Towards the Identification of Semantic Points in Trajectories of Moving Objects with Weighted Averages

Jarbas Nunes Vidal-Filho[1,2], Valéria Cesário Times[1], Jugurta Lisboa-Filho[3]

[1] Informatics Center (CIn), Federal University of Pernambuco (UFPE) – Recife – Brazil

[2] LAPIS – Federal Institute of Ceara (IFCE) - Tabuleiro do Norte – Brazil

[3] Informatics Department (DPI), Federal University of Viçosa (UFV) – Viçosa – Brazil

jarbas.vidal@ifce.edu.br, vct@cin.ufpe.br, jugurta@ufv.br

Abstract. *In this paper, we explore a technique of clustering GPS points to: (1) extract a single point from each candidate cluster for stop point, named semantic point for this research, and (2) analyze similarities of semantic points identified in trajectories using three different algorithms. We propose a new algorithm based on a weighted average for the identification of the semantic point in the cluster - that which is the simplest, most efficient, and possesses the least computational cost when compared to other state-of-the-art solutions. We identified 1050 semantic points in trajectories of the Geolife project and compared the distances between them from the semantic points. The algorithm proposed was compared to the central point and K-Medoid algorithms. From the results, we concluded that the semantic points are at an acceptable distance from one another as defined by the literature.*

1. Introduction

The discovery of stop points in raw trajectories of moving objects has been an important research topic for data mining [Lehmann et al. 2019; Fu et al. 2016]. Due to great heterogeneity, lack of accuracy, and differing sampling levels in the collection of trajectories, one encounters some uncertainty in the detection of stop points (that is to say, the place visited by the user) in raw trajectories [Lehmann et al. 2019; Furtado et al. 2018]. Uncertainty in the identification of stop points causes difficulties in the association of semantic information (i.e., information about the place visited) with semantic points, inference of activities performed by the object in movement, discovery of patterns, among others. In this work, we define the semantic point as a spatial-temporal point representing the physical stop point in the cluster (i.e., the clustering of points seen as a candidate for the stop point). The semantic point is identified after the formation of clusters candidate stop points in trajectories.

The mobile individual has their localization (longitude/latitude) registered over time, represented by a sequence of spatial-temporal points. This is also known as the raw trajectories of moving objects [Furtado et al. 2018]. From the raw trajectory, it is possible to extract diverse parameters such as, for example, velocity and direction of the moving object (i.e., semantic information). With the use of these parameters, it is possible to identify the similarity between semantic points by way of diverse techniques such as, for example, classification, clustering, and matching patterns. For [Furletti et al. 2013], even though trajectories are the representation of stop points and movements

(i.e., the sequence of spatial-temporal points between stop points), it is important to identify stop points as the place visited by the individual.

Stop points are normally represented by semantic points or by a geographical region. The literature reports few works that approach the identification of semantic points in clusters, principally due to the uncertainties that exist in current approaches. Let us imagine that a determined method of semantic annotation uses the approach of associating the closest place to visit concerning the semantic point. For the same cluster, the location of a semantic point returned by different methods will hardly be the same, resulting in uncertainties in the definition of the place visited. Therefore, the uncertainties of current methods for returning semantic points can give occasion to flaws in the inference of the place visited, in the inference of activities performed by the moving object, amongst other things.

The main goal of this paper is to propose a new algorithm based on weighted average to infer semantic points. Besides this, the paper seeks to discuss, analyze, and compare identification methods for semantic points in raw trajectories. The motivation to investigate new approaches for identification of semantic points took place because of disadvantages in the traditional methods, highlighted by the literature - for example, the computational cost of iterations and the necessity for defining values of parameters of entry for the choice the semantic point. The approaches found in the literature initially require the definition of the number of clusters, the number of interactions, and the number of medoids (among other parameters) to realize the processing of algorithms. Besides this, the results depend on a good calibration of these parameters.

