ICIC Express Letters
Part B: Applications
Volume **11**, Number **10**, October **2020**

ICIC International ©2020 ISSN 2185-2766
pp. 921–928

TRAJECTORY CLUSTERING BY GPS TRACKING DATASET USING QUICKBUNDLES

Nattapon Kumyaito[1] and Kreangsak Tamee[1,2]

[1]Department of Computer Science and Information Technology
[2]Research Center for Academic Excellence in Nonlinear Analysis and Optimization
Naresuan University
Muang, Phitsanulok 65000, Thailand
{ nattaponk; kreangsakt }@nu.ac.uk

Received March 2020; accepted June 2020

Abstract. *Trajectory clustering is considered as analytical methods in trajectory data mining for many modern applications such as a tourist recommendation system. Our research project is a route recommending system for cycling tourists, and we aim to retrieve all their route patterns by adopting GPS-based trajectories dataset. This dataset is informative and high volume; thus, we need clustering methods that are effective and fast in computational time. This paper proposes methods to collect, preprocess, and analyze the GPS tracking data from the existing online service for cyclists. QuickBundles is the primary technique we use to allocate route items into clusters, and also propose cluster validation methods based on silhouette coefficient. The results have shown that QuickBundles has the capability of clustering the trajectory data in terms of computational speed and validity. This approach can benefit online systems, by recommending personalized routes for the cycling tourists.*

Keywords: Trajectory model, Data processing, Pattern classification, Recommendation system

1. **Introduction.** Nowadays, location-tracking devices have evolved and are widely adopted. These devices are capable of precisely tracking locations and record them on-the-fly to cloud services, with the most famous one being the Global Positioning System (GPS). GPS provides the advantage of both global and accuracy, via a group of satellites that orbit the earth broadcasting signals that are picked up by a system of receivers.

Li et al. [1] proposed that the use of GPS for tourism research was very beneficial for tourism research which matured from feasibility and usefulness, tourist behavior and tourism recommendation. Since our research project aimed to classify routes by analyzing GPS data, we reviewed other related works that analyzed routes for tourism. For instance, Du et al. [2] used the topic hierarchy of scenic spots and the features of scenic spots for travel route mining system. Jiang et al. [3] identified scenic hotspot by proposing density based clustering algorithm using the data of geographical coordinates and tour route data. Malik and Kim [4] proposed travel route optimization techniques that use tourist vehicle data. Hamid and Croock [5] predicted tourists' interested place using their smartphone's GPS trajectory. Nardini et al. [6] analyzed the trajectories of mobile users to initiate the recommender system for tourists.

This research aims to propose the route recommendation system for cyclists who need to discover riding routes that have similar characteristic to their riding history. This system must recognize the routes pattern from cyclists' activities data source that may be considered as big data. Therefore, we need effective methods to cluster vast amount of data into groups of route patterns. We adopted the QuickBundles (QB) [7] algorithm, due to its efficient computational complexity and simplification, which was successfully

DOI: 10.24507/icicelb.11.10.921

applied to tractography clustering analysis in Diffusion MRI dataset that overcame the complexity of extensive data, and provided informative clusters in seconds. The details of the QB technique are described in Section 2.3.

The further details of trajectory clustering in this paper are presented in the following sections. Section 2, Problem Statements and Preliminaries, describes the main purpose of trajectory clustering problems and the methods required to prepare and analyze the dataset. Section 3, Experiment, describes the techniques for constructing trajectory clustering models and its results. Section 4, Discussion, presents the findings of this study. Finally, Section 5, Conclusions, will sum up the important issues and research opportunities.

2. Problem Statements and Preliminaries.

Trajectory data is recorded in different forms according to the types of sensors, objects movement, or tracking purposes. In this paper, we focused on geographical trajectory data, which is represented as a sequence of n points in geographical space as the following equation:

$$T = (t_1, t_2, \ldots, t_n) \tag{1}$$

where t_i is a tracking point which is represented as a combination of coordination and timestamp that forms the sub equation:

$$t_i = (x_i, y_i, z_i, ts_i) \tag{2}$$

where x_i, y_i, z_i, and ts_i are latitude, longitude, altitude, and timestamp, respectively. This data sequence is an essential standard for several data schemes in GPS tracking devices, while some of the data schemes add other relevant information, such as velocity and cumulative distance.

2.1. Dataset.

The cycling workout data that contains GPS track points were explored and selected from online data services, https://ridewithgps.com/. This research used selection criteria as follows.

