n_{rm} = 10$ on the left and of $n_{rm} = 20$ on the right. The value of the total number of points removed from each trajectory, which depends on the number of clusters K identified in each trajectory, is reported on the x-axis; the value of the average distance (m) among the true and the imputed trajectory is reported on the y-axis, as determined from equation (1).

With reference to the first scenario, where $n_{rm} = 10$ and $k \in [1, 22]$, the minimum is $d_\phi = 0.13$ m, the median is 0.46 m and the maximum is 1.83 m. In the second scenario with $n_{rm} = 10$ and $k \in [1, 22]$, the minimum is $d_\phi = 0.32$ m, the median is 1.41 m and the maximum is 3.36 m. Although the distance between the true and the imputed trajectory generally increases as the number of points removed increases, other factors (as expected) may affect the accuracy of the imputation procedure; this is mainly related to the path followed by the considered unit. Nonetheless, the very low average distances between the true and imputed trajectories for all 50 units under consideration suggest that the algorithm proposed is a high performer in the estimation of missing observations.

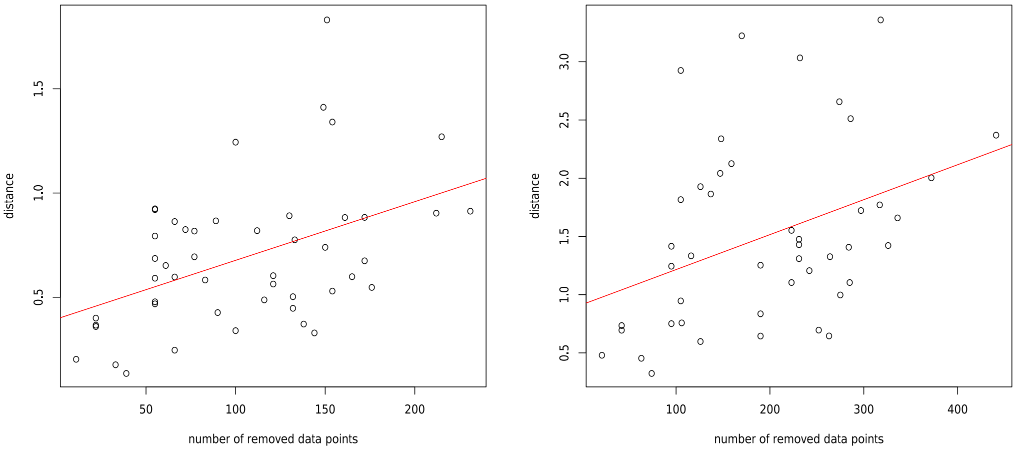

Figure 7. On the left side the first scenario, $n_{rm} = 10$; and on the right side, $n_{rm} = 20$. The straight line shows the linear increasing of the distance d_ϕ when the number of removed points increases.

Sensitivity of the DBSCAN algorithm with respect to the choice of the tuning parameter

In order to evaluate the sensitivity of the DBSCAN algorithm with respect to the tuning parameters, a set of trajectories with a certain (known) number of clusters was simulated for cluster identification in GPS tracking data. A trajectory can be generated with two theoretical models: *correlated random walks* in which the turning angle of each step is the direction of the previous step ± some error; and *directed walks*, or *compass-based navigation*, in which the angular errors at each step are added to the 'ideal' or compass direction. In the simulation study we assumed correlated random walks with landscape resistance raster from the R *SiMRiv* package. The former takes into account the characteristics of the city in terms of a street configuration. Furthermore, the clusters were simulated from a mixture of Gaussian distributions and added to the correlated random walk path. The pseudo-algorithm for generating trajectories with clusters is as follows:

- Generate a correlated random walks with landscape resistance raster from the package R *SiMRiv*. Denote the generated trajectory with D.
- Take K positions from the trajectory D. These positions indicate the position of the K clusters.
- Generate a vector of $\mu = (\mu_1, \ldots, \mu_K)$ and an array of $\Sigma = \Sigma[1], \ldots, \Sigma[K]$.
- Use the function gen.mix to generate K clusters from a mixture of bivariate Gaussian distributions with mean μ and variance–covariance Σ.
- Reorder the cluster points such that the first point of the cluster will be close to the last point of the trajectory before the cluster and the last point of the cluster will be closer to the first point of the trajectory after the cluster. The R function solve_TSP with method 'nearest_insertion' was used to obtain the reordering.
- Add these K ordered clusters in the K position of the trajectory D.

Algorithm 3: Generate trajectories with clusters.

The simulations were conducted making use of R, by considering a 2 × 2 factorial design resulting from the combination of two factors: number of clusters $K = 5$, 20, and autocorrelation of the random walk $\rho = 0.70$, 0.90. The numbers of points for each cluster were kept fixed at $n\_clust = 40$. Each scenario was repeated 100 times. We used the DBSCAN algorithm described in the third section in which the path of $\epsilon \in [10, 70]$ and the minimum number of points for each cluster was fixed at $minpts = 30$. The sensitivity of the tuning parameter ϵ to the set of values was evaluated by two indices: the distance between the centroid of the true clusters and the centroids of the estimated clusters (averaged across the 100 replicates); and the difference in the absolute value between the number of true clusters and those estimated (averaged across the 100 replicates).