The approaches available identify the semantic point based on the central point method and the K-Medoid and K-Means algorithms. The problems with these approaches are: (1) the central point will always be a point which does not belong to the clustering; (2) the centroids returned by K-Means will hardly belong to the cluster; (3) these algorithms require entry parameters such as, for example, number of clusters and probable random semantic points; and (4) K-Medoid and K-Means have computational cost with iterations and return semantic points closer to the center. According to [Steinbach et al. 2005], the results of the K-Medoid and K-means algorithms depend on the calibration of initialization parameters. However, undue calibration causes more uncertainty in the definition of the semantic point.

We propose an algorithm based on a weighted average that seeks to infer the semantic points closest to the points of lowest velocity, has a lower computational cost, and always chooses a semantic point belonging to a candidate cluster the stop point. We compared the similarity based on the distance between the semantic points returned by the weighted average, central point and K-Medoid algorithms, in order to identify the proximity between the semantic points returned by these approaches. By way of this comparative analysis, we identify that the choice of the semantic point also affects the services of the inference of activities.

The rest of this paper is organized into four sections. Section 2 describes the works related to the problem of identification of stop points. Section 3 presents the contributions of the paper. Section 4 exhibits a comparative evaluation between the method proposed and the central point and K-Medoid methods. In Section 5, conclusions and possible future studies are brought up to discussion.

2. Theoretical Foundation

Before delineating the objectives of this paper, we present the most important basic concepts to understand this work and some works about the identification of stop points.

2.1. Basic concepts

The process of semantic enrichment has the finality of semantically annotating raw trajectories with information on the place visited or the activity that the user performs during movement [Fu et al. 2016]. The moving object can be represented by a single localization in the region that is a candidate for the stop point (i.e., semantic point), which would then be used to label the place visited by the user. Besides the definition of the process of semantic enrichment, the definitions of raw trajectory, stop point and clusters are important for understanding the work proposed in this paper, and as such, they are listed here.

Definition 1 (*Raw trajectory*): is a sequence $<p_1, p_2, ..., p_n>$ of points p = ((x, y), t), where ($\underline{x}$, $\underline{y}$) is the localization of the object and $\underline{t}$ is the moment of collection.

Definition 2 (*Stop point*): is the place visited by a user during time interval T = ($t_{(start)}, t_{(end)}$), represented by a sequence of GPS points belonging to time interval T.

Definition 3 (*Clusters*): is a subset of the raw trajectory formed by points that possess similarity with each other based on a certain trajectory collection parameter.

2.2. Related work

K-Means is a clustering algorithm that receives a predefined number of clusters to be formed and randomly selects centroids iteratively to be grouped into clusters [Zhou et al. 2007]. The algorithm iterates by way of the centroid and the rest of the points in the cluster, calculating the distance between all the points, computing the centroid and attributing each point to the centroid of least distance. Finally, the centroid of each cluster is found by way of the average of all the instances associated with the cluster. The K-Medoid is similar to K-Means, however, instead of choosing the centroid that never corresponds to a true data point, it randomly selects medoids (i.e., points of clusters with the best computational cost) [Steinbach et al. 2005]. The K-Medoid receives, as parameters, the number of clusters to be formed, the number of medoids and the instances of the cluster to calculate the cost function. The problem of the K-Medoid is found in the definitions of values for the entry parameters which can generate less optimized medoids and computational cost with iterations. The K-Medoid has greater relevance for choosing a medoid point that belongs to the cluster and is more robust in the face of noise and outliers.

[Kang et al. 2005] defined a clustering approach based on time where the user stays stopped in one place. If the distance between the starting point and the finishing point of the cluster is less than a determined threshold, new clusters are formed. Smaller clusters, having less time in their formation, are discarded. Finally, if the centroid of the cluster is at a certain distance from an existing POI, this approach merges the centroid with the POI to represent the semantic point.

[Fu et al. 2016] proposed an approach to clustering based on two stages, considering the similarity of values for the parameters of time and distance. The algorithm initially groups points based on the time of stay. Following this, it verifies the

distance between the points to reduce problems with the loss of signal. Finally, the algorithm does a scan to identify the points' peaks of density and extract stop points. [Zhou et al. 2017] presented a genetic clustering algorithm based on density and K-Means variation that does not need to inform the number of clusters. Finally, this variation in the K-Means uses quality indicators to reduce the size of the clusters generated.

