



## Characterization of Individual Mobility and Society Using CDR Data

Mohammed Zohdy Abdulhady<sup>1</sup> , Loay E. George<sup>2</sup>

College of Science, Computer Science Department, Baghdad University,  
Baghdad, Iraq

Email: <sup>1</sup>[mohammedassafly@gmail.com](mailto:mohammedassafly@gmail.com), <sup>2</sup>[loayedwar57@yahoo.com](mailto:loayedwar57@yahoo.com)

<https://doi.org/10.26782/jmcms.2019.10.00002>

---

### Abstract

*Through the previous years, a large number of cell phones information has become in the hand for the analysis patterns of people movements. This information's carry a massive assurance for realizing behavior of human on a very large scale, as well as with an accuracy and precision never happened before can be allowed with surveys, censuses or other available data selection techniques. There are a number of researches that has open key advance into analyzing mobility of human utilizing this available recent data source, as well as there have been multiple various calculations of mobility applied. Mobility of human, or motion over large or short distances for narrow or vast durations of time, is an essential until continuous study for occurrence in the sciences of demographic and social systems. Meanwhile there have been harmonious progresses in compassionate migration (consider continuous pattern of mobility) as well as its effect on people happiness, social organizations, economic, and political organization, progresses in researches of mobility have been embarrass by complexity in measuring and recording how people move on a second and in detailed range. In this paper, the ability of using mobile network records will be described for analyzing the mobility of people and society for various objectives such as monitoring the mobility in cities and builds the suitable infrastructure for them. The mobility of individuals will be very benefit for observation the behavior of peoples and their effect in security issues. In order to test the system performance, a set of tests was applied on Zain calls dataset. The results indicates for the society mobility has been exported for the Baghdad Karkh area peoples. The results have been exported for two phases, one phases when the number of people's routes where only 10 movement and the second phase when the people routes where 3 routes.*

**Keywords :** Call phones, Mobile network records, Mobility, Human's behavior, Zain.

---

### I. Introduction

One of the very important and essential streams of information applied in telecommunication field is Call Detail Record. It offer modern chances and selections for industry of telecommunication as well as boost its incomes as well as CDR gives

the population ability to increase its living standards in various approaches. In addition, we require understanding operation of call detail record in fact to exclude its huge value that will help to get modern business chances. Processing data stream in real time consider modern directions in processing of call detail record [II]. Processing gives chances to understanding records of call detail with truth time as well as assist in getting location of real time of any user as well as the actions of network within truth time. It is important to notice those call detail records uses a big volume of information, types of data as well as various rate of data, in spite of recent system of telecommunications are created without taking into account these problems in mental. Records of call detail can be shown as source of huge data; as a consequence, records of call detail is allowable to use technologies of big data (for analysis, processing as well as storage) as example Hadoop in analysis of record of call detail. Too many large researches implemented to indicate the records of call detail understanding difficulties faced in industry of telecommunications [XVI].

The previous work in [XIII] introduced the implementation of Big data mechanism in call details records analysis by utilizing CDR analytics depend on implementation examples, concentrating on their structure, the application of Big Data material and mechanisms, as well as the CDR utilize case outline. The practical results noticed very well performance as well as cost effective in processing of CDR as well as processing of stream and real time. In [XV], they used data of CDR to understand distance of travel enclosed in traffic areas. Data of CDR may as well apply to classify zones of Traffic Analysis (TAZ) as well as exclude effective data required for generation of trip. Additionally, they can be applied to verify paths of travelling and displacement of cell phone customers. As a whole, this research delivers a recent worthy pattern exclude approach for large data driven planning of transportation, the displacement travelled as well as behavior of travel study. The method depends on utilizing K-means clustering approach, GIS, and Mahalanobis displacement to excluding the ground. As there is advancement in 4G signal of mobile, Cell phone with station of base delivers more easy information for studies. In the future researches contains understand the customer law of travel speed, determining speed of street traffic motions. By combining the geographical position details as well as the division of administrative, for the locative analysis transforming law. In [I], they discussed conclusion the social habits from the behavior of call. This work done by evaluating amount of time the person spent on his friends, his family, and his business. The number of users selected about 749 customers for three months of CDR. The behavior of each person in this work for each day classified into forty eight sections with each 30 minutes as a standard element where is addressed with one mode activity. By using hierarchal conditional random fields, it can be conclude there are about eight activity modes that compounds of 2 states consider the standard parameter for behavior of calling, 2 states for chat purpose, and four states for purpose of work. In [IX], they depicted a structure that determines passengers among whole population by studying mobile phone position information from records of call details. The structure also compounds of an effective yet efficient information methods of scan from massive call details records database of the whole mobile phone customers. This work structure can verify passengers by utilizing CDR information, a not active source of mobile phones position information. The method

