A Novel Method for Smoothing Raw GPS Data with Low Cost and High Reliability

(Invited Paper)

Xun Zhou[1], Changle Li[1,2,*], Xiaoming Yuan[1], Bing Xia[1], Guoqiang Mao[3], and Lei Xiong[2]
[1]State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an, Shaanxi, 710071 China
[2]State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, 100044 China
[3]School of Computing and Communications, University of Technology Sydney, NSW, 2007 Australia
*clli@mail.xidian.edu.cn

***Abstract*—The precise spatio-temporal position data of vehicles is useful for most studies, such as wireless link lifetime and node degree in vehicular ad hoc networks. However, due to the system errors and random errors, the existing Global Positioning System (GPS) only provides the positional accuracy about 10m or even worse. In this paper, to address the issue of positional accuracy, a Clustering and Approximating (C-A) algorithm is proposed. We first divide each road into several small parts which are described by linear functions. Then a linear regression algorithm is utilized to approximate traces under system errors, which is reliable for reducing GPS errors. Particularly, when two roads are very close, GPS points may be mapped on adjacent roads. A clustering algorithm is taken to separate GPS points and their positions are revised by the iterative utilization of the linear regression algorithm. In the end, the method mentioned above smoothes raw GPS data of buses in Taiwan to make it available for further researches. Compared with existing methods, the method described in this paper characterized with low cost and high reliability in different situations. Besides, its simple model will make the process of revising data more convenient.**

***Index Terms*—Global Positioning System, raw GPS data, Clustering, Approximating.**

I. Introduction

GPS has been widely used in the traffic system. Most cars are equipped with GPS receivers in order to obtain relevant real-time traffic data while driving [1]. Based on GPS data, more information can be acquired, such as node degree [2] and the connectivity of vehicle network [3]. However, the majority of research teams can only gain raw GPS data which has a positional accuracy of 10 meters in normal conditions. The major reason is about ionospheric refraction, multipath effect and the inner noise of receivers [4]. Therefore, vehicles cannot be mapped on the electronic map accurately. For instance, vehicles are sometimes on adjacent roads, or even out of roads. In order to make raw data available, it is necessary to revise errors of it.

Currently, Differential Global Positioning System (DGPS) is the most advanced system which increases the accuracy of GPS data to 1m [5]. By establishing GPS fiducial stations and broadcasting correction factors to specific receivers, DGPS can revise GPS errors by themselves. Taking use of the correlation among GPS points can also adjust their positions. In the paper [6], correlation is indicated by potential energy wells. When the first point is fixed, the position of the second one can be revised. In order to simplify the model, the spread of GPS points is considered with the Gaussian distribution [7]. Meanwhile, data processing is another important way to reduce errors and Kalman filtering can acquire better result [8].

Although some excellent works have been done to reduce GPS errors, two main factors can influence the efficiency of these methods, namely, constraints of geographical conditions and human factors. Human factors such as the height of buildings in different areas and human behavior can lead to different levels of multipath effect, shadow effect and noise for GPS signal. Therefore the methods based on distribution functions cannot always be reliable. Some other methods, such as DGPS are cooperated with fiducial stations which are restricted by geographical conditions can also increase the cost. In order to improve the efficiency about error correction, the more simple algorithm is still required urgently, which is supposed to be reliable with low cost in different situations.

In this paper, we propose a novel method named C-A which is based on the knowledge about linear regression and K-means algorithm to conquer difficulties mentioned above. We first establish rectangular coordinate system to describe roads with linear functions in a small range. Then, a linear regression algorithm is utilized to extract traces of buses from GPS points. If GPS data is collected from two close roads, it should be separated by the clustering algorithm at first. Finally, by combining with the information about roads, GPS points are assigned to the correct lanes. Comparing with methods mentioned above, we make use of the fact the behavior of moving objects are restricted by objective conditions, such as space and speed [9] to establish a linear model, which is more simple. Without depending on special devices or distribution functions, our method cannot be influenced by constraints of geographical conditions and human factors in different situations.

The rest of this paper is organized as follows. Section II introduces relevant significant works. Then Section III displays the process of data preprocessing and error analysis. In Section IV, we introduce the C-A algorithm in detail. Furthermore, in Section V, simulation results are discussed. Finally, Section VI concludes this paper.