The related works focus on improvements in the existing clustering techniques, taking time, speed, direction, density and distance thresholds into consideration to minimize the problems with the collection of trajectories or loss of GPS signal. We perceive that few works focus on the identification of semantic points in clusters, and those that do so are based on centroids or medoids. Therefore, little is yet discussed about semantic points, being valid the investigation and implementation of new approaches that use semantic points to improve the performance of localization applications.

3. The Weighted Average Algorithm

GPS data represented by movements and stop points with semantic information are called semantic trajectories. Recently, the literature has been expanding upon the concept of semantic trajectories to multiple aspect trajectories. According to [Petry et al. 2019], this new type of trajectory has as its objective to receive semantic annotations from different sources and formats such as, for example, places visited by the user [Furletti et al. 2014] and data from sensors and social networks. The objective is semantically enriching the raw trajectory gathered in real-time. We identified that semantic and multiple aspects trajectories constantly use real-time resources from different Application Programming Interface (API). APIs supply information on places, health, climate, social networks, and other Web services. In this context, the quality of semantic addition to the stop point is related to the identification of the semantic point.

The algorithm proposed in this work has as its objective the contribution to the identification of semantic points in clusters using weighted average as a way of improving the inference of activities and semantic annotations of trajectories. Initially, to understand the proposal in a better manner, we formalize the concept of semantic point as the contribution of our study. Following this, we present the weighted average algorithm as the main contribution of this work.

Definition 4 (*Semantic point*): is a tuple $((x, y), t, as)$, where $(\underline{x}, \underline{y})$ are the geographical coordinates, $\underline{t}$ is the time register and $\underline{as}$ the semantic annotation, representing the localization of a single point within a candidate cluster for a stop point and its semantic information. The semantic point can be obtained from the algorithm proposed for this paper, Algorithm 1.

Algorithm 1 receives as its entry parameter a list of clusters that can be defined automatically by way of clustering algorithms and based on some kind of similarity criteria. From this point on, the central idea of the algorithm is to prioritize points belonging to the cluster that have low speed. Therefore, the algorithm will check the instantaneous speed of each point belonging to the cluster. We understand that, the user using a means of transport tends to reduce his speed during the stopping process, consequently, being able to arrive until reaching a complete halt - approximately, velocity null.

Algorithm 1. Identification of semantic point based on the weighted average

```
List<Stop> StopsDiscoveringByWeightedAverage (clusters)

Input: clusters: clusters list
Output: SemanticPoint
1: stops = new List <Stop>;
2: FOR EACH cluster IN clusters DO
//sort trackpoints by instantaneous velocity
3:              ordered = new List<TrackPoint>
4:              Collections.sort(ordered);
// weighted average calculation of the latitude and longitude of each cluster
5: FOR (i = 0, weight = ordered.size(); i < ordered.size(); i++, weight--) DO
6:               totalWeight += weight;
7:               totalLatitude += weight*ordered.get(i).getCoordinate().getLat();
8:               totalLongitude += weight*ordered.get(i).getCoordinate().getLng();
9:              totalTime += weight*ordered.get(i).getInstant().getTime();
10:              stopLatitude = totalLatitude/totalWeight;
11:              stopLongitude = totalLongitude/totalWeight;
// the value totalTime/totalWeight corresponds to the instant in millis
12:              instant = new Date((totalTime/totalWeight));
13:              stopCoordinate = new Coordinate (stopLatitude, stopLongitude);
// stop definition
14:              stop = new Stop (stopCoordinate, instant, cluster);
15:              stops.add(stop);
16:              cluster.setStop(stop);
17: return stops;
```

To give priority to the points of low velocity, we use the concept weighted arithmetic average to attribute weights to the points belonging to the clusters. In this way, the points of low velocity receive larger weightings while high-velocity points receive lower weightings. After receiving a list of clusters, the algorithm orders the points based on the velocity of each candidate cluster for a stop point (lines 2 – 4). Having done this, the algorithm calculates the weighted average of latitude, longitude, time, and instant of the points of each cluster (lines 5 – 12). The weights used are defined based on the number of points that exist in the cluster. For example, if the cluster possesses 60 points the weights attributed are numbered 0 to 59. In this way, points at low velocity will be given priority in the generation of the semantic point. Finally, the algorithm defines the localization of the semantic point as the coordinate of the stop point (line 13), instantiates and returns a new semantic point related to the cluster (lines 14 – 17). The localization of the returned semantic point is utilized as a parameter to access and seek API services, as well as semantic information, to enrichen trajectories.