The best ϵ is that which minimizes the distance between the true and estimated centroids, and the difference between the true and estimated number of clusters. This indicates that both lines should be as low as possible. The two plots in the upper part of Figure 8 indicate that the DBSCAN is quite robust with a choice of ϵ between [10, 20]. The two plots in the lower part of Figure 8 indicate that the DBSCAN is quite robust with a choice of ϵ between [10, 30].

EMPIRICAL APPLICATION

The data sets collected comprise 303, 51 and 73 GPS tracks relating to cruiser passengers disembarking from Palermo (2014), Dubrovnik (2015) and Copenhagen (2018), respectively. The details of data collection procedures and other survey characteristics may be found in (Ferrante et al., 2018). After the pre-processing step, the *ClusNet* algorithm is applied to the three samples of cruise passengers who disembarked in Palermo, Copenhagen and Dubrovnik. Since this

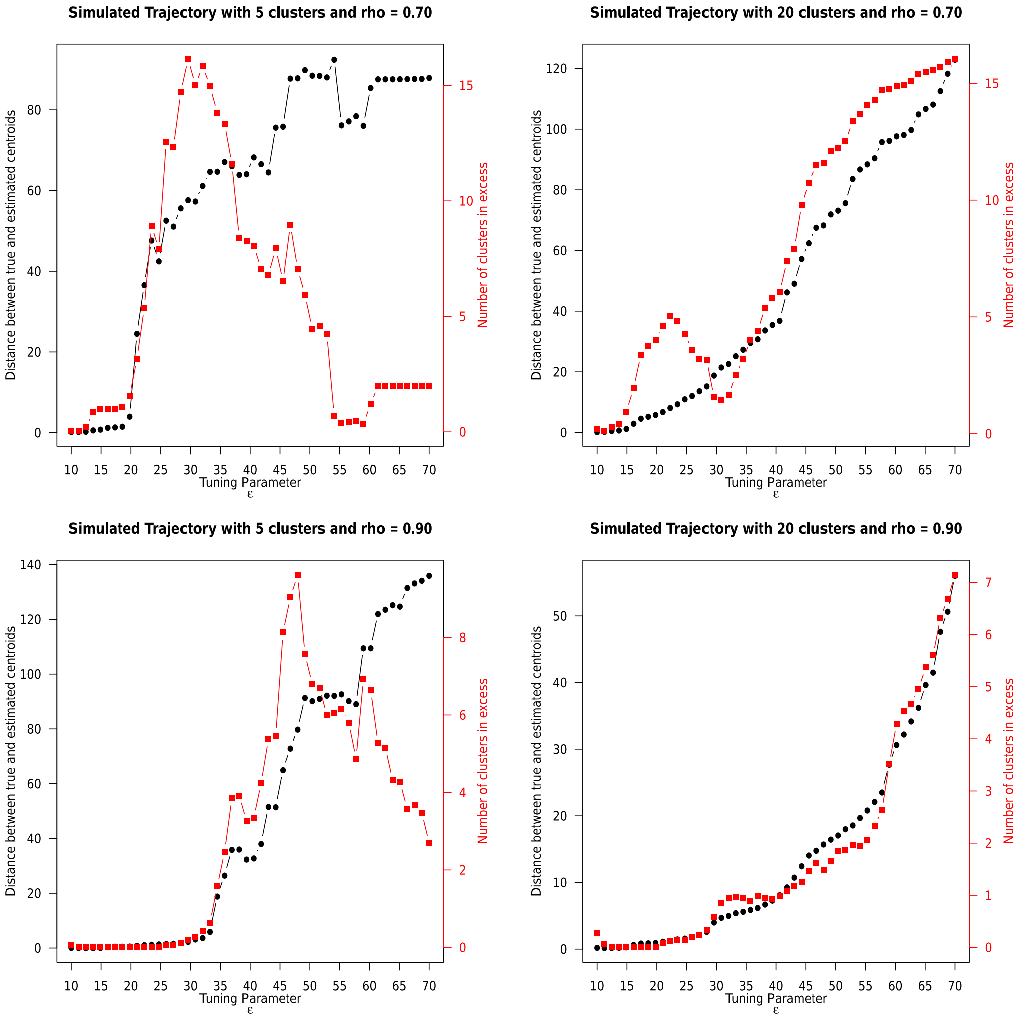

Figure 8. Results from simulated trajectories with correlated random walks 0.7 and 0.9, a landscape resistance raster, and 5 and 10 clusters.

application relates to tourism data sets, it can be assumed that the clusters are point of interests visited by tourists, and the moves form a virtual trajectory used to travel from one attraction to another. The application of the *ClusNet* algorithm, as described in this paper, can reveal the paths of individual tourists through the directed network (defined in Definition 3.3) or a global tourist path, through the weighted directed network, as defined in Definition 3.4. As discussed in the third section, the tuning parameters were heuristically fixed by *minpts* and ϵ. Specifically, the *minpts* was fixed at 30 and $\epsilon = 20$ in order to form individual clusters. *minpts* $= 10$ and $\epsilon = 15$ were fixed in the second step of the algorithm. Note that in this second step, the *minpts* represents the minimum number of subjects required to form a cluster.

Palermo

This section will outline the application of the proposed *ClusNet* algorithm to GPS tracks relating to a sample of cruise passengers disembarking in Palermo. Based on the *ClusNet* algorithm, 27 clusters were estimated. Figure 9 shows the centroids (circles) for 23 over 27 detected clusters on a map. Note that a zoom feature has been used over the city centre of Palermo in order to improve the clarity of the points of interest, even though four more places resulted from the algorithm: Piazza di Monreale, Monreale Cathedral, Santuario di Santa Rosalia and the Catacombs. The bigger the circle, the greater the number of tourists who visited the point of interest.