978-1-5090-1701-0/16/$31.00 ©2016 IEEE

II. Related Works

In the literature of moving object databases, spatio-temporal positions of moving objects are often considered to be precise. In addition, most projects, such as trajectory query processing and indexing techniques, are processed on the assumption. However, because of the measurement and sampling errors, datasets collected by mobile sensors and GPS are usually inaccurate [10]. In order to meet the demand of the quality of datasets, data cleaning cannot be overlooked.

The spread of trace nodes for a road is usually modeled as a Gaussian distribution [11] and the width of a road is related to the standard deviation of the Gaussian distribution [7]. According to the real data, the distribution can be modeled and the number of lanes can be estimated by the standard deviation. Correlation among GPS points is also discussed to revise errors and Cao et al. [6] use a novel aggregation technique which adjusts positions of GPS points in response to simulate potential energy wells created around each trace.

Kalman filtering [8] is a common and available method to smooth data, and sometimes, it results in the least difference between the true movement and the representation. In fact, the behavior of moving objects are restricted by objective conditions, such as space and speed [9]. For instance, vehicles move on the road network and cannot be driven at a limitless speed . If these conditions are fully used, datasets also can be smoothed by simple method.

On the other hand, improving the quality of the original data is an available way, for example, DGPS [12] uses a network of fixed ground-based reference stations to broadcast the difference about pseudoranges between the measured positions and real positions. Receivers can correct their pseudoranges based on the difference they get.

III. Data Preprocessing

In this paper, the GPS data contains 3000 buses in several cities. Buses upload data every 20 seconds, including the main information which is full used in this paper, such as speed, direction, position and time stamps of buses.

A. Electronic Map

Google Earth is a free electronic map. The GPS data can be imported in it and GPS points can be shown on the map. Information such as the width of roads and distance between any two points can also be known from it. In this paper, we import GPS data in Google Earth and correct the error on it.

B. Extracting Data

Datasets are preprocessed before doing the integration. Firstly, we choose the related data which owns the specific attributes. In this paper, the data of a bus is selected randomly. Then, the selected data is converted to the image based on the information of position. Finally, the image data can be imported in the Google Earth and GPS points of buses can be shown on the map. In Fig. 1, we display GPS points of the bus XB-048.

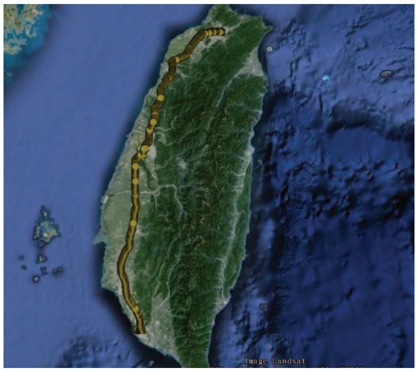


Fig. 1: Distribution about GPS points of the bus XB-048 in April 1st

C. Error Analysis

We mainly analyze errors of position information. Because of external interference, such as ionospheric refraction, multipath effect and the inner noise of the receivers, longitude and latitude of GPS data do not coincide with the actual data. External interference contributes same errors to every receiver within certain distance range. It leads all GPS points, which are within certain distance range, moving towards the same direction for about 10 meters. It will provide system errors.

Inner noise of the receivers contributes to GPS points distributed randomly around real traces and errors caused by it is named random errors. In the paper [6], the expectation of random errors is 0 and GPS points are symmetrically distributed around their traces. Errors of GPS data and the expectation of random errors can be shown as follows:

$$\theta = N + \alpha \tag{1}$$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \Delta y_i = 0 \tag{2}$$

In general condition, θ denotes the value of GPS errors and it is about 10 meters. N denotes the value of system errors and it is a constant about 10 meters. α denotes the value of random errors and it is a random variable with the expectation of zero. Δy_i denotes the value of random errors of the point i. Because the traces can be seen as the straight line reasonably in a small range, system errors lead traces to move towards one direction about 10m and GPS points are distributed around them with the distance from 0 to 3m resulted from random errors.