4. Experiments and Evaluation

In this section, we present two types of similarity analysis between semantic points of trajectories. The first considers the similarity of the distance parameter between points and the second approaches the inference of activities between semantic points. For all algorithms used in the experimental analysis of this work, the CB-SMoT algorithm proposed by [Palma et al. 2008] was utilized for the formation of clusters, because it uses similarity based on time and speed to form clusters.

The simulations occurred from the catalog of services from SDI4Trajectory (www.sdi4trajectory.ifce.edu.br), and the parameters defined for CB-SMoT were *stop time = 300 s, Speed limit = 4 m/s,* and *Average speed = 3 m/s,* chosen for the definitions utilized by [Furletti et al. 2013]. The algorithms compared in this study are: weighted average proposed for this study, central point and K-Medoid [Steinbach et al. 2005]. The choice of algorithms is based on some reasons, such as: (1) The literature cites works using methods based on medoids and centroids [Zhou et al. 2017; Kang et al. 2005]; (2) The K-Medoid is frequently discussed in the literature because it selects

semantic points belonging to the points of the clusters [Steinbach et al. 2005]; (3) The central point is a simple approach that does not select semantic points belonging to the cluster and always uses the center of the cluster to represent the semantic point. The definition of the number of clusters for K-Medoid is done after the execution of the CB-SMoT algorithm, not manually as that which frequently occurs.

4.1. Data Acquisition

To validate and verify the efficiency of the proposed algorithm, we used two sets of data. Geolife is a project proposed by Microsoft Research Asia that collected 17,621 raw trajectories by 178 users between 2008 and 2012. The data from this project are organized into folders, where each pasta represents a determined user. Some folders possess a text archive informing the period of collection of the trajectory and the transport method used. The data represent routine activities of users, for example, in the same folder there are trajectories of a user who used the train between 10 a.m. and 11 a.m., and then walked on foot between 11 a.m. and 11:30 a.m., on different days. These trajectories can correspond to the activity of commuting to work. Therefore, among the folders that have labels, we chose 15 to perform a comparative analysis of similarity based on the parameter of distance. Table 1 exhibits the distribution of data analyzed by means of transport, clusters and identified semantic points.

Table 1. Set of experimental data from the Geolife project

Geolife Dataset	Number of Trajectories	Number of Clusters	Number of Semantic Points
Car	55	95	95
Taxi	47	202	202
Subway	20	90	90
Train	33	200	200
Bike	201	345	345
Walking	53	118	118

We also analyzed the similarity between points based on the inference of activities. To do this, we used the second set of data collected by 7 volunteer users for 1 month. Of the 7 users, 5 were using a smartphone with the My Tracks application and 2 were using bicycles and carrying the TomTom watch for data collection. All users involved in the experiment were lawyers, teachers, and amateur cycling athletes. They collected the trajectories on foot, bicycle, and by car. Table 2 exhibits the distribution of the set of data gained by the volunteers.

Table 2. Set of experimental data collected by volunteers

Voluntary dataset	Number of Trajectories	Number of Users	Number of Clusters	Number of Semantic Points
Bike	2	2	2	2
Car	20	4	26	26
Walking	3	1	5	5

The quantitative clusters and semantic points presented in Tables 1 and 2 were the same for all the algorithms used in the analyses of similarity. In total, 409 trajectories of different users were analyzed and 1,050 semantic points for the set of Geolife data. For the voluntary data set, 33 semantic points were analyzed for the inference of activities. Even though being a limited set of data, the idea using the second

set of data is to investigate if the inferences of activities that were returned can be related to the identification algorithm of the chosen semantic point and the means of transport used by the moving object. In this way, in exploring the manual semantic annotations and inferences of activities, it is possible to gain results that aid in deciding upon the methods of identification of semantic points to be used, based on the means of travel.