1) The selected data must be recorded as TCX file.

2) Every track point in the selected files must have altitude data.

3) The route length in the selected files must range between 3 and 180 km.

After previous selection criteria were applied, selected files were converted into raw dataset using a GPSBabel application [8]. However, this data had defects such as missing values and/or sensor reading errors. Therefore, preprocessing is required to deal with the defects in the raw dataset.

2.2. Data preprocessing.

The dataset usually contains some errors and/or outliers caused by the hardware. Therefore, to rectify these problems, the dataset was preprocessed by the following methods.

1) Unify the GPS track point's time intervals to 1 second.

2) To correct the outlier and missing values, we used the linear Kalman filter technique [9].

3) All elevation data was altered by resetting the initial track points to zero.

2.3. Cluster analysis using QuickBundles.

QB is a clustering technique for classifying routes by their elevation profile, and its tractography simplification of a complex Diffusion MRI dataset provides informative clusters in seconds. The tractography that is sequential of points in R^3 can be categorized as a form of trajectory data. Therefore, QB is a clustering technique that may also be able to cluster trajectory data. The primary reason why QB successfully classifies trajectory data is that it uses the simple symmetric distance function, which is called the minimum average direct-flip (MDF) distance as

Equations (3) to (5).

$$d_{direct}(s,t) = d(s,t) = \frac{1}{K}\sum_{i=1}^{K} |s_i - t_i| \tag{3}$$

$$d_{flipped}(s,t) = d\left(s,t^F\right) = d\left(s^F,t\right) \tag{4}$$

$$MDF(s,t) = \min(d_{direct}(s,t), d_{flipped}(s,t)) \tag{5}$$

where a trajectory $s = [s_1, s_2, \ldots, s_k]$ consists of k sequential tract points. s^F is a flipped version on s where $s^F = [s_k, s_{k-1}, \ldots, s_2, s_1]$. $|x - y|$ denotes the Euclidean distance between two points x and y. The direct distance $d_{direct}(s,t)$ between two trajectories s and t is the mean of Euclidean distance between corresponding track points. The main advantages of the MDF distance supports QB to have substantial high performance by taking account of both direct and flipped trajectories.

2.4. **Cluster validation.** Since our work is unsupervised, clustered data without labels are unknown, cluster validation is considered an essential method to confirm the clustering technique by using the silhouette coefficient [10]. Let i be any object in the clustering and A its corresponding cluster. Then

$$a(i) = \frac{1}{|A| - 1}\sum_{j\in A, j\neq i} \Delta(i,j) \tag{6}$$

measures the average distance of i to all other objects in cluster A. We then compute each cluster $C \neq A$

$$d(i,C) = \frac{1}{|C|}\sum_{j\in C} \Delta(i,j) \tag{7}$$

to quantify the distance to other ones. The minimum value is

$$b(i) = \min_{C\neq A} d(i,C) \tag{8}$$

that gives the distance of i to the second-best cluster; therefore, the silhouette value $s(i)$ of i is then defined as

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{9}$$

This value in the range of $[-1; 1]$ quantifies how well object i fits into cluster A. If the $s(i)$ is negative, the object is placed in the wrong cluster, and if it is around 0, the object is between the clusters, while if it is positive, the object is placed in the correct cluster.

3. **Experiment.** In this section, the preprocessed dataset that is a result of methods in prior sections is used for route clustering. The experiment goal is to classify the routes dataset according to their elevation profiles by applying QB.

Since the length of the trajectory item possibly influences the QB's result [7], the route dataset was pre-classified by length. Therefore, the differentiation of trajectory had less effect. We use k-Means to classify routes into 3 groups: short-distance, medium-distance, and long-distance ones. After trajectory data items were grouped by their route length, each item in each group had their altitude profiles classified using QB clustering.

However, the primary parameters that influence the validity of QB clustering are their distance threshold by varying their range by 10, 20, 50, 100, 200 mts, and vary the number of clusters from 2 to 6. To evaluate the validity of clustering, we applied the silhouette coefficient [10] to each group with similar route lengths. We aimed to find the best parameter schemes which presented the best results by clustering validity. We focused on the parameter scheme with a high average silhouette coefficient.