As indicated in Table 2, the top-10 tourist points of interest (by tourist visits) in Palermo are: the Cathedral, Teatro Massimo, Piazza Pretoria, Piazza Bellini, Piazza San Domenico, Santuario di Santa Rosalia, Teatro Politeama, the Catacombs, Cappella Palatina and San Matteo. But the area of disembarkation also appears, that is, where tourists start and end their visit to the city.

The *ClusNet* algorithm recovers the weighted directed network of the clusters (attraction points), which have been visited by the units of the sample (Figure 10). Note that the complete adjacency matrix (on the left of Figure 10) and a reduced network (where the directed links with fewer than 10 moves from one cluster to another have been removed; on the right of Figure 10) are both shown. The circle size of each node is proportional to the number of units detected at each point of interest. The link size between clusters i and j is proportional to the number of units that have been transferred from i to j. The links should be read in pairs and not as a sequence of movements. It can, for example, be stated that there have been numerous moves from the Cathedral to Teatro Massimo, but it is not possible to specify a path between the points of interests.

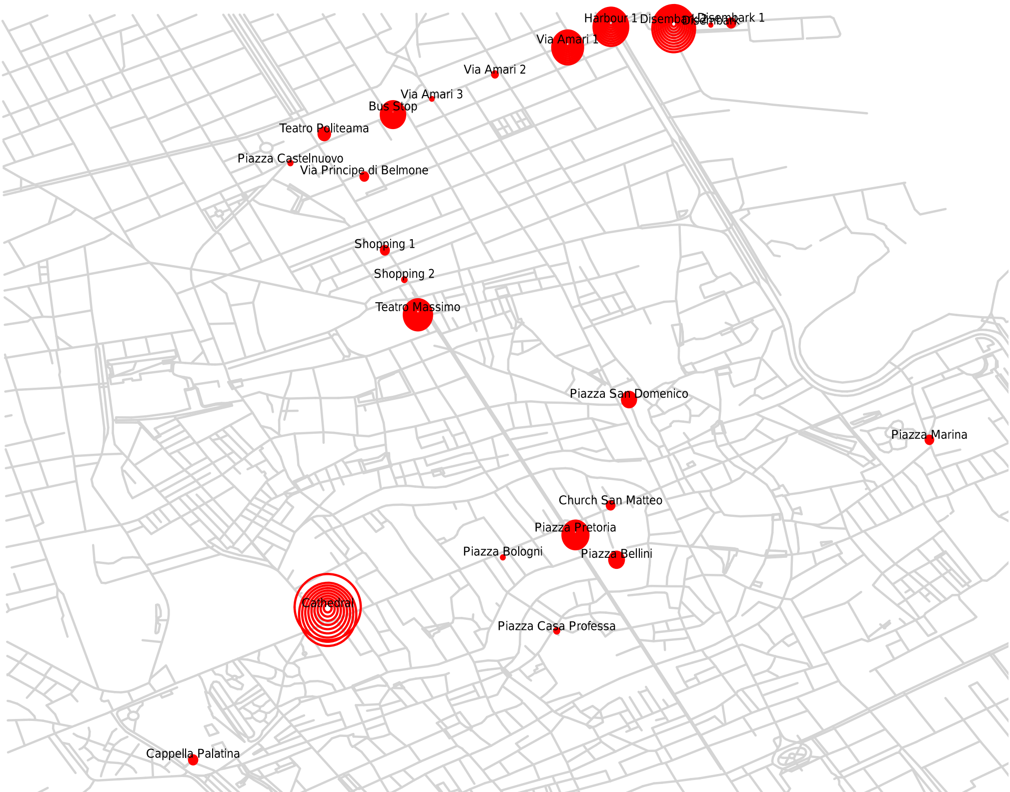

Figure 9. Circled points indicate the clusters recovered from the DBSCAN algorithm for the city of Palermo with a zoom shot of the town centre. The larger the circle, the higher the number of tourists who visited the point.

Table 2. Tourist points of interest in Palermo recovered by the DBSCAN algorithm.

Points of attraction names	Frequency	%Frequency
Cathedral	196	66.44%
Disembarkation 2	132	44.75%
Harbour 1	107	36.27%
Via Amari 1	95	32.20%
Teatro Massimo	87	29.49%
Piazza Pretoria	80	27.12%
Bus Stop	75	25.42%
Piazza Bellini	45	15.25%
Piazza San Domenico	43	14.58%
Santuario di Santa Rosalia	37	12.54%
Teatro Politeama	35	11.86%
Catacombs	34	11.53%
Disembarkation point 1	25	8.47%
Cappella Palatina	25	8.47%
Shopping 1	24	8.14%
Church San Matteo	23	7.80%
Piazza Marina	23	7.80%
Via Principe di Belmonte	23	7.80%
Via Amari 2	17	5.76%
Piazza Casa Professa	15	5.08%
Shopping 2	13	4.41%
Piazza Castelnuovo	12	4.07%
Cattedrale di Monreale	12	4.07%
Piazza Bologni	11	3.73%
Piazza di Monreale	10	3.39%
Via Amari 3	10	3.39%
Disembarkation point 3	10	3.39%

Note: Frequencies represent the number of subjects in each cluster. The names of the points of attraction were obtained by adding the centroid coordinates of each cluster in Google maps.