4.2. Analysis of similarity between semantic points based on the parameter of distance

According to [Furletti et al. 2013], a distance covered by a user between the stop point and the place visited can be flexible up to 500m. For a candidate cluster for a stop point being formed within the region of a shopping center, a semantic point can be identified in the parking lot or the establishment visited by the user. In the face of this, we consider that 500m is an acceptable distance to indicate similarity between semantic points. For [Smith and Butcher 2008], in the different types of traveled environments by a user, the distance between stop points can vary between 300m to 500m. In this section, we initially analyze the distance between semantic points returned by the weighted average, central point, and K-medoid algorithms.

Besides this, we assume that the K-Medoid algorithm should be used as the baseline in the analysis of similarity. This is because K-Medoid has already undergone many experiments and is well defined in the literature [Velmurugan et al. 2010; Arora et al. 2016], becoming the object of study through a diversity of works that explore clustering algorithms [Steinbach et al. 2005]. Therefore, the closer a semantic point identified by another method is to the semantic point returned by K-Medoid, it is considered that there is a similarity between the methods. Table 3 shows the number of semantic points - returned by weighted average and central point algorithms – that were closest to the semantic points returned by K-Medoid.

Table 3. Quantity of semantic points closest to K-Medoid

Method/Transport	CAR	TAXI	SUBWAY	TRAIN	BIKE	WALKING
Weighted average	43	82	44	72	148	56
Central point	51	117	46	111	182	60

We used the Google Earth tool to measure the distance between semantic points *in loco*. Of the 1050 semantic points analyzed from the Geolife data, we identified that 3.62% (38 semantic points) are practically in the same localization, so we assumed that they were at the same point. Of the other 96.38% (1012 semantic points), it was verified that the central point algorithm presented 54% (567 semantic points) closer to K-Medoid, while the weighted average algorithm presented only 42.38% (445 semantic points). The *in loco* analysis verified that the K-Medoid identifies medoids closer to the center of the clusters which can justify a greater quantity of semantic points identified by the central point. This is probably associated with its similarity to the K-Means, which can identify a medium point in the cluster.

Besides this, we explore in greater detail the distances between the semantic points returned by the three algorithms under study. Firstly, we compared the distance between the semantic points of the K-Medoid and the central point. The results can be seen in Table 4, which presents the number of semantic points by ranges of distance.

The idea is to be able to understand how these points by ranges of distance are distributed, seeing that the literature defines parameters of distance that an object can travel after a stop. For [Smith and Butcher 2008], climate and time also influence the activity of a mobile object. We consider the means of transport and the distance between semantic points as parameters that can influence the inference of activities, due to the variations in localizations of semantic points.

Table 4. Quantity of semantic points by ranges of distances (Central Point vs K-Medoid)

Distance/ Transport	CAR	TAXI	SUBWAY	TRAIN	BIKE	WALKING
0 – 10 m	10	72	29	84	117	30
10 – 50 m	37	65	27	55	115	34
50 – 100 m	15	31	15	31	52	32
100 – 500 m	23	31	17	24	59	19
above 500 m	6	3	2	6	2	3

We know that shopping centers possess various physical locations with distances that exceed 500m. Therefore, if the user went to the shopping center, whatever semantic point in the region of the shopping center is considered a hit in the process of identification of stop points. In this context, we believe that the distance between semantic points can vary up to 500m. Therefore, the closer they are, the more similar they can be considered. In Table 4 only 22 semantic points exist which are at a distance greater than 500m. Approximately 97.6% of the semantic points within an acceptable distance, defined as the distance between the stop point and the place visited by the user. The problem with the central point is the fact of not selecting the points of clusters. This causes uncertainties in the definition of activities performed at the stopping point of the moving object, due to the points tend to be distant from the points of interest.

In Table 5, we also compare distances between semantic points returned by the weighted average and K-Medoid algorithms. The idea is also to analyze how the semantic points are distributed by ranges of distance. Table 5 exhibits the quantitative view of semantic points by ranges of distance concerning the K-Medoid. One can verify that only 20 semantic points are at a distance superior to 500m. Therefore, 98.1% of the semantic points are at an acceptable distance that consists of a distance between the stop point and the place visited by the user. This means saying that the algorithms possess similarity concerning the definition of semantic points and considering the parameter of distance. The weighted average algorithm becomes important in the process of identification of stop points represented by semantic points. The algorithm prioritizes points of low velocity and that belongs to the candidate cluster for the stop point and possesses similarity to the K-Medoid.