3.1. **Route-length grouping using k-Means.** Since this research is part of the route recommendation system for cyclists, the route distances are a primary factor that they need to consider according to their physical fitness. Instead of recommending routes without distances, to recommend ones that meet the cyclists preferred distances is more useful. In addition, the QB technique is sensitive to the varying lengths of the trajectories in the dataset. Therefore, they need to be classified by their total distances. Our classification criteria were applied due to our expert's suggestion that the appropriate distance categories should be termed as, short, medium, and long. Therefore, this paper divides the data of 481 trajectory items into 3 categories of route length by applying k-Means where $k = 3$, and the results are illustrated in Table 1.

Table 1. Classification dataset by route length

Group	**Shortest route (km)**	**Longest route (km)**	**Number of routes**
Short	1.078	59.130	303
Medium	60.005	133.918	124
Long	134.609	270.926	54

3.2. **Cluster analysis.** In the QB clustering technique, the primary parameters that influence clustering validity are the QB thresholds and the defined number of clusters. In Table 2 to Table 4, we demonstrate the average silhouette coefficient for the classification

Table 2. Average silhouette coefficient for classification of a short-distance group

QB threshold	Number of clusters				
	2	3	4	5	6
20	0.7455	0.5144	0.2650	0.1527	0.1635
30	0.7455	0.5144	0.2722	0.2792	0.2786
50	0.8601	0.6572	0.6282	0.6263	0.6308
100	0.8751	0.8346	0.8207	0.7925	0.7925
200	0.8834	0.8834	0.8834	0.8834	0.8834

Table 3. Average silhouette coefficient for classification of a medium-distance group

QB threshold	Number of clusters				
	2	3	4	5	6
20	0.6814	0.5826	0.4564	0.3317	0.2557
30	0.6814	0.5826	0.4564	0.3317	0.3347
50	0.6814	0.5826	0.4564	0.4709	0.4459
100	0.6814	0.5935	0.5268	0.5425	0.5740
200	0.6991	0.7034	0.7034	0.7034	0.7034

Table 4. Average silhouette coefficient classification of a long-distance group

QB threshold	Number of clusters				
	2	3	4	5	6
20	0.3772	0.2860	0.3406	0.2590	0.2212
30	0.3772	0.2860	0.3406	0.2590	0.2212
50	0.3772	0.2860	0.3406	0.2590	0.2212
100	0.3772	0.2860	0.4229	0.2590	0.2212
200	0.3772	0.4407	0.4229	0.4229	0.4229

of short, medium, and long-distance routes which corresponds to the QB thresholds and the number of clusters.

For the short-distance group, the silhouette coefficient scores corresponding to the variated QB thresholds and the variated number of clusters are demonstrated in Table 2. When the QB threshold is 200, the average silhouette coefficient value for each cluster is higher than all other parameters. The lowest number of clusters is preferred because it classifies the items effectively, while minimizing computation time by comparing items distance against a less number of centroids. Therefore, we considered the QB threshold = 200 and the number of clusters = 2 as the optimal parameters for the short-distance group.

For the medium-distance group, the silhouette coefficient scores corresponding to the variated QB threshold and the variated number of clusters are demonstrated in Table 3. The average silhouette coefficient for QB threshold = 200 in all clusters is the highest among other parameter configurations. We selected QB threshold = 200 and the number of clusters = 3 as the best parameter configuration because it is the lowest number of clusters that have the highest average silhouette coefficient values.

For the long-distance group, as illustrated in Table 4, the parameter configuration that has the highest average silhouette coefficient consists of the QB threshold at 200 and 3 clusters.

3.3. **Route clustering.** After we achieved the optimal parameter for QB clustering, we applied it to the dataset for all groups, and routes were divided according to their length and elevation profile.

For the short-distance group, 303 routes were divided into 2 clusters according to their elevation profiles, and the centroids of the clusters that correspond to each category are illustrated in Figure 1(a).

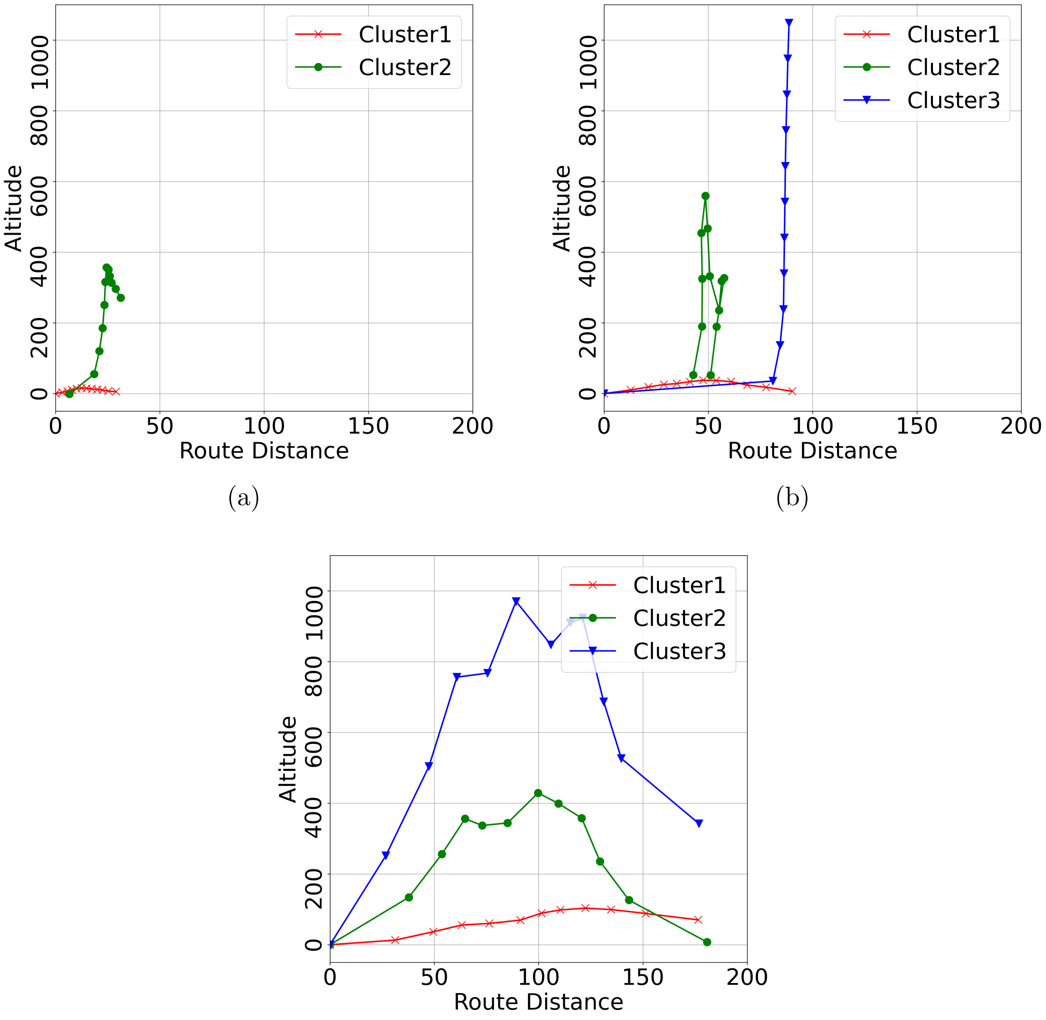
Short routes elevation in the 1st cluster ranges between −227.88 to 316.43 meters above sea-level while the other cluster has their routes elevation range between 9.10 to 836.69 meters above sea-level, as illustrated in Figure 2(a). It shows that most of routes elevation data in the 1st cluster are relatively closer to each other at sea-level while routes elevation in the 2nd cluster is distributed among above sea-level altitude.

For the medium-distance group, 124 routes were divided into 3 clusters according to their elevation profiles, and the centroids of the clusters corresponding to each category are illustrated in Figure 1(b).

The 1st cluster of the medium-distance group consists of 89 routes with their altitudes ranging from −417.80 to 429.40 meters. The 2nd cluster has 34 routes and altitudes ranging from −181.56 to 765.33 meters above sea-level. The last cluster consists of only 1 route, and its altitude ranges between −9.11 to 836.69 meters. As illustrated in Figure 2(b), routes in the 1st cluster mostly remain in sea-level while routes in the 2nd cluster distribute on higher altitude. Finally, the route elevation in the last cluster distributes among limited range of above sea-level but consists of the highest altitude point in comparison to other routes.

In the long-distance group, 54 routes were divided into 3 clusters according to their elevation profiles, and the centroids that corresponded to each category, are illustrated in Figure 1(c).

The 1st cluster consists of 24 routes with their altitudes ranging from −319.14 to 818.4 meters. The 2nd cluster consists of 23 routes with varying altitudes between −191.80 to 481.48 meters. The 3rd cluster consists of 7 routes with altitude ranges from 0 to 1576.68 meters. Route elevation distribution in each cluster was illustrated in Figure 2(c). The 2nd cluster consists routes that mostly locate at sea-level while others consists of routes that locate above sea-level. However, routes in the 1st cluster locate at lower altitude than routes in the 3rd cluster.



(a)

(b)

(c)

Figure 1. (a) Cluster centroids of short-distance groups, (b) cluster centroids of medium-distance groups, and (c) cluster centroids of long-distance groups

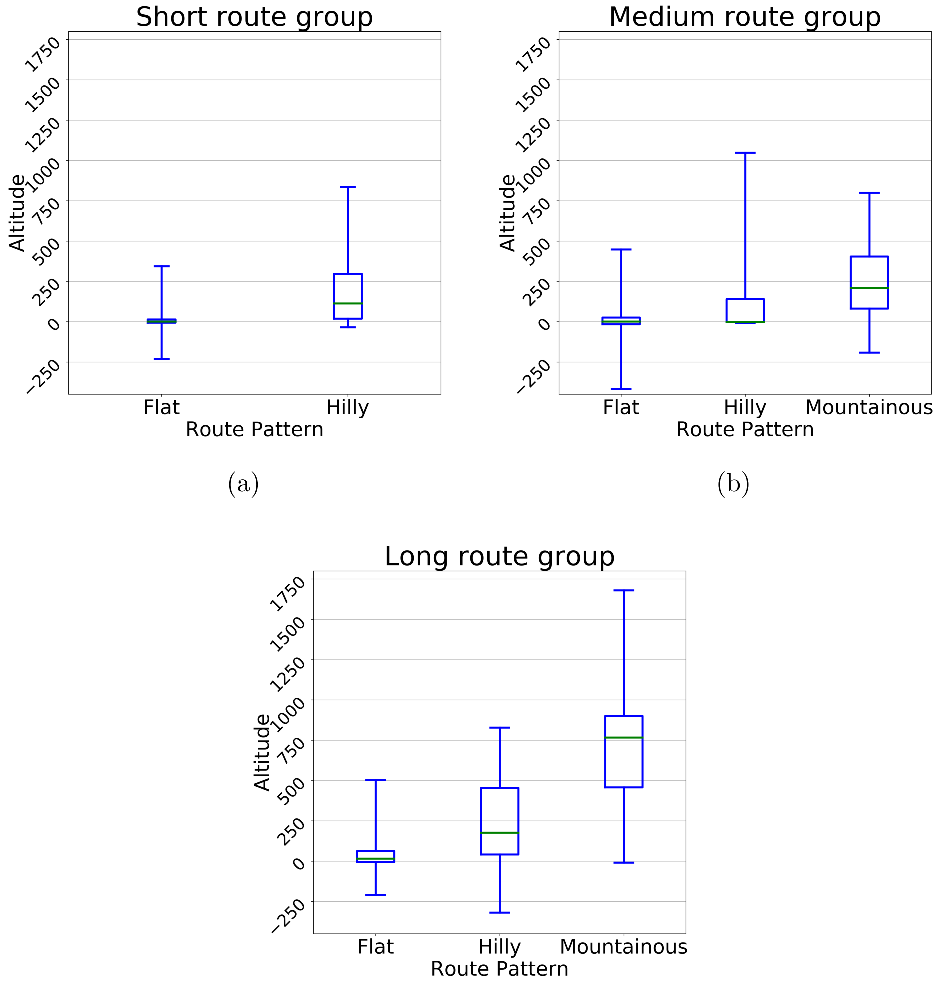
4. **Discussion.** The QB technique has shown that it is an effective clustering method for GPS datasets. QB threshold and the number of clusters are the primary parameters that influence the validity of route clustering based on their elevation profile. The centroid of each cluster showed some distinctive patterns that can be described as the following.

1) The flat route pattern. Route elevation profiles in this cluster usually show sustained horizontal levels with some minor changes that raises the lowest level in comparison to other clusters.

2) The hilly route pattern. The elevation profile of routes in this cluster shares a similar pattern which is a combination of flat and hilly terrain, and the raised altitude is higher than the 1st cluster but lower than the 3rd one.

3) The mountainous route pattern. Overall, routes in this pattern consist of high altitudes that are greater than either of the others. This trajectory pattern usually starts at a low altitude, but does not define the finish elevation.

The description of classified routes conforms to the cycling expert's opinion that categorizes the routes into 3 types: flat, hilly, and mountainous. Therefore, these proposed techniques have the capabilities to retrieve route patterns by classifying them according to elevation automatically. Our methods are considered as a useful tool for recommending routes to cyclists that are similar to their preferences.



(a)

(b)

(c)

Figure 2. (a) The box plot for routes elevation by short-distance group, (b) the box plot for routes elevation by medium-distance group, and (c) the box plot for routes elevation by long-distance group

In addition, the reason for QB's high performance is the algorithm's complexity. The best case is when QB's complexity is linear time $O(N)$ with the number of items N while the worst-case is $O(N2)$ when every cluster contains only 1 item [7]. The average case is $O(MN)$, where M is the number of clusters, but M is typically lower than N; thus, we can neglect it and denote it as $O(N)$ because it is common in the complexity theory. In this experiment, QB takes 250 milliseconds to analyze and classify the route elevations and total distance of the entire dataset. The processing was done on Intel®Core$^{\text{TM}}$ i5-8350U CPU at 1.8GHz, 8GB of RAM, and the Windows 10 Home Edition 64bit.

5. **Conclusion.** This study proposed methods to cluster trajectory data by using QuickBundles as the primary technique. The selected data was recorded by GPS-devices and stored on open-access services, which we collected using the defined criteria. The collected data was preprocessed to correct the missing values, outliers, and error sensor readings.

We classified the routes by performing the following 2 steps: 1) grouping routes into 3 groups by their total distance using k-Means; 2) classifying routes by their elevation profiles using QuickBundles. The clustering validity was measured by using the silhouette coefficient method. QuickBundles successfully clusters route data according to their elevation profiles. The results of the proposed techniques agreed with the expert's opinions.

These results show that QuickBundles is an effective method for retrieving route patterns from GPS stored data, and recommending them to cyclists.

The suggested further research can be to enrich route patterns with additional features, such as the route‘s turns and route’s terrain. These enriched route patterns can be used to propose tourism recommendation systems for different tourists who emphasize the routes by different routes’ characteristics.

Acknowledgement. This research was funded by the National Research Council of Thailand under Grant R2561B074. In addition, many thanks to Mr. Peter G. Barton of the Naresuan University Language Centre for his editing assistance and advice on English expression in this document.

REFERENCES

[1] J. Li, L. Xu, L. Tang, S. Wang and L. Li, Big data in tourism research: A literature review, *Tour. Manag.*, vol.68, pp.301-323, doi: 10.1016/j.tourman.2018.03.009, 2018.

[2] S. Du, H. Zhang, H. Xu, J. Yang and O. Tu, To make the travel healthier: A new tourism personalized route recommendation algorithm, *J. Ambient Intell. Humaniz. Comput.*, vol.10, no.9, pp.3551-3562, doi: 10.1007/s12652-018-1081-z, 2019.

[3] Z. A. Jiang, M. Wang and Y. Chen, Path recommendation based on geographic coordinates and trajectory data, *J. Commun.*, vol.38, no.5, pp.165-171, 2017.

[4] S. Malik and D. Kim, Optimal travel route recommendation mechanism based on neural networks and particle swarm optimization for efficient tourism using tourist vehicular data, *Sustainability*, vol.11, no.12, doi: 10.3390/su11123357, 2019.

[5] R. A. Hamid and M. S. Croock, A developed GPS trajectories data management system for predicting tourists’ POI, *Telkomnika*, vol.18, no.1, pp.124-132, 2020.

[6] F. M. Nardini, S. Orlando, R. Perego, A. Raffaetà, C. Renso and C. Silvestri, Analysing trajectories of mobile users: From data warehouses to recommender systems, in *A Comprehensive Guide through the Italian Database Research over the Last 25 Years*, S. Flesca, S. Greco, E. Masciari and D. Saccà (eds.), Cham, Springer International Publishing, 2018.

[7] E. Garyfallidis, M. Brett, M. M. Correia, G. B. Williams and I. Nimmo-Smith, QuickBundles, a method for tractography simplification, *Front. Neurosci.*, vol.6, doi: 10.3389/fnins.2012.00175, 2012.

[8] R. Lipe, *GPSBabel, Free Software for GPS Data Conversion and Transfer*, 2010.

[9] P. L. Houtekamer and H. L. Mitchell, Data assimilation using an ensemble Kalman filter technique, *Mon. Weather Rev.*, vol.126, no.3, pp.796-811, 1998.

[10] P. J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, *J. Comput. Appl. Math.*, vol.20, pp.53-65, doi: 10.1016/0377-0427(87)90125-7, 1987.