Dubrovnik and Copenhagen

This section will outline the application of the proposed *ClusNet* algorithm to GPS tracks, which are related to two samples of cruise passengers disembarking in Dubrovnik and Copenhagen. Based on the *ClusNet* algorithm, six and 10 clusters were estimated for Dubrovnik and Copenhagen, respectively.

The right part of Figure 11 shows the weighted directed networks of the main points of interest, which were visited by the units of the sample. The size of each cluster in this case is also proportional to the number of units visiting that particular point of interest. Table 3 shows the frequency of the cluster units of the top-six attractions in Copenhagen (left-hand side) and Dubrovnik (right-hand side).

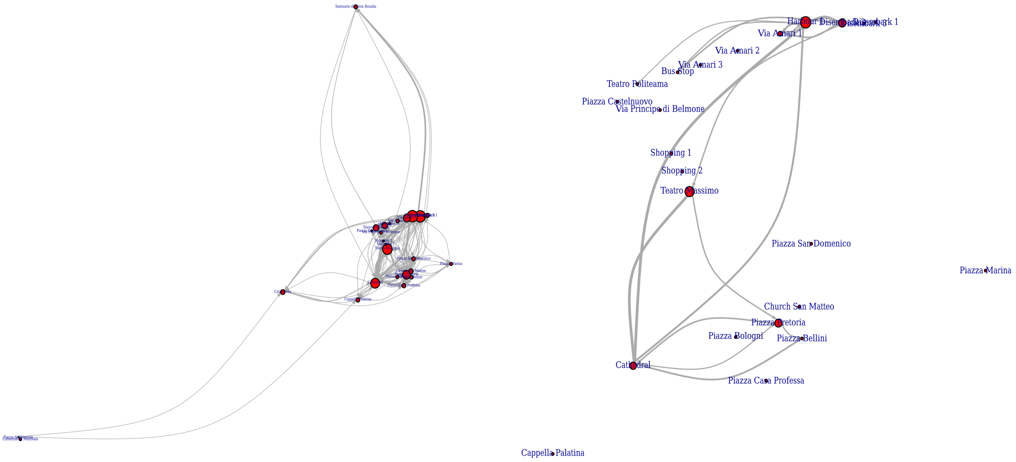

Figure 10. Weighted directed network recovered from the *ClusNet* algorithm. The graph derived by the complete adjacency matrix is shown on the left; and the reduced network (where the directed links with fewer than five moves have been removed) is shown on the right.

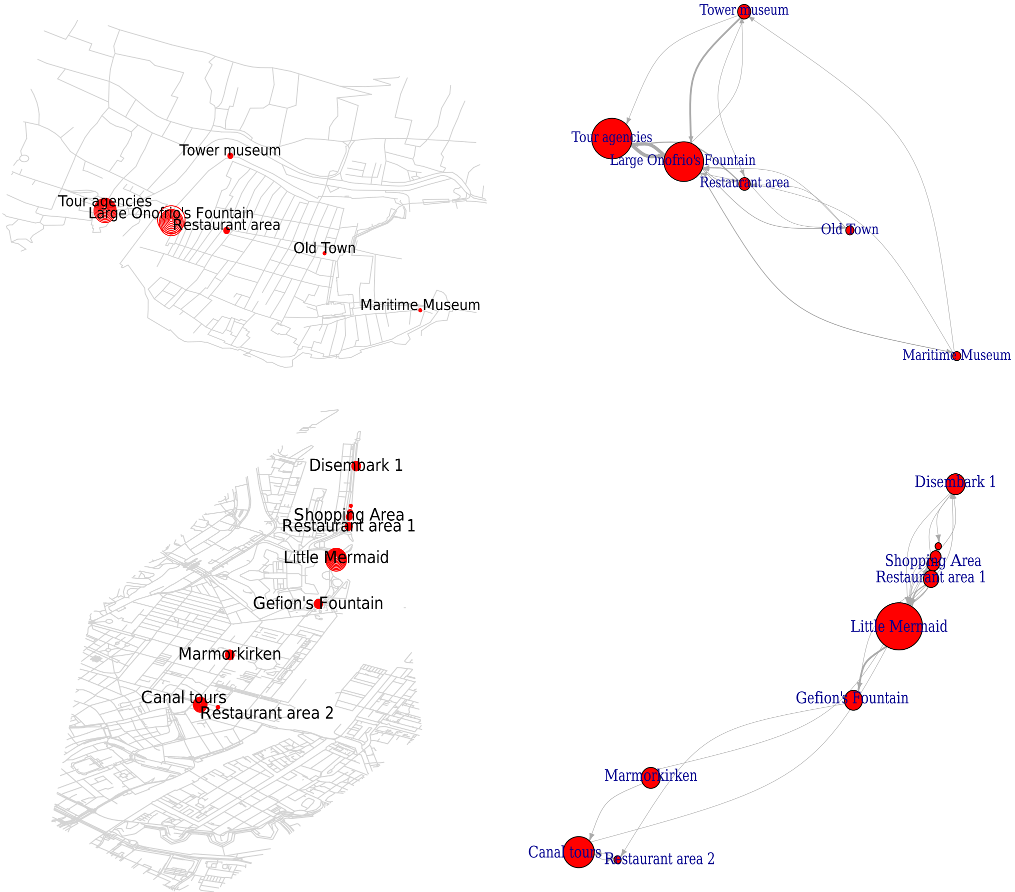

Figure 11. Centroid cluster and weighted directed networks recovered by using the *ClusNet* algorithm. On the upper part are the clusters and network of the city of Dubrovnik; and on the lower part are the clusters and network of the city of Copenhagen.