Table 5. Quantity of semantic points by ranges of distances (Weighted Average vs K-Medoid)

Distance/ Transport	CAR	TAXI	SUBWAY	TRAIN	BIKE	WALKING
0 – 10 m	9	47	16	64	96	21
10 – 50 m	31	76	35	75	114	43
50 – 100 m	27	41	14	33	68	25
100 – 500 m	22	35	23	22	65	28
above 500 m	6	3	2	6	2	1

Finally, we consider the distances between the semantic points returned by the proposed algorithm and the central point, since Table 3 presented a greater quantity of semantic points extracted by the central point method and that are closer to K-Medoid. The medoids tend to be closer to the center of the clusters, principally due to similarity to K-Means. However, we use the weighted average algorithms and the central point to identify the similarity between semantic points based on the parameter of distance. The idea is to construct ranges of distance to investigate how the semantic points of Table 3 are distributed.

For example, based on Table 3 and considering the Train and Bike means of transport, the central point method presented, respectively, 39 and 34 semantic points more than the proposed algorithm and listed semantic points closer to K-Medoid. In Table 6 shows the number of semantic points by distance ranges for the proposed method and central point. However, for the Train as a mode of transport, only 5 semantic points are at a distance greater to 500m, while the Bike as a mode of transport did not present any semantic point with a distance superior to 500m. In Table 6 one perceives that approximately 99% of the semantic points returned by the weighted average algorithm are at a distance of up to 500m from the central point, in conformity with the acceptable distance between semantic points adopted by this work.

Table 6. Quantity of semantic points by ranges of distances (Weighted Average vs Central Point)

Distance/ Transport	CAR	TAXI	SUBWAY	TRAIN	BIKE	WALKING
0 – 10 m	21	60	17	89	103	34
10 – 50 m	36	90	41	67	177	44
50 – 100 m	18	30	8	24	44	27
100 – 500 m	18	20	22	15	21	13
above 500 m	2	2	2	5	0	0

From Tables 4, 5 and 6 we verify that more than half of the semantic points are at a distance of 0 – 50m between each other. Increasing the distance buffer between points, we identify that 99% of the semantic points analyzed are at a distance of up to 500m. [Yang et al. 2012; Smith and Butcher 2008] affirm that the distance covered on foot is related to activity, use of transport, recreation, and other things. The authors affirm that the acceptable measures of distance acceptable for a determined user who needs to stop and then walk to the place to be visited can vary from 300m to 1.5km, depending on the objective. [Millward et al. 2013] defined time and distance values for users who go on foot. Thus, we identified that 500m is an acceptable value for a user to go on foot.

The proposed algorithm possesses similarity when compared to the approaches defined in the literature when the parameter of the distance between semantic points is taken into consideration. The analyses based on ranges in distance help in the validation of similarity between methods. The proposal presented uses the concept of weighted average. It is based on the parameter of velocity to give priority to points in the cluster, is easy to implement, it always chooses low-velocity points from the cluster, and the semantic points that are returned tend to be closer to the places visited.

4.3. Analysis of similarity between semantic points based on the inference of activities

In this section, we use the data set of volunteers. The maximum distance covered after a stop was defined by basing ourselves on the work of [Yang et al. 2012; Smith and Butcher 2008]: 500m for users who are in movement by foot or by bicycle, and for those using a car, 400m. We used the K-Medoid algorithm as the baseline.

For the set of data utilized, each user manually noted down the set of places visited during the collection of the trajectory. Table 7 presents for each one of the methods discussed, the ID Track and the number of semantic points identified by trajectory, followed by the activity and the probability of the activity occurring. We instantiated the definition of the gravitational model utilized by [Furletti et al. 2013], which uses the concept of attraction of bodies. This model returns the probability of the occurrence of a determined activity associated with the stop point.