states to scan massive database of CDR is as well effective in resource terms usage with no offer up the basic goal quality. This work can be worthy to persons who are in touch or not in touch related to the sector of tourism.

The purpose of this work is using the records of mobile networks for analyzing the mobility of human and society for different purposes. These analyses concerned the monitoring of cities and building the proper model for it as well as the effect of time and location. The results of these movements for societies have many advantages for peoples who manage these cities such as town planning and management. Analysis of data which classified as big data will be implemented and developing the required tools for fast dealing with it and minimizing the storage requirements for saving the data. After that, Geo data analytics operations performed for showing paths of human. The next stage will be using some methods data mining to find the models of movements for society and individuals. Finally, it can be invest results of this operation for purpose of modeling or prediction or characterization.

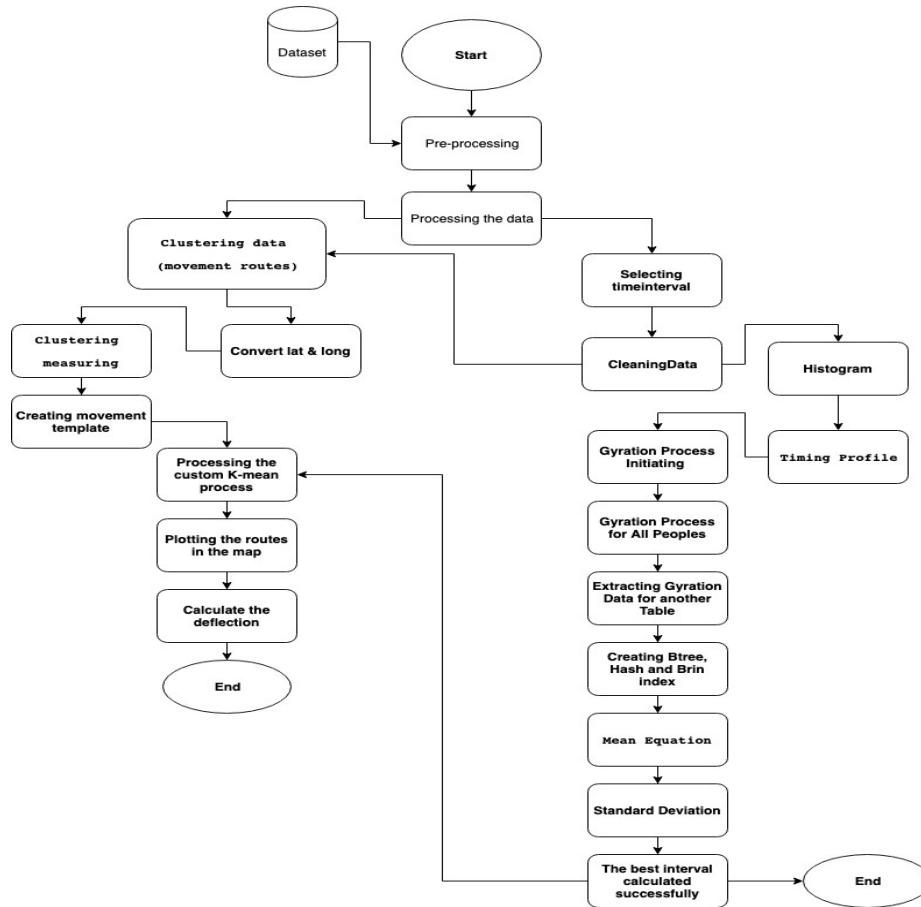
## **II. Materials and Methods**

### **II.i. Call Detail Record (CDR)**

The word of telecommunication consider as basic for telephony, communication, internet as well as video, telecommunication has very great variety of information as example, data of CDR such as data of user, data of network, as well as data of customers private as well billing. Records of call detail consider a report compound of complete information about transaction of telecommunication, as example call beginning time, time of call ending, call period in seconds, parties of call, ID of cell, websites requested, data type through calling or using the internet. As well as it delivers details of every event that happen via the network. Records of call detail cycle of life commonly starts with record of call detail creation of a call, as well as it is finished relating to the proceeding happens in the telephone (end of call, join of call, so on.), then it is selected by various elements of network [XVI].

### **II.ii. System Layout**

The proposed system in this paper, used mobile network records for analyzing the mobility of people and society, the overall design of the proposed system shown in figure 1.



**Figure 1:** The General Structure of the Proposed System

In this paper, the implemented stages in the suggested system offered. The workflow of the suggested system consists of three major phases: (a) Pre-processing phase, (b) Selecting time interval phase, (c) Clustering data (movement routes) phase, each phase consists of many stages.

### II.iii. Per-processing Phase

The preprocessing is important step, the data applied for this operation was data of Zain Company as well as this company uses standardization of Nokia binary. For manipulate this problem, a mediator program constructed that utilize standardization of Nokia row documentation so that the data will be able to read as well as the outcome will be mediated file.