Table 3. Top-six attractions identified through the *ClusNet* algorithm in Copenhagen and Dubrovnik.

Copenhagen			Dubrovnik		
Attraction	**Frequency**	**% Frequency**	**Attraction**	**Frequency**	**% Frequency**
Little Mermaid	50	70.42%	Tour agencies	47	25.76%
Canal tours	33	46.48%	Onofrio's Fountain	46	92.16%
Disembarking point	21	29.58%	Tower museum	16	31.37%
Marmorkirken	21	29.58%	Restaurant area	14	27.45%
Real Gefion's Fountain	20	28.17%	Old Town	10	19.61%
Restaurant area 1	17	23.94%	Maritime Museum	10	19.61%

CONCLUSIONS

Any attempt to understand human mobility is a challenging topic notwithstanding its economic and social relevance; the latter includes the provision of transportation services, the spatial distribution of facilities and, more generally, regional and urban planning and destination management (Siła-Nowicka et al., 2016). The widespread availability of GPS tracking data, which can be derived from various location-aware sensors (such as smartphones, GPS tracking devices and web-based information), offer great potential in social science research. Nonetheless, the variety and complexity of this type of data poses new challenges in terms of research questions and related methods, which may arise in various fields of application. It has been demonstrated in this paper that the analysis and synthesis of GPS tracking data involve several steps that require the implementation of appropriate analytical techniques and statistical methods. The complexity and performance of the various techniques to be implemented depends on the quality and the structure of data available, the context under consideration and the specific research aim.

It has been shown in the literature that GPS tracking data are generally affected by the challenges of missing data and outlier observations. Such irregularities could introduce bias into any analysis of GPS tracking data, and this needs to be rigorously addressed. Moreover, there is a need to provide appropriate statistical methods capable of extracting relevant information from copious quantities of information relating to individual trajectories.

Departing from these considerations, we contend that the contribution of this paper to the literature is threefold. First, it proposed two algorithms regarding outlier identification and missing data imputation, with which to tackle the pre-processing of GPS tracking data. The proposed algorithms make use of various known techniques, which have been applied in other fields of study. An example is changepoint detection for the analysis of time series, here adapted for the identification of outlier observations in the context of GPS tracking data. On the other hand, the missing data imputation algorithm proposes an innovative procedure for imputing missing observations when these are lost in a segment of the trajectory and in proximity of a point of interest. The simulation results described in this paper have provided satisfactory results regarding the proposed algorithm, thereby enhancing its utility in solving common problems in the pre-processing stage of this type of data.

Second, an existing spatial clustering algorithm, namely the DBSCAN, was implemented for the identification of clusters in GPS tracking data. The aim here is to identify relevant locations at the individual and aggregate levels. Thereafter, the analysis was integrated in an original way by identifying the relationships between the identified cluster in terms of the network at the individual and aggregate levels. The sensitivity of the algorithm to the tuning parameters was also analysed by means of a simulation study. The latter provided insights regarding the robustness of the results according to the value of the tuning parameter. The simulation study also indicated ways of implementing this procedure, which, however, will depend on the research aim and the empirical context being studied. We believe the proposed methodology to be easily replicable in different contexts.

Third and from an empirical perspective, the policy implications of the knowledge to be obtained from an analysis of GPS tracking data in the tourism context are of relevance to destination management. The determination of the number of visits and information regarding the length of stay at each attraction or other areas of the destination is fundamental for service management. This information may be used to improve the administration of existing attractions, enhance the efficacy of their marketing, or plan to introduce new attractions to the tourist market (Lew & McKercher, 2006). Of great promise, knowledge about network relationships between the various places of interest (visitor flow) may orient transportation planning and promotion strategies, to mention but one of the potential applications of analysing GPS tracking data.

Future research could include the modelling of network relationships among points of interest regarding individual and contextual information, to include the characteristics of the various locations. Furthermore, the effectiveness of policy actions may be measured in terms of their capability of changing the structure of relationships of a given network. The increasing availability of spatio-temporal information describing human movements in general and tourist behaviour specifically make imperative the identification of new research questions. The latter will demand relevant methodological solutions, a challenging topic for future research indeed.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the invaluable comments of the editor and reviewers. They are also grateful to Professor Szilvia Gyimóthy at the Copenhagen Business School, Department of Marketing; the Department Internationalisation and Tourism at the Municipality of Copenhagen; and the Copenhagen Cruise Network at VisitCopenhagen, Copenhagen, for data collection operations in Copenhagen.

DISCLOSURE STATEMENT

No potential conflict of interest was reported by the authors.