IV. Revised Method

This paper proposes a novel C-A method to smooth raw GPS data. Firstly, we judge the number of roads which our data is collected from corresponding scenarios. Then a linear regression algorithm is used to extracted traces, which are named fitting results, from GPS points which are named fitting points. If GPS data is collected from two close roads, data should be separated by the clustering algorithm at first. Finally, combined with roads and lanes information, GPS points are assigned to corresponding lanes.

A. Judging the Number of Roads

In this paper, each road is one-way and vehicles driven on adjacent road are toward opposite direction. On each road, time stamps of GPS points are continuous. In our experimental scenarios, if time stamps of points are all continuous, the data can only be gotten from one road. If points can be separated for two or even more clusters by time stamps and directions are different, the number of roads is two.

B. Linear Regression Algorithm in C-A Method

The part will proof the bus trace can be gained from its GPS points based on the least square method.

Proof: The trace can be defined as Eq. (3) in a small range and Eq. (4) is the linear function fitted with n GPS points collected on the road, which meets the condition in Eq. (5).

$$y = a_0 \times x + b_0 \tag{3}$$

$$Y_i = a_1 \times x_i + b_1 \tag{4}$$

$$\begin{aligned} f(a_1, b_1) &= \min \sum_{i=1}^{n} (Y_i - y_i)^2 \\ &= \min \sum_{i=1}^{n} (a_1 * x_i + b_1 - y_i)^2 \end{aligned} \tag{5}$$

x_i and y_i respectively denote the longitude and latitude of the ith node in n points. It means partial derivatives a_1 sub $f(a_1, b_1)$ and b_1 sub $f(a_1, b_1)$ should meet Eq. (6) and Eq. (7).

$$\frac{f(a_1, b_1)}{\partial a_1} = \sum_{i=1}^{n} x_i (a_1 * x_i + b_1 - y_i) = 0 \tag{6}$$

$$\frac{f(a_1, b_1)}{\partial b_1} = \sum_{i=1}^{n} (a_1 * x_i + b_1 - y_i) = 0 \tag{7}$$

Because GPS points are symmetrically distributed around the trace defined by Eq. (3), when a_1 equals to a_0 and b_1 equals to b_0, the Eq. (6) can be workable. The Eq. (2) also can be shown as follow:

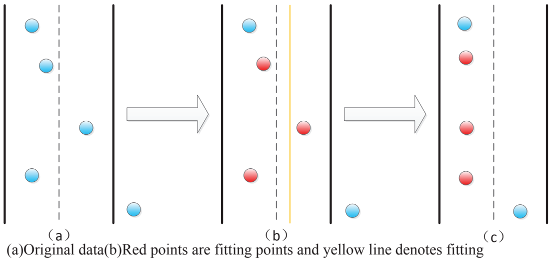
$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} (a_0 * x_i - y_i) = -b_0 \tag{8}$$

And the Eq. (7) can be converted to follow that

$$\frac{1}{n} \sum_{i=1}^{n} (a_1 * x_i - y_i) = -b_1 \tag{9}$$

So, when the n is endless or big enough, the b_0 and a_0 also can meet the Eq. (7). ■

However, in fact, the number of points is limited, and in some conditions such as bad weather, errors of data may be serious. In order to improve the accuracy of fitting results and choose the best result as the real trace, we adjust the process of the linear regression and compare the slope of fitting results with the slope of the real road. The most similar one is considered as the trace. The method can be shown as follows.



(a)Original data(b)Red points are fitting points and yellow line denotes fitting trace(c)Based on the distance between the points and trace, the points are assigned on the corresponding lanes.

Fig. 2: Process of linear regression algorithm

1) Step 1: Assuming the number of GPS points which are shown on the map is n, such as in Fig. 2(a), the n equals 5. The candidate set which is used to fit the trace is established as follows:

$$\{m_2, m_3.......m_n\} = c_n^i \tag{10}$$

i represents the number of points which are used in linear regression algorithm, m_i represents the permutation and combination of the relevant number i. For example, if n equals 3 and i equals 2, then m_2 is calculated as follow:

$$m_2 = \{[1, 2], [1, 3], [2, 3]\} \tag{11}$$

It represents point No. 1 and point No. 2, or point No. 2 and point No. 3, or point No. 1 and point No. 3 , which will be used to fit the traces and these are the three results.