As discussed in the previous section, 96.38% of the semantic points analyzed from the Geolife project presented variations in localizations. Under these circumstances, the objective of investigating similarity based on the inference of activities is due to the algorithms discussed returning semantic points with different localizations, causing uncertainties in the inference of activities performed by the moving object. Therefore, we explore the similarity between semantic points based on the inference of activities, also seeking to identify the correlation between the methods of identification of semantic points and the means of transport used by the moving object. In a general way, for the model used by [Furletti et al. 2013], if a stop point is closer to a set with a greater number of places visited from the same category, the probability that the activity is the same in this set is greater.

Table 7. Comparing similarities between semantic points based on the inference of activities.

ID TRACK	Semantic Point	Weighted Average Activity (%)	Central Point Activity (%)	K-Medoid Activity (%)
1	1	SERVICES: 62.49	SERVICES: 61.18	SERVICES: 54.60
	2	OTHERS: 88.36	SHOPPING: 99.98	SERVICES: 77.49
2	1	SHOPPING: 40.19	SHOPPING: 28.02	SHOPPING: 39.32
3	1	FOOD: 18.23	FOOD: 25.75	FOOD: 40.00
	2	OTHERS: 26.28	OTHERS: 54.33	OTHERS: 52.08
4	1	LEISURE: 72.50	LEISURE: 72.12	LEISURE: 72.50
5	1	OTHERS: 86.82	OTHERS: 42.99	SHOPPING: 75.32
6	1	SERVICES: 18.54	SERVICES: 15.43	SERVICES: 20.08
	2	SERVICES: 31.67	OTHERS: 30.34	SERVICES: 32.84
	3	SERVICES: 70.50	SHOPPING: 43.35	SERVICES: 89.42
7	1	OTHERS: 62.79	OTHERS: 56.59	OTHERS: 57.44
8	1	OTHERS: 59.05	OTHERS: 58.62	OTHERS: 55.17
9	1	SERVICES: 49.58	OTHERS: 44.98	OTHERS: 30.97
10	1	OTHERS: 62.53	OTHERS: 68.35	OTHERS: 59.42
11	1	SERVICES: 49.43	SERVICES: 51.25	SERVICES: 49.61
	2	OTHERS: 40.57	OTHERS: 39.56	OTHERS: 40.27
12	1	OTHERS: 45.13	OTHERS: 45.13	OTHERS: 47.98

continues on the next page

13	1	OTHERS: 31.84	OTHERS: 41.9	OTHERS: 28.89
	2	SERVICES: 3.40	OTHERS: 10.02	OTHERS: 33.98
14	1	SERVICES: 58.53	SERVICES: 82.37	SERVICES: 41.19
15	1	SHOPPING: 52.44	SHOPPING: 55.86	SHOPPING: 47.76
16	1	SHOPPING: 42.13	SHOPPING: 64.73	SHOPPING: 64.03
17	1	SERVICES: 42.77	SERVICES: 28.04	SERVICES: 62.13
18	1	OTHERS": 73.77	OTHERS": 58.41	OTHERS": 95.75
19	1	SERVICES: 99.91	SERVICES: 49.42	SERVICES: 17.45
20	1	OTHERS: 58.08	OTHERS: 45.20	OTHERS: 45.20
21	1	SERVICES: 29.65	SERVICES: 23.10	SERVICES: 14.38
	2	FOOD": 26.38	FOOD": 28.26	FOOD": 28.35
22	1	SHOPPING: 40.60	SERVICES: 39.30	SHOPPING: 72.76
	2	FOOD: 49.25	FOOD: 48.76	FOOD: 49.22
23	1	OTHERS: 50.40	SERVICES: 70.34	OTHERS: 55.23
24	1	OTHERS: 27.47	SERVICES: 97.32	SERVICES: 64.17
25	1	SERVICES: 36.87	SERVICES: 24.36	SERVICES: 51.46

In Table 7, of the 25 trajectories analyzed, 33 semantic points were identified. These were compared to each other concerning inference and the probability of the activity occurring. From the semantic points analyzed, only for the trajectory of Id 12 was there inference of the same activity and the same probability for the central point and weighted average. Another 4 semantic points analyzed had different activities for the three methods. For example, the semantic point for Id 02 (Id 01 track) presented a different activity for each method discussed.