ORCID

Antonino Abbruzzo http://orcid.org/0000-0003-2196-3570
Mauro Ferrante http://orcid.org/0000-0003-1287-5851
Stefano De Cantis http://orcid.org/0000-0002-5068-6421

REFERENCES

Ahas, R., Silm, S., Järv, O., Saluveer, E., & Tiru, M. (2010). Using mobile positioning data to model locations meaningful to users of mobile phones. *Journal of Urban Technology*, *17*(1), 3–27. https://doi.org/10.1080/10630731003597306

Alvares, L. O., Bogorny, V., Kuijpers, B., de Macedo, J. A. F., Moelans, B., & Vaisman, A. (2007, November 7–9). *A model for enriching trajectories with semantic geographical information*. Proceedings of the 15th International Symposium on Advances in Geographic information Systems ACM GIS 2007 A, Seattle, Washington, USA (pp. 1–8).

Andrade, T., & Gama, J. (2019). Identifying Points of Interest and Similar Individuals from Raw GPS Data. In: Cagáňová D., Horňáková N. (Eds) Mobility Internet of Things 2018. Mobility IoT 2018. EAI/Springer

Innovations in Communication and Computing. Springer, Cham. https://doi.org/10.1007/978-3-030-30911-4_21.

Atluri, G., Karpatne, A., & Kumar, V. (2018). Spatio-temporal data mining: A survey of problems and methods. *ACM Computing Surveys (CSUR)*, *51*(4), 1–41. https://doi.org/10.1145/3161602

Bermingham, L., & Lee, I. (2017). A framework of spatio-temporal trajectory simplification methods. *International Journal of Geographical Information Science*, *31*(6), 1128–1153. https://doi.org/10.1080/13658816.2017.1290250

Blanchard, R. A., Myers, A. M., & Porter, M. M. (2010). Correspondence between self-reported and objective measures of driving exposure and patterns in older drivers. *Accident Analysis & Prevention*, *42*(2), 523–529. https://doi.org/10.1016/j.aap.2009.09.018

Brakatsoulas, S., Pfoser, D., & Tryfona, N. (2004, July 9). *Modeling, storing and mining moving object databases*. Proceedings: International database Engineering and applications Symposium, 2004. IDEAS'04, Coimbra, Portugal.

Butz, W. P., & Torrey, B. B. (2006). Some frontiers in social science. *Science*, *312*(5782), 1898–1900. https://doi.org/10.1126/science.1130121

Campello, R. J., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *7819 LNAI*(PART 2), 160–172. https://doi.org/10.1007/978-3-642-37456-2_14

Castro, M., Iglesias, L., Rodríguez-Solano, R., & Sánchez, J. A. (2006). Geometric modelling of highways using global positioning system (GPS) data and spline approximation. *Transportation Research Part C: Emerging Technologies*, *14*(4), 233–243. https://doi.org/10.1016/j.trc.2006.06.004

Chazal, F., Chen, D., Guibas, L., Jiang, X., & Sommer, C. (2011, November 1–4). *Data-driven trajectory smoothing*. Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Chicago, IL, USA.

Chen, C., Gong, H., Lawson, C., & Bialostozky, E. (2010). Evaluating the feasibility of a passive travel survey collection in a complex urban environment: Lessons learned from the New York city case study. *Transportation Research Part A: Policy and Practice*, *44*(10), 830–840. https://doi.org/10.1016/j.tra.2010.08.004

Desgraupes, B. (2013). Clustering indices. *University of Paris Ouest-Lab Modal?X*, *1*, 34. Retrieved June 2020, from https://cran.biodisk.org/web/packages/clusterCrit/vignettes/clusterCrit.pdf

Du Mouza, C., & Rigaux, P. (2005). Mobility patterns. *GeoInformatica*, *9*(4), 297–319. https://doi.org/10.1007/s10707-005-4574-9

Elgethun, K., Fenske, R. A., Yost, M. G., & Palcisko, G. J. (2003). Time–location analysis for exposure assessment studies of children using a novel global positioning system instrument. *Environmental Health Perspectives*, *111* (1), 115–122. https://doi.org/10.1289/ehp.5350

Erenoglu, R., & Hekimoglu, S. (2010). Efficiency of robust methods and tests for outliers for geodetic adjustment models. *Acta Geodaetica et Geophysica Hungarica*, *45*(4), 426–439. https://doi.org/10.1556/AGeod.45.2010.4.3

Ester, M., Kriegel, H.-P., Jorg, S., & Xu, X. (1996, August 2–4). A density-based clustering algorithms for discovering clusters. In E. Simoudis, J. Han, and U. Fayyad (Eds.), *KDD-96: Proceedings of the second international conference on knowledge discovery and data mining, 96(34), Portland, Oregon*, (pp. 226–231). Retrieved June 2020, from https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf

Ferrante, M., De Cantis, S., & Shoval, N. (2018). A general framework for collecting and analysing the tracking data of cruise passengers at the destination. *Current Issues in Tourism*, *21*(12), 1426–1451. https://doi.org/10.1080/13683500.2016.1194813

Gilad, E.-T. (2001). Graph theory applications to gps networks. *GPS Solutions*, *5*(1), 31–38. https://doi.org/10.1007/PL00012874

Giorgino, T. (2009). Computing and visualizing dynamic time warping alignments in R: The dtw package. *Journal of Statistical Software*, *31*(7), 1–24. https://doi.org/10.18637/jss.v031.i07

Gong, H., Chen, C., Bialostozky, E., & Lawson, C. T. (2012). A GPS/GIS method for travel mode detection in New York city. *Computers, Environment and Urban Systems*, *36*(2), 131–139. https://doi.org/10.1016/j.compenvurbsys.2011.05.003