### II.iv. Processing Data Phase

This phase divided into two parts, first part where the clustering conducted which each single tower where defined as centroids. In the second part the centroids where defined as a group of towers which they represented as a route of people.

#### **II.iv.a Selecting Time Interval**

Initial phase of the data processing, this phase will be connected by number of steps. The basic target of this phase is to determine the good timing period for the mean of single month that represent a span of the history in the data will be processed. This stage will base on pair basic formulas for determining the better period using the standard deviation and mean equations. As well as, the operation of optimization will be implementing for the behavior basing on the method of indexing for the dataset basic elements. The steps of this part will be summarized as:

- **Cleaning Data**

Cleaning of the data is the operation of calculating and rectifying mistakes or inexact reads from a set of records, database, or tables as well as correspond to verifying inaccurate, incorrect, incomplete or not related portions of the information and after that justifying, replacing, or removing the bad or rough data.

- **Histogram**

The histogram mechanism considers a precise exemplification of the numerical distribution for the data. The histogram represents an evaluation of the distribution of probability of a non-stopping element (CORAL) as well as histogram initially submitted by Karl.

- **Timing Profile**

Timing period a technique used for developing a constant template of dataset for record details of individual's who's involving in the manipulating operation. This portion of manipulate implemented by various options corresponding to the intervals number in every day for the initial [6 periods] each day for every month and the result was not precise causing that every period approximately larger from the obtainable details of timing of the individuals. Following that a twenty four timing period which added where every hour indicate for a one period in every day for one month.

- **Gyration Process Initiating**

It can be defined the body gyration radius or gyradius about an rotation axis as the point radial length from the rotation axis of a body by assuming the total body mass is to be fixed as well as its inertia moment around the available axis will be similar with its real mass distribution and it is indicated with  $R_g$ .

- **Gyration Process for All Peoples**

By depend on the former mechanism of step as well as procedures, the operation will contain overall individuals details of call as well the result will be sent and saved in second dataset.

- **Extracting Gyration Data for another Table**

The excluded gyration for a distinct table considers the initial portion of the behavior optimization. To maximize the processing behavior a recent table where created for the data of gyration only in the current step. Following that the period's data determined and added into the recent table.