As a way to improve the inferences of our results, we consider the category of the place visited by the user and the manual semantic annotation. The objective is to know the activity executed by the user and then compare it to the return of the activity done in the methods. Of the 4 semantic points analyzed with different activities, the K-medoid algorithm hit the activity of a semantic point. Analyzing the 28 remaining semantic points, 64.29% of the points presented greater similarity for inference and probability of activities between the weighted average and K-Medoid, while only 35.71% presented better similarity between the central point and K-Medoid. This leads us to define that the weighted average algorithm presents itself as a viable solution for the context of identification of semantic points, for the second set of examined data.

5. Conclusions

The comparisons between the proposed algorithm and the central point and K-Medoid methods present similarity based on the parameter of distance for the semantic points analyzed. Approximately 85% of the semantic points are at an acceptable distance within 500m. One also perceives that the inference of activities for returned semantic points by the proposed algorithm has better precision since it possesses the biggest hits concerning K-Medoid and better correspondence with manual semantic annotation. Therefore, the weighted average algorithm becomes an interesting approach for the identification of semantic points in trajectories.

It is also visible that the inferences of activities are sensitive to variations in localization of semantic points. These often tend to localize themselves closer to a set of Points of Interest with an activity that is different from that noted down manually. The

variations in the inference of activity bring to the fore the importance of identifying semantic points. We identify the better similarity between the proposed method and K-Medoid, mainly for when the car is transport. However, it requires better investigations.

For future works, it is planned to amplify the analysis of the similarity between semantic points based on the inference of activities, utilizing more voluminous data such as those returned by Geolife and exploring new modes of transport used.

References

Arora, P.; Varshney, S. (2016). Analysis of k-means and k-medoids algorithm for big data. *Procedia Computer Science*, 78, 507-512.

Fu, Z., et al. (2016). A two-step clustering approach to extract locations from individual GPS trajectory data. *ISPRS International Journal of Geo-Information*, *5*(10), 166.

Furletti, B., et al. (2013). Inferring human activities from GPS tracks. In *Proceedings of the 2nd ACM SIGKDD international workshop on urban computing* (pp. 1-8).

Furtado, A. S., et al. (2018). Unveiling movement uncertainty for robust trajectory similarity analysis. *Int. Journal of Geographical Information Science*, *32*(1),140-168.

Kang, J. H., et al. (2005). Extracting places from traces of locations. *ACM SIGMOBILE Mobile Computing and Communications Review*, *9*(3), 58-68.

Lehmann, A. L., et al. (2019). SMSM: a similarity measure for trajectory stops and moves. *Int. Journal of Geographical Information Science*, *33*(9), 1847-1872.

Millward, H.; Spinney, J.; Scott, D. (2013). Active-transport walking behavior: destinations, durations, distances. *Journal of Transport Geography*, *28*, 101-110.

Palma, A. T., et al. (2008). A clustering-based approach for discovering interesting places in trajectories. In *Proceedings of the ACM Symposium on Applied Computing*, pages 863-868.

Petry, L. M., et al. (2019). Towards semantic-aware multiple-aspect trajectory similarity measuring. *Transactions in GIS*, *23*(5), 960-975.

Smith, M. S., et al. (2008). How far should parkers have to walk?. *Parking*, *47*(4).

Steinbach, M.; Kumar, V.; Tan, P. (2005). Cluster analysis: basic concepts and algorithms. *Introduction to data mining, 1st ed. Pearson Addison Wesley*.

Velmurugan, T.; Santhanam, T. (2010). Computational complexity between K-means and K-medoids clustering algorithms for normal and uniform distributions of data points. *Journal of Computer Science*, 6(3), 363-368.

Yang, Y.; Diez-Roux, A. V. (2012). Walking distance by trip purpose and population subgroups. *American Journal of Preventive Medicine*, *43*(1), 11-19.

Zhou, C., et al. (2007). Discovering personally meaningful places: An interactive clustering approach. *ACM Transactions on Information Systems*, *25*(3), 12.

Zhou, X., et al. (2017). An automatic K-Means clustering algorithm of GPS data combining a novel niche genetic algorithm with noise and density. *ISPRS Int. Journal of Geo-Information*, *6*(12), 392.