Gong, L., Sato, H., Yamamoto, T., Miwa, T., & Morikawa, T. (2015). Identification of activity stop locations in GPS trajectories by density-based clustering method combined with support vector machines. *Journal of Modern Transportation*, *23*(3), 202–213. https://doi.org/10.1007/s40534-015-0079-x

Gong, L., Yamamoto, T., & Morikawa, T. (2018). Identification of activity stop locations in GPS trajectories by dbscan-te method combined with support vector machines. *Transportation Research Procedia*, *32*, 146–154. https://doi.org/10.1016/j.trpro.2018.10.028

Güting, R. H., de Almeida, T., & Ding, Z. (2006). Modeling and querying moving objects in networks. *The VLDB Journal—The International Journal on Very Large Data Bases*, *15*(2), 165–190. https://doi.org/10.1007/s00778-005-0152-x

Hahsler, M., Piekenbrock, M., Arya, S., & Mount, D. (2018). Package 'dbscan': Density Based Clustering of Applications with Noise (DBSCAN) and Related Algorithms. *R package version*. https://github.com/mhahsler/dbscan.

Haynes, K., Eckley, I. A., & Fearnhead, P. (2017a). Computationally Efficient changepoint detection for a range of Penalties. *Journal of Computational and Graphical Statistics*, *26*(1), 134–143. https://doi.org/10.1080/10618600.2015.1116445

Haynes, K., Fearnhead, P., & Eckley, I. A. (2017b). A computationally efficient nonparametric approach for changepoint detection. *Statistics and Computing*, *27*(5), 1293–1305. https://doi.org/10.1007/s11222-016-9687-5

Hu, W., Xie, D., & Tan, T. (2004). A hierarchical self-organizing approach for learning the patterns of motion trajectories. *IEEE Transactions on Neural Networks*, *15*(1), 135–144. https://doi.org/10.1109/TNN.2003.820668

Jun, J., Guensler, R., & Ogle, J. (2006). Smoothing methods designed to minimize the impact of GPS random error on travel distance, speed, and acceleration profile estimates. *Transportation Research Record*, *1972*, 141–150. https://doi.org/10.1177/0361198106197200117

Khetarpaul, S., Gupta, S., Subramaniam, L. V., & Nambiar, U. (2012, August 8–10). *Mining GPS traces to recommend common meeting points*. Proceedings of the 16th International Database Engineering & Applications Sysmposium, Prague (pp. 181–186). https://doi.org/10.1145/2351476.2351497

Knapen, L., Hartman, I. B.-A., & Bellemans, T. (2020). Using path decomposition enumeration to enhance route choice models. *Future Generation Computer Systems*, *107*, 1077–1088. https://doi.org/10.1016/j.future.2017.12.053

Knapen, L., Hartman, I. B.-A., Schulz, D., Bellemans, T., Janssens, D., & Wets, G. (2016). Determining structural route components from gps traces. *Transportation Research Part B: Methodological*, *90*, 156–171. https://doi.org/10.1016/j.trb.2016.04.019

Lew, A., & McKercher, B. (2006). Modeling tourist movements: A local destination analysis. *Annals of Tourism Research*, *33*(2), 403–423. https://doi.org/10.1016/j.annals.2005.12.002

Lin, M., & Hsu, W.-J. (2014). Mining gps data for mobility patterns: A survey. *Pervasive and Mobile Computing*, *12*, 1–16. https://doi.org/10.1016/j.pmcj.2013.06.005

McCabe, S., Kwan, H., & Roorda, M. J. (2013). Comparing GPS and non-GPS survey methods for collecting urban goods and service movements. *International Journal of Transport Economics/Rivista Internazionale di Economia dei Trasporti*, *40*(2), 183–205. Retrieved June 2020, from https://www.jstor.org/stable/42748309

Murakami, E., & Wagner, D. P. (1999). Can using global positioning system (GPS) improve trip reporting? *Transportation Research Part C: Emerging Technologies*, *7*(2–3), 149–165. https://doi.org/10.1016/S0968-090X(99)00017-0

Palma, A. T., Bogorny, V., Kuijpers, B., & Alvares, L. O. (2008, March 16–20). *A clustering-based approach for discovering interesting places in a single trajectory*. SAC '08: Proceedings of the 2008 ACM symposium on Applied computing, Fortaleza, Ceará, Brazil (pp. 863–868), https://doi.org/10.1145/1363686.1363886

Raanan, M. G., & Shoval, N. (2014). Mental maps compared to actual spatial behavior using GPS data: A new method for investigating segregation in cities. *Cities*, *36*, 28–40. https://doi.org/10.1016/j.cities.2013.09.003

Rabiner, L. R., & Juang, B.-H. (1999). *Fundamentals of speech recognition*. Tsinghua University Press.