- **Creating Btree, Hash and Brin index**

The structure of data that enhances the information retrieval speed processes on a table of database at which the cost of appended writes as well as space of storage to preserve the structure of data index. In each time we accessed the table of database we did not need search in every row of the database table because the indexes

deliver feature of quick location for the data we need. By using single column or more than one column the indexes can be created and delivers the fast basis for both rapid arbitrary lookups as well as effective access of requested records. This operation considers the second portion of the optimization of performance. The indexes which built for those columns are (calling date, ID, the number of calling) available in database for developing the behavior of operation.

- Mean Equation

The Median, Mean, as well as Mode are the average of arithmetic of the set of data. This is developing by appended the telephone numbers in a set of data as well as dividing by the number of these telephone numbers available. The term median is the center number in a set of data when the telephone numbers are recorded in either descending or ascending in order. The term mode is the number of occurrences that happen generally in a set of data, as well as the term range consider the variation between the lowest and highest amplitudes in a set of data. For this operation the process of gyration for data was determined utilizing the mean equation [VII].

$$\text{mean} = \frac{\sum d}{M} (1)$$

- Standard Deviation

The term of standard deviation determines consistency and variability of the population or sample. In general applications of real world, uniformity is a great benefit. In analysis of statistical data, little variation is generally better [VII].

standard

$$= \sqrt{\frac{\sum (d - \text{mean})^2}{M}} \quad (2)$$

#### **II.iv.b. Clustering Data (Movement Routes)**

The basic goal of this part to exclude the routes of people motions, that's will be determined based on the two major portions of operation which building a constant interval of template for individuals initially as well as after that manipulate the result in function of clustering. The portion will be implemented by number of stages.

- Convert Latitude and Longitude

From the sphere into the plane there is no map of isometric. As we transport the coordinates from latitude to longitude that we mean from the sphere to coordinates of x and y in the plane, there will be no ensuring that all distances will be maintained in this process. It must be we accept some type of what is named deformation. A lot of various map dropping there will be, which can perform various agreement between lengths preservations, areas, as well as angles. For smaller portions of surface of earths, Mercator of transverse is completely general.

- Clustering Measuring

The method of Elbow is an interpretation approach as well as consistency validation within analysis of cluster used to assist for determining the suitable clusters number in a set of data. The term elbow is determined by the circle with red color and the clusters number selected must be 4. This approach concentrate at the variance

percentage explained as clusters number function: One might select a clusters number so that appending more cluster will not make better data modeling.

- **Creating Movement Template**

Template of movement (96 periods) based on straight line equation. That mean the developed template will be indicated as a constant distance registration for all individuals for every day. Every fifteen minutes in a one day gives a single period and the overall of fifteen minutes (96 period) gives the motions registers for the individuals in single day. The process essential, due to sequential operation of specific k-mean in the following stage based on a templated registers.

- **Processing the Custom K-mean Process**

Clustering with K-means is an approach generally applied to dynamically divide a set of data into groups. It manipulate by choosing k primary center of cluster and after repeatedly purification it. Developing ten clusters which gives the better paths of the individuals for determining the movement of the individuals.

- **Plotting the Routes in the Map**

The platform of Bing Maps delivers number of API features for the implementations such as control of web, control of windows with Store apps, control of WPF, services of Spatial Data, as well services of REST. Utilize the data below and at documents of Microsoft to assist calculate which maps of Bing of API better proper for the needs.