2) Step 2: In order to keep points with serious errors from influencing the result, the fitting result is rejected, if the distance between fitting points and the fitting straight line is greater than lane width, which is 3.5m in the general condition.

3) Step 3: Based on Step 2, each slope of the result is compared with the slope of real road and the most similar one is chosen as the trace under system errors. In Fig. 2(b), three red points are chosen as fitting points and the yellow line is the best result which denotes the trace of the bus under system errors.

4) Step 4: The fitting trace denotes the driving trace in normal condition, however, in special situations the bus will change the lane, such as stopping and overtaking. The GPS points which are recorded at that time may have a greater distance from the fitting trace and the distance d is shown as follows:

$$d = \alpha + \delta \tag{12}$$

α denotes the value of random errors. δ denotes width of the lane and it is a constant about 3.5m. Therefor, if the distance between the point and fitting trace is lower than 3.5m, the point will be on the fitting trace. If it is greater than 3.5m and the point is at the left of the trace, the point is on the lane at the left of the trace, on the contrary, it is on the lane at the right of the trace. As a result, the more accurate result can be acquired. The result is shown in Fig. 2(c), based on the information of distance and the number of lanes, points are assigned to the corresponding lanes.

C. Clustering Algorithm in C-A Method

When two roads are close, GPS points may be shadowed on the adjacent road, because of errors of GPS data. Therefor, it is difficult to make sure which road GPS points really belong to and the revised method mentioned above is invalid. Before revising GPS data, the points should be separated. In this paper, an algorithm which is based on the K-means algorithm is used to undertake the task [13]. Based on our data, the method is adjusted as follow:

1) Step 1: k denotes the number of roads in the certain range and k also denotes the number of centroids of clusters in the K-means algorithm. The scale of scenario can be controlled to make sure there are only two roads are close.

2) Step 2: Attribute vector A[time, speed, angle] of the point will be established.

3) Step 3: On the road, two GPS points are chosen as the original centroids randomly. Based on the vector A, the Euclidean distance can be calculated as follow:

$$L_{ik} = a_1(T_i - T_k)^2 + a_2(V_i - V_k)^2 + a_3(\theta_i - \theta_k)^2 \quad (13)$$

$$\sum_{j=0}^{m} a_i = 1 \quad (14)$$

L_{ik} denotes the distance between the point i and the centroid k. It also denotes the similarity between them and the value of L_{ik} is lower, the value of similarity is higher. If L_{ik} is minimum, the point i will belong to the cluster k. m denotes the number of attributes. T_i, V_i and θ_i indicate the value of different attributes of ith points.

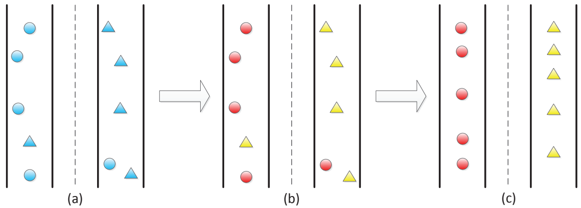
4) Step 4: Based on Step 3, after each point being assigned to the nearest cluster, the centroid of each cluster is calculated again by the arithmetic mean of each attribute. In terms of new centroids, the Euclidean distance is calculated and points are assigned to the new cluster again.

5) Step 5: The Step 4 will be iterated until members in each cluster do not change.

6) Algorithm modification: In order to improve the quality of clustering, the K-means algorithm should be mended. In Fig. 3(a), there are two close roads and different shapes of points denote GPS points which are collected from various roads. Because of GPS errors, some points are mapped on adjacent road. In Fig. 3(b), firstly, time stamp is used in the K-means algorithm. As the result, GPS points are assigned for two clusters and the different colors denote the different clusters. In each cluster, time stamps of points are continuous. Secondly, the distance between these points and the center liner of two roads is calculated. In order to use the correlation among members in the same clusters, the "distance of cluster" is defined and it is as the attribute in clustering. It can be calculated as follows:

$$D_k = \frac{\sum_{i=0}^{n} d_{ki}}{n} \quad (15)$$

D_k denotes the distance of cluster, and the d_{ki} denotes the distance between the point i which is in cluster k and the center line of the two roads. In order to cover the disadvantage of K-means algorithm that it is sensitive for acnodes, the initial value of two centroids is about -10 and +10m. The negative sign denotes that the point is at the left of the center line, on the contrary, the positive sign denotes the point is at the right of the center line. Finally, the result is shown in Fig. 3(c), the points are separated completely.