R Development Core Team. (2008). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org

Schubert, E., Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (2017). DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems*, *42*(3), 1–21. https://doi.org/10.1145/3068335

Schuler, K. L., Schroeder, G. M., Jenks, J. A., & Kie, J. G. (2014). Ad hoc smoothing parameter performance in kernel estimates of GPS-derived home ranges. *Wildlife Biology*, *20*(5), 259–266. https://doi.org/10.2981/wlb.12117

Sester, M., Feuerhake, U., Kuntzsch, C., & Zhang, L. (2012). Revealing underlying structure and behaviour from movement data. *KI-Künstliche Intelligenz*, *26*(3), 223–231. https://doi.org/10.1007/s13218-012-0180-9

Shen, L., & Stopher, P. R. (2014). Review of GPS travel survey and GPS data-processing methods. *Transport Reviews*, *34*(3), 316–334. https://doi.org/10.1080/01441647.2014.903530

Shoval, N. (2008). Tracking technologies and urban analysis. *Cities*, *25*(1), 21–28. https://doi.org/10.1016/j.cities.2007.07.005

Shoval, N., & Ahas, R. (2016). The use of tracking technologies in tourism research: The first decade. *Tourism Geographies*, *18*(5), 587–606. https://doi.org/10.1080/14616688.2016.1214977

Siła-Nowicka, K., Vandrol, J., Oshan, T., Long, J. A., Demšar, U., & Fotheringham, A. S. (2016). Analysis of human mobility patterns from GPS trajectories and contextual information. *International Journal of Geographical Information Science*, *30*(5), 881–906. https://doi.org/10.1080/13658816.2015.1100731

Silm, S., & Ahas, R. (2010). The seasonal variability of population in Estonian municipalities. *Environment and Planning A: Economy and Space*, *42*(10), 2527–2546. https://doi.org/10.1068/a43139

Spaccapietra, S., Parent, C., Damiani, M. L., de Macedo, J. A., Porto, F., & Vangenot, C. (2008). A conceptual view on trajectories. *Data and Knowledge Engineering*, *65*(1), 126–146. https://doi.org/10.1016/j.datak.2007.10.008

Stopher, P., FitzGerald, C., & Zhang, J. (2008). Search for a global positioning system device to measure person travel. *Transportation Research Part C: Emerging Technologies*, *16*(3), 350–369. https://doi.org/10.1016/j.trc.2007.10.002

Theo, L. (2011). Simplifying central place theory using GIS and GPS. *Journal of Geography*, *110*(1), 16–26. https://doi.org/10.1080/00221341.2010.511244

Tiwari, S., & Kaushik, S. (2013, March 25–27). Mining popular places in a geo-spatial region based on GPS data using semantic information. In A. Madaan, S. Kikuchi, & S. Bhalla (Eds.), *Databases in networked information systems, 8th international workshop DNIS 2013, Aizu-Wakamatsu, Japan* (pp. 262–276). https://doi.org/10.1007/978-3-642-37134-9_20

Torrens, P., Li, X., & Griffin, W. A. (2011). Building agent-based walking models by machine-learning on diverse databases of space–time trajectory samples. *Transactions in GIS*, *15*, 67–94. https://doi.org/10.1111/j.1467-9671.2011.01261.x

Třasák, P., & Štroner, M. (2014). Outlier detection efficiency in the high precision geodetic network adjustment. *Acta Geodaetica et Geophysica*, *49*(2), 161–175. https://doi.org/10.1007/s40328-014-0045-9

Tsui, J. B.-Y. (2005). *Fundamentals of global positioning system receivers: A software approach* (Vol. 173). John Wiley & Sons.

Tukey, J. (1977). *Exploratory data analysis* (Vol. 2). Addison-Wesley.

Van Craenendonck, T., & Blockeel, H. (2015). Using internal validity measures to compare clustering algorithms. *Benelearn 2015 Poster presentations (online)*, 1–8.

Van der Spek, S., Van Schaick, J., De Bois, P., & De Haan, R. (2009). Sensing human activity: GPS tracking. *Sensors*, *9*(4), 3033–3055. https://doi.org/10.3390/s90403033

Wolf, J., Guensler, R., & Bachman, W. (2001). Elimination of the travel diary: Experiment to derive trip purpose from global positioning system travel data. *Transportation Research Record: Journal of the Transportation Research Board*, *1768*(1), 125–134. https://doi.org/10.3141/1768-15

Wolf, J., Oliveira, M., & Thompson, M. (2003). Impact of underreporting on mileage and travel time estimates: Results from global positioning system-enhanced household travel survey. *Transportation Research Record: Journal of the Transportation Research Board*, *1854*(1), 189–198. https://doi.org/10.3141/1854-21

Wolfson, O., Xu, B., Chamberlain, S., & Jiang, L. (1998, July 3). *Moving objects databases: Issues and solutions*. Proceedings of the Tenth International Conference on Scientific and Statistical Database Management (cat. no. 98tb100243) (pp. 111–122). https://doi.org/10.1109/SSDM.1998.688116

Xiao, G., Juan, Z., & Zhang, C. (2015). Travel mode detection based on GPS track data and Bayesian networks. *Computers, Environment and Urban Systems*, *54*, 14–22. https://doi.org/10.1016/j.compenvurbsys.2015.05.005

Zenk, S. N., Schulz, A. J., Matthews, S. A., Odoms-Young, A., Wilbur, J., Wegrzyn, L., Gibbs, K., Braunschweig, C., & Stokes, C. (2011). Activity space environment and dietary and physical activity behaviors: A pilot study. *Health & Place*, *17*(5), 1150–1161. https://doi.org/10.1016/j.healthplace.2011.05.001

Zheng, Y., Zhang, L., Xie, X., & Ma, W.-Y. (2009, April 20–24). *Mining interesting locations and travel sequences from GPS trajectories*. WWW'09: Proceedings of the 18th international conference on World Wide Web, Madrid, Spain (pp. 791–800). https://doi.org/10.1145/1526709.1526816