- **Calculate the Deflection**

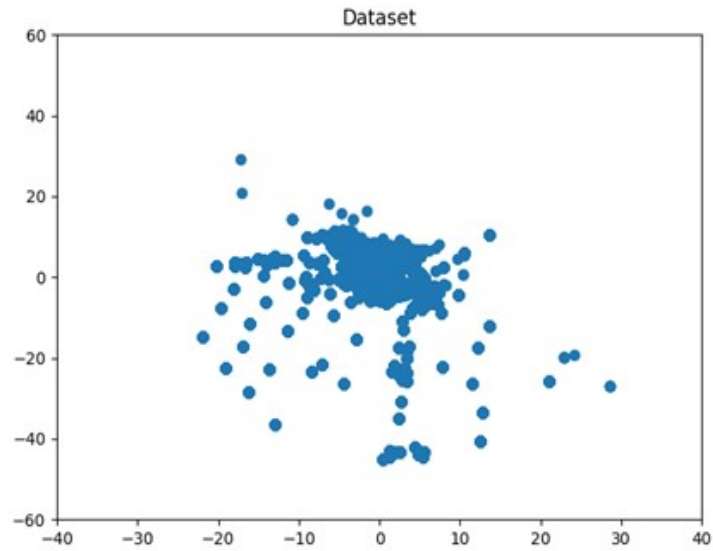
### **III. Results**

In this section, the results of some conducted tests are presented and discussed to evaluate the performance of the established system. The dataset used for the proposed system is delivered from Zain company and it was structured as binary files (RAW file) collected in 4 directory's each directory includes 7000 binary file and each file size between 4.6 MB to 4.8 MB. The data includes the call detail records of the company clients for one month collected from MSC servers, the total size of the data 130 GB. After plotted the locations into a map it's noticed that only the data of Karkh area included in the Baghdad data.

After delivery the data and the documentation for converting the data to readable files. The data is a group of raw files (custom binary structured files). Depending on the documentation demonstrate that the data structured as a group of segmented fields, each files represents a specific parameter and many of the fields indicates for the header of each record and other segments indicate for end of each record. Each segment has a specific format which encoded as a raw (binary) and it has a different decoding mechanism from other segments.

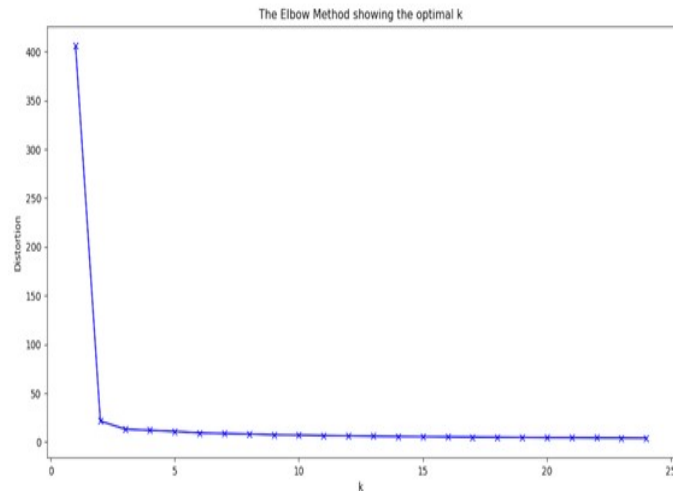
Figure 2 describes the clusters points plotting, after using Elbow for clustering to calculating the best number of clusters.





**Figure 2: Clusters Points Plotting**

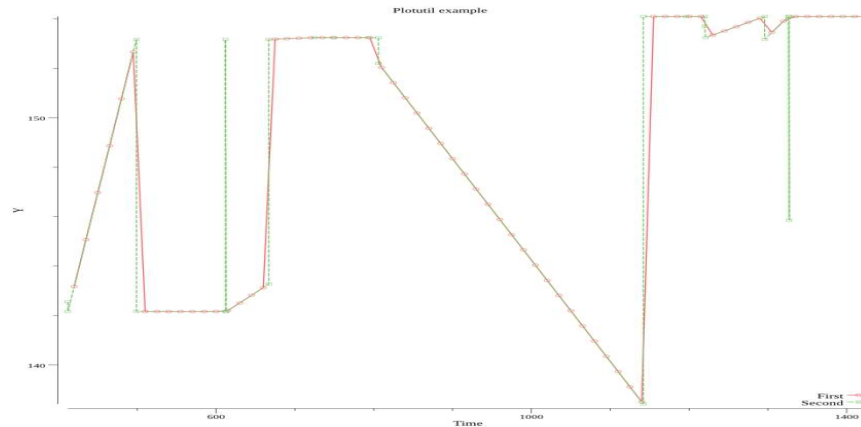
**Figure. 3** describes the clustering measuring and the best number of K in K-mean clustering algorithm.