(a)Original data(b)Clustering result based on the time stamp(c)Finally result, the GPS points are separated completely.

Fig. 3: Process of clustering algorithm

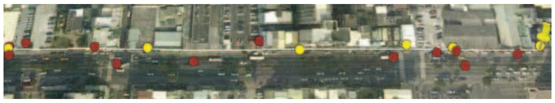


Fig. 4: Fitting trace and fitting points

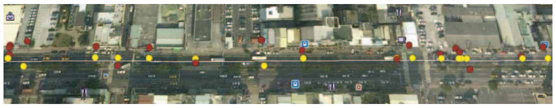


Fig. 5: Fitting result

V. Experimental Results

In the section, we use the linear regression algorithm to smooth the raw data mentioned above and separate the points when two roads are close. The moving object is chosen randomly and results are shown in part A as well as part B respectively. In part C, the accuracy of results and impact factors are discussed. Finally, we also discuss the relationship between the number of fitting points and fitting results.

A. Trace Approximated and Points Assigned

A one-way stretch is cut out from the map randomly and GPS points collected on it is gotten. In Fig. 4, there are 18 GPS points and based on the line regression algorithm 6 yellow points are used to approximate the trace. The white line denotes the trace under system errors. According to the distance between points and the fitting trace, the lanes which points belong to are ensured. The result is shown in Fig. 5. Red points denote original positions of GPS points and yellow points denote revising results.

B. Separate Points

In Fig. 6, there is a piece of area taken out randomly and GPS are mapped on the two close roads. Firstly, time stamp is used in the clustering algorithm. As the result, GPS points are assigned for three clusters. In Fig. 7, red points denote the uploading information from 04:57:29 to 04:57:49. Purple points denote the uploading information from 20:55:19

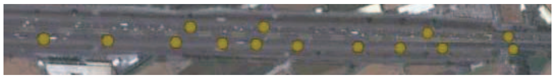


Fig. 6: Process of clustering and result(a)

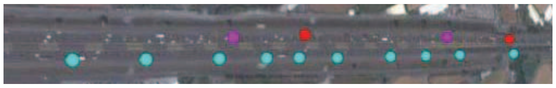


Fig. 7: Process of clustering and result(b)

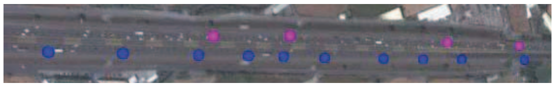


Fig. 8: Process of clustering and result(c)

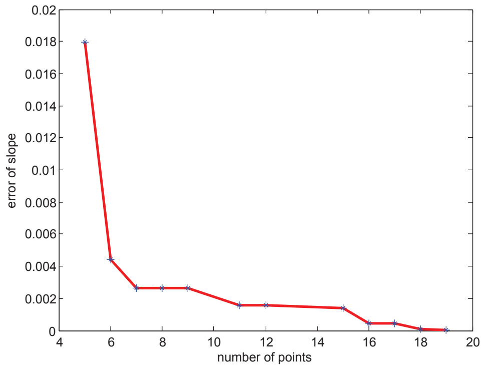


Fig. 9: Relationship between the number of fitting points and fitting results

to 20:55:39. Cyan points denote the uploading information from 12:03:54 to 12:06:54. Based on the distance of cluster the result is in Fig. 8, purple points denotes the data which is recorded when the bus is driving on the left road and blue points denotes the data which is recorded when the bus is driving on the right road.

C. *Discussions*

The result seems reasonable when compared with original datasets, however, GPS errors cannot be eliminated completely and some points may be worse than the original data. We analyze the phenomenon and there are two main reasons:

1. Errors in the original road map may lead to errors in the result.

2. Different GPS receivers own the different accuracy and some accuracy may be worse than 10m. It will influence the results.

We also discuss the relationship between the number of fitting points and fitting results. In Fig. 9, the X-axis denotes the number of points and the Y-axis denotes the error of slope between the real road and the fitting result. From the figure, with increasing of the number, the error is decreasing. It denotes the accuracy of the fitting result is positive correlation with the number of fitting points. The trend of the math curve demonstrates the mentioned formula. When GPS points which are used in the linear regression algorithm are endless, the real trace will be found in the fitting results.

VI. Conclusion

In the paper, we propose a novel method to smooth raw GPS data and make use of it to reduce GPS errors and GPS points are matched with the real lanes information according to the distance from the fitting traces. Compared with the traditional methods, the method described in this paper is lower cost without special equipment, such as fiducial stations. It can also be reliable and simple in different situations. At the end of this paper, we demonstrate these characteristics with the increase of datasets, the model is not getting complex and the results is getting more accurate. According to our work in this paper, the raw GPS data can be available and reliable for further research.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grant No. 61271176, No. 61401334 and No. 61571350, the Fundamental Research Funds for the Central Universities (BDY021403), the State Key Laboratory of Rail Traffic Control and Safety (Contract No. RCS2016K011), Beijing Jiaotong University, and the 111 Project (B08038).

References

[1] P. J. Tseng, C. C. Hung, Y. H. Chuang, K. Kao, W. H. Chen, and C. Y. Chiang, "Scaling the Real-Time Traffic Sensing with GPS Equipped Probe Vehicles," in *Proc. of IEEE VTC Spring*, pp. 1-5, May 2014.

[2] J. Qin, H. Zhu, Y. Zhu, L. Lu, G. Xue, and M. Li, "POST: Exploiting Dynamic Sociality for Mobile Advertising in Vehicular Networks," in *Proc. of IEEE INFOCOM*, pp. 1761-1769, May 2014.

[3] W. Zhang, Y. Chen, Y. Yang, X. Wang, Y. Zhang, X. Hong, and G. Mao, "Multi-Hop Connectivity Probability in Infrastructure-Based Vehicular Networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 4, pp. 740-747, May 2012.

[4] S. Miura, L. T. Hsu, F. Chen, and S. Kamijo, "GPS Error Correction with Pseudorange Evaluation Using Three-Dimensional Maps," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 6, pp. 3104-3115, Dec. 2015.

[5] C. Parent, S. Spaccapietra, C. Renso, G. Andrienko, N. Andrienko, V. Bogorny, and Z. Yan, "Semantic Trajectories Modeling and Analysis," *ACM Comput. Surv.*, vol. 45, no. 4, pp. 115-123, Aug. 2013.

[6] L. Cao, and J. Krumm., "From GPS Traces to a Routable Road Map," in *Proc. of ACM GIS*, pp. 3-12, Nov. 2009.

[7] L. Zhang, F. Thiemann, and M. Sester, "Integration of GPS Traces with Road Map," in *Proc. of ACM SIGSPATIAL*, pp. 17-22, Nov. 2010.

[8] R. Lopez, J. P. Malarde, F. Royer, and P. Gaspar, "Improving Argos Doppler Location Using Multiple-Model Kalman Filtering," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 8, pp. 4744-4755, Aug. 2014.

[9] R. H. Gting, T. de Almeida, and Z. Ding, "Modeling and Querying Moving Objects in Networks," *VLDB J.*, vol. 15, no. 2, pp. 165-190, 2006.

[10] Z. Yan, C. Parent, S. Spaccapietra, and D. Chakraborty, "A Hybrid Model and Computing Platform for Spatio-Semantic Trajectories," in *Proc. of 7th ESWC*, pp. 60-75, 2010.

[11] Y. Chen, and J. Krumm, "Probabilistic Modeling of Traffic Lanes from GPS Traces," in *Proc. of ACM SIGSPATIAL*, pp. 81-88, 2010.

[12] J. Jun, R. Guensler, and J. H. Ogle, "Smoothing Methods to Minimize Impact of Global Positioning System Random Error on Travel Distance, Speed, and Acceleration Profile Estimates," *Transp. Res. Rec.: J. Transp. Res. Board*, vol. 1972, pp. 141-150, Aug. 2006.

[13] T. Velmurugan, "Performance Based Analysis between K-Means and Fuzzy C-Means Clustering Algorithms for Connection Oriented Telecommunication Data," *Appl. Soft Comput.*, vol. 19, pp. 134-146, Jun. 2014.