**Figure 3: Best Number of K-mean Clustering**

Figure 4 describes the movement template from the information of movement which calculated for specific person for one day. The green line describes movement information of specific people for one day. The red line describes the 96-movement template depending on straight-line equation with Lambda threshold.





**Figure 4:** Movement template and normal movement information

**Figure 5** represents the map plotting for the 10 clusters who calculated from clustering 6 phases represented as a 10 different colors line on the map for timing interval between (12:00 PM and 06:00 PM) for one month.



**Figure 5:** 10 Cluster map plotting for timing interval between (12:00 PM and 06:00 PM)

## References

- I. Chen Zhou, Xu, Z., & Huang, B., "Activity Recognition from Call Detail Record: Relation Between Mobile Behavior Pattern And Social Attribute Using Hierarchical Conditional Random Fields", International Conference on Green Computing and Communications & IEEE/ACM International Conference on Cyber, Physical and Social Computing, 2010.

- II. Ghotekar, N., "Analysis and Data Mining of Call Detail Records using Big Data Technology", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 5, December 2016.
- III. M.Donato, K., "Current trends and patterns of female migration: Evidence from Mexico". International Migration Review, 27(4), 748-771, 1993.
- IV. Massey DS., "Social structure, household strategies, and the cumulative causation of migration". Population Index. 56:3–26. 10.2307/3644186, 1990.
- V. Massey, D. S., & Espinosa, K. E., "What's driving Mexico-US migration? A theoretical, empirical, and policy analysis", American journal of sociology, 102(4), 939-999, 1997.
- VI. Massey, D. S., Williams, N., Axinn, W. G., & Ghimire, D. J., "Community services and out-migration". International Migration. 48(3), 1-41, 2010.
- VII. Martin B., "Mean and Standard Deviation", report of applied statistics, 2006
- VIII. R., Harris, J. & Todaro, M. P., "Migration, unemployment and development: a two-sector analysis", The American economic review, 126-142, 1970.
- IX. Ratul Sikder, Uddin, M. J., & Halder, S., "An Efficient Approach of Identifying Tourist by Call Detail Record Analysis" International Workshop on Computational Intelligence (IWCI) 12-13 Dhaka, Bangladesh, December 2016.
- X. S. Massey, D., Arango, J., Hugo, G., Kouaouci, A., Pellegrino, A., & Taylor, J. E., "Theories of international migration: A review and appraisal", Population and development review, 431–466, 1993.
- XI. Stark, O., & Bloom, D. E., "The new economics of labor migration". The American Economic review, 75(2), 173-178. 1985.
- XII. Stark O, Taylor JE. "Migration incentives, migration types: The role of relative deprivation". The Economic Journal. 101:1163–1178. 10.2307/2234433, 1985.
- XIII. Sara B. Elagib, Hashim, A. H. A., & Olanrewaju, R. F. "CDR Analysis using Big Data Technology", International Conference on Computing, Control, Networking, Electronics and Embedded Systems Engineering, 2015
- XIV. W. Kandel, J. Durand, Parrado, E. A., & Massey, D. S., "International migration and development in Mexican communities". Demography, 33(2), 249-264, 1996.
- XV. Xuzhao Wang, Dong, H., Zhou, Y., Liu, K., Jia, L., & Qin, Y., "Travel Distance Characteristics Analysis Using Call Detail Record Data", 29th Chinese Control And Decision Conference (CCDC), 2017.
- XVI. Zhang, S., Yin, D., Zhang, Y., & Zhou, W., "Computing on Base Station Behavior Using Erlang Measurement and Call Detail Record", IEEE transactions on emerging topics in computing, 3(3), 444-453